Universidade Federal da Paraíba
Centro de Ciências Sociais Aplicadas
Programa de Pós-Graduação em Economia

Ensaios sobre Economia Aplicada: Doações eleitorais, compras públicas, análise de políticas afirmativas e reprovação no Ensino Superior

ANDRÉA FERREIRA DA SILVA

João Pessoa - PB
2019

ANDRÉA FERREIRA DA SILVA

Ensaios sobre Economia Aplicada: Doações eleitorais, compras públicas, análise de políticas afirmativas e reprovação no Ensino Superior

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Economia da Universidade Federal da Paraíba - UFPB como parte dos requisitos necessários à obtenção do título de Doutorado em Economia.

Universidade Federal da Paraíba

Centro de Ciências Sociais Aplicadas

Programa de Pós-Graduação em Economia

Orientador: Dr. Aléssio Tony Cavalcanti de Almeida

Coorientador: Dr. Hilton Martins de Brito Ramalho

João Pessoa - PB

2019

Catalogação na publicação
Seção de Catalogação e Classificação

S586e Silva, Andréa Ferreira da.
Ensaios sobre Economia Aplicada: Doações eleitorais, compras públicas, análise de políticas afirmativas e reprovação no Ensino Superior / Andréa Ferreira da Silva. - João Pessoa, 2019.
167 f. : il.

Orientação: Aléssio Tony Cavalcanti de Almeida.
Coorientação: Hilton Martins de Brito Ramalho.
Tese (Doutorado) - UFPB/CCSA/PPGE.

1. Doações Eleitorais. 2. Políticas afirmativas. 3. Avaliação de impacto. 4. Machine Larning. 5. Reprovação escolar. I. Almeida, Aléssio Tony Cavalcanti de. II. Ramalho, Hilton Martins de Brito. III. Título.

UFPB/BC

ANDRÉA FERREIRA DA SILVA

Ensaios sobre Economia Aplicada: Doações eleitorais, compras públicas, análise de políticas afirmativas e reprovação no Ensino Superior

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Economia da Universidade Federal da Paraíba como parte dos requisitos necessários à obtenção do título de Doutor em Economia, Submetida e APROVADO pela Comissão Examinadora abaixo assinada.

Defesa realizada no Campus I da UFPB em João Pessoa, Paraíba, no dia 6 de Setembro de 2019, às 14:00.

Dr. Alessio Tony Cavalcanti de Almeida
Orientador

Dr. Hilton Martins de Brito Ramalho
Coorientador

Dr. Jevuks Matheus de Araújo
Examinador Interno

Dr. Wallace Patrick S. de Farias Souza
Examinador Interno

Dr. Luciano Menezes Bezerra Sampaio
Examinador Externo (UFRN)

Por Carlos da N. Bezerra

Dr. Wellington Ribeiro Justo
Examinador Externo (URCA)

A Deus, minha família, meu noivo e amigos.

Agradecimentos

A Deus por me ouvir todas as vezes que me pus de joelhos para pedir e agradecer. Por me conceder saúde e ter os sonhos mais lindos em todas as etapas minha vida.

A minha pequena grande família. Minha Mãe, Francisca de Sousa, aos meus Avós (maternos) e Padrinhos, Antônia de Sousa e Raimundo Ferreira. Ao meu Padrasto, Mauro Cavalcante, por estar sempre perto da pessoa mais importante da minha vida. A meu pai, Antônio Paulino (*in memória*), que mesmo estando no céu sempre se manteve presente, fazendo minha proteção e me guiando para os melhores caminhos.

Ao meu noivo Pedro Lima, presente desde o inicio dessa caminhada. Agradeço pela compreensão, proteção e incentivo, e também à sua família pelos cuidados e atenção.

Aos amigos e colegas dos cursos de pós-graduação em economia, obrigada pelo companheirismo e força nos últimos anos: Janaildo, Charles, Vanessa, Guilherme, Otoniel, Laércio, Ana, Wallace, Celina, Stélio e Edinéia. Em especial, faço um agradecimento a Eryka, minha companheira de todas as noites de preocupação, de todas as lágrimas de tristeza e de felicidade. Grata pela paciência e pelos aprendizados. Aos amigos de uma vida, Netinho, Rhayane, Kátia, Leice, Ralyne, Edilene, Natália, Thays e Fernanda. Aos amigos do Iguatu, de João Pessoa e de todos os outros estados que fui ganhando no decorrer dessa caminhada.

Ao meu orientador Aléssio Tony Cavalcanti de Almeida por todas as oportunidades dadas ao longo deste curso. Pela paciência, dedicação, conhecimento e confiança não só durante a realização desta pesquisa, mas desde o primeiro momento que tive a oportunidade de ser sua aluna. Agradeço também a todos os meus professores como Jair Araujo (MAER/UFC), Aydano Leite (URCA), Adriano Paixão (UFPB), Ignácio Tavares (UFPB), Cássio Besarria (UFPB), Paulo Águiar (UFPB), Sinézio Maia (UFPB), Edilean Aragón (UFPB), José Luis (UFPB), Magno Vamberto (UFPB), entre tantos outros da graduação, mestrado e doutorado.

Aos professores membros da banca examinadora, meu co-orientador o Prof. Dr. Hilton Martins de Brito Ramalho, Prof. Dr. Jevuks Matheus de Araújo, Prof. Dr. Wallace Patrick S. de Farias Souza, Prof. Dr. Luciano Menezes Bezerra Sampaio e Prof. Dr. Wellington Ribeiro Justo, pela disponibilidade e valiosas contribuições para a elaboração final desta tese.

A Universidade Federal da Paraíba (UFPB) e, em especial ao Programa de Pós-Graduação em Economia (PPGE) pela dedicação na formação dos seus alunos. Aos funcionários do PPGE/UFPB, Risomar, Ricardo e Waleska, pela ajuda, competência e

profissionalismo em cada período.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro recebido com a concessão da bolsa de estudos, sem o qual não poderia ter iniciado e concluído o curso de doutorado.

Por fim, a todos aqueles, não diretamente citados, que contribuíram de alguma forma para a elaboração deste trabalho e encerramento de mais uma etapa da minha vida.

Resumo

Esta tese é composta por três ensaios não relacionados em microeconomia aplicada. O primeiro avalia o impacto de doações eleitorais sobre um possível favorecimento em compras públicas. Foram utilizados dados longitudinais de empresas e prestadores de serviços para as gestões municipais da Paraíba durante o período de 2004 a 2016, cuja as estimativas de impacto sobre os valores de contratos foram realizadas a partir do estimador de diferenças em diferenças, com controle para a heterogeneidade específica das empresas e prestadores de serviços com recortes amostrais para corrigir o viés de autosseleção. Os resultados centrais da pesquisa validam a hipótese que os financiamentos de campanhas políticas por agentes privados geram um retorno para os doadores de candidatos eleitos, em média, de 42% nos valores contratados, sendo essa taxa de retorno maior para as empresas do que para prestadores de serviços. Por sua vez, o segundo ensaio avalia os efeitos de uma ação afirmativa de reserva de vagas no ensino superior sobre indicadores educacionais de abandono e desempenho acadêmico. Para tanto, utilizou-se informações dos estudantes que ingressaram na Universidade Federal da Paraíba (UFPB), nos anos de 2010 e 2011. A metodologia adotada consistiu em duas etapas: (i) primeiramente, foram adotadas três técnicas de pareamento, *Propensity Score Matching* (PSM), *Mahalanobis Distance Matching* (MDM) e *Classification Tree Analysis* (CTA), para avaliar os efeitos da intervenção sobre o desempenho, captado pelo coeficiente de rendimento acadêmico (CRA) relativo; (ii) em seguida, fez-se uso de dados longitudinais dos estudantes, contemplando os anos de 2011 até 2018, para estimar modelos de duração de risco proporcional de Cox, ponderado pelo PSM, a fim de avaliar o efeito do aluno ser cotista sobre a probabilidade de sobrevivência na UFPB. Os resultados apontam que a existência do sistema de cotas reduziu o nível de desempenho dos discentes, independente do modelo de pareamento empregado, principalmente na distribuição que capta as melhores médias do CRA relativo. Já a estimação do modelo *survival analysis* aponta que a probabilidade de sobrevida dos alunos não cotistas é inferior aos dos alunos cotistas, o que permite concluir que estes últimos tendem a persistir mais no ensino superior. Por fim, o terceiro ensaio propõe identificar o risco de reprovação de discentes do ensino superior usando algoritmos de *Machine Learning* (ML). Com base nos registros administrativos e acadêmicos da UFPB e da Plataforma *Lattes*, para o período de 2010 a 2016 da disciplina de cálculo diferencial e integral I, foi verificado que os modelos com a melhor performance de previsão foram *Penalized Methods Lasso* e Regressão Logística. A partir da modelagem sobre os dados de treinamento (2010 a 2014), os resultados encontrados explicitam que, das 1.532 observações que compõem um novo conjunto de dados (2015 e 2016), a frequência dos alunos com status (reprovados e aprovados) previstos corretamente pela *Accuracy* foi de 67%, em ambos os modelos. Por sua vez, 72,5% dos discentes foram previstos corretamente como reprovados (*Sensitivity*). Esses achados ratificam que os algoritmos de ML podem ser instrumentos viáveis para auxiliar ações pedagógicas e gerenciais preventivas que visem a redução dos índices de reprovações no ensino superior.

Palavras-chave: Doações Eleitorais. Políticas afirmativas. Avaliação de impacto. *Machine Larning*. Reprovação escolar.

Abstract

This thesis encompasses three unrelated essays in applied microeconomics. The first one assesses the impact of electoral donations on possible favoritism in public procurement. We use longitudinal data from companies and service providers for the municipal administration of Paraíba during the period from 2004 to 2016, whose estimates of impact on contract values were carried out by the differences in differences estimator, with control for the specific heterogeneity of companies and service providers with subsamples to correct self-selection bias. The main results of the research validate the hypothesis that political campaign funding by private agents generates a return for donors of elected candidates, on average, of 42% in the contracted values, where this return rate is higher for the companies than for the service providers. In turn, the second essay evaluates the effects of an unreserved affirmative action in higher education on dropout and academic performance educational. For this, we used information from students who were admitted were admitted the Federal University of Paraíba (UFPB), in the years 2010 and 2011. The adopted methodology consisted of two steps: (i) first, we use three matching techniques, Propensity Score Matching (PSM), Mahalanobis Distance Matching (MDM) e Classification Tree Analysis (CTA), in order to evaluate the effects of the intervention on performance, captured by the relative *Coeficiente de Rendimento Acadêmico* (CRA), (ii) then, we use longitudinal data from the students, contemplating the years from 2011 to 2018, to estimate models of Cox proportional risk duration, weighted by the PSM, in order to evaluate the effect of the student being a quota holder on the probability of survival in the UFPB. The results indicate that the existence of the quota system reduced the performance level of students, regardless of the matching model employed, especially in the distribution that captures the best relative CRA averages. The estimation of the survival analysis models points out that the unlikely probability of non-quota students is lower than that of quota students, which allows us to conclude that the latter tend to persist more in higher education. Finally, the third essay proposes to identify the risk of failing higher education students using Machine Learning (ML) algorithms. Based on the administrative records of the UFPB and Plataforma Lattes, for the period 2010-2016 of the discipline of differential and integral calculus I, we verify that the models with the best performance of forecasting were Penalized Methods Lasso and Logistic Regression. From the modeling on training data (2010-2014), the results show that of the 1,532 observations that make up a new data set (2015 and 2016), the frequency of students with status (failed and approved) correctly predicted by Accuracy was 67 % in both models. In turn, 72.5 % of students were correctly predicted to fail (Sensitivity). These findings confirm that ML algorithms can be viable instruments to assist preventive pedagogical and managerial actions aimed at reducing the failure rates in higher education.

Keywords: Electoral donations. Affirmative policies. Impact assessment. Machine Learning. School failure.

Lista de tabelas

Lista de ilustrações

Lista de abreviaturas e siglas

AUC ROC	*Area Under the ROC Curve*
ATT	*Average Treatment Effect on Treated*
BNDES	Banco Nacional de Desenvolvimento Econômico e Social
CCAE	Centro de Ciências Aplicadas e Educacional
CCEN	Centro de Ciências Exatas e da Natureza
CCSA	Centro de Ciências Sociais e Aplicadas
CEAR	Centro de Energias e Alternativas e Renováveis
CI	Centro de Informática
CEFETS	Centros Federais de Educação Tecnológica
CT	Centro de Tecnologia
CNPq	Conselho Nacional de Desenvolvimento Cientifico e Tecnológico
CTA	*Classification Tree Analysis*
CRA	Coeficiente Rendimento Acadêmico
CONSEPE	Conselho Superior de Ensino, Pesquisa e Extensão
COPERVE	Comissão Permanente do Concurso Vestibular
DD	Diferenças em Diferenças
EMN	Esforço Mínimo Necessário
ENADE	Exame Nacional de Desempenho dos Estudantes
FDA	Função de Distribuição Acumulada
FN	Falso Negativo
FP	Falso Positivo
HR	*Hazard Ratio*
IBGE	Instituto Brasileiro de Geografia e Estatística

INEP	Instituto Nacional de Estudos e Pesquisa Anísio Teixeira
IES	Instituições de Ensino Superior
IF	Instituto Federal
KNN	*K-Nearest Neighbors*
LASSO	*Least Absolute Shrinkage and Selection Operator*
ML	*Machine Learning*
MDM	*Mahalanobis Distance Matching*
MQO	Mínimos Quadrados Ordinários
MPL	Modelo de Probabilidade Linear
PSS	Processo Seletivo Seriado
PSM	*Propensity Score Matching*
PT	Partido dos Trabalhadores
RDD	Regressão Descontinua
RQ	Regressão Quantílica
ROC	*Receiver Operating Characteristic*
SiSU	Sistema de Seleção Unificada
STI	Superintendência de Tecnologia da Informação
SVM	*Support Vector Machines*
TSE	Tribunal Superior Eleitoral
TCE-PB	Tribunal de Contas do Estado da Paraíba
TVP	Taxa de Verdadeiro Positivo
TFP	Taxa de Falso Positivo
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
WPS	*Workplace Panel Survey*

Sumário

1 Introdução

Em economia aplicada, é consenso na literatura que ocorra a conciliação entre os instrumentos abordados em métodos quantitativos e a formação em todas as linhas de pesquisa em economia, como por exemplo em análise de políticas públicas, desenvolvimento econômico, macroeconomia, microeconomia, teoria econômica, economia da saúde e outras. A ideia é que por meio das ferramentas dos métodos quantitativos (estatística, matemática e econometria), os pesquisadores das mais diversas áreas de estudo possam testar as hipóteses como também fazer predições de modelos teóricos.

Métodos quantitativos aplicados nas mais diversas áreas da economia se tornam fundamentais para o crescimento de cada respectiva linha de pesquisa. Estudos como Nash Jr (1950), com um dos seus trabalhos para a teoria dos jogos em microeconomia, Solow (1956), e sua contribuição para a teoria do crescimento econômico em macroeconomia, Arrow (1978), em economia da saúde, Hanushek (1986), para a economia da educação, assim como vários outros (SCHULTZ, 1961; MINCER, 1974), passaram a ser referências base para o desenvolvimento de novos estudos.

Dessa maneira, sob a perspectiva da importância de métodos quantitativos em economia aplicada, muito embora o Brasil já possua uma produção quantitativa considerável nesta vertente, esta tese desenvolve três ensaios, estruturadas em capítulos, na busca por suprir algumas lacunas existentes em cada área estudada. Alicerçados nas teorias de financiamento de campanhas políticas (economia da corrupção) e teorias em economia da educação, com foco no ensino superior, a presente pesquisa explora e investiga novas aplicações de métodos e técnicas que buscam melhores inferências e desempenhos (para o caso dos modelos de previsão) em problemas econômicos já existentes.

No primeiro ensaio, o escopo da pesquisa é conduzido pela avaliação do impacto de doações eleitorais sobre um possível favorecimento em compras públicas através do estimador de diferenças em diferenças, com controle para a heterogeneidade específica das empresas e prestadores de serviços com recortes amostrais para corrigir o viés de autosseleção. Esta pesquisa pretende avançar nas discussões das vantagens advindas de vitórias eleitorais e conexões políticas em nível municipal. Frente ao poder discricionário das gestões nas compras governamentais, a hipótese a ser testada segue a linha de que o financiamento em campanhas eleitorais resulta em favorecimentos nos processos de contratos públicos pelas empresas e prestadores de serviços que doaram recursos financeiros aos prefeitos e vereadores eleitos em 2004, 2008 e 2012.

Por sua vez, o segundo ensaio evidencia os efeitos de uma ação afirmativa

de reserva de vagas no ensino superior sobre indicadores educacionais de abandono e desempenho acadêmico dos estudantes que ingressaram na Universidade Federal da Paraíba (UFPB) nos anos de 2010 e 2011. Este capítulo adota duas abordagens metodológicas: (i) três técnicas de pareamento não-experimental para avaliar os efeitos da intervenção sobre o desempenho, captado pelo coeficiente de rendimento acadêmico (CRA) relativo; (ii) para os anos os anos de 2011 até 2018, foram adotados os modelos de duração de risco proporcional de Cox, combinado com o *propensity score matching* (PSM), a fim de avaliar o efeito do aluno ser cotista sobre a probabilidade de sobrevivência na UFPB. Ressalta-se que uma análise completa de ação compensatória requer, além de se observar o diferencial dos rendimentos acadêmicos, deve-se avaliar também o comportamento dos discentes cotistas ao longo do curso escolhido.

Finalmente, o terceiro e último ensaio objetiva prever o risco de reprovação de discentes de curso superior a partir da aplicação dos algoritmos de *Machine Learning* (ML) para dados da UFPB. A principal motivação baseia-se em desenvolver uma estratégia eficiente que consiga identificar o comportamento dos discentes, de modo a permitir a intervenção de professores e coordenações com o objetivo de recuperar o aluno antes que o corra a efetiva reprovação. A identificação de discentes com maior risco de reprovação nas disciplinas com os maiores índices de retenção pode estar diretamente relacionado a necessidade de intervenções na educação no ensino superior, as quais buscam reduzir não apenas a retenção, mas, consequentemente, a evasão, para que assim possa evitar desperdício de recursos públicos.

Esta tese é composta por cinco capítulos. Além desta introdução, o Capítulo 2 aborda o primeiro ensaio a respeito da relação entre doações eleitorais e compras públicas. O Capítulo 3, segundo ensaio, apresenta os efeitos de políticas afirmativas sobre desempenho e abandono ensino superior. O Capítulo 4, equivalente ao terceiro e último ensaio, inova ao aplicar uma nova caixa de ferramentas a um problema de previsão de risco de reprovação em uma universidade pública federal no Brasil. Por fim, as considerações finais são feitas no Capítulo 5.

2 Doações eleitorais e favorecimento em compras públicas: avaliação de impacto aplicada ao caso dos municípios da Paraíba

2.1 Introdução

O debate sobre os financiamentos de campanhas eleitorais no Brasil ficou mais intenso a partir da redemocratização na década de 80, onde ocorreram vários cenários envolvendo práticas ilegais entre agentes políticos e privados. Frente a uma legislação que proibia as doações de pessoas jurídicas nos processos eleitorais, ainda era evidente que o setor privado tinha se tornado, mesmo que de forma ilícita, a principal fonte de recursos para o financiamento de campanhas políticas[1] (KRIEGER; RODRIGUES; BONASSA, 1994).

Em reação ao exposto, o financiamento de campanhas no Brasil teve uma significativa mudança a partir da Lei nº 8.713/1993, a qual permitiu as doações eleitorais de pessoas jurídicas do setor privado e estabeleceu a transparência da prestação de contas dos candidatos por parte do Tribunal Superior Eleitoral. Por sua vez, a Lei nº 9.504/1997 estabelece normas para as eleições, onde os recursos para as campanhas eleitorais podem ser derivados de diversas fontes, a saber: pessoas jurídicas; pessoas físicas; os próprios candidatos; fundo partidário; rendimentos de aplicações financeiras; e pela comercialização de bens e realização de eventos (BRASIL, 1993; BRASIL, 1997).

A partir de então, as doações da iniciativa privada às campanhas políticas passaram a ter um papel fundamental no processo eleitoral. Conforme as prestações de contas dos candidatos, após a reforma eleitoral em 1993 as doações de empresas responderam por mais de 50% dos recursos arrecadados nas eleições seguintes a tal data, chegando a representar, por exemplo, 74,4% dos recursos investidos nas eleições de 2010 (Tribunal Superior Eleitoral, 2012). Em virtude da importância do setor privado no processo democrático, surgiram questionamentos sobre os possíveis favorecimentos econômicos que os doadores possam vir a obter ao contribuir nas campanhas políticas.

[1] Segundo Krieger, Rodrigues e Bonassa (1994), escândalos como o do ex-presidente Collor, em 1991/1992, e o dos "anões do orçamento", em 1993 (KRIEGER; RODRIGUES; BONASSA, 1994) são exemplos da relação do financiamento eleitoral por agentes privados em troca de vantagens em contratos públicos.

A literatura existente documenta que o financiamento de campanhas políticas se divide em duas linhas centrais. A primeira aponta que doações de campanha podem ser entendidas com um investimento por parte do doador. Com um estudo aplicado para os EUA, Snyder Jr (1990) concluiu que as empresas que doam para os candidatos no período eleitoral veem suas contribuições na forma de um investimento, esperando futuros benefícios após as eleições. A outra linha destaca que as doações eleitorais podem ser vistas como uma forma de empoderamento político, e não como um financiamento (ANSOLABEHERE; FIGUEIREDO; Snyder Jr, 2003). Contudo, no contexto apresentado por Ansolabehere, Figueiredo e Snyder Jr (2003), a maior parte dos recursos recebidos pelos candidatos é oriundo de pessoas físicas, as quais esperam o retorno na qualidade da prestação dos serviços públicos, pois são ideologicamente motivadas.

No entanto, em diversos outros contextos políticos (RUMBA; JASČIŠENS, 2009; MIRONOV; ZHURAVSKAYA, 2012; COVIELLO; GAGLIARDUCCI, 2017), inclusive no Brasil, onde os valores das doações provenientes de pessoas jurídicas respondem por uma parte significativa das receitas dos candidatos[2] (Tribunal Superior Eleitoral, 2012), é necessário mensurar o impacto das doações sobre possíveis favorecimentos futuros, principalmente quando consideram-se doações da iniciativa privada, empresas e prestadores de serviços. Nesse caso, investimento com vantagens mútuas, ou conexões políticas, são consideradas como trocas entre doadores e candidatos ou a disponibilização de recursos por benefícios. Tais benefícios, ou retornos, podem ser auferidos nas mais diversas formas, como a valorização das empresas por meio de suas ações, realização de gastos públicos que favoreçam interesses próprios, obtenção de regulamentações favoráveis, aquisições de contratos com o poder público, etc.

No tocante aos estudos sobre a relação entre as doações eleitorais e os resultados dos processos de compras públicas no Brasil, a literatura nacional nessa temática evidencia que as empresas privadas com alguma conexão política obtém maiores benefícios em relação ao valor contratado com o poder público, sendo esses originados de emendas parlamentares (ARVATE; BARBOSA; FUZITANI, 2013; BOAS; HIDALGO; RICHARDSON, 2014; FONSECA, 2017). Contudo, essa investigação é escassa considerando os aspectos das doações em nível descentralizado e envolvendo o financiamento de campanhas para o poder executivo. Dessa maneira, para preencher essa lacuna, visto que os trabalhos existentes no Brasil são para candidatos a cargos estaduais do legislativo (ARVATE; BARBOSA; FUZITANI, 2013) e cargos federais do legislativo e do executivo (BOAS; HIDALGO; RICHARDSON, 2014; FONSECA, 2017), o presente estudo pretende avançar e contribuir nas discussões de conexões políticas em nível

[2] Destaca-se que a partir das eleições municipais de 2016, o Supremo Tribunal Federal (STF) vetou o financiamento de empresas. Dessa forma, essa conjectura é válida para as eleições anteriores a 2016, que de fato, são o foco deste estudo.

municipal.

Embora a descentralização, implantada na Constituição de 1988, permitiu aos municípios estruturas administrativas capazes de gerir suas próprias políticas, Brueckner (1999) afirma que esta pode acarretar em uma série de imperfeições, destacando que nas instituições locais é mais provável a existência de corrupção do que em níveis mais centralizados, visto que estas são mais desenvolvidas e há um maior poder de fiscalização por órgãos de controle.

Por conseguinte, compreender a existência de um possível efeito das contribuições eleitorais sobre potenciais benefícios aos doadores de campanhas em nível municipal é relevante para o desenho de mecanismos que visem desestimular os atos ilícitos entre os agentes públicos e privados. Nesse contexto, o presente estudo busca avaliar o impacto de doações eleitorais sobre um possível favorecimento em compras públicas em nível de governo municipal no estado da Paraíba. Mais especificamente, objetiva-se testar se as empresas e prestadores de serviços, doadoras para candidatos eleitos nas eleições municipais de 2004, 2008 e 2012, apresentaram uma taxa de retorno positiva em termos dos contratos públicos com os municípios.

Assim, frente ao elevado poder discricionário do prefeito no município, a hipótese a ser testada segue a linha de que o financiamento em campanhas eleitorais resulta em favorecimentos nos processos de contratos públicos pelas empresas e prestadores de serviços que doaram recursos financeiros aos prefeitos e vereadores eleitos em 2004, 2008 e 2012. A justificativa para o período temporal em estudo se dá pois o processo eleitoral em 2012 foi o último, em nível municipal, antes de ocorrer uma nova reforma na Lei que trata das normas para as doações em campanhas eleitorais, sendo determinada pela Resolução nº 23.463, de 15 de dezembro de 2015 (BRASIL, 2015).

Metodologicamente, para o cumprimento do objetivo proposto foi usada a estratégia empírica de *Diferenças em Diferenças* (DD) considerando um painel longitudinal de empresas e prestadores de serviços, que possuem contratos públicos nos municípios paraibanos. Na estimação, os dados longitudinais do Tribunal Superior Eleitoral (TSE) de 2004, 2008, e 2012, referente as informações sobre os candidatos eleitos e não eleitos e os doadores de campanhas, são utilizados em conjunto com as informações do Tribunal de Contas do Estado da Paraíba (TCE-PB) referentes aos dados de contratos públicos municipais de 2005 a 2016.

Este ensaio está dividido em cinco partes, o que inclui além desta introdução, a Seção 2.2 que descreve os principais estudos da literatura nacional e internacional baseados em conexões políticas, enfatizando os principais benefícios auferidos pelas empresas no setor público. A Seção 2.3 e a Seção 2.4 detalham a estratégia empírica e a base de dados com a descrição das variáveis, respectivamente. Por fim, a Seção 2.5 e

a Seção 2.6 tratam dos principais resultados e conclusões do estudo.

2.2 Conexões Políticas e o Processo de Concessão de Benefícios

Considerado um dos primeiros estudos nessa temática, Snyder Jr (1990) analisa a relação entre as contribuições de campanhas eleitorais e a vitória dos candidatos na Câmara dos Deputados dos Estados Unidos entre os anos de 1980 e 1986. Se por um lado, os doadores de campanha veem as contribuições como um investimento, ao qual esperam algum retorno futuro (isenções fiscais, contratos para fornecer bens ou serviços ao governo, entre outros), por outro lado, os candidatos esperam que os investimentos recebidos os ajudem a ganhar as eleições. Dessa maneira, o autor desenvolve um modelo de forma intuitiva e empírica o qual discute que o valor do retorno esperado pelos doadores é uma função de probabilidade que depende da vitória do candidato vencer as eleições, ou seja, consiste em um mercado em que os candidatos estão disputando pelas doações. Ante isso, a pesquisa corrobora positivamente as duas hipóteses impostas sobre a demanda por contribuições dos investidores: o primeiro é que todos os candidatos buscam maximizar as contribuições de modo a prometer o maior número de favores possíveis; já o segundo refere-se aos candidatos que têm uma grande probabilidade de vencer as eleições e não buscam maximizar suas contribuições, assim não se atentam a prometer o número máximo de favores, diferente dos candidatos com pequenas chances de vencerem, pois vão prometer a quantidade de favores suficientes para receberem as contribuições.

Isto posto, é possível analisar a associação entre as conexões políticas, sendo estas mensuradas por meio de doações eleitorais na grande parte dos estudos, e os retornos alcançados pelos agentes privados. No campo da literatura que sublinha tal ligação, serão apresentados a seguir alguns trabalhos que investigaram essa temática e confirmaram a influência das relações políticas no processo de concessão de benefícios para empresas. Os trabalhos a seguir estão agrupados por evidências de benefícios recebidos e por diferentes definições adotadas de conexões políticas. Tem-se inicialmente estudos em que os doadores são favorecidos no acesso ao financiamento em bancos públicos e no impacto significativo sobre o valor da empresa.

Com base nesta discussão, Fisman (2001) inicia seus estudos afirmando que as conexões políticas podem desempenhar um papel importante em muitas das maiores e mais importantes economias do mundo. Por meio de um índice de percepção de corrupção utilizado como *proxy* para identificar se existem conexões políticas, o autor buscou analisar a relação entre os valores das empresas e a saúde do presidente Suharto da Indonésia, assumindo que a conexão política era o principal determinante

da rentabilidade/lucratividade das empresas. Estimando um modelo para dados em painel com 79 empresas, e após construir um Índice de Dependência de Suharto no período entre 1995 e 1997, as principais descobertas dizem respeito a uma estrutura política altamente centralizada, estável e um forte nível de corrupção. Dessa forma, na Indonésia, o valor das ações das empresas varia de acordo com o nível de estabilidade das conexões políticas. Em particular, os preços das ações de empresas que estão ligadas ao Governo de Suharto caíram mais do que os preços das empresas menos conectadas no momento em que ocorreram instabilidades na saúde do presidente.

Por sua vez, Faccio (2006) parte com o objetivo de investigar qual a principal característica em comum entre os países com conexões políticas, e debate ainda se fazer as referidas conexões com o poder público adicionam algum valor às empresas. Para responder os objetivos propostos, o autor montou uma base de dados composta por 20.202 empresas de capital aberto de 47 países. Diferentemente de outros estudos, a definição de conexões políticas adotada segue de acordo com a relação entre o "estar" político e ou ser acionista ou ser um dos principais diretores de alguma das empresas em analise. Sendo esta definição justificada por ser uma conexão mais de longo prazo do que apenas as contribuições de campanhas entre as empresas e os candidatos. Assim, através de uma pesquisa descritiva, sem a intenção de implicar qualquer causalidade, a pesquisa apresenta evidências de que 541 empresas estão politicamente ligadas em 37 países, representando quase 8% da capitalização do mercado mundial. Faccio (2006) concluiu que as conexões são particularmente comuns em países com os mais altos níveis de corrupção, em países que impõem restrições sobre investimentos estrangeiros e em países com sistemas mais transparentes. Nesse contexto, as empresas detêm a oportunidade de lucros maiores, mais acesso ao crédito bancário, pagar menos impostos, e auferir uma maior fatia do mercado. No que se refere ao valor das ações, aumentaram significativamente, no entanto, o valor da empresa aumenta mais quando um empresário é eleito como primeiro-ministro e não como membro do parlamento.

Em uma pesquisa aplicada ao Brasil, Claessens, Feijen e Laeven (2008) examinam que em ambientes onde são realizadas contribuições destinadas às campanhas eleitorais existem fortes conexões políticas. Uma vez que o Brasil é um dos poucos países a registrarem as doações de campanha à nível candidato, o autores testaram empiricamente duas hipóteses principais: a primeira refere-se que as empresas mais ativas politicamente, ou seja, oferecem maiores contribuições nas campanhas, são mais propensas a receber maiores retornos políticos, principalmente no que tange os financiamentos dos bancos públicos nos quatro anos seguintes as eleições dos candidatos eleitos. Já a segunda hipótese, por sua vez, trata-se que conexões políticas propiciam retornos significativos em relação ao mercado de ações das empresas contribuintes. Com base nos dados do Tribunal Superior Eleitoral (TSE) para as eleições de 1998 e 2002 referentes aos candidatos a deputados federais no Brasil, os autores, através de

um painel de efeito fixo, aferem as diferenças do acesso ao financiamento entre as empresas que possuem conexões políticas e as empresas que não possuem conexões nesse período. Os resultados da pesquisa indicam que o acesso ao capital bancário é um importante canal onde as ligações operam, sobretudo para as empresas que contribuem para três grupos de candidatos específicos, a saber: candidatos à reeleição, candidatos da base aliada do presidente da república, e candidatos eleitos. Como também, contribuir para candidatos a deputados federais que ganham as eleições está associado a um impacto positivo na valorização das ações das empresas.

Do mesmo modo, Lazzarini *et al.* (2014) testaram, dentre outras, a hipótese de que as empresas com conexões políticas (medidas por meio das doações de campanha) são beneficiadas com uma maior alocação de capital (financiamentos) fornecido pelo Banco Nacional de Desenvolvimento Econômico e Social (BNDES). Construíram um painel de dados com base nas informações de 286 empresas de capital aberto na Bolsa de Valores de São Paulo (BM&F Bovespa) entre 2002 e 2009, e das informações de todos os valores doados em campanha oriundas do Tribunal Superior Eleitoral (TSE) para cada candidato concorrendo a Presidente, Senador, Deputado Estadual ou Deputado Federal no Brasil, como também com base das informações de um conjunto de variáveis de controle. Encontraram que as empresas que doaram para os candidatos que ganharam as eleições têm uma maior probabilidade de receber o financiamento sob a forma de empréstimos do BNDES do que as empresas que doaram para os candidatos que não ganharam as eleições.

Outros benefícios encontrados na literatura sobre conexões políticas é a correlação positiva entre as empresas e o número de contratos no setor público. Estudos como os de Goldman, Rocholl e So (2008), Arvate, Barbosa e Fuzitani (2013), Boas, Hidalgo e Richardson (2014), Coviello e Gagliarducci (2017) e Fonseca (2017) apontaram um aumento no número de contratos entre as empresas conectadas politicamente.

O estudo de Goldman, Rocholl e So (2008) tem por escopo aferir se as conexões políticas entre empresas e compras governamentais afetam a alocação de contratos nos Estados Unidos. Como também, esclarecer se tais conexões agregam valor nas empresas sob a luz do contexto político. Dessa maneira, para definir o conceito de conexões políticas na pesquisa, os autores classificaram a filiação política de 412 empresas que fazem parte do mercado S&P 500 (mercado de bolsas de valores e títulos nos EUA), ou seja, colheram informações detalhadas sobre cada membro do conselho administrativo e a sua posição política, como forma de mensurar a obtenção de acesso ao governo. Goldman, Rocholl e So (2008) então calcularam a mudança no valor dos contratos de aquisição de cada empresa em torno das eleições de 1994 e 2000. Uma vez que estão classificadas nas seguintes categorias: (1) Empresas conectadas ao partido Republicano; (2) Empresas conectadas aos Democratas; e (3) outros, para cada amostra de 1994

(amostra 2000) foi comparado o valor total dos contratos públicos recebidos entre 1990 e 1993 (1996 e 1999) com o valor total desses contratos recebidos entre 1995 e 1998 (2001 e 2004). Por meio de um modelo *cross-section*, as principais conclusões são que, tanto para a amostra de 1994 quanto de 2000, as empresas que estavam ligadas ao partido republicano (democrata) eram as mais prováveis de receber um aumento (redução) no valor dos seus contratos, visto que houve uma mudança no cenário político no controle da Câmara e do Senado dos democratas para os republicanos. Portanto, o artigo destaca uma das formas em que as conexões políticas podem ter uma influência no valor dos contratos das empresas por meio da alocação de recursos do Governo.

Com o objetivo de examinar o retorno liquido de uma doação de campanha sob a possibilidade do candidato a deputado estadual perder ou ganhar as eleições em oito estados brasileiros no ano de 2006, Arvate, Barbosa e Fuzitani (2013) investigaram a relação entre as empresas que fizeram doações para deputados estaduais em troca de contratos de compras como definições de conexões políticas. Os autores fizeram uso de uma Regressão Descontinua (RDD) e de um banco de dados composto por 1.795 candidatos, dos quais 321 foram eleitos e 1.474 não foram eleitos, e observam que o retorno liquido esperado pelas empresas doadoras, considerando todos os contratos, é alto e positivo, reforçando a ideia da forte relação entre candidato e doadores. Contudo, os benefícios esperados pelas empresas doadoras diferem entre si, o retorno oriundo do financiamento para o candidato de qualquer partido é mais elevado do que para os candidatos da aliança do governador.

Ainda no Brasil, Boas, Hidalgo e Richardson (2014) estimaram o efeito de uma vitória eleitoral dos candidatos sobre o valor dos contratos públicos auferidos pelas doadoras de campanha. Com base nas informações dos resultados das eleições para deputados federais em 2006; do Tribunal Superior Eleitoral (TSE) a respeito das características demográficas dos candidatos e das doações de campanha registradas; das informações sobre as empresas oriundas do Ministério da Fazenda (ano de fundação, código da classificação da empresa, e o estado sede); e da Controladoria Geral da União, para medir os contratos entre empresas doadoras e Governo Federal, os autores montaram um banco de dados com todas as despesas do Governo Federal para 7.375 empresas doadoras no período entre 2004 e 2010 e 1504 candidatos a deputados federais. As estimativas feitas por Regressão Descontinua (RDD), para captar o retorno das empresas doadoras, utilizam como variável de resultado o valor médio dos contratos das empresas, agregadas por benefícios recebidos entre janeiro de 2004 a setembro de 2006 e entre janeiro de 2008 a setembro de 2010. Dessa maneira, a partir de uma análise realizada a nível candidato, os resultados empíricos encontraram que o investimento político das grandes empresas, principalmente no que tange as do setor de construção civil, esperam um aumento substancial em torno de R$ 138,60 e R$ 346,26 adicionais nos contratos governamentais - o que equivale a um valor entre 14 e 39 vezes o valor

de suas contribuições - quando doam para um candidato a deputado federal do Partido dos Trabalhadores (PT). Por fim, o estudo não encontrou qualquer efeito entre outras conexões políticas em relação aos outros partidos, incluindo também os aliados legislativos do PT.

Após uma revisão feita nos estudos de Arvate, Barbosa e Fuzitani (2013) e Boas, Hidalgo e Richardson (2014), Fonseca (2017) traçam uma nova estratégia para averiguar se as empresas que efetuaram doações para partidos de coalizão do governo federal receberam maiores valores contratuais no Brasil, mas diferentemente dos outros estudos apresentados até aqui, a pesquisa investiga a relação antes e após as eleições. Com base nos dados do TSE, sobre doações de campanha de 2006, e de valores contratuais extraídos do Portal da Transparência do Governo Federal para os anos de 2004 a 2006 e 2008 a 2010, a pesquisa adota duas estratégias, a saber: a primeira verifica o viés do efeito do tratamento por meio de uma regressão linear; e a segunda consiste em uma estimação que tem como variável dependente o valor dos contratos recebidos antes das eleições. Dessa maneira, Fonseca (2017) estima o efeito mínimo de doar para integrantes da coalizão sobre os valores contratuais assumindo o problema do viés de seleção oriundo tanto das variáveis não observadas quanto de variáveis omitidas. Os resultados centrais apontaram um baixo retorno contratual após as eleições, por outro lado, valores mais expressivos recebidos dos contratos foram observados antes. No que tange as coalizões, não foi encontrado fundamentação empírica para afirmar que as conexões políticas entre doadores do partido do governo federal poderiam receber maiores retorno quando comparados aos doadores da oposição. Posto isto, o autor concluiu afirmando que o benefício recebido pelas doadoras é recebido antes das eleições.

Partindo de um estudo a nível municipal, Coviello e Gagliarducci (2017) baseiam-se nas informações de todos os prefeitos italianos eleitos entre 1985 e 2010 (*Ministero degli Interni*) alinhadas com as informações sobre aquisições para obras públicas oriundas de licitações/leilões por município entre 2000 e 2005 (*Autorità per la Vigilanza sui Contratti Pubblici di Lavori, Servizi e Forniture*) para analisar empiricamente a influência do cargo do prefeito nas contratações públicas nos seus municípios (consideradas como conexões políticas). De modo especifico, para cada município foi relacionado o mandato do prefeito com vários resultados dos processos de compras: consideraram o número de licitantes, o desconto da empresa vencedora, a probabilidade de o vencedor ser local e a probabilidade de que a mesma empresa receba licitações/leilões repetidos. Com base em duas estratégias de identificação diferentes: RDD e um quase-experimento, os autores revelaram que uma reeleição dos prefeitos está relacionada a perdas do setor público nas contratações, pois um prazo adicional reduz significativamente o número de licitantes que participaram dos leilões (-10,4%), como também reduz o desconto do vencedor (-5,2%), ou seja, há uma deterioração no

sistema de licitações. Realçam também uma maior probabilidade de que o vencedor da licitação seja local, e que a mesma empresa ganhe licitações diversas vezes. Assim, sugerem que quanto maior o tempo do prefeito em seu exercício maior seria a probabilidade de existir conluio entre os agentes públicos e as empresas locais, ou de outro modo, desenvolver um conjunto de relações corruptas.

2.3 Estratégia Empírica

A hipótese levantada por esta pesquisa diz respeito ao papel do financiamento de campanhas eleitorais dos prefeitos e vereadores eleitos em 2004, 2008 e 2012 sobre os favorecimentos nos processos de compras públicas pelas empresas e prestadores de serviços doadores nas eleições municipais no estado da Paraíba entre 2005 e 2016. Para testar esta hipótese, as estimações dos parâmetros de interesse foram desenvolvidas por meio do método de Diferenças em Diferenças (DD) (GALIANI; GERTLER; SCHARGRODSKY, 2005; ATHEY; IMBENS, 2006) com painel quadrienal. Por meio do DD é possível lidar com o viés de seleção associado às características não observáveis invariantes no tempo das empresas e prestadores de serviços.

A literatura existente para o efeito causal dessa problemática, tanto em nível nacional quanto internacional, não adotou o método DD em seus estudos, dada às peculiaridades das análises (avaliação na ótica do político). O modelo empírico mais comum aplicado entre as pesquisas apresentadas na Seção 2.2 é a Regressão Descontínua (RDD) (ARVATE; BARBOSA; FUZITANI, 2013; BOAS; HIDALGO; RICHARDSON, 2014; COVIELLO; GAGLIARDUCCI, 2017). No que diz respeito a validade externa de tal modelo, este apresenta uma baixa validade externa, pois a sua aplicabilidade é através de um ponto de corte (*cutoff*) para definir tratamento e controle. De outro modo, a investigação delimita-se apenas a uma análise local, onde, por exemplo, os autores Arvate, Barbosa e Fuzitani (2013) e Boas, Hidalgo e Richardson (2014) utilizaram a margem de vitória como o *cutoff*.

Dessa maneira, a estratégia de identificação proposta nessa pesquisa consiste em comparar o grupo de tratamento (composto pelos agentes privados que financiaram os candidatos a prefeitos e vereados eleitos em 2004, 2008 e 2012 e que, ao mesmo tempo, possuíam contratos de compras e serviços com o respectivo município após os resultados eleitorais) com o grupo de controle (composto pelos agentes privados que não tiveram seus candidatos eleitos, mas que também obtinham contratos públicos advindos das compras e serviços com os municípios após as eleições). Assim, para contornar um possível viés de autosseleção – fatores associados a decisão das empresas e prestadores de serviços decidirem serem financiadoras de campanhas – foi usado como estratégia central de identificação a comparação entre os resultados contratuais de

doadores para candidatos eleitos e os valores recebidos por doadores para candidatos não eleitos.

Por sua vez, a variável de resultado da análise é o montante do valor empenhado dos contratos públicos por empresas e prestadores de serviços durante o quadriênio referente ao mandato dos candidatos eleitos. Isto é, para as eleições municipais de 2004, o mandato dos candidatos vitoriosos seria de 2005 a 2008, e para as eleições de 2008 e 2012 os quadriênios de interesse seriam, respectivamente, 2009 a 2012 e 2013 a 2016. Este recorte temporal justifica-se pelo fato de que as informações dos doadores são fixas no tempo até ocorrer uma nova eleição. Posto isto, o objetivo é identificar o efeito médio das doações eleitorais para candidatos eleitos sobre os retornos auferidos nos contratos públicos, ou seja, comparar os contratos que são firmados no mesmo momento no tempo, contudo na presença e na ausência do tratamento.

A princípio, a melhor estratégia metodológica seria um método experimental, onde o impacto seria observado apenas ao comparar as médias dos dois grupos, ou seja, a diferença de médias entre tratamento e controle. Contudo, na ausência de um estudo randomizado, recorre-se aos métodos não-experimentais que refletem condições robustas ao do método experimental. Assim, uma grande preocupação existente é encontrar um grupo de controle que represente o verdadeiro contrafactual.

Ante aos vários tipos de problemas que podem confundir a estratégia de identificação, principalmente no que se refere a associação entre a variável de resultado (valor empenhado), e participação no programa (doadores de candidatos eleitos) e as características não observáveis das unidades de observação que sejam invariantes no tempo, o método DD com efeito fixo em conjunto com um recorte amostral de empresas e prestadores de serviços doadoras minimiza possíveis problemas de seleção. Logo, o método empírico proposto pode oferecer indícios importantes sobre a taxa de retorno das doações eleitorais.

2.3.1 O método de Diferenças em Diferenças (DD)

O modelo de DD é um dos métodos mais empregados na área de avaliação de impacto. Sua principal hipótese, das trajetórias paralelas, refere-se a trajetória temporal, de forma que a variável de resultado do grupo de controle represente o que ocorreria para o grupo tratado, caso não ocorresse a intervenção do programa (BERTRAND; DUFLO; MULLAINATHAN, 2004; GALIANI; GERTLER; SCHARGRODSKY, 2005; ATHEY; IMBENS, 2006).

Formalmente, o modelo empírico de DD baseia-se em um modelo de regressão linear de efeito fixo, tendo em vista que o valor empenhado nos contratos de compras e serviços públicos é função dos doadores de campanhas e das variáveis de controle

(características das empresas e prestadores de serviços), a Equação 2.1, a seguir, é a base para as primeiras estimativas desta pesquisa:

$$y_{it} = \delta(T_{it}t) + \mathbf{X'}_{it}\gamma + c_i + \lambda'_t + u_{it} \tag{2.1}$$

em que y_{it} é o *log* do valor de empenho das empresas e prestadores de serviços, i, no período de tempo t, definido por quadriênios (2005 a 2008; 2009 a 2012; e 2013 a 2016). O parâmetro δ mede o impacto do candidato que recebeu a doação ser eleito, onde ocorre a interação das variáveis T e t. Sendo T e t são variáveis binárias que representam, respectivamente, se é tratado (doadores para candidatos eleitos) ou caso contrário, se é no período pós-programa ou caso contrário. $\mathbf{X'}_{it}$ é o vetor de variáveis observáveis que variam no tempo. Por sua vez, c_i captura os efeitos fixos observados e não observadas da i-ésima empresa, isto é, a sua inclusão permite controlar para a heterogeneidade existente entre as unidades observadas com características que sejam invariantes no tempo. E, por fim, λ'_t é o vetor das *dummies* de tempo referentes a cada quadriênio.

O termo de erro, u_{it}, varia entre as unidades e no tempo, e assume-se distribuído independentemente de todos os c_i e λ'_t. A persistência de fatores exógenos não observáveis e variantes no tempo podem induzir a correlação de séries temporais a nível das empresas e prestadores de serviços doadoras, ou seja, os termos de erros podem ser correlacionados ao longo do tempo e espaço.

Assim, a especificação da regressão linear empregada no modelo de dados em painel pode ser utilizada para verificar como o método de DD controla para a influência das características não observáveis que não se alteram no tempo. Os efeitos fixos individuais passam a ser incluídos na variável de resultado no modelo, os quais entram na expressão como um conjunto de variáveis binárias.

Não obstante, apesar do modelo de DD possuir uma série de vantagens, existem alguns casos em que o mesmo não consegue lidar, principalmente no que se refere a mudança temporária em um fator não observável dos indivíduos que pode vir a afetar a decisão de participar do programa. De outra forma, qualquer situação em que as características não observáveis que varie no tempo ocorrer, acarretará em um viés que afetará, simultaneamente, a variável de resultado e a participação no programa (GALIANI; GERTLER; SCHARGRODSKY, 2005). Por esse motivo foram utilizadas especificações que consideram tanto tratados quanto controles empresas com registro de doações, com a única diferença entre elas o resultado do candidato financiado nas eleições.

2.4 Base de Dados e Descrição das Variáveis

Para a execução desta pesquisa foi necessária a construção de uma base de dados envolvendo o cruzamento de informações de duas fontes, sendo elas: os microdados do Tribunal Superior Eleitoral (TSE) e do Tribunal de Contas do Estado da Paraíba (TCE-PB). O painel longitudinal construído é composto por 964.672 empresas e prestadores de serviços não doadoras e doadoras que financiaram campanhas políticas, nos 223 municípios do estado da Paraíba, para eleger candidatos a prefeitos e vereadores nas eleições de 2004, 2008 e 2012. De forma clara, foram coletadas informações apenas para as empresas (pessoas jurídicas) e prestadores de serviços (pessoas físicas) não doadoras e doadores que evidenciaram alguma conexão com os agentes políticos, mensuradas por meio de contratos de compras e serviços públicos firmados com os municípios nos quadriênios após os respectivos períodos eleitorais, a saber: 2005 a 2008, 2009 a 2012 e 2013 a 2016.

As informações referentes as doações de campanhas recebidas pelos candidatos dos municípios paraibanos são originadas da base de "prestações de contas" disponibilizadas pelo TSE, mais precisamente, na base de "receitas dos candidatos". Do mesmo modo, obtidos através do TSE, as informações sobre os resultados eleitorais de 2004, 2008 e 2012, ou seja, os candidatos eleitos nos respectivos processos eleitorais, são coletadas da base de "resultados" para fazer o cruzamento entre os agentes privados (empresas e prestadores de serviços) que doaram para candidatos eleitos (prefeitos e vereadores), e agentes privados que doaram para candidatos não eleitos.

No que tange as informações dos valores contratados nas compras e serviços públicos na Paraíba, estes são coletados no TCE-PB referentes aos quadriênios após as eleições municipais supracitadas anteriormente. Adotou-se no modelo a variável valor empenhado como *proxy* dos valores contratados, uma vez que esta variável significa de fato o pagamento liquidado no processo de contração pública. Ou seja, representa efetivamente o valor recebido pelo agente privado no setor público.

A Tabela 2.1 apresenta, de forma estratificada, as estatísticas descritivas das variáveis de tratamento e de resultado[3] por período eleitoral e por quadriênio antes e após as eleições municipais no estado da Paraíba. Conforme as estatísticas descritivas da variável de tratamento, o valor médio doado pelas 13.839 empresas e prestadores de serviços durante os três períodos eleitorais a nível municipal é de R$ 1.926,2 com um desvio padrão de R$ 7.101. Observando o comportamento por período, observa-se que os candidatos do processo eleitoral de 2012 receberam o maior valor médio de doações, R$ 2.221,7.

[3] Foi feito a correção monetária de ambas as variáveis para dezembro de 2016 usando o Índice Nacional de Preços ao Consumidor Amplo – IPCA.

Em um primeiro momento, na análise estratificada por doações à candidatos eleitos (grupo de tratamento) e não eleitos (grupo de controle), verifica-se que, em média, o valor doado para os candidatos eleitos é superior aos não eleitos, respectivamente, R$ 2.370,5 e R$ 1.403,8. Sendo também na campanha de 2012 o período com o maior valor arrecadado de doações de candidatos não eleitos de pessoas jurídicas e físicas (R$ 2.117,9), por outro lado, as doações para candidatos eleitos concentraram-se em 2008 (R$ 2.460,2). No segundo momento, ao observar para qual cargo os agentes privados preferem direcionar os seus financiamentos, nota-se que os valores doados concentram-se para os candidatos a prefeitos, a saber: R$ 3.202,3 para os vitoriosos e R$ 3.093,4 para os perdedores. Já para os candidatos a vereadores, os eleitos receberam R$ 1.344,2 e os não eleitos R$ 876,6, em média.

No que se refere as estatísticas descritivas, por grupos e quadriênios, da variável de resultado dos agentes privados que tem, ou não, conexão política com os agentes públicos municipais são apresentados também na Tabela 2.1. A princípio, as evidências inicias apresentaram que o grupo, que financiou as campanhas políticas municipais na Paraíba, recebeu um valor contratado médio de R$ 127.219,6 para o período em análise. Em contrapartida, o grupo de não doadores de campanhas eleitorais no estado apresentou um valor recebido médio de apenas R$ 67.051,9 nos contratos referentes às compras e serviços públicos nos municípios, ou seja, inferior ao valor médio recebido pelo grupo de doadores.

Por sua vez, os valores contratados para os agentes privados que financiaram as campanhas de candidatos eleitos foram, em média, superiores aos dos agentes que também financiaram os candidatos a nível municipal, mas que não foram eleitos. De outro modo, os doadores para candidatos eleitos receberam em média um valor de R$ 154.464,9 nos contratos de compras e serviços nos municípios, frente a R$95.180,6 recebido pelos doadores de campanhas de candidatos não eleitos.

Em relação aos valores médios contratados por doadores a prefeitos e vereados eleitos e não eleitos por quadriênio, tem-se indícios que as empresas e prestadores de serviços que financiaram apenas os candidatos a prefeitos, sendo eles eleitos (R$ 238.989,5) ou não (R$202.537,7), receberam, em média, um valor maior nos contratos do que os agentes privados que doaram só para os candidatos a vereadores, eleitos (R$ 65.693,8) ou não (R$ 66,160,0) (Tabela 2.1). Esse comportamento também é observado ao longo dos quadriênios em análise.

Tabela 2.1 – Estatística descritiva da variável de tratamento e resultado, estratificado por tipo de doador e período: Média e Desvio Padrão.

Valor Doação (R$)	**Período Eleitoral**			**Total**
	2004	**2008**	**2012**	
Doações	1.842,4	1.815,7	2.221,7	1.926,2
	(4.628)	(8.556)	(6.751)	(7.101)
Doações - Eleitos	2.327,5	2.460,2	2.282,9	2.370,5
	(5.086)	(11.533)	(7.079)	(8.846)
Prefeito	3.587,6	3.514,6	2.504,5	3.202,3
	(6.785)	(15.671)	(7.781)	(605)
Vereador	1.188,2	1.204,8	1.639,6	1.344,2
	(2.037)	(2.615)	(4.681)	(255)
Doações - Não Eleitos	1.353,6	1.127,3	2.117,9	1.403,8
	(4.057)	(2.905)	(6.155)	(4.150)
Prefeito	3.823,4	2.338,7	3.118,1	3.093,4
	(8.376)	(4.992)	(8.409)	(742)
Vereador	735,5	801,9	1.092,4	876,6
	(1.107)	(1.700)	(1.956)	(189)
N	4.230	6.119	3.490	13.839
Valor Contrato (R$)	**Quadriênio**			**Total**
	2005-2008	**2009-2012**	**2013-2016**	
Doadores	119.203,4	114.421,1	159.375,2	127.219,6
	(1.572.386)	(1.031.382)	(1.345.441)	(1.297.169)
Não Doadores	30.919,9	67.593,6	101.512,7	67.051,9
	(1.663.610)	(3.246.346)	(5.470.906)	(3.810.800)
Doadores Eleitos	166.415,2	130.385,3	177.562,0	154.464,9
	(2.174.156)	(979.593)	(1.576.688)	(1.573.803)
Prefeitos	272.716,7	201.770,5	242.481,3	238.989,5
	(3.244.468)	(1.292.737)	(1.943.169)	(35.601)
Vereadores	81.264,7	47.219,4	68.597,2	65.693,8
	(368.970)	(3445.518)	(499.092)	(17.207)
Doadores Não Eleitos	71.633,0	97.372,3	128.510,8	95.180,6
	(444.268)	(1.083.866)	(814.227)	(864.318)
Prefeitos	223.029,7	223.612,7	160.970,7	202.537,7
	(934.229)	(1.527.885)	(944.023)	(35.999)
Vereadores	33.923,6	66.962,5	97.594,0	66.160,0
	(150.375)	(950.565)	(671.612)	(31.842)
N	313.543	328.391	322.737	964.672

Fonte: Elaboração própria a partir dos dados do SAGRES, Tribunal de Contas do Estado-PB, e TSE.
Nota: Os valores entre parenteses são os desvios padrões das variáveis.

A Tabela 2.3 apresenta de forma sumarizada o total de empresas e prestadores de serviços que financiaram candidatos a prefeitos e vereadores eleitos e não eleitos por campanhas eleitorais municipais no estado da Paraíba. Pelo o comportamento da amostra, os prestadores de serviços concentravam-se mais em doar do que as empresas, principalmente para os vereadores. Contudo, deve-se levar em conta a frequência no número de candidatos, uma vez que os candidatos para o cargo legislativo do município são em maior número do que os candidatos para os cargos do executivo.

Tabela 2.3 – Total de empresas e prestadoras de serviços doadoras por candidatos e períodos eleitorais nos municípios da Paraíba.

Tipo de Pessoa	Tipos de Doação	Candidatos	Total de Doadores por Período Eleitoral		
			2004	2008	2012
Empresa	Doação Eleitos	Prefeito	114	159	112
		Vereador	44	40	66
	Doação Não Eleitos	Prefeito	79	87	98
		Vereador	65	108	87
Prestadores de Serviços	Doação Eleitos	Prefeito	823	1.468	1.283
		Vereador	1103	1399	639
	Doação Não Eleitos	Prefeito	341	487	539
		Vereador	1614	2.244	552

Fonte: Elaboração própria a partir dos dados do TSE.

Entretanto, mesmo os prestadores de serviços sendo observados com maior assiduidade no processo de doações de campanhas, quando se trata do valor doado, as empresas se impõem, detendo a maior parte das doações, em média. Chegando a apresentar um valor até quatro vezes maior do que os prestadores de serviços, a saber: pessoas físicas doam R$ 1.605,5 e pessoas jurídica R$ 5.699,8, em média. Outro ponto que deve ser observado na Tabela 2.3 é o comportamento estável dos doadores, sinalizando que provavelmente são os mesmos em todos os períodos eleitorais, exceto para os grupos de doadores para vereadores eleitos e não eleitos em 2012, em que foi observado uma redução considerável.

2.5 Resultados

Com a intenção de apresentar algumas evidências iniciais a respeito do comportamento da variável de impacto, a Tabela 2.5 reporta os resultados dos testes de médias e intervalos de confiança do valor dos contratos para os grupos de controle e tratamento da análise. Para as informações dos grupos que foram utilizados nos modelos de DD, os intervalos afirmam, com 95% de confiança, que as médias são diferentes com sobreposição. Os grupos de doadores e não doadores apresentam uma persistente heterogeneidade ao longo do tempo na média dos valores contratados.

Comportamento similar é observado também nos grupos de doadores para candidatos eleitos e não eleitos, contudo, essas diferenças vão reduzindo-se ao longo do tempo. Desse modo, tais diferenças entre os grupos consistem com a aplicação do procedimento do modelo, pois refletem as diferentes influências das características observáveis e não observáveis dos indivíduos. Logo, a média da variável "valor dos contratos" não precisa coincidir entre os grupos de controle e tratamento. Por outro lado, o que o modelo requer é que a variação temporal do que ocorre com o grupo de

controle antes e após o programa represente de fato a variação temporal na situação do contrafactual, na ausência do tratamento.

Tabela 2.5 – Testes de médias e intervalo de confiança da variável de resultado, estratificado por tipo de doador e período eleitoral.

Valor dos Contratos	Média	IC	*p-value*
		2005-2008	
Doadores	119.203,4	71.805 - 166.601	0,000
Não Doadores	30.919,9	25.058 - 36.781	0,000
Doadores Eleitos	166.415,2	73.879 - 258.951	0,000
Empresas	1.385.477,9	206.627 - 2.564.328	0,021
Prestadores de Serviços	63.685,2	52.736 - 74.634	0,000
Doadores Não Eleitos	71.633,0	52.652 - 90.613	0,000
Empresas	656.084,7	403.240 - 908.929	0,000
Prestadores de Serviços	28.759,3	23.668 - 33.850	0,000
		2009-2012	
Doadores	114.421,0	88.573 - 140.268	0,000
Não Doadores	67.593,6	56.385 - 78.801	0,000
Doadores Eleitos	130.385,3	96.217 - 164.455,3	0,000
Empresas	1.610.965,4	1.128.326 - 2.093.604	0,000
Prestadores de Serviços	27.135,7	24.342 - 29.928	0,000
Doadores Não Eleitos	97.372,3	58.303 - 136.441	0,000
Empresas	1.211.294,4	646.806 - 1.775.781	0,000
Prestadores de Serviços	17.489,4	14.668 - 20.310	0,000
		2013-2016	
Doadores	159.375,1	144.722 - 204.028	0,000
Não Doadores	101.512,7	82.534,9 - 120.490	0,000
Doadores Eleitos	177.562,0	111.581 - 243.542	0,000
Empresas	1.882.914,4	1.128.757 - 2.637.071	0,000
Prestadores de Serviços	22.530,0	19.408 - 25.651	0,000
Doadores Não Eleitos	128.510,8	84.105 - 172.916	0,000
Empresas	786.351,3	496.304 - 1.076.398	0,000
Prestadores de Serviços	16.689,8	14.148 - 19.231	0,000

Fonte: Elaboração própria a partir dos dados do SAGRES, Tribunal de Contas do Estado-PB.
Nota: Intervalos com 95% de confiança para as médias calculadas.

Conforme evidenciado nas Tabelas 2.1 e 2.5, informações sobre os possíveis grupos de doadores e não doadores, como tratamento e controle, foram expostos na intenção de apresentar tanto a clareza dos dados, quanto a defesa de o porquê esses não são os grupos adotados no presente estudo. Devido a grande amostra de doadores, e principalmente, de não doadores, a heterogeneidade existente nesses grupos poderia acarretar em vieses de seleção no modelo.

Mesmo assim, a Tabela 2.7 apresenta uma estimativa ingênua para o modelo descrito. Apesar de existir um potencial problema de viés, os parâmetros foram positivamente associados e significativos. Contudo, avaliar a decisão das empresas

entre doar ou não em campanhas eleitorais pode determinar um viés de seleção, já que o que determina as doações não é um fator exógeno ao modelo.

Nesse caso, possíveis erros de especificação podem acarretar em erros de medição nas estimativas ocasionados por fatores não observados, como, por exemplo, a principal motivação para a ligação política entre os agentes. O que por si só, não poderia ser corrigido com o modelo de DD com efeito fixo. Assim, para corrigir o viés de autosseleção, foi feito um recorte amostral apenas para os doadores de campanhas, corrigindo o problema de endogeneidade, que será apresentado na Tabela 2.9.

Tabela 2.7 – Estimação inicial do modelo Diferenças em Diferenças por doadores e não doadores de campanha. Variável Dependente: Valor de Contratos (*log*).

Variáveis	Coeficiente	Erro-padrão	p-valor
Doadores para prefeitos e vereadores			
Base (não doadores)			
Doador	0,137***	0,037	0,000
Tendência			
2009-2012	0,098***	0,010	0,000
2013-2016	0,342***	0,011	0,000
Observações		298.649	
Efeito Fixo		Sim	
Doadores para prefeitos e vereadores eleitos			
Base (não doadores)			
Doador	0,173***	0,050	0,001
Tendência			
2009-2012	0,110***	0,010	0,000
2013-2016	0,354***	0,011	0,000
Observações		292.409	
Efeito Fixo		Sim	

Fonte: Elaboração própria a partir dos dados do SAGRES, TCE-PB, e TSE.
Nota: Níveis de significância: *10%, **5% e ***1%.

A partir de agora, os resultados discutidos nesta seção são com base no modelo apresentado na seção 2.3, estratégia empírica, referentes ao efeito que a doação para os candidatos eleitos pode ter causado sobre o valor contratado das empresas e prestadores de serviços no quadriênio após o período eleitoral em que foi realizado a doação. Para averiguar este efeito, a abordagem não experimental do modelo de DD é utilizado para identificar a magnitude dessa conexão política no processo de compras e serviços públicos nos municípios paraibanos entre 2004 e 2016.

A Tabela 2.9 expõe os principais resultados do modelo DD para três painéis quadrienais distintos: Doadores para prefeitos e vereadores (Modelo 1); Doadores para prefeitos (Modelo 2), e; Doadores para vereadores (Modelo 3), controlando a heterogeneidade por candidatos eleitos e não eleitos para todos os modelos. Além da

separação por tipo de impacto quando se doa para diferentes grupos de candidatos, foi feito também o recorte para os tipos de doadores (empresas e prestadores de serviços, apenas empresas e apenas prestador de serviço) com o objetivo de mensurar o efeito para diferentes perfis de doações.

Tabela 2.9 – Estimação do modelo de Diferenças em Diferenças por grupo doado e por tipo de doador. Variável Dependente: Valor de Contratos (*log*).

	Empresas e Prestadores		*Empresas*		*Prestador de Serviço*	
	(1)	(2)	(3)	(4)	(5)	(6)
Doadores para prefeitos e vereadores						
Doador para eleitos	0,423***	0,427***	0,662**	0,704**	0,401***	0,406***
	(0,095)	(0,096)	(0,284)	(0,303)	(0,101)	(0,102)
2009-2012	−0,546***	−0,545***	−0,514**	−0,533**	−0,544***	−0,542***
	(0,069)	(0,069)	(0,255)	(0,262)	(0,072)	(0,072)
2013-2016	−0,537***	−0,533***	−0,776***	−0,761***	−0,499***	−0,496***
	(0,099)	(0,103)	(0,273)	(0,279)	(0,107)	(0,11)
Observações	13.839	13.839	1.084	1.084	12.755	12.755
Efeito Fixo	Sim	Sim	Sim	Sim	Sim	Sim
Covariadas	Não	Sim	Não	Sim	Não	Sim
Doadores para prefeitos						
Doador para eleitos	0,639***	0,640***	1,737**	1,719**	0,548***	0,554***
	(0,171)	(0,172)	(0,740)	(0,784)	(0,173)	(0,175)
2009-2012	−0,736***	−0,734***	−0,671	−0,687	−0,736***	−0,733***
	(0,126)	(0,127)	(0,520)	(0,534)	(0,129)	(0,129)
2013-2016	−0,620***	−0,612***	−0,037	−0,064	−0,684***	−0,671***
	(0,160)	(0,164)	(0,615)	(0,655)	(0,162)	(0,169)
Observações	5.459	5.459	613	613	4.846	4.846
Efeito Fixo	Sim	Sim	Sim	Sim	Sim	Sim
Covariadas	Não	Sim	Não	Sim	Não	Sim
Doadores para vereadores						
Doador para eleitos	0,285*	0,257	0,495	0,486	0,284	0,258
	(0,168)	(0,172)	(0,437)	(0,441)	(0,177)	(0,180)
2009-2012	−0,545***	−0,552***	−0,691*	−0,671*	−0,538***	−0,545***
	(0,102)	(0,103)	(0,397)	(0,392)	(0,105)	(0,106)
2013-2016	−0,488**	−0,547**	−1,168**	−1,407***	−0,390*	−0,433*
	(0,215)	(0,223)	(0,393)	(0,415)	(0,235)	(0,241)
Observações	7.900	7.900	384	384	7.516	7.516
Efeito Fixo	Sim	Sim	Sim	Sim	Sim	Sim
Covariadas	Não	Sim	Não	Sim	Não	Sim

Fonte: Elaboração própria a partir dos dados do SAGRES, TCE-PB, e TSE.
Nota: Níveis de significância: *10%, **5% e ***1%. Covariadas inseridas nos modelos: (1) Total de municípios que as empresas e prestadores de serviços fizeram doações em campanhas nos períodos eleitorais de 2004, 2008 e 2012; (2) *log* dos valores das doações.

As colunas (1), (3) e (5) de cada modelo apresentam as estimativas do efeito apenas com a variável de tratamento, já os resultados das colunas (2), (4) e (6) são controladas por duas variáveis explicativas no modelo, a saber: (1) Total de municípios que as empresas e prestadores de serviços fizeram doações em campanhas nos períodos eleitorais de 2004, 2008 e 2012; e (2) *log* dos valores das doações. Contudo, em nenhum dos modelos que as referidas covariadas foram inseridas, suas estimativas mostraram-se significantes ou com o sinal esperado.

No entanto, também não foi observado nenhuma alteração significativa na variável de impacto quando são comparados aos modelos sem as variáveis de controle, evidenciando uma possível robustez nos resultados dos modelos. Sendo importante notar também, que com a inserção das covariadas, foi observado um pequeno aumento no erro-padrão das estimações, (2), (4) e (6). Em que, de fato, a inclusão das variáveis de controle citadas não melhorou a precisão da estimação, uma vez que o erro-padrão aumentou em todos os modelos.

Os resultados encontrados no primeiro modelo mostram uma associação positiva e estatisticamente significativa de que as empresas e prestadores de serviços que doam para candidatos eleitos a prefeitos e vereadores geram um retorno, em média, de 42,3% nos valores contratados quando comparados aos valores contratados dos doadores para candidatos não eleitos. Por sua vez, com um recorte apenas para as empresas doadoras para candidatos eleitos, o benefício médio auferido é mais intenso do que o impacto observado nos valores contratados dos prestadores de serviços. De outra maneira, as pessoas jurídicas que evidenciaram conexões políticas através de doações para candidatos eleitos receberam um retorno de 66% nos valores contratados no serviço públicos, contra 40% do retorno recebido pelas pessoas físicas.

Adotando o mesmo recorte amostral, eleitos e não eleitos, Claessens, Feijen e Laeven (2008), Arvate, Barbosa e Fuzitani (2013), Boas, Hidalgo e Richardson (2014) encontraram em suas pesquisas que as empresas que doaram para candidatos que ganharam as eleições tem uma maior probabilidade de receber o financiamento sob a forma de benefícios nos valores contratados do que os candidatos que não ganharam as eleições. Em virtude de que o poder discricionário pertencente ao agente público eleito tem maior probabilidade de manipular os processos de contratações públicas, e assim, retribuir os favores recebidos nos períodos de campanhas.

Na busca por entender qual o principal canal de maior influência nas doações, estimou-se os modelos dois e três também expostos na Tabela 2.9. Pois, embora a relação positiva entre as doações e o retorno nos contratos públicos municipais ter sido mensurada no primeiro modelo, buscar evidências sobre quais as doações estariam guiando esse resultado motivou-se para estimar os modelos de forma separados por doadores para prefeitos eleitos e doadores para vereadores eleitos.

Dessa maneira, quem está induzindo o impacto de fato para obtenção de maiores benefícios nos contratos é o chefe do executivo. Apresentando também uma relação positiva e significante, o retorno para os doadores de candidatos a prefeitos eleitos é, em média, 63,9%, nos valores a receber dos serviços públicos. Impacto maior do que o observado para os valores recebidos para ambos os candidatos. No que tange as empresas que financiaram suas respectivas campanhas políticas, estas recebem uma taxa de retorno maior do que os prestadores de serviços. Por outro lado, o modelo três, o qual apresenta as estimações dos doadores para vereadores eleitos, foi significante apenas para empresas e prestadores de serviços, os quais obtém um retorno nos valores contratados de 28,5%.

Por fim, Coviello e Gagliarducci (2017) discute na sua pesquisa que a relação entre a influência do cargo do prefeito nas contratações públicas nos seus municípios gera perdas no setor público. Uma vez que o chefe do executivo direciona maiores benefícios para as empresas que apresentam conexões políticas com os mesmos, evidenciando não só um possível elo de ligação entre os agentes públicos e privados, mas também desenvolvendo relações corruptas, deteriorando tanto o sistema de compras, quanto a eficiência dos gastos públicos.

Ante ao exposto, os resultados encontrados corroboram a hipótese de que as empresas e os prestadores de serviços, doadores para candidatos eleitos nas eleições municipais de 2004, 2008 e 2012, apresentaram uma taxa de retorno positiva em termos dos contratos públicos com os municípios. Por outro lado, vão contra as evidências já publicadas para o Brasil, contudo a níveis estaduais e federais, pois afirmam que as doações destinadas aos candidatos para cargos legislativos resultam em maiores benefícios para as financiadoras (ARVATE; BARBOSA; FUZITANI, 2013; BOAS; HIDALGO; RICHARDSON, 2014; FONSECA, 2017).

2.6 Conclusões

Com base nos dados longitudinais de empresas e prestadores de serviços doadores de campanhas eleitorais nos municípios do estado da Paraíba, este ensaio verificou se tal conexão política entre os agentes privados e públicos geraria algum possível favorecimento nos contratos de compras e serviços públicos, em nível de governo municipal, nos períodos posteriores às vitorias eleitorais.

Utilizando de dados provenientes do TCE-PB e do TSE para o período entre 2004 e 2016, os resultados encontrados, por meio do modelo de DD com efeito fixo, validam a hipótese adotada na presente pesquisa, a qual afirma que o financiamento em campanhas eleitorais resulta em benefícios oriundos nos processos de contratos públicos pelas empresas e prestadores de serviços que doaram recursos financeiros

aos prefeitos e vereadores eleitos em 2004, 2008 e 2012.

Com controle para a heterogeneidade específica das empresas e dos prestadores de serviços e com recortes amostrais para doadores de candidatos eleitos e não eleitos na busca de corrigir o viés de autosseleção, os principais resultados apontam que as doações efetuadas em campanhas políticas por agentes privados geram um retorno, em média, de 42% nos valores contratados para os doadores de candidatos eleitos. Sendo essa taxa de retorno maior para as empresas do que para os prestadores de serviços.

A partir de tais sinalizações, pode-se inferir que as empresas mais ativas politicamente, ou seja, aquelas que oferecem, em média, maiores contribuições nas campanhas (particularmente para os candidatos eleitos) são mais propensas a receber maiores retornos políticos, principalmente no que tange os contratos de compras e serviços públicos nos quatro anos seguintes das eleições dos candidatos eleitos. Pois, apesar da quantidade de prestadores de serviços doadores ser maior, devido a frequência de candidatos, por outro lado, o valor médio doado é maior oriundo de empresas, devido ao montante total direcionado para os financiamentos.

De outro modo, os resultados da pesquisa indicam que os contratos públicos de compras e serviços a nível municipal é um importante canal onde as conexões políticas operam, sobretudo para as empresas e prestadores de serviços que contribuem para candidatos específicos, a saber: os que doam para ambos os candidatos, para prefeitos e para vereadores eleitos, mas, principalmente aos que doam para o poder executivo, nesse caso, candidatos a prefeitos eleitos.

Por fim, a partir de uma aplicação inovadora para o Brasil, com um recorte para o estado da Paraíba, as conclusões auferidas nesse estudo reconhecem que apesar dos resultados apresentarem apenas uma relação entre os valores das doações e o valores contratados, as evidências encontradas podem ser consideradas como fortes elementos para subsidiar as discussões sobre práticas de corrupção originando em vantagens indevidas nos processos de contratações públicas.

3 Efeitos de Políticas Afirmativas sobre Desempenho e Abandono

3.1 Introdução

O termo "ações afirmativas" foi utilizado pela primeira vez na década de 30, nos Estados Unidos (EUA), para se referir à políticas governamentais trabalhistas no combate às diferenças entre raças, com o objetivo principal de corrigir desigualdades acumuladas ao longo dos anos e ainda presentes na sociedade (MOEHLECKE, 2002). Na educação, tem sido utilizada como forma de permitir um maior acesso às instituições de ensino por parte de grupos historicamente discriminados e mantidos marginalmente no processo de formação de capital humano. Estas ações normalmente se manifestam no ensino superior, por meio da admissão via reserva de vagas, cotas ou sistema de bonificação. Ademais, tais políticas também visam a permanência nas instituições de ensino, o que ocorre por meio de auxílio financeiro.

Tem sido recorrente na literatura a necessidade de se compreender os reais efeitos das políticas de ações afirmativas, principalmente por gerarem discussões tão divergentes e com fortes implicações sociais e econômicas. Se por um lado existe o argumento de que tais medidas seriam necessárias para a reparação de injustiças históricas e sociais (KANE, 1998; DEE, 2004; FRANCIS; TANNURI-PIANTO, 2012), por outro, há críticos que condenam essa política com base no discurso de que não ocorre a redução da desigualdade, uma vez que os ingressantes no ensino superior por meio do sistema de cotas não seriam capazes de amenizar o *gap* de formação escolar, acarretando em uma incompatibilidade (*mismatch*) entre os grupos (SANDER, 2004; SOWELL, 2004; WINSTON; ZIMMERMAN, 2004). Adicionalmente, ao baixar os requisitos necessários para a admissão, a política incentiva o menor desempenho e desincentiva a competição, além ter como consequência em potencial o indesejado efeito de baixar a qualidade dos profissionais que serão formados.

Hickman (2009) elenca dois principais argumentos em favor da adoção de políticas afirmativas. Primeiramente, é notório a baixa participação de estudantes pertencentes às minorias em universidades de ponta, ao mesmo tempo em que se observa uma imensa participação destes em universidades de baixa qualidade. Em segundo lugar, tem-se que estes estudantes apresentam rendimento médio inferior em testes padronizados, em relação aos mais privilegiados, o que seria consequência das injustiças e desigualdades históricas que ainda não foram plenamente sanadas. Estes dois argumentos, conhecidos na literatura como *enrollment gap* (*gap* de matrículas) e

achievement gap (*gap* de resultados), justificariam a necessidade de se realizar políticas afirmativas. Desse modo, tais ações teriam o desejado efeito de incentivar as minorias a alcançarem os mais altos níveis educacionais, pois torna possível a aceitação nas melhores universidades do país.

No esteio dessa discussão, o sistema de cotas no ensino superior surge como uma forma de se reservar um determinado número de vagas, em relação à quantidade total, para grupos com características definidas e que se encontram em desvantagens sociais (MOEHLECKE, 2002). No Brasil, a desigualdade racial e social, aliada à demorada expansão na oferta de educação superior, resultou em um acesso restrito às Instituições de Ensino Superior (IES). As primeiras experiências de cotas em universidades públicas ocorreram na Universidade Estadual do Rio de Janeiro (UERJ) e na Universidade Estadual do Norte Fluminense (UENF), em 2001.

Contudo, o sistema de reserva de vagas passou por uma importante mudança em 2012. Até então voluntária, a implantação desse sistema por parte das universidades passou a ser obrigatória por meio da Lei nº 12.711, de 29 de agosto de 2012, a qual dispõe sobre o ingresso nas universidades federais e nas instituições federais de ensino (BRASIL, 2012). A Lei estabelece que 50% das vagas nas instituições públicas de ensino sejam destinadas para o sistema de cotas, adotando-se critérios raciais, socioeconômicos e de conclusão do ensino médio na rede pública.

Em um contexto onde ainda não existia a Lei supracitada, o presente estudo busca contribuir com esta discussão e verificar, a partir de dados da Universidade Federal da Paraíba (UFPB), os efeitos diretos e indiretos de políticas afirmativas (raça e renda) de reserva de vagas no ensino superior sobre indicadores educacionais dos discentes. Os efeitos diretos foram captados ao se comparar os indicadores de impacto entre os cotista e não cotista elegíveis para o ingresso na IES em 2011 (período de início da reserva de vagas na UFPB). Não obstante, objetivou-se captar também os efeitos indiretos ao comparar os indicadores entre os cotistas e os não cotistas em dois períodos: um período em que não existia a política de reservas de vagas (2010) e um período em que existia (2011).

Vale destacar que a política de cotas não assegura que os discentes cotistas melhorarão o seu desempenho, mas permite um maior acesso dos grupos menos favorecidos às IES. Como o debate sobre a aplicação desta política em admissões universitárias tem se mostrado contraditório do ponto de vista de suas possíveis externalidades, principalmente no que se refere à questão do *mismatch*, a hipótese a ser testada neste ensaio é a de que a política de reserva de vagas no ensino superior, por meio do sistema de cotas, pode afetar os incentivos individuais de forma direta e indireta, gerando impactos sobre resultados educacionais.

Entender como os grupos de cotistas e não cotistas se comportam durante o

ensino superior é de suma importância para os gestores de políticas públicas educacionais, não só pelo aumento da representatividade, mas também em relação aos indicadores educacionais. No tocante aos estudos empíricos acerca de políticas afirmativas no Brasil, ainda são escassas as pesquisas que trabalharam essas ações direcionadas para o sistema de reserva de vagas no ensino superior levando em conta aspectos locais, socioeconômicos e de habilidades cognitivas sobre o desempenho (rendimento) escolar dos discentes.

Para preencher essa lacuna na literatura nacional, este ensaio pretende avançar nas discussões dos efeitos das políticas de reservas de vagas no ensino superior. Ressalta-se que uma análise completa da ação compensatória requer, além de se observar o diferencial dos rendimentos acadêmicos decorrente do aumento da diversificação de grupos dentro da instituição, avaliar também o comportamento dos discentes cotistas ao longo do curso escolhido. Tal procedimento consiste em analisar a sobrevivência dos mesmos na IES por meio das taxas de abandono.

Para o cumprimento dos objetivos elencados, em um primeiro momento foram adotadas técnicas de pareamento não-experimental, o *Propensity Score Matching* (PSM), o *Mahalanobis Distance Matching* (MDM) e o *Classification Tree Analysis* (CTA), para avaliar os efeitos da intervenção sobre o desempenho dos discentes, captado pelo Coeficiente de Rendimento Acadêmico (CRA) relativo. Em seguida, de caráter inédito, estimou-se modelos de duração de risco proporcional de Cox, ponderado pelo PSM, para avaliar o efeito do aluno ser cotista sobre a probabilidade de sobrevivência na UFPB. As informações utilizadas nas análises são oriundas dos microdados disponibilizados pela Superintendência de Tecnologia da Informação (STI) da UFPB, para os anos de entrada de 2010 e 2011.

Além desta introdução, a pesquisa é composta por mais seis seções. A Seção 3.2 apresenta os modelos teóricos sobre as políticas afirmativas e uma breve discussão da literatura empírica. A Seção 3.3, por sua vez, expõe a descrição das características do sistema de cotas na UFPB. As Seções 3.4 e 3.5 detalham, respectivamente, a estratégia empírica e a base de dados da pesquisa. Por fim, as Seções 3.6 e 3.7 apresentam os principais resultados e as conclusões do ensaio.

3.2 Revisão da Literatura

3.2.1 Abordagens Teóricas

Os modelos teóricos utilizados para quantificar a relação entre políticas afirmativas no ensino superior e indicadores educacionais baseiam-se nos estudos discutidos por Hickman (2009) e Zylberstajn e Souza (2010). Inicialmente, Hickman (2009) desen-

volveu um modelo teórico para analisar as implicações decorrentes da adoção de ações afirmativas nas universidades americanas. Conforme ressaltado pelo autor, é possível elencar dois principais argumentos em favor da adoção dessas políticas. Primeiramente, é notório a baixa participação de estudantes pertencentes às minorias em universidades de ponta, ao mesmo tempo em que se observa uma imensa participação destes em universidades de baixa qualidade. Em segundo lugar, tem-se que estes estudantes apresentam desempenho médio inferior em testes padronizados em relação aos mais privilegiados, o que seria consequência das injustiças e desigualdades históricas que ainda não foram plenamente sanadas. Estes dois argumentos, conhecidos na literatura como *enrollment gap* (*gap* de matrículas) e *achievement gap* (*gap* de resultados), justificariam a necessidade de se realizar políticas afirmativas.

Os defensores desse tipo de intervenção afirmam que ela incentiva as minorias a alcançarem os mais altos níveis educacionais, pois torna possível a aceitação nas melhores universidades do país. Já os críticos argumentam que, ao baixar os requisitos necessários para a admissão, a política incentiva o menor esforço e desincentiva a competição. Além disso, ela pode ter como consequência o indesejado efeito de baixar a qualidade dos profissionais que serão formados. Levando esta discussão em consideração, Hickman (2009) considera a implementação de três tipos de ações afirmativas no seu modelo, a saber: preferências de admissão (*admission preferences*), política de cotas e política de admissão neutra em relação à raça (*race-neutral admissions* ou *color-blind rule*). Os critérios utilizados para avaliar seus efeitos são: a performance acadêmica (*achievement gap*), o diferencial de resultado entre raças (*racial achievement gap*) e o diferencial de matrículas nas universidades (*enrollment gap*).

Vale ressaltar que o modelo assume que o formulador de política valoriza o desempenho acadêmico, e busca reduzir o *racial achievement gap* e o *enrollment gap*. Assim, os agentes tomadores de decisão no modelo são representados por um conjunto $\mathcal{K} = \{1, \ldots, K\}$ de estudantes competindo para serem aceitos na universidade. Cada estudante possui um tipo de custo de estudo $\theta \in [\underline{\theta}, \overline{\theta}]$ e escolhe um nível de nota $s \in \mathbb{R}_+$. A escolha por determinando nível de nota incorre em um custo de utilidade $\mathcal{C} = \{s; \theta\}$, sendo que as seguintes condições devem ser satisfeitas: $\frac{\partial C}{\partial s} > 0$, $\frac{\partial C}{\partial \theta} > 0$, $\frac{\partial^2 C}{\partial s^2} \geq 0$, $\frac{\partial^2 C}{\partial s \partial \theta} \geq 0$. A estrutura de custos pode ser pensada como resultante de um *tradeoff* subjacente entre trabalho e lazer, enquanto que os tipos de custos privados englobam fatores externos que afetam o rendimento acadêmico dos alunos, tais como as condições de moradia, riqueza, a qualidade da escola, o acesso a cuidados de saúde, etc.

Existe um conjunto de prêmios $\mathbf{P}_{\mathcal{K},K} = \{p_k\}_{k=1}^{K}$, em que p_k denota a utilidade de se consumir o k-ésimo prêmio. Estes prêmios representam as vagas na universidade pela qual os alunos competem. Há vagas para todos os estudantes que desejam cursar a

faculdade, mas nem todas as vagas são igualmente desejáveis: $p_k \neq p_j, k \neq j$. Os valores dos prêmios são modelados como amostras aleatórias independentes dentro de um intervalo $\mathcal{P} = [\underline{p}, \overline{p}]$, de acordo com uma distribuição conhecida $F_P(p)$, com densidade $f_P(p)$ estritamente positiva em $\mathcal{P}$. Hickman (2009) ressalta que não é necessário para o modelo que todos os estudantes atribuam o mesmo valor a uma vaga em uma dada universidade. Porém, é importante assumir que os estudantes ranqueiam os valores dos prêmios da mesma maneira, de modo que eles possuam preferências similares acerca dos atributos das universidades.

Cada estudante pertence a um dos dois grupos observáveis: o grupo das minorias, $\mathcal{M} = \{1, 2, \ldots, M\}$, e o grupo das não minorias, $\mathcal{N} = \{1, 2, \ldots, N\}$. Tem-se que $M + N = K$. Cada competidor ver os custos privados de seu oponente como variáveis aleatórias independentes, seguindo distribuições de custos conhecidas $F_{\mathcal{M}}(\theta)$ e $F_{\mathcal{N}}(\theta)$, com densidades estritamente positivas $f_{\mathcal{M}}(\theta)$ e $f_{\mathcal{N}}(\theta)$. Assume-se também que o número de competidores do grupo das minorias é modelado por um número $\mu \in (0, 1)$. Logo, a natureza aloca cada estudante ao grupo $\mathcal{M}$ com probabilidade μ, depois do qual um custo privado é obtido a partir da distribuição apropriada [4]. Uma vez que uma das três regras possíveis (cotas, preferências de admissão e neutralidade de raça) é especificada, então o problema de decisão do agente define um jogo estratégico Bayesiano. Sob os *payoffs* induzidos por uma regra de admissão particular, os estudantes escolhem suas notas de forma ótima, levando em conta seus próprios custos privados e o comportamento ótimo de seus oponentes. De acordo com Hickman (2009), este modelo é estrategicamente equivalente a um tipo de jogo conhecido na literatura como *all-pay auction*[5].

Sob uma regra do tipo *color-blind*, o comitê recompensa as realizações mapeando os quantis da distribuição populacional de notas nos quantis correspondentes da distribuição de prêmios. Já sob uma política de cotas, o comitê mapeia os quantis das distribuições de notas de cada grupo nos quantis correspondentes dos prêmios. Finalmente, sob uma política de preferências de admissão, o comitê recompensa as minorias mapeando os quantis da distribuição de notas desse grupo, penalizando as notas dos alunos que não pertencem às minorias nos quantis de prêmio correspondentes. Para as não minorias, o comitê mapeia os quantis da distribuição de notas desse grupo, subsidiando as notas das minorias nos quantis de prêmio correspondente.

Conforme dito, o objetivo final das políticas de ações afirmativas é, no equilíbrio, induzir o bom desempenho acadêmico, minimizar o *racial achievement gap* e reduzir o

[4] A lógica por trás desse pressuposto de assimetria nas chances de participar do grupo das minorias é que, na média, estes estudantes devem gastar maiores esforços pessoais para superar dificuldades, tais como a pobreza e a educação de baixa qualidade.

[5] Um tipo de leilão em que cada licitante tem de pagar o valor do lance independentemente de ter vencido ou não.

enrollment gap. Seja $G_{\mathcal{M}}$ e $G_{\mathcal{N}}$ a distribuição de notas de cada grupo e G a distribuição de notas da população, então o *achievement gap* pode ser representado por uma função $\mathcal{A} : [0,1] \rightarrow \mathbb{R}$, definida por $\mathcal{A}(q) \equiv G_{\mathcal{N}}^{-1}(q) - G_{\mathcal{M}}^{-1}(q)$. Ou seja, $\mathcal{A}$ caracteriza a diferença de resultados entre estudantes de minorias e não minorias em cada quantil das distribuições de notas. Em relação ao *enrollment gap*, seja $F_{P_i}(p)$, $i = \mathcal{M}, \mathcal{N}$, a distribuição de prêmios recebida por cada grupo no equilíbrio. Então o *enrollment gap* é uma função $\mathcal{E} : [0,1] \rightarrow \mathbb{R}$, definida por $\mathcal{E}(q) \equiv F_{P_N}^{-1}(q) - F_{P_M}^{-1}(q)$. Por fim, o perfil geral de resultados acadêmicos é representado pela distribuição populacional de notas $G(s) = \mu G_{\mathcal{M}}(s) + (1-\mu)G_{\mathcal{N}}(s)$. O formulador de políticas busca implementar ações que maximize G e minimize $\mathcal{A}$ e $\mathcal{E}$.

Por sua vez, Zylberstajn e Souza (2010) estudaram de forma teórica e empírica as consequências que a introdução de uma política de cotas pode ter no esforço dos estudantes em idade escolar, uma vez que estes estão se preparando para realizar o exame de admissão para o ensino superior nas universidades. Conforme ressaltado pelos autores, ao alterar a estrutura de incentivos, as políticas afirmativas podem ter efeitos na eficiência econômica da sociedade. Sendo assim, é importante entender de forma detalhada seus impactos sobre o acúmulo de capital humano por parte dos alunos beneficiados e não beneficiados pela ação.

O modelo teórico elaborado pelos autores parte de uma economia simples em que todos os indivíduos têm as mesmas características natas, à exceção da cor da pele, que pode ser negra (N) ou branca (B), e da habilidade (a), que pode ser alta (a_H) ou baixa (a_L). A proporção de indivíduos negros na população total é dada por $\theta_N \in [0,1]$, enquanto que a proporção de indivíduos brancos é $(1-\theta_N)$. Por sua vez, a proporção de indivíduos da cor $i \in [N,B]$ de baixa habilidade é $\alpha_i^L \in [0,1]$, e a de alta habilidade $(1-\alpha_i^L)$. O modelo assume que a cor da pele não tem nenhum impacto na produtividade ou em qualquer outra característica individual, sendo observável por qualquer indivíduo.

Todos os indivíduos do modelo ofertam trabalho, recebendo salários com base na sua qualificação. Assim, os salários são expressos por w_H e w_L. Supõe-se que todos os indivíduos conseguem emprego, independentemente de sua qualificação, e que não há discriminação de qualquer forma. Para poderem se qualificar, os indivíduos de habilidade a precisam realizar um esforço e para conseguir acesso a uma instituição qualificadora. Dado o esforço, a probabilidade de acesso é $\pi(e,a)$, com $0 \leq \pi(e,a) \leq 1$. Adicionalmente, supõe-se que $\pi(0,a) = 0$ e que $lim_{e\rightarrow\infty}\pi(e,a) = 1$. O esforço tem um custo associado, $c(e,a)$, que inclui custos diretos (monetários) e custos indiretos (como custos psicológicos, por exemplo).

Uma das hipóteses centrais do modelo é que, para um mesmo nível de esforço $\tilde{e}$, uma habilidade maior implica em um custo total e um custo marginal menor ou

igual ao enfrentado por indivíduos com habilidade baixa. Ou seja, $c(\tilde{e}, a_H) \leq c(\tilde{e}, a_L)$ e $\frac{\partial c(e,a_H)}{\partial e} \leq \frac{\partial c(e,a_L)}{\partial e}$ $\forall$ $\tilde{e} \in [0, \infty[$. Outra importante hipótese do modelo é que, para um mesmo nível de esforço, indivíduos de alta habilidade têm probabilidade de acesso a uma instituição qualificadora maior ou igual à probabilidade dos indivíduos de baixa habilidade, isto é, $\pi(\tilde{e}, a_H) \geq \pi(\tilde{e}, a_L)$ $\forall$ $\tilde{e} \in [0, \infty]$.

Em suma, a única variável de escolha dos indivíduos é o esforço. Portanto, o problema de cada agente é expresso da seguinte forma: $\max_e \{w_H.\pi(e,a) + w_L.(1 - \pi(e,a)) - c(e,a)\}$. Para obter a solução do problema, deve-se encontrar as condições de primeira e segunda ordem para o máximo. Estabelecidas as condições iniciais do modelo, parte-se agora para a introdução da política de cotas. Esta tem como objetivo facilitar o acesso dos negros à universidade pública, definida como uma instituição qualificadora. Seja $0 < \varphi_N < 1$ a proporção de vagas reservadas para os alunos do tipo N. Agora, a probabilidade de acesso passa a depender da cor da pele, sendo dada por: $\pi_i(e, a, \varphi_N)$.

O modelo, então, assume os seguintes efeitos da adoção da política de cotas, dado um nível de esforço $\tilde{e} \in [0, \infty[$ e um nível de habilidade $\tilde{a} \in [a_L, a_B]$: i) $\pi_N(\tilde{e}, \tilde{a}, \varphi_N) \geq \pi(\tilde{e}, \tilde{a}, 0)$; ii) $\pi_B(\tilde{e}, \tilde{a}, \varphi_N) \leq \pi(\tilde{e}, \tilde{a}, 0)$; e iii) $c(\tilde{e}, \tilde{a}, \varphi_N) = c(\tilde{e}, \tilde{a}, 0)$. Dessa forma, tem-se que a probabilidade de acesso, com cotas, é maior (menor) para alunos negros (brancos), ao passo em que as cotas não alteram o custo do esforço. A discussão considera que uma política de cotas pode tanto incentivar quanto desincentivar o esforço dos alunos em idade escolar. Os alunos beneficiados (não beneficiados) pelas cotas que enfrentam custos elevados tendem a aumentar (diminuir) o esforço, enquanto que aqueles que enfrentam custos mais baixos tendem a diminuí-lo (aumentá-lo).

Outra questão abordada pelo modelo é o diferencial de salários, que desempenha um papel importante na determinação do esforço ótimo. De forma geral, o modelo considera que quanto maior o diferencial de salários $(w_H - w_L)$, maior o esforço realizado por qualquer tipo de indivíduo (desde que não seja nulo). Para tornar o modelo mais geral, os autores também incorporaram a existência de um Esforço Mínimo Necessário (EMN) para que seja possível tentar o acesso à qualificação. Uma vez que o modelo trata do ingresso no Ensino Superior, então os alunos têm de necessariamente ter, no mínimo, o diploma do Ensino Médio.

De acordo com os autores, quando existe diferença de custos é possível que, na ausência de cotas, os candidatos com maior custo realizem esforço nulo ou esforço sobreótimo. A adoção das cotas poderia, portanto, fazer com que aqueles alunos que inicialmente desistiriam de realizar qualquer esforço passassem a optar por tentar obter uma vaga ou, pelo menos, que o esforço realizado alcançasse o nível ótimo e, assim, a eficiência do sistema fosse aumentada. Em outras palavras, com as cotas é possível que os alunos negros que antes realizavam esforço nulo passem a fazer o

esforço mínimo. Por outro lado, para os alunos brancos as cotas podem ter o efeito contrário. Por fim, é possível elencar outros estudos que abordaram de forma teórica as implicações de formas variadas de ações afirmativas. Nessa linha, merecem destaque os trabalhos de Chan e Eyster (2003), Su (2005) e Bishop (2006).

3.2.2 Efeitos de Políticas Afirmativas - Abordagens Empíricas

É possível encontrar na literatura vários estudos que se propuseram a avaliar os efeitos de ações afirmativas sobre indicadores educacionais e de mercado de trabalho. Nesse contexto, Ayres e Brooks (2005) analisam os efeitos de políticas afirmativas sobre o número de advogados negros formados pelas universidades americanas. A motivação do estudo parte da constatação de que o desempenho desses estudantes beneficiados pela ação é consideravelmente inferior ao dos estudantes brancos. Desse modo, surge a questão de se a inclusão desses alunos em universidades incompatíveis com o seu perfil não estaria, na verdade, ocasionando uma redução da taxa de concluintes negros nas faculdades de Direito. Utilizando modelos de regressão, os autores identificaram que a eliminação da ação afirmativa reduziria o número de advogados negros. Entretanto, também se observou que estudantes negros possuem chances quase 50% menores de concluir o curso, em relação aos alunos brancos. Assim, os autores concluem que a faculdade de Direito pode ser considerada de alto risco para os discentes negros beneficiados pela política.

Sob à luz das proibições de políticas de ações afirmativas por raça no ensino superior nos EUA, Dickson (2006) investigou como o fim da política afirmativa que adotava o critério de raça na seleção de universidades do Texas impactou no percentual de candidatos pertencentes às minorias étnicas. Além disso, a autora também analisou os efeitos do programa de bolsas de estudos oferecido pela Universidade do Texas, o *Longhorn Opportunity Scholarship Program*, sobre o percentual de estudantes do ensino médio que realizaram o teste de admissão para a universidade. Empregando o estimador de efeitos fixos em um painel de escolas acompanhadas no período 1994-2001, os resultados apontaram que o fim da ação afirmativa reduziu o percentual de latino-americanos e negros que fizeram a seleção para a faculdade, não tendo efeito sobre o percentual de brancos. Em relação ao programa de bolsas de estudo, este se mostrou bem-sucedido em ampliar a participação das minorias na realização do teste de admissão.

Seguindo a mesma linha, Cortes (2010) busca fornecer novas evidências sobre os efeitos de políticas alternativas de admissão na persistência e na conclusão dos grupos de estudantes de minorias nas universidades do Texas. Após a política de reserva de vagas no ensino superior por raça ser banida, o estado do Texas implantou o *Texas Top 10% Plan*. Este plano garante que os 10% dos melhores alunos das suas turmas

do ensino médio garantem automaticamente a admissão para qualquer universidade pública do estado, incluindo as mais seletivas. Após aplicar o método de *difference-in-differences* na análise nos anos 90, os autores discutem a "*mismatch hypothesis*". Pois, afirmam que os alunos mais afetados por esta mudança na política de admissão foram os estudantes classificados a partir do segundo decil, uma vez que os do primeiro estavam aptos para serem admitidos, aumentando mais ainda a lacuna racial entre os grupos de estudantes. Ou seja, a mudança da ação afirmativa afetou negativamente tanto a taxa de retenção quanto as taxas de graduação universitária do grupo de alunos das minorias.

Bertrand, Hanna e Mullainathan (2010) estudaram os efeitos de uma política de ação afirmativa existente na Índia. Tal política consiste em reservar mais de 50% das vagas em universidades públicas a estudantes oriundos de castas inferiores, ou seja, historicamente marginalizados e discriminados. O objetivo dos autores é investigar a efetividade desta ação em termos de resultados no mercado de trabalho e redistribuição de oportunidades, bem como comparar os ganhos econômicos obtidos pelos beneficiados com a perda potencial sofrida por aqueles excluídos pela política. A amostra selecionada considera apenas indivíduos que se candidataram a uma vaga em faculdades de engenharia, no ano de 1996, em um estado indiano. Foram coletados dados censitários sobre todos estes candidatos que realizaram a seleção em 1996, juntamente com informações domiciliares referentes a estes mesmos indivíduos 8 a 10 anos após a seleção. Em termos de estratégia empírica, os autores estimaram regressões por Mínimos Quadrados Ordinários (MQO).

Os resultados encontrados sugerem que a ação é bem-sucedida em redistribuir recursos educacionais para os menos favorecidos economicamente. Os autores ressaltam ainda que, apesar dos beneficiados pela política apresentarem renda superior à média das minorias de seu estado, ainda sim eles são relativamente mais pobres do que os estudantes que eles substituíram. Um efeito indesejado observado foi a redução da participação das mulheres nas faculdades de engenharia. Desse modo, fica evidente que a política não gerou benefícios a todas as minorias. Por fim, em relação aos resultados no mercado de trabalho, observou-se que os beneficiados obtiveram substantivos ganhos em termos de renda. Todavia, estes ganhos foram consideravelmente menores que a perda sofrida pelos indivíduos de castas superiores prejudicados pela intervenção.

No que concerne a pesquisa de Li e Weisman (2011), estes discutem sobre as implicações de eliminar políticas de ação afirmativa no processo de admissões universitárias nos EUA. A partir da construção de um modelo teórico buscam mostrar que é possível que uma mudança na política de admissão a qual reduza as preferências por raça leve a um corpo discente "menos capaz". Contrariando os críticos

das políticas afirmativas, pois estes afirmam que a eliminação das preferências raciais nas admissões universitárias levaria a um corpo discente "mais capaz". Levando em consideração alguns pressupostos e uma análise composta por três classes de admissão nas universidades – mérito, raça e legado (alunos atletas) – os autores afirmam que é possível que a eliminação das preferências raciais em combinação com a retenção de preferências legadas poderia levar a um corpo discente "menos capaz". Por fim, o estudo ainda conclui que se a real intenção da mudança na política é produzir um "corpo estudantil capacitado", a melhor ação talvez deva ser dada à eliminação de todas as preferências, e não apenas por raça, ou seja, um conjunto de admissões sem preferências. Além destes estudos citados anteriormente, estudos como Chan e Eyster (2003) e Hinrichs (2014) também analisam as consequências da eliminação de tais ações afirmativas na diversificação do ensino superior.

Jung, Sung e Kim (2012) estimaram os impactos econômicos de ações afirmativas na Coreia sobre o nível de emprego das mulheres e a performance das empresas, juntamente com o seu potencial efeito sobre o crescimento econômico. Conforme ressaltado pelas autoras, este tipo de política entrou em vigor na Coreia pela primeira vez em 2006, como uma medida para expandir o emprego das mulheres e remediar práticas discriminatórias profundamente enraizadas. Por meio dos dados da *Workplace Panel Survey* (WPS), referentes aos anos de 2005, 2007 e 2009, as autoras empregaram o método de *difference-in-differences* para aferir o efeito da intervenção. Os principais resultados encontrados apontam que a política ainda não elevou significativamente a parcela de mulheres no emprego total ou em cargos gerenciais. Além disso, também não se encontrou qualquer impacto significativo sobre o desempenho das empresas. Em relação aos potenciais efeitos macroeconômicos, estimados por intermédio de uma versão aumentada do modelo de Solow, identificou-se que a ação afirmativa pode acelerar o crescimento econômico se ela for eficaz em reduzir a diferença salarial de gêneros.

Utilizando dados administrativos institucionais de quatro universidades israelenses que implementaram, voluntariamente, o programa de ação afirmativa entre 2001 e 2008, Alon e Malamud (2014) analisaram se o objetivo da política baseada em classes – cega, daltônica e estudantes de meios desfavorecidos – na admissão nas universidades estava sendo alcançado. Ou seja, se o corpo discente das universidades de elite estava sendo diversificado em consequência da mobilidade social e econômica da população mais necessitada. Dessa maneira, para examinar o efeito da elegibilidade dos discentes no programa de reserva de vagas sob os resultados de admissão e matricula, bem como no desempenho acadêmico, foi estimado o modelo de regressão descontinua (RDD). Então, comparando candidatos e alunos com níveis semelhantes de desvantagens econômicas e outras características, mas probabilidades muito diferentes de receber a ação afirmativa, os autores concluíram que: (i) os estudantes elegíveis para a ação, após

a matricula nas universidades, não estão ficando para trás academicamente, mesmo nos cursos mais seletivos; (ii) o programa também levou a altas taxas de admissões de alunos elegíveis nos cursos mais seletivos.

Por sua vez, em um estudo aplicado para a Índia, Bagde, Epple e Taylor (2016) estudam o efeito de um programa de admissão que fixa cotas percentuais em mais de 200 faculdades de engenharias, para castas desfavoráveis e para mulheres. A discussão inicial parte da ideia de que a ação afirmativa possa, ao invés de beneficiar, prejudicar os beneficiários do programa ao longo dos seus estudos, evidenciando que estão mal preparados. O argumento é que os alunos com preparação deficitária podem ser sair mal nas faculdades mais seletivas, sendo assim, estes alunos se encontrariam em melhor situação nas instituições onde sua preparação acadêmica se encaixa melhor, evitando a "*mismatch hypothesis*". Com um painel formado por 131.290 alunos classificados para ingressarem nas 215 faculdades, os autores encontram evidencias de que a política funciona basicamente como pretendida, aumenta a frequência entre os alunos visados, especialmente para aqueles das castas mais desfavorecidas. Em princípio, políticas de ação afirmativa podem prejudicar os beneficiários, colocando-os em situações acadêmicas para quais não estão preparados, contudo, não foram encontradas evidências para o *mismatch*.

Utilizando dados chilenos para o processo de admissão nas universidades em 2009, Grau (2018) inova ao avaliar empiricamente os efeitos das políticas de admissão sob o desempenho acadêmico dos estudantes do ensino médio. O ponto de partida da pesquisa é o modelo teórico construído baseado em um torneio de ordem de classificação com habilidades heterogêneas, onde os estudantes do ensino médio decidem seu nível de esforço e se querem ou não fazer o teste de admissão, considerando como essas decisões afetam suas futuras chances de admissão em universidades. A partir da simulação de duas políticas afirmativas – sistema de cotas e aumentar o peso das notas dos estudantes do ensino médio por distribuição de grupos socioeconômicos – foi estimado um modelo em duas etapas, a saber: (1) Função de produção por mínimos quadrados em dois estágios; (2) Máxima verossimilhança. O estudo mostra a importância da ação, pois a política de admissão na universidade pode aumentar o volume de esforço acadêmico exercido pelos estudantes do ensino médio. Conclui ainda que embora o aumento do peso da nota no ensino médio seja mais eficaz, o sistema de cotas é mais eficiente na alocação dos melhores alunos para as melhores universidades.

No Brasil, a intensa discussão sobre as políticas de cotas e as características particulares do contexto brasileiro proporcionam aos pesquisadores importantes oportunidades de examinarem as questões políticas e acadêmicas sobre as ações afirmativas de reserva de vagas no ensino superior. Dessa maneira, no que tange o trabalho de

Francis e Tannuri-Pianto (2012), estes estudaram a experiência da Universidade de Brasília com a implantação do sistema de reserva de vagas no ensino superior a partir da política de cotas raciais no ano de 2004. Esta reserva 20% do total de vagas disponíveis para os alunos que se autodeclaram negros. Para determinar até que ponto a ação afirmativa de raça promoveria a equidade por meio de uma redistribuição no processo de admissão, os autores, a partir de um modelo de primeira diferença entre 2003 e 2005, compararam indivíduos que não foram admitidos, mas teriam sido se o sistema de cotas não tivesse existido com aqueles que foram admitidos. A evidências encontradas sugerem que a raça, o *status* socioeconômico e o gênero continuam a impor barreiras substanciais à frequência e à realização de cursos universitários, e que as cotas baseadas na raça ajudam a promover a equidade até certo ponto. Pois, apesar da implementação de cotas raciais, a grande maioria dos brasileiros ainda tem poucas chances de frequentar a faculdade.

Já Pereira, Bittencourt e Silva Junior (2013) buscaram avaliar os impactos das cotas sociais e raciais no sistema de ensino superior brasileiro sobre as notas do Exame Nacional de Desempenho dos Estudantes (ENADE). Considerando somente os alunos de universidades federais, estaduais e municipais, totalizando 74.080 observações para o ano de 2008, a variável de interesse selecionada foi a nota obtida na prova de conhecimentos específicos. O arcabouço teórico utilizado foi Modelo de Reserva de Vagas de Su (2005) e o Modelo de Decisão de Esforço de Bishop (2006), implicitamente, ambos os modelos convergem para o fato de que um grupo de pessoas excluídas, que não entrariam em um curso superior sem a reserva de vagas, são induzidas a um maior esforço graças à política. Para aferir o efeito causal da política de cotas, o método adotado foi o *difference-in-differences* combinado com o Pareamento por Escore de Propensão (PSM), enquanto o grupo de tratamento foi composto pelos estudantes cotistas e os não cotistas fizeram parte do grupo de controle. Os resultados encontrados apontaram efeitos negativos da política de cotas sobre as notas no ENADE de 2008 para os cursos de Pedagogia, História e Física, e efeitos positivos sobre a nota do curso de Agronomia. Para os demais cursos, não foram encontrados efeitos significativos.

Mais recentemente, Mendes Junior, Souza e Waltenberg (2016) estudaram os efeitos do sistema de cotas na Universidade do Estado do Rio de Janeiro (UERJ), a primeira universidade brasileira a adotar a referida ação afirmativa. A política consiste em reservar 20% das vagas para indivíduos das raças preta e indígena, 20% para estudantes de escolas públicas e 5% para pessoas com deficiência. Adicionalmente, para poder se candidatar a uma dessas vagas é preciso ainda satisfazer um critério de desvantagem econômica, que é definido a cada ano pela comissão de elaboração do vestibular, baseado na renda domiciliar *per capita*. A partir dos dados referentes ao vestibular da UERJ de 2010, levando em conta informações socioeconômicas detalhadas de todos os candidatos que prestaram o exame de seleção, a estratégia empírica adotada

consistiu em estimar modelos de regressão por MQO e modelos de escolha binária *logit*. Os resultados dos modelos indicaram que as principais diferenças de desempenho no exame de seleção estão associadas à fatores socioeconômicos dos candidatos. Além disso, o estudo também identificou que as notas mínimas de admissão são consideravelmente menores entre os cotistas e que poucos desses alunos conseguiriam obter uma vaga caso esse sistema de reserva de vagas não existisse.

Estevan, Gall e Morin (2016) examinaram os efeitos de uma ação afirmativa implementada na Universidade Estadual de Campinas (UNICAMP). Tal ação, implementada em 2005, consiste em conceder um bônus de pontos no exame de admissão da universidade (vestibular) aos alunos que cursaram o ensino médio exclusivamente em escolas públicas. Estes alunos recebem 30 pontos de bonificação em sua nota, mais um adicional de 10 pontos se eles se autodeclararem da raça preta, parda ou indígena. O objetivo da pesquisa foi verificar como essa política de bônus sociorracial impacta a composição dos estudantes selecionados, a decisão do aluno de participar do vestibular e a sua escolha de esforço de preparação. Usando uma base de dados administrativa da UNICAMP, entre 2003 e 2008, e aplicando o modelo de *difference-in-differences* para aferir o impacto da intervenção, os resultados das estimações indicaram que a política é bem-sucedida em aumentar a probabilidade de admissão de estudantes oriundos de escolas públicas. Logo, observa-se uma mudança de composição nos candidatos selecionados com o aumento da participação daqueles em pior situação socioeconômica. Ademais, não foram encontrados efeitos relevantes em relação às decisões de esforço e participação no exame. Ou seja, a política aparentemente não tem efeito significativo de aumentar a proporção de candidatos procedentes de escolas públicas, ao mesmo tempo em que não se verifica mudanças nas decisões de esforço.

3.3 Características do Sistema de Cotas na UFPB

O Conselho Superior de Ensino, Pesquisa e Extensão (Consepe) da Universidade Federal da Paraíba aprovou, em março de 2010, o projeto com as propostas de ações afirmativas correspondentes à reserva de vagas. Então, por meio da resolução nº 09/2010 foi instituída a modalidade de ingresso por reserva de vagas para o acesso aos cursos de graduação da UFPB. Dessa maneira, os processos seletivos a partir de então já ocorreriam com a previsão de reserva de vagas tanto com recorte social quanto o étnico-racial.

Posto isto, através do Edital nº 40/2010, do Processo Seletivo Seriado 2011 (PSS 2011), regulamentado pela resolução nº 027/2009, foi publicado o primeiro edital com reserva de vagas de políticas afirmativas na UFPB. A ocupação das vagas oferecidas para os cursos de graduação em 2011 ocorreria por meio das seguintes formas: (i) por

concorrência geral; (ii) por reserva de vagas. Do total de vagas ofertadas no PSS 2011, 25% são reservadas aos candidatos que cursaram todo o ensino médio e, pelo menos, três séries do ensino fundamental em escolas da rede pública de ensino.

Dentro desses 25% foram redistribuídos aos demais recortes referentes às cotas. As vagas foram reservadas para pessoas com alguma deficiência, como também para candidatos negros e indígenas. No que se refere a determinação da proporção das vagas destinadas para o critério de raça (negros, pardos e índios) foi utilizado como parâmetro a participação desses grupos na população do estado da Paraíba, de acordo com o Censo Demográfico do ano 2000, publicado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), e estabelecido no Edital nº 40/2010.

Assim, dentro dos 25% reservados para os candidatos egressos do ensino público, 56% foram reservados para negros e pardos, 0,29% para indígenas e, 5% para portadores de necessidades especiais. No ato da inscrição do PSS 2011, o estudante tinha que informar a opção pelo tipo de ocupação de vaga que iria concorrer, se fosse por reserva de vagas, teria as seguintes opções: egresso de escola pública; egresso de escola pública autodeclarado negro ou pardo; egresso de escola pública autodeclarado indígena; e egresso de escola pública portador de deficiência.

O processo de seleção no PSS 2011, no que se refere a aplicação das provas, ocorreu de maneira tradicional para ambos os tipos de ocupação das vagas, onde os candidatos foram submetidos em duas etapas de avaliações. Inicialmente, o concorrente realizou uma prova de múltipla escolha abrangendo diferentes áreas de conhecimentos, logo em seguida a seleção ocorreu por meio de uma prova específica de redação. Após aprovados, no ato na inscrição, os alunos cotistas deveriam apresentar a documentação necessária relacionada a opção escolhida na reserva de vagas.

Para confirmar a informação prestada, os documentos que deveriam ser apresentados são: certificados de conclusão e históricos escolares do ensino fundamental e médio em escolas públicas para os classificados pela modalidade de reserva de vagas; laudo médico, atestado pela comissão médica da UFPB, para o caso do candidato ter sido classificado em vagas reservadas para pessoas com deficiência; e por fim, o candidato aprovado em vaga destinada negros, pardos ou indígena, deveria entregar um documento de autodeclaração étnico-racial.

No âmbito nacional, só a partir de agosto de 2012, foi sancionada a Lei nº 12.711/2012 a qual dispõe sobre o ingresso nas universidades federais e nas instituições federais de ensino técnico de nível médio e dá outras providências. Tornando obrigatório a adoção do sistema de reserva de vagas em todas as instituições federais de educação superior do Brasil, com um total de 59 universidades e 38 institutos.

A lei determina que em cada concurso seletivo para ingresso nos cursos de

graduação, por curso e turno, no mínimo 50% de suas vagas deverão ser reservadas para estudantes que tenham cursado integralmente o ensino médio em escolas públicas e a outra metade para ampla concorrência. No preenchimento destas vagas, foram reservadas para os estudantes oriundos de famílias com renda igual ou inferior a 1,5 salário-mínimo (um salário-mínimo e meio) per capita, como também para estudantes autodeclarados pretos, pardos e indígenas e por pessoas com deficiência.

3.4 Estratégia Empírica

3.4.1 *Matching*

A hipótese a ser testada nesta pesquisa é que a política de reserva de vagas no ensino superior por meio de cotas pode afetar os incentivos individuais, de forma direta e indireta, gerando impactos sobre os resultados educacionais. Inicialmente, para avaliar essa hipótese formulada, as estimações dos parâmetros de interesse para o cálculo do efeito médio das cotas são feitas considerando três diferentes algoritmos de pareamento: (i) *Propensity Score Matching* (PSM), (ii) Mahalanobis *Distance Matching* (MDM) e (iii) *Classification Tree Analysis* (CTA). Como teste de robustez dos resultados, são feitos recortes amostrais por grande área de conhecimento e cursos, a fim de se controlar os níveis de habilidades e determinados grupos de preferências dos indivíduos.

Existem diferentes abordagens de pareamento na literatura[6], com destaque para os três supracitados. O PSM nos últimos anos é o método de pareamento com uma grande frequência nas pesquisas de inferência causal, contudo, como demonstrado por King *et al.* (2011) e King e Nielsen (2016), esta abordagem aumenta o viés, a ineficiência dos estimadores, a dependência do modelo e o desbalanceamento quando confrontado com outros métodos de pareamento. Tendo em conta esses pontos, este trabalho usa ambos os métodos, mas enfatiza as análises do MDM e CTA.

Para verificar a diferença do resultado da variável de interesse, nível de desempenho acadêmico (capturado pelo coeficiente de rendimento escolar relativo) dos estudantes, em decorrência das cotas, precisa-se comparar indivíduos com atributos semelhantes. A variável de tratamento é a forma de ingresso dos estudantes, se por meio da política afirmativa de Cotas (grupo de tratamento) ou por meio de Concorrência Ampla (grupo de controle), em que Cotas $= 1$ denota a situação de entrada na universidade por meio de reservas de vagas e Cotas $= 0$ para a situação de não tratamento.

[6] Em Stuart (2010), por exemplo, é feita uma revisão e sumarização dos principais modelos de pareamento.

O cálculo do efeito médio por meio do método de pareamento, como ressalta Rosenbaum e Rubin (1983), requer informações sobre as características observáveis ($\mathbf{X}_i$) dos indivíduos no período pré-tratamento. A construção do contrafactual hipotético pelo pareamento é realizada a partir do desenvolvimento de um grupo de controle que contenha indivíduos com características observáveis idênticas, em média, aos indivíduos pertencentes ao do grupo de tratamento.

De acordo com Caliendo e Kopeinig (2005), Rosenbaum (2010), Stuart (2010) e King e Nielsen (2016), o método de pareamento baseia-se em duas hipóteses centrais. A primeira delas é conhecida como seleção nos observáveis (ignorabilidade ou independência condicional): ao controlar todos fatores $\mathbf{X}_i$ que podem afetar o resultado potencial e também a decisão do indivíduo de participar ou não das Cotas, o resultado potencial torna-se independente do indicador de atribuição de tratamento, i.e., $[Y_i(0), Y_i(1)] \perp \text{Cotas}_i|\mathbf{X}_i$, em que $Y_i(0)$ e $Y_i(1)$ representam, respectivamente, o resultado de um dado indivíduo na ausência e na presença do tratamento. A segunda é chamada de hipótese de sobreposição ou suporte comum: para assegurar grupos de controle e tratamento semelhantes é necessário que haja um espaço da região do vetor $\mathbf{X}_i$ que represente bem tanto pessoas do grupo de tratamento como do grupo de controle que poderiam ter sido tratadas, i.e., $0 < Pr(\text{Cotas}_i = 1|\mathbf{X}_i) < 1$.

Por essas suposições, os indivíduos são pareados segundo as suas características observáveis, de tal modo que a diferença entre eles pode ser atribuída ao fato de um ter sido tratado e o outro não. O primeiro estágio do PSM, MDM e CTA requer um vetor $\mathbf{X}_i$ que contempla tais características observáveis no período pré-tratamento.

Tendo o suporte desses fatores, o PSM requer o cálculo da probabilidade condicional de participação da Política de Cotas dado o vetor de características observáveis (ver equação 3.1).

$$Pr(\text{Cotas}_i = 1|\mathbf{X}_i) = G(\mathbf{X}_i\boldsymbol{\lambda} + \epsilon_i), \tag{3.1}$$

onde: $G(\cdot)$ representa a função de distribuição acumulada, seguindo por hipótese uma distribuição de probabilidade logística. Assim, neste ensaio $Pr(\mathbf{X}_i)$, o escore de propensão, é estimado pelo modelo de resposta qualitativa *logit*.

Por sua vez, o MDM também requer uma métrica para definir o grau de similaridade entre os grupos. Neste caso, conforme Rubin (1979), Rubin (1980), Stuart e Rubin (2007) e King e Nielsen (2016), a métrica de Mahalanobis (MDM) é dada por:

$$m(X_{i,0}, X_{j,0}) = [(X_{i,0} - X_{j,0})^T \mathcal{S}^{-1}(X_{i,0} - X_{j,0})]^{1/2}, \tag{3.2}$$

em que: $\mathbf{X}_i$ e $\mathbf{X}_j$ representam os vetores de características observáveis dos estudantes tratados i e dos estudantes não tratados j; $\mathcal{S}$ = matriz de covariância das variáveis $\mathbf{X}$,

que são covariadas, segundo Stuart (2010), associadas com a atribuição de tratamento e os indicadores de impacto.

Por fim, o terceiro e último método para o cálculo do escore de propensão, o *Classification Tree Analysis* (CTA), é considerado um método, não-paramétrico, de árvore de decisão em que a amostra será dividida eficientemente em grupos homogêneos. Para Linden e Yarnold (2018), os subgrupos criados são chamados de "estrato de amostra", uma vez que o modelo CTA utiliza dos próprios atributos da amostra para fazer a estratificação. São homogêneos intra grupo, e heterogêneos entre os estratos.

Após o cálculo da métrica de distância por PSM, MDM e CTA, o efeito médio do tratamento sobre os tratados (τ), conhecido também por ATT (*average treatment effect on treated*), é estimado a partir da Equação 3.3.

$$\hat{\tau} = E[Y_i - \hat{Y}_i(0)|\text{Cotas} = 1], \text{ com } \hat{Y}_i(0) = J^{-1}\sum_j Y_j \tag{3.3}$$

em que J representa o número de indivíduos que compõem o contrafactual estimado de i, com $j \in \{j||f(X_j) - f(X_i)| \leq \delta, \text{Cotas}_i = 1, \text{Cotas}_j = 0\}$, sendo δ uma constante, com $\delta \to 0$. Como $f(\cdot)$ é uma variável contínua e isto torna difícil observar dois ou mais indivíduos com igual valor de métrica da distância de Mahalanobis, $f(\cdot) = m(\cdot)$, e do escore de propensão, $f(\cdot) = p(\cdot)$, o cálculo da equação 3.3 só é possível a partir da utilização de algum algoritmo de pareamento. Seguindo as recomendações da literatura, como Caliendo e Kopeinig (2005), Rosenbaum (2010) e Gertler, Martinez e Premand (2011), este trabalho usa duas diferentes métricas de distância, sem reposição e sem raio, a saber: o método tradicional do vizinho mais próximo (*Nearest*) e o método de Hansen (2004) de identificação ótima do pareamento (*Optimal Matching*).

Destaca-se que o método tradicional do vizinho mais próximo escolhe, um por vez, o correspondente do grupo de controle mais próximo para cada unidade tratada, sem tentar minimizar uma medida de distância global. Já o *optimal matching* encontra as amostras correspondentes com a menor distância absoluta média entre todos os pares correspondentes. Gu e Rosenbaum (1993) discutem que as abordagens do vizinho mais próximo e da identificação ótima geralmente escolhem os mesmos grupos de controles para as amostras combinadas gerais, mas o *optimal* faz o melhor trabalho no sentido de minimizar a distância dentro de cada par. Afirmam ainda que a identificação ótima do pareamento pode ser útil quando não há muitos indivíduos semelhantes para formar o grupo de controle apropriados para os indivíduos do grupo de tratados.

3.4.2 *Análise de Sobrevida*

A segunda metodologia empregada neste ensaio consistiu na abordagem de análise de sobrevida (*Survival Analysis*), também conhecida como análise de duração, para medir os possíveis efeitos da política de reserva de vagas sobre a probabilidade de sobrevivência dos alunos da UFPB. Os modelos dedicados aos estudos de questões de sobrevivência temporal já tem uma literatura estabelecida e consolidada dentro da estatística e da econometria (COX; OAKES, 1984; KALBFLEISCH; ROSS, 1980; LANCASTER, 1992; MENEZES-FILHO; PICCHETTI, 2000; CAMERON; TRIVEDI, 2005).

O modelo de sobrevivência possui como variável de interesse o tempo transcorrido até a ocorrência de um determinado evento, de maneira que um indivíduo transita de um estado para outro após a ocorrência deste evento, sendo denominado como "falha". De outro modo, as análises de duração estimam o período transcorrido entre o evento inicial, no qual o indivíduo assume um estado em particular, até correr a falha, ou seja, quando o agente deixa o estado inicial, e ocupa um novo. Normalmente, essa modelagem avalia a relação entre o tempo de sobrevida de indivíduos (WOOLDRIDGE, 2010; CAMERON; TRIVEDI, 2005; FOX; WEISBERG, 2011)

O período temporal transcorrido entre a origem e a falha é denominado por $T \geq 0$, que pode assumir um valor particular t. Nesse estudo, a falha adotada será o abandono do curso. As conclusões obtidas a partir do modelo são fornecidas através da função de sobrevivência e da função de risco, dadas em função de T. A função de sobrevivência, $S(t)$, indica a probabilidade do aluno, ingressante na UFPB em 2010 e 2011, cotista e não cotista, de sobreviver no seu curso além do tempo t, dado que ele permaneceu até aquele instante. Já a função de risco, $\lambda(t)$, fornece a probabilidade de que o evento de falha (abandono) ocorra, em um dado intervalo de tempo, considerando o fato de que o abandono não ocorreu até aquele momento.

A função de distribuição cumulativa da variável aleatória T é denotada por $F(t)$, e é definida como a probabilidade de um evento ocorrer até o tempo t, $F(t) = Pr[T \leq t]$ para $t \geq 0$ (CAMERON; TRIVEDI, 2005). Logo, a função de sobrevivência, $S(t)$, é um conceito complementar da função de distribuição cumulativa, $F(t)$:

$$S(t) = 1 - Pr[T \leq t] = 1 - F(t) = Pr[T > t] \tag{3.4}$$

Nos termos da análise proposta, a função de sobrevivência, $S(t)$, informa a probabilidade de que um estudante continue no curso por um período maior que t, o período de abandono, sendo escrito por:

$$S(t) = Pr[T > t] \tag{3.5}$$

Por sua vez, a função de risco é expressa pela razão entre a probabilidade que a falha (abandono) ocorra no intervalo $[t, t+h)$ e a variação de tempo h, condicionada ao fato que o aluno permaneceu estudando até t, quando o limite $t \to 0$. Menezes-Filho e Picchetti (2000) e Wooldridge (2010) destacam que a função de risco representa uma probabilidade condicional avaliada em cada período (instante) no tempo. Em suma, a função de risco corresponde a probabilidade de que o estudante abandone o curso o qual ingressou na UPFB, condicionado ao fato de ter sobrevivido até t. A função de risco pode ser representada como:

$$\lambda(t) = \lim_{h \to 0} \frac{Pr[t \leq T < t+h | T \geq t]}{h} \tag{3.6}$$

3.4.2.1 *Cox proporcional hazard model*

Ante as diversas abordagens do modelo de duração, o modelo de sobrevivência adotado para estimar a probabilidade de sobrevida dos alunos da UFPB foi o Modelo de Risco Proporcional de Cox (*Cox proporcional hazard model*). O referido modelo é considerado como um dos principais instrumentos de análise de duração (COX; OAKES, 1984; ZHOU, 2001; FOX; WEISBERG, 2011), e é constantemente utilizado nas mais diversas áreas de estudo para modelar a distribuição do tempo de sobrevivência.

Dessa maneira, o objetivo principal do modelo é investigar a relação entre a distribuição do tempo de duração e as covariáveis. Uma vez que o modelo de Cox é considerado uma abordagem semi-paramétrica, este foi escolhido por ser um método mais flexível e por se adequar melhor as relações complexas existentes nas escolhas dos estudantes quando estes estão cursando o ensino superior na UFPB.

Modelo de Risco Proporcional de Cox admite uma função de risco constante, o que implica em uma função de distribuição exponencial do tempo de sobrevivência. Em geral, a análise da distribuição do tempo de sobrevivência e a relação entre as covariáveis, x_i, resulta em uma relação linear do logaritmo do risco nas covariáveis, ou em um modelo multiplicativo do risco:

$$log\lambda_i(t) = \alpha(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \tag{3.7}$$

ou, de forma equivalente:

$$\lambda_i(t) = \lambda_0(t) exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}) \tag{3.8}$$

em que $\lambda_i(t)$ é o modelo ser estimado; os parâmetros (β_k) medem o efeito individual de cada covariável sobre a distribuição do tempo de sobrevivência, e são estimados pelo método de máxima verossimilhança; e, por fim, λ_0 representa a função base de

risco (*baseline hazard*), em razão que $\lambda_i(t) = e^{\alpha} \rightarrow log\lambda_i(t) = \alpha$, com $\alpha(t) = \lambda_0(t)$. Este último termo refere-se ao valor do risco se todos os x_{is} são iguais a zero, de modo similar aos modelos de regressão linear convencionais, ou seja, representa o termo de intercepto da função. Tomando duas observações, i e i', como valores diferentes de x_i, os respectivos regressores lineares são dados por:

$$\begin{aligned} \eta_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \\ \eta_{i'} &= \beta_1 x_{i'1} + \beta_2 x_{i'2} + ... + \beta_k x_{i'k} \end{aligned} \tag{3.9}$$

Logo, independente de t, a razão de risco proporcional para ambas as observações é descrita por:

$$\frac{\lambda_i(t)}{\lambda_{i'}(t)} = \frac{\lambda_0(t)e^{\eta_i}}{\lambda_0(t)e^{\eta_{i'}}} = \frac{e^{\eta_i}}{e^{\eta_{i'}}} \tag{3.10}$$

Nesse sentido, pode ser inferido que diferentes indivíduos têm funções de risco proporcionais entre si, e que a razão entre elas não varia ao longo do tempo. Como exemplo, pode-se supor que a taxa de risco do indivíduo i é duas vezes maior em relação ao indivíduo i', então esta proporção deve se manter constante para todo t. Por isso, o modelo de risco de Cox é denominado como um modelo de risco proporcional.

Por fim, o coeficiente exponencial estimado, $exp(\beta_i)$ é denominado de taxa de risco (*Hazard Ratio*, HR), e representa o efeito individual do fator correspondente, i, sobre a distribuição do tempo de sobrevivência. O β_i maior que zero, equivale a uma taxa de risco maior que um, e infere que quando o valor da i-ésima covariável aumenta, eleva a probabilidade de ocorrência da falha, estando associada a uma menor taxa de sobrevida.

3.5 Base de dados e descrição das variáveis

A base de dados utilizada nesta pesquisa é oriunda da Superintendência de Tecnologia da Informação (STI) da Universidade Federal da Paraíba (UFPB) para os anos de entrada de 2010 e 2011. Ressalta-se que em 2010, conforme o Edital Nº 40/2010 da Comissão Permanente do Concurso Vestibular (Coperve) da UFPB, institui no processo seletivo para o ingresso na instituição em 2011 a reserva de vagas em cursos da UFPB. A justificativa para não usar apenas os ingressantes de 2011 se atém ao objetivo de minimizar o viés de seleção.

A estratégia é adotar um grupo de discentes que se matricularam na universidade em 2010 para compor o grupo de controle, uma vez que não existia o programa, para minimizar o viés causado pelo processo de autoseleção. Assim, após a expansão

de opções de anos para a construção do grupo de controle, este é composto por discentes não cotistas que matricularam-se na UFPB em 2010 e 2011, enquanto o grupo de tratamento é formado pelos dicentes cotistas que ingressaram na universidade apenas em 2011.

O banco de dados fornece informações dos discentes na UFPB, contemplando as seguintes variáveis: sexo, raça, naturalidade, informações socioeconômicas (como renda familiar, estado civil e escolaridade dos pais), se fez ensino médio em escola pública, nota obtida no vestibular, matrícula na universidade, curso, grande área, turno, período do ingresso, forma de entrada, situação do aluno, ano de abandono, ano de conclusão, rendimento escolar, naturalidade, CRA relativo e outras.

Tendo em vista as informações disponíveis, o vetor de características observáveis, $\mathbf{X}$, utilizados nos métodos de pareamento, será formado por variáveis que estejam associadas aos critérios de elegibilidade da política de cotas da UFPB e a fatores que possam explicar os indicadores de impacto. O vetor $\mathbf{X}$ será construído com o intuito de desenvolver o contrafactual do resultado dos discentes que ingressaram na instituição por meio da política afirmativa de reserva de vagas. A seguir, algumas explicações sobre as covariadas que compõem o vetor utilizado neste estudo.

- Tratamento: *Cota* é uma *dummy* que busca comparar os indicadores entre o grupo de cotistas – grupo de tratamento que assume valor um quando o estudante ingressa na UFPB através das políticas afirmativas (escola pública e raça) – e o grupo de não cotistas – grupo de controle ou contrafactual que assume valor zero quando os estudantes são elegíveis ao programa mas vinculam-se a universidade por meio de ampla concorrência – em dois períodos distintos: um período em que não existia a política de reservas de vagas (2010) e um período em que existia (2011).

- Características individuais para a elegibilidade da política de cotas: como já foi descrito na Seção 3.3, poderiam se inscrever no programa de cotas na UFPB apenas os candidatos egressos do ensino público, e egressos do ensino público autodeclarados negros ou pardos. Logo, as variáveis que captam esses comportamentos são *Escola Pública* e *Raça*, respectivamente. Ambas são variáveis binárias, *dummys*, no caso de serem iguais a um, os discentes estudaram em escolas públicas e se autodeclaram brancos.

- Variáveis que captam o comportamento pós ensino médio: com o intuito de perceber a relação entre o tempo que o indivíduo termina o ensino médio e a demanda em relação ao ensino superior, principalmente no que se refere a forma seriada de admissão local, através do PSS, foi adotado a variável *Tempo de conclusão do ensino médio*. Como também é utilizada a variável *Média do vestibular*

para compreender o desempenho inicial dos alunos ingressantes e a relação com a variável de tratamento.

- Naturalidade: a variável *Natural da Paraíba* é uma *dummy* que busca mensurar se há dependência entre o indivíduo que é natural do estado, quando assume valor igual a um, e a demanda por participar da política de reserva de vagas na UPFB.

- Fatores socioeconômicos: dado que os estudantes que possuem acesso à plataformas online tenderiam a participar de outros processos seletivos, a variável *Acesso a internet* é adotada para mensurar o impacto de acesso as estruturas virtuais sobre o estudante participar ou não do programa de cotas. Por sua vez, a variável *Renda até dois salários mínimos* é um forte indicador da estrutura econômica dos estudantes e da sua família. Uma vez que é presente a relação entre os estudantes de escolas públicas (característica de elegibilidade do programa) serem oriundos de famílias de baixo poder aquisitivo.

- *Background* familiar e atributos individuais: seguindo a literatura sobre economia da educação, mais especificamente no ensino superior, Sampaio *et al.* (2011) afirmam que características que mensuram a estrutura familiar são fundamentais para entender o comportamento dos estudantes nas universidades. Neste estudo, serão adotados as variáveis *dummys Pai com ensino superior* e *Mãe com ensino superior* relacionados ao *background* familiar – no caso de serem iguais a um, os pais possuem ensino superior – e a variável binária *Sexo* para os atributos individuais, quando assume valor um o estudante é do sexo feminino.

- Fatores relacionados as escolhas no ensino superior: *Turno* é uma variável *dummy* que considera se o estudante que cursa o ensino superior a noite teve maiores chances de optar pelo ingresso na UFPB pela política de cotas. Do mesmo modo, são adotas *dummys* para cada grande área de conhecimento ofertada pela universidade, as quais buscam captar a relação que se o estudante busca o curso mais demandado têm maior possibilidade de escolha concorrer como aluno cotista. As grandes áreas abordadas são: (1) Ciências agrárias; (2) Ciências biológicas, área base na análise; (3) Ciências da saúde; (4) Ciências exatas de da terra; (5) Ciências Humanas; (6) Ciências sociais aplicadas; (7) Engenharias; (8) Linguística, letras e artes; e (9) Outras.

3.5.1 Características do mecanismo de admissão, desempenho e abandono no ensino superior na UFPB

Nesta subseção são expostas as informações sobre o comportamento da variável de tratamento, alunos cotistas e não cotistas, e das variáveis de resultados, a saber:

taxas de abandono e o Coeficiente de Rendimento Acadêmico (CRA) relativo, relacionadas com outros indicadores educacionais importantes para uma análise inicial dos objetivos propostos da presente pesquisa. Os quais buscam verificar os efeitos, direitos e indiretos, das políticas afirmativas de cotas na UFPB sobre as taxas de abandono e desempenho dos discentes. Dessa maneira, a Tabela 3.1 apresenta o número de alunos ingressantes separados por ano, 2010 e 2011, por cota apenas para 2011, e os testes de médias e intervalos de confiança das variáveis quantitativas da média de idade, média no vestibular, taxas de abandono e CRA relativo médio por sistema de ingresso na UFPB.

Tabela 3.1 – Evidências iniciais dos indicadores de resultado dos alunos ingressantes por ano de ingresso e cota na UFPB - Testes de médias e intervalo de confiança.

Ano	Cota	Ingres.	Idade	Média	Tx de Abandono			CRA
			Média	Vestibular	Total	1º Ano	2º Ano	Relativo
			23,6	524,2	0,52	0,19	0,29	35,1
2010	0	8.426	23,5 - 23,8	521,9 - 526,4	0,50 - 0,52	0,18 - 0,19	0,28 - 0,30	34,1 - 36,1
			0,000	0,000	0,000	0,000	0,000	0,000
			23,3	533,2	0,51	0,22	0,34	37,0
	0	6.946	23,2 - 23,5	531,4 - 535,1	0,49 - 0,52	0,21 - 0,23	0,32 - 0,35	35,7 - 38,2
			0,000	0,000	0,000	0,000	0,000	0,000
			22,9	494,6	0,45	0,16	0,28	28,7
2011	1	1.564	22,6 -23,3	491,4 - 497,8	0,43 - 0,47	0,14 - 0,18	0,26 - 0,30	26,5 - 30,9
			0,000	0,000	0,000	0,000	0,000	0,000
			23,3	524,7	0,49	0,21	0,32	35,4
	Total	8.510	23,1 - 23,4	523,0 - 526,4	0,48 - 0,50	0,20 - 0,21	0,31 - 0,34	34,3 - 36,5
			0,000	0,000	0,000	0,000	0,000	0,000

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.
Nota: Intervalos com 95% de confiança para as médias calculadas.

A partir de uma análise inicial, a UFPB admitiu um total de 8.426 alunos em 2010, onde todos ingressaram pelo sistema de ampla concorrência, visto que em 2010 ainda não existia a política de cotas. Por sua vez, em 2011, ano em já existia a política, e como já se sabe 25% do total das vagas ofertadas pela instituição deveriam ser reservadas para os candidatos vindos de escolas públicas, contudo, apenas 18% do total desse ano foram ocupadas pelos alunos elegíveis ao programa, ou seja, em 2011 de um total de 8.510 vagas, apenas 1.564 foram preenchidas por alunos cotistas. Acredita-se que tal situação tenha ocorrido por ter sido o primeiro ano após a implantação da política e os estudantes ainda estavam em processo de adaptação, principalmente em virtude de que eram exigidos vários documentos comprobatórios no ato da matrícula.

As informações contidas na Tabela 3.1 apontam que, com 95% de confiança, os intervalos afirmam que as médias entre os grupos são iguais com sobreposição. Logo,

a partir de uma análise ingênua, os estudantes têm, em média, uma realidade similar entre os anos, 2010 e 2011 no geral. Por outro lado, o comportamento dos indicadores começa a mudar a partir da análise da heterogeneidade entre os grupos, cotistas e não cotistas, em 2011. Tal situação acontece para todas as variáveis que mensuram o rendimento educacional dos estudantes, seja antes de entrar no ensino superior, como a média do vestibular, até o CRA relativo médio em que o aluno já está na universidade. No que concerne as médias do vestibular, a diferença entre os grupos é maior para o grupo de controle, possuindo uma média de 533,2, contra apenas 494,6 para o grupo de alunos cotistas.

Fica claro que os alunos cotistas oriundos de escolas públicas, que ingressaram na UFPB em 2011, têm menor desempenho ou informação para obterem notas semelhantes no vestibular com o grupo de alunos não cotistas. Evidência discutida também por Francis e Tannuri-Pianto (2012), onde afirmam que estudantes beneficiários de políticas afirmativas por raça no Brasil tiveram uma nota no vestibular significativamente menor do que os não beneficiários. Quanto a variável taxa de abandono, esta possui três indicadores, taxa de abandono total, abandono no primeiro ano e no segundo ano do curso. O que se observa em ambas situações é que o abandono é menor no grupo de tratamento, ou seja, os alunos cotistas tendem a desistir do curso a uma taxa menor do que os alunos não cotistas. Assim, tal inferência inicial pode sinalizar uma externalidade positiva do sistema de reserva de vagas, onde a política de cotas reduz o abandono no ensino superior.

Por fim, ainda na Tabela 3.1, é apresentada a variável de resultado adotada para captar o desempenho dos alunos no ensino superior na UFPB, o CRA relativo médio. É válido destacar que este capta o nível de rendimento médio dos discentes no decorrer de cada semestre durante sua permanência no curso. Do mesmo modo como foi visto na variável média do vestibular, o comportamento entre os grupos no CRA relativo médio é superior para o grupo de alunos que não ingressaram na universidade por meio das reservas de vagas. Embora elegíveis para o programa, os alunos desse grupo possuem habilidades que se sobressaem em seus rendimentos quando comparados ao CRA relativo médio dos discentes cotistas. Em face do exposto, a Tabela 3.3 apresenta o mesmo comportamento visto na Tabela 3.1, mas por grande área de conhecimento e cursos na UFPB em 2011, com o objetivo de observar os desempenhos educacionais dos discentes por diferentes níveis de concorrência. A Tabela B.1 no apêndice apresenta os testes de médias e intervalos de confiança detalhados para as médias expostas na Tabela 3.3.

Tabela 3.3 – Evidências iniciais dos indicadores de resultado dos alunos ingressantes por cota por grande área de conhecimento e cursos na UFPB, 2011.

	Cota	Ingres.	Média Idade	Média Vestibular	Taxa de Abandono			CRA Relativo
					Total	1º Ano	2º Ano	
Grande Área de Conhecimento								
C. Sociais Aplicadas	0	1574	22,6	512,3	0,52	0,21	0,34	44,8
	1	409	23,3	470,7	0,53	0,19	0,30	35,7
Ling., Letras e Artes	0	1023	23,9	527,3	0,47	0,22	0,32	31,2
	1	249	23,3	490,7	0,44	0,16	0,30	27,6
Engenharias	0	831	20,6	574,2	0,56	0,18	0,30	47,9
	1	223	21,0	525,5	0,54	0,15	0,26	43,9
C. Humanas	0	978	24,4	530,8	0,41	0,18	0,28	24,5
	1	217	25,7	497,1	0,35	0,08	0,20	20,4
C. da Saúde	0	821	21,5	583,1	0,33	0,17	0,25	16,5
	1	202	22,1	536,9	0,21	0,10	0,14	12,9
C. Exatas e da Terra	0	796	25,7	525,7	0,69	0,27	0,46	87,7
	1	126	22,4	487,3	0,69	0,31	0,50	72,2
C. Agrárias	0	681	24,8	493,4	0,55	0,26	0,40	48,9
	1	112	21,5	465,3	0,41	0,19	0,31	38,4
Outras	0	121	23,5	497,4	0,44	0,25	0,37	23,7
	1	26	22,6	472,2	0,23	0,11	0,23	22,8
C. Biológicas	0	121	26,6	-	0,81	0,47	0,59	50,8
	1	-	-	-	-	-	-	-
Cursos com maiores entradas de cotistas								
Direito	0	328	21,4	624,9	0,31	0,09	0,16	16,3
	1	97	26,3	551,0	0,26	0,09	0,19	13,6
Administração	0	203	21,5	547,6	0,37	0,16	0,26	37,2
	1	58	22,5	504,8	0,29	0,06	0,12	31,7
Pedagogia	0	315	25,4	475,3	0,43	0,21	0,31	27,2
	1	51	25,1	450,9	0,41	0,01	0,19	26,8
Cursos com menores entradas de cotistas								
Química Industrial	0	33	20,2	550,9	0,60	0,18	0,36	79,5
	1	11	20,6	496,7	0,63	0,18	0,63	63,4
Antropologia	0	45	25,1	472,3	0,66	0,26	0,53	82,9
	1	10	25,6	454,2	0,90	0,60	0,70	-
Sist. de Informações	0	48	21,4	501,9	0,62	0,22	0,39	53,7
	1	9	23,7	447,3	0,66	0,22	0,44	44,8

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.

A ordem das grandes áreas e dos cursos estão de forma decrescente no que se refere a quantidade de vagas preenchidas por cotas, como por exemplo, a área das ciências sociais aplicadas teve 409 alunos cotistas ingressantes em seus cursos. Isto justifica-se porque a referida grande área de conhecimento ofertou o maior número de vagas no processo seletivo em 2010, para a entrada em 2011. Por outro lado, ciências

biológicas não teve nenhuma vaga preenchida para alunos beneficiados pelo programa. A mesma relação pode ser vista nos cursos com maiores e menores entradas de cotistas. O curso de direito, por exemplo, apresentou uma maior entrada de alunos cotistas visto que possuía o maior número de vagas nas suas turmas. No outro extremo, o curso de sistema de informações deteve apenas nove discentes selecionados por meio da política. Pela quantidade total de vagas, o curso supracitado deveria ter apenas uma, ou no máximo duas turmas em 2011.

Assim como na Tabela 3.1, o comportamento entre os grupos, mas agora estratificados por áreas e cursos, são distintos entre si, como pode ser visto na Tabela 3.3. Diferenças significativas também podem ser visualizadas entre as grandes áreas, como por exemplo, as maiores médias no vestibular são dos discentes que optaram pelos cursos da área das ciências da saúde, em que alunos não cotistas tiveram uma média de 583,1, enquanto os alunos cotistas 536,9. Enquanto isso, no mesmo período, os discentes que ingressaram nos cursos das ciências agrárias detiveram médias de entradas inferiores as demais grandes áreas. No que corresponde aos cursos, diferenças entre os grupos de tratamento e controle também podem ser vistas, com destaque para o curso de direito, o qual obtém alunos com as maiores médias de entrada na instituição.

As informações relacionadas as taxas de abandono por grupos variam por grandes áreas. Enquanto as taxas de abandono são maiores para os alunos não cotistas nos cursos das ciências da saúde, nas engenharias, nas ciências humanas, e outros, nos cursos das ciências exatas e da terra os alunos cotistas abandonam mais do que os não cotistas. Tal fato pode ser relacionado a dificuldade de entrada por parte dos alunos cotistas nos cursos mais concorridos da UFPB. Pois, em cursos que apresentam maiores índices de concorrência, os discentes que conseguem uma vaga, seja por cotas ou não, se dedicam mais a não abandonar suas escolhas. De acordo com as informações da Coperve em 2010, referente aos candidatos inscritos ao Edital Nº40/2010, os cursos com maiores candidatos inscritos por vagas são das áreas engenharias, ciências humanas e ciências da saúde [7].

3.6 Resultados

Os resultados apresentados nesta seção referem-se aos possíveis efeitos da política afirmativa de reserva de vagas sobre o indicador de desempenho acadêmico, medido pelo CRA relativo, e sobre as taxas de abandono durante o período de permanência no curso dos alunos cotistas da UFPB. Para investigar este efeito, as abordagens não-experimentais, PSM, MDM e CTA, foram utilizadas para identificar

[7] ver <http://www.coperve.ufpb.br/>

a magnitude e a direção dessa intervenção política no processo de desempenho educacional dos alunos no ensino superior.

Pelos resultados incondicionais expostos na Tabela 3.1, o desempenho médio dos alunos cotistas, medidos pelo CRA relativo, é de 28,7, enquanto o dos alunos não cotistas é de 37,0. Já as taxas de abandono são mais favoráveis aos alunos cotistas, onde apresentaram taxas menores do que as taxas de alunos não cotistas. Indicando que os alunos que são admitidos pela nova forma de ingresso no ensino superior, em 2011, tendem a apresentar um nível de desempenho menor, em contrapartida, têm um nível de desistência menor nos cursos escolhidos.

Para fins de uma mensuração menos tendenciosa, é de suma importância a identificação de um grupo de alunos ingressantes com características similares àqueles que entraram na universidade pelo novo sistema de concorrência de vagas. A não participação na política não é suficiente para que os estudantes possam ser alocados como grupo de controle, dado que estes podem ter características muito distintas das dos cotistas, de modo a não representarem uma situação de cotrafactual.

Dessa maneira, esta seção apresenta os principais elementos que subsidiam as discussões dos efeitos do aluno ser cotista ou não sobre o desempenho e abandono no ensino superior na Paraíba, no período em estudo. Assim, as subseções a seguir exibem os resultados sobre o grau de ajustamento dos modelos de pareamento, o efeito médio das cotas e o seu impacto nas taxas de abandono a partir de um modelo de sobrevivência.

3.6.1 Análise do grau de ajuste do pareamento

Dentre as características observáveis que podem influenciar a decisão de participar do sistemas de cotas e os indicadores de resultados, a Tabela 3.5 apresenta os resultados obtidos para o modelo *logit*, tendo como variável dependente *dummy* de cota, e como fatores observáveis que podem intervir ou não a escolha do aluno, justificadas na Seção 3.5, se o estudante veio de escola pública, a média do vestibular, tempo de conclusão do ensino médio e outras características socioeconômicas do aluno e da sua família considerando 2011 como o ano de entrada. Dentre os métodos de pareamento adotados para a construção do grupo de controle observável, a metodologia PSM demanda parametrizar o cálculo do escore de propensão.

As características observáveis e adotadas no modelo, e descritas na Tabela 3.5, revelam que a variável que indica se o aluno veio de escola pública apresentou uma maior chance do candidato, em 2011, aderir ao sistema de concorrência de vagas sendo cotista. Este resultado já era esperado, uma vez que os critérios para participar da política afirmativa na UFPB em 2011 estabelece que os alunos que cursaram o

ensino médio e, pelo menos, três anos do ensino fundamental, em escolas de rede pública de ensino, concorram separadamente para 25% do total de vagas ofertadas pela Universidade. Sendo assim, é considerada o principal fator de motivação para que o aluno tenha escolhido ser cotista na análise.

Tabela 3.5 – Estimação do *propensity score* para o ano de entrada 2011. Modelo de probabilidade *logit*.

Variável	Coeficiente	Erro-padrão	*Odds Ratio*
Intercepto	-0,369	0,551	0,69
1. Raça	-0,127	0,071	0,88
2. Renda até dois SM	0,237**	0,083	1,26
3. Escola pública	2,254***	0,128	9,52
4. Média do vestibular	-0,004***	0,000	0,99
5. Turno	-0,143	0,085	0,86
6. Tempo de conclusão do EM	-0,080***	0,013	0,92
7. Acesso à internet	-0,374***	0,072	0,68
8. Pai com ensino superior	-0,526***	0,155	0,59
9. Mãe com ensino superior	-0,702***	0,131	0,49
10. Sexo	-0,273***	0,073	0,76
11. Natural da Paraíba	-0,103	0,133	0,90
Grandes Áreas de Conhecimento			
Ciências Agrárias	-0,347	0,287	0,70
Ciências da Saúde	0,680*	0,279	1,97
Ciências Exatas e da Terra	-0,377	0,282	0,68
Ciências Humanas	-0,287	0,271	0,75
Ciências Sociais Aplicadas	-0,188	0,263	0,82
Engenharias	0,910**	0,279	2,48
Linguística, Letras e Artes	0,050	0,268	1,05
N		7.194	
Count R^2		46%	

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.
Nota: Níveis de significância: *10%, **5% e ***1%.

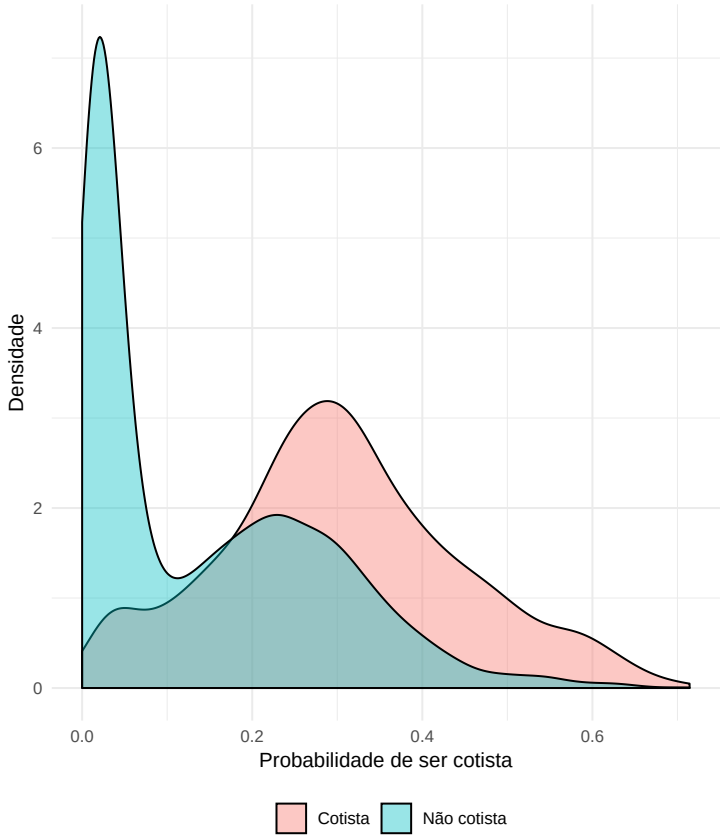
Outra informação que reforça esta análise advém da variável renda com até dois salários mínimos, observa-se que quanto menor a renda do domicílio maiores as chances dos alunos estudarem em escolas púbicas, e, consequentemente, demandarem as vagas direcionadas para cotistas. Logo, a família possuir uma renda de até dois salários mínimos aumenta em 1,26 vezes as chances de participar do vestibular nas vagas reservadas para a política na UFPB. Ainda que pelo modelo de estimação do escore de propensão a variável raça não tenha se mostrado significativa na análise, a sua utilização é de grande importância para determinar o desempenho acadêmico, pois 56% dos 25% das vagas no processo seletivo para candidatos vindos de escolas públicas são reservadas para negros e pardos no vestibular da UFPB em 2011. Nesse caso, com um sinal negativo, o indivíduo autodeclarar-se branco reduziria as chances em 0,88 vezes de optar pelo sistema de cotas.

Por sua vez, o tempo médio de conclusão do ensino médio, assim como a variável média do vestibular, observar-se menores chances de opção pelo sistema de reserva de vagas, principalmente por parte daqueles egressos do ensino médio a pouco tempo. Algumas variáveis que captam o *background* familiar também se mostram negativamente relacionadas com à escolha da política de cotas na universidade. As variáveis correspondentes a escolaridade dos pais, pai com ensino superior e mãe com ensino superior apresentam uma razão de chances 0,59 e 0,49 vezes menor, respectivamente, de optar pelo sistema de reserva de vagas para os filhos. No mais, os indivíduos do sexo feminino são menos propensos a demandarem a política afirmativa, enquanto ser natural do estado da Paraíba e ter feito o ensino médio no período da noite apresentaram coeficientes estatisticamente iguais a zero a um nível de significância de 5%.

As *Dummies* relacionadas as grandes áreas de conhecimento, embora não seja possível constatar significância estatística pelo modelo estimado na grande maioria delas, são de grande importância na análise para entender o comportamento nas preferências dos alunos por cursos e os seus desempenhos em cada um deles. O que se pode constatar ante aos resultados é que os estudantes que demandam cursos das grandes áreas de engenharias e ciências da saúde apresentam uma razão de chances de 2,48 e 1,97 vezes maior de optar pela concorrência no sistema de cotas. Uma possível explicação se atém ao fato dessas áreas possuírem os cursos mais concorridos, como por exemplo, o curso de medicina que teve uma concorrência de 27,6 para uma vaga no PSS da UFPB em 2010, com entrada em 2011, como também o curso de engenharia civil, com uma concorrência de 9,1 alunos para uma vaga, de acordo com as informações da Coperve em 2010.

Após a análise dos coeficientes da Tabela 3.5, se torna importante analisar o poder preditivo do modelo de resposta binária por meio da estimação do *propensity score*. A proporção prevista de casos corretamente foi de 46%, indicando que o número de casos classificados corretamente é inferior as ocorrências classificadas erroneamente. No intuito de melhor explorar algumas evidências já apresentas, a Figura 3.1 ilustra os resultados a partir da distribuição do escore de propensão entre os ingressantes por cotas e por livre concorrência, ou não cotistas. A Figura revela que há diferenças estatísticas significantes entre as duas distribuições estimadas para o período. Em resumo, é possível observar uma maior concentração de massa de probabilidade à esquerda na distribuição para os alunos não cotistas.

Figura 3.1 – Sobreposição das curvas de densidade do escore de propensão



Fonte: Elaboração própria a partir dos microdados do STI/UFPB 2010-2011.

Por outro lado, a característica bimodal da distribuição para o grupo de controle também revela a presença de um grupo de alunos que apresenta um comportamento mais à direita da distribuição, ou seja, um grupo semelhante ao grupo de alunos cotistas, representada pela curva que evidencia uma maior concentração de massa de probabilidade para a direita. Dessa maneira, há uma sobreposição considerável entre as duas curvas do escore de propensão possibilitando uma justificativa para a comparação entre os dois grupos. Nesse cenário, após a análise dos resultados do modelo *logit*, é de suma importância testar o balanceamento das características observáveis dos alunos ingressantes pelos dois sistemas de concorrência de vagas no ensino superior na Paraíba.

Nesse sentido, a Tabela 3.7 vem apresentar a média para cada um dos fatores observáveis dos dois grupos e o teste de hipótese (teste t), o qual afirma que o valor médio de cada variável é igual entre os grupos de controle e tratamento. É válido ressaltar que foi adotado o PSM com vizinho mais próximo. Como ressaltado por Caliendo e Kopeinig (2005), Rosenbaum (2010), Gertler, Martinez e Premand (2011), é fundamental que a hipótese de balanceamento seja atendida em sua totalidade, na qual os grupos de controle e tratamento devam ter as mesmas características observáveis

em média, para que assim, o modelo de pareamento seja considerado efetivo.

Tabela 3.7 – Teste de balanceamento das covariadas antes e após o pareamento por PSM,2011.

Variável	Não Pareado			Pareado		
	Cotista	Não Cotista	Diferença	Cotista	Não Cotista	Diferença
1, Raça	0,41	0,49	-0,07*	0,41	0,44	-0,02
2, Renda até dois SM	0,72	0,42	0,30*	0,72	0,70	0,02
3, Escola Pública	0,93	0,48	0,45*	0,93	0,92	0,01
4, Média do vestibular	502,87	532,21	-29,34*	502,87	503,72	-0,85
5, Turno	0,34	0,31	0,03	0,34	0,32	0,01
6, T. de conclusão do EM	2,73	2,71	0,01	2,73	2,65	0,08
7, Acesso à Internet	0,53	0,73	-0,20*	0,53	0,57	-0,04
8, Pai com E. Sup.	0,04	0,25	-0,20*	0,04	0,03	0,02
9, Mãe com E. Sup.	0,07	0,31	-0,24*	0,07	0,07	0,00
10, Sexo	0,58	0,63	-0,05*	0,58	0,59	-0,00
11, Natural da Paraíba	0,88	0,88	0,00	0,88	0,87	0,01
G. A. de Conhecimento						
C. Agrárias	0,07	0,07	0,01	0,07	0,07	0,00
C. da Saúde	0,22	0,18	0,03	0,22	0,23	-0,01
C. Exatas e da Terra	0,04	0,05	-0,01	0,04	0,03	0,00
C. Humanas	0,16	0,21	-0,05*	0,16	0,18	-0,01
C. Sociais Aplicadas	0,20	0,22	-0,03	0,20	0,19	0,01
Engenharias	0,10	0,09	0,00	0,10	0,12	-0,03
Ling., Letras e Artes	0,19	0,16	0,03	0,19	0,17	0,02

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.
Nota: *p-valor<0,05. Teste de diferenças de média entre cotista e não cotista (hipótese nula: diferença de média igual a zero).

Ante ao exposto, conforme apresentado na Tabela 3.7, antes do pareamento, os dois grupos, cotistas e não cotistas, evidenciaram diferenças entre si, em média. Contudo, para um nível se significância de ao menos 5%, após o pareamento, todas as covariadas apresentam a mesma média, indicando que a hipótese nula não pode ser rejeitada para nenhuma das variáveis explicativas do modelo.

3.6.2 Efeito da Política Afirmativa, Cotas

A partir de agora, os resultados dos modelos de pareamentos serão analisados, tendo como variável de interesse o CRA relativo, como variável de rendimento no curso escolhido e concluído pelos alunos ingressantes na UFPB nos processos seletivos de 2010 e 2011. A Tabela 3.9 apresenta o efeito do discente ter ingressado na universidade por meio do sistema de cotas considerando diferentes algoritmos de pareamento, nela estão expostos os resultados por PSM, MDM e CTA, e por método de pareamento: vizinho mais próximo (*Nearest*) e identificação ótima (*Optimal*) .

As técnicas de pareamentos são estimadas através do método de Mínimos Quadrados Ordinários (MQO) ou (OLS), representados pelos modelos (1) e (4) na

Tabela 3.9, e através do modelo de Regressão Quantílica (RQ), considerado o método econométrico que possibilita compreender os efeitos das variáveis explicativas sobre a variável que capta o desempenho dos alunos para os n-ésimos quantis da distribuição condicional, para cada método de pareamento (*Nearest* e *Optimal*). Nessa análise, serão enfatizados apenas os quantis 25 e 75, os quais representam os coeficientes dos alunos da UFPB que apresentam os piores e os melhores, respectivamente, rendimentos nos na universidade, representados pelos modelos (2), (3), (5) e (6), simultaneamente.

Os ATTs são estatisticamente significativos em todas as técnicas, e a direção dos resultados permanecem os mesmos independentemente do método de pareamento, contudo apresentam comportamentos de impacto distintos entre si. Principalmente quando se observa por estratificação da variável de desempenho (RQ). A relação negativa em todos os parâmetros para esses modelos evidencia inicialmente que o aluno ser cotista na UFPB reduz o seu rendimento no curso escolhido, impactando negativamente em até nove vezes, em média e dependendo do método de pareamento (*Nearest* ou *Optimal*), no CRA em comparação aos rendimentos dos alunos que ingressaram pelo sistema de livre concorrência, nos modelos (1) e (4). A partir desse resultado, tem-se indícios para fundamentar que a política afirmativa de cotas adotada na UFPB no período em análise agravou o nível de qualidade dos ingressantes na universidade, medido por uma variável que capta o rendimento do aluno.

Tabela 3.9 – Efeitos da política de cotas (ATT) sobre a variável de esforço (CRA relativo)

	Nearest			***Optimal***		
	OLS	**Quantile Regression**		**OLS**	**Quantile Regression**	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Propensity score matching						
	-6,025***	-5,143***	-7,524**	-7,090***	-3,359***	-7,800**
	(1,717)	(1,247)	(2,967)	(1,685)	(1,268)	(3,269)
Mahalanobis matching						
	-4,275**	-3,133**	-6,631**	-11,086***	-5,926***	-13,737***
	(1,789)	(1,487)	(2,902)	(1,745)	(1,587)	(3,294)
Classification trees						
	-5,762***	-5,790***	-6,297**	-8,125***	-4,815***	-10,068***
	(1,687)	(1,293)	(2,605)	(1,687)	(1,265)	(3,127)

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.Nota: Níveis de significância: *10%, **5% e ***1%.

No extremo do quantil que representa o desempenho dos estudantes com piores notas, RQ(25), modelos (2) e (3), o discente ser cotista reduz, em média das três técnicas de pareamento e dos dois métodos, o CRA relativo em até 4,6 vezes no comparativo com o rendimento de alunos não cotistas que também se encontram nesse extremo. No que tange ao extremo oposto, RQ(75), modelos (3) e (6), onde capta as melhores

médias de rendimento, ser cotista também provoca uma queda na performance no rendimento do aluno, chegando a uma redução, em média, de até 10,5 vezes menor do que o CRA dos alunos que não aderiram a política [modelos (2), (3), (5) e (6) da Tabela 3.9].

Em suma, ante aos resultados encontrados nos modelos supracitados, o aluno ser ingressante na UFPB, em 2011, e ter optado pelo sistema de reservas de vagas, ou seja, ser cotista reduz o CRA em um grau menor no quantil inferior, RQ(25), exposto nos modelos (2) e (5), do que no quantil superior, RQ(75), exposto pelos modelos (3) e (6). Essa evidência permite refletir sobre a relação entre a média de entrada dos discentes no ensino superior, mensurada por meio da variável média do vestibular, e o nível de rendimento dentro da universidade, medido pelo CRA relativo.

Como pode ser observado na Seção 3.5, na Tabela 3.1, a nota média de entrada no PSS dos alunos cotista é inferior aos dos alunos não cotistas da UFPB, ou seja, se torna mais provável que o impacto seja menor em médias mais baixas no CRA, uma vez que o discentes então mais homogêneos em termos de conhecimento, do que nas médias mais altas. Como pode ser visualizado na Tabela 3.9, onde no quantil superior ser cotista reduz o CRA em um grau superior do que no quantil inferior. Logo, o diferencial visto na nota de entrada afeta o desempenho dos alunos cotistas progressivamente dentro na universidade em qualquer estrato de nota.

Os achados deste trabalho acompanham as indicações de outros estudos empíricos, em um caso internacional Hickman (2009), conhecido na literatura como o *achievement gap*, em estudos nacionais, Ferman e Assunção (2005) e Mendes Junior, Souza e Waltenberg (2016), como também em Sander (2004), Sowell (2004), Bertrand, Hanna e Mullainathan (2010) e Arcidiacono *et al.* (2011) e a *mismacth hypotheses*. Arcidiacono *et al.* (2011) frisam em sua pesquisa que as políticas de ações afirmativas no ambiente educacional selecionam grupos minoritários para o qual não estão preparados, de modo que há uma incompatibilidade (*mismacth*) entre os beneficiários do programa, discentes cotistas, dos demais colegas, ficando claro a desigualdade no desempenho acadêmico entre os dois grupos.

A Tabela 3.11 apresenta o efeito do tratamento agora estratificado por cursos e por grandes áreas de conhecimento da UFPB no período em estudo. Como já foi discutido anteriormente na Seção 3.5, na Tabela 3.3, a justificativa para a discussão sobre o comportamento dos cotistas que escolheram os cursos que estão expostos na tabela 3.11 se atém tanto ao fato dos mesmos possuírem, proporcionalmente, os maiores números de alunos cotistas do que os demais cursos. Como também verificar o impacto sobre a variável de resultados por diferentes níveis de concorrências nos cursos da UFPB, tentando relacionar com as habilidades cognitivas dos alunos. Uma vez que para entrar na universidade, o aluno deve ter uma nota no vestibular mínima

para conseguir a vaga, sendo elas variantes entre as grandes áreas e os seus respectivos cursos.

Na Tabela 3.11, agora apenas com a técnica de pareamento do PSM, ocorre o mesmo comportamento do efeito do tratamento sobre a variável de resultado que foi observado na Tabela 3.9, com uma ressalva para as variações nos níveis de significância estatística dos estimadores por modelo. No que se refere a análise por cursos com maior número de cotista em suas turmas, a saber, direito, administração e pedagogia, referentes as estimações por OLS e RQ, embora não tenham sidos significantes, pode ser visto um comportamento similar aos resultados auferidos para o modelo geral (para todos os cursos) encontrados na Tabela 3.9, onde o discente ser cotista reduz o seu rendimento, nos respectivos cursos, quando comparados aos alunos não cotistas.

Por outro lado, na análise por grandes áreas de conhecimento, a magnitude dos impactos são bastantes heterogêneos, e muitas delas são estatisticamente diferentes de zero à pelo menos 10%. Com destaque para quatro grandes áreas: linguística, letras e artes, engenharias, ciências humanas e ciências da saúde. Como também já foi discutido anteriormente, de acordo com as informações da Coperve em 2010, algumas dessas áreas de conhecimento detém os cursos que receberam os maiores níveis de concorrência na UFPB. Nas ciências da saúde, com destaque para os cursos de medicina, nutrição, odontologia e enfermagem, nas engenharias, destaque para a engenharia civil e engenharia de produção, nas ciências humanas, o curso de direito detém a maior concorrência.

Já no caso de linguística, letras e artes, apesar de ser a segunda grande área de conhecimento a ingressar a maior quantidade de alunos cotistas, visto na Tabela 3.3, o nível de concorrência entre seus cursos é tido como um dos menores, como por exemplo em uma das turmas do curso de pedagogia a concorrência era de dois alunos para uma vaga. Sendo assim, a análise dentro dessa área, Tabela 3.11, parte da relação em que estudante ser cotista reduz o CRA em até 10 vezes em comparação ao nível de rendimento dos estudantes não cotistas, resultado que pode ser visualizado no modelo (1). A situação se agrava quando se observa no quantil com os alunos com as melhores médias do CRA, ou seja, optar pelo sistema de cotas reduz o nível de rendimentos em até 24 vezes quando comparados aos discentes ingressantes pelo sistema de livre concorrência [modelo (3)]. Tal resultado vai ao encontro das discussões impostas por Hickman (2009), em que a partir do momento que se abre vagas para as grandes minorias em grandes universidades, pode se observar uma maior concentração na participação de alunos cotistas nos cursos de menores concorrências.

Tabela 3.11 – Efeitos da política de cotas (ATT) sobre a variável de desempenho acadêmico por cursos e grandes áreas do conhecimento

	Nearest			*Optimal*		
	OLS	**Quantile Regression**		**OLS**	**Quantile Regression**	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Propensity score matching						
Cursos						
Direito	-0,143	0,678	-3,472	-1,938	0,178	-3,611
	(2,019)	(3,948)	(3,899)	(1,832)	(2,679)	(2,655)
Administração	-5.630	-3,435	-6,690	0,880	-3,435	-1,627
	(6.759)	(9,127)	(9,002)	(5,625)	(13,504)	(4,921)
Pedagogia	-5,847	-5,607	-4,679	-4,841	-5,910	-3,082
	(3,479)	(6,630)	(4,161)	(2,960)	(7,340)	(3,488)
Áreas do Conhecimento						
Ciências Sociais Aplicadas	-5.872	-1,392	-12,229	-5,760	-1,487	-10,288
	(4.211)	(3,273)	(8,678)	(3,664)	(3,076)	(7,302)
Linguística, Letras e Artes	-10,026***	-0,100	-24,468***	-5,607*	0,853	-16,226***
	(3,442)	(2,531)	(6,717)	(2,985)	(2,634)	(4,996)
Engenharias	-15,926*	-20,657***	-15,712*	-13,069**	-19,460***	-18,716**
	(8,081)	(6,356)	(8,904)	(6,490)	(7,232)	(8,727)
Ciências Humanas	-6,877***	-4,453	-8,465**	-4,456**	-4,851**	-4,631
	(2,318)	(2,750)	(3,247)	(2,150)	(2,305)	(3,396)
Ciências da Saúde	-6,226***	-4,561	-9,464***	-6,014***	-6,652**	-8,091***
	(1,594)	(3,144)	(2,018)	(1,483)	(2,643)	(1,967)
Ciências Exatas e da Terra	2,028	-4,414	-10,353	9,082	6,343	-5,668
	(21,490)	(15,907)	(53,694)	(20,194)	(15,568)	(65,606)

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.Nota: Níveis de significância: *10%, **5% e ***1%.

É interessante perceber também que a política de cotas evidencia um impacto negativo, e estatisticamente significativo, para os estudantes dos cursos das engenharias em todos modelos estimados. Principalmente quando se atém ao impacto sobre o CRA relativos dos alunos com as piores notas, em que o aluno ser cotista impacta negativamente em até 20 vezes em comparação ao rendimento dos alunos não cotistas. Uma possível justificativa para esse resultado se vincula ao fato de que como as engenharias possuem os cursos mais concorridos, os alunos ingressantes por livre concorrência tenham um nível de habilidade cognitiva superior aos dos alunos cotistas (como pode ser visualizado na Tabela 3.3, por meio da variável média do vestibular), onde estes vão precisar de um nível de desempenho maior para acompanhar o restante da turma.

Dessa maneira, é válido destacar que os mecanismos causais que eventualmente estão por trás dos resultados encontrados nas estimações podem ser diferentes entre os cursos das grandes áreas de conhecimento de maior e menor concorrência em termos de notas de entrada, que podem ser sinalizadores da dedicação dos alunos dentro da universidade. Por fim, em relação as áreas das ciências humanas e da saúde, a magnitude do impacto, negativo, sobre a variável de resultado que mede o rendimento

dos discentes cotistas foram similares, com foco para os modelos (1), (3), (4), (5) e (6), os quais se apresentaram significantes.

3.6.3 Efeitos das cotas a partir da análise de sobrevivência

Após a realização da análise do grau de ajustamento dos modelos de pareamento e do efeito médio das cotas sobre a variável de desempenho acadêmico vistos nas subseções 3.6.1 e 3.6.2, e para uma melhor identificação sobre as evidências encontradas nas Tabelas 3.1 e 3.3 em relação às taxas de abandono, foi utilizado o modelo de duração de risco proporcional de Cox, ponderado pelo PSM, pra avaliar o efeito da política de cotas sobre a probabilidade de sobrevivência na UFPB.

O efeito da política afirmativa de reserva de vagas na UFPB sobre a variável de resultado taxa de abandono poderia ser estimada da mesma forma como foi com a variável de desempenho, CRA relativo. Contudo, taxa de abandono é considerada uma variável de natureza diferente, no sentindo de que o abandono pode acontecer em diferentes períodos, ou seja, de semestre a semestre, e a cada semestre cursado, abandonar tem custos diferentes para o discente. Por exemplo, um aluno que abandonar o curso no primeiro semestre tem um custo diferente do estudante que decidir abandonar no sétimo período. Logo, não pode-se inferir simplesmente como abandonou ou não.

O PSM poderia ser aplicado de forma isolada, porém mascararia o efeito da variável de interesse, taxa de abandono, como já foi discutido no parágrafo anterior. Dessa maneira, o presente estudo inova ao investigar o efeito da política de cotas em uma universidade pública brasileira com a seguinte proposta: integrar o PSM, utilizando amostra pareada entre os grupos de cotistas e não cotistas, ao modelo de duração de risco proporcional de Cox.

Os modelos de sobrevida são considerados como as técnicas mais adequadas para serem aplicadas em estudos longitudinais para o caso de análise de resposta binária (neste estudo, abandono ou não abandono) em que se observa os períodos de observações diferentes entre os indivíduos, devido a ocorrência, ou não, da "falha" (CAMERON; TRIVEDI, 2005). De outro modo, neste estudo, os modelos de duração avaliam a relação do tempo de sobrevida dos estudantes desde o seu ingresso na UFPB através do sistema de cotas até abandonar o curso escolhido. O período da análise longitudinal dos discentes vai desde o primeiro semestre de 2011 (2011.1) até o primeiro semestre de 2018 (2018.1).

Sendo assim, os coeficientes das regressões e a *hazard ratio* (razão de risco), estimados sobre a *dummy* de abandono dos discentes, em relação ao efeito do tratamento, ou seja, ter entrado na UFPB por meio da política afirmativa de cotas, estão apresentados na Tabela 3.13. O efeito das cotas sobre o risco de abandono do aluno, ou

sobre a taxa de sobrevivência, foi avaliado através de duas estratégias, representadas pelos modelos (1) e (2). O modelo básico, (1), é estimado sem considerar a dependência temporal do conjunto das variáveis de controle, analisando apenas o efeito da política.

Já a estratégia de dependência temporal, do modelo (2), analisa o efeito das cotas sobre a margem estensiva da probabilidade de sobrevida dos discentes da UFPB considerando o conjunto de covariadas no modelo de regressão de Cox, a saber: o CRA relativo, a média de entrada no vestibular, e as variáveis binárias de sexo, raça e trancado. Este, por sua vez, permite aproveitar da melhor maneira o conjunto dos dados longitudinais para investigar os efeitos das cotas, ou seja, se o fato do aluno ser cotista na UFPB poderia afetar a taxa de abandono no ensino superior.

Inicialmente, considerando ambos os modelos, os resultados mostram-se estatisticamente significantes, uma vez que é observada à sobreposição dos intervalos de confiança das estimações, supondo um grau de 95% de confiança. No que se refere ao modelo básico, ingressar na UFPB por meio das vagas reservadas para a política de cotas reduz o risco de abandono do estudante em aproximadamente 7,3% $(= (0,927 - 1) * 100)$. Por outro lado, no que concerce a estratégia de dependência temporal, o qual apresenta a especificação mais completa da análise, aponta que a adesão a política afirmativa reduz em 17,7% $(= (0,823 - 1) * 100)$ o risco do aluno desistir do ensino superior naquele período.

Tabela 3.13 – Resultados do modelo de regressão de Cox.

Tratamento	**Modelo Básico**	**Dependência Temporal**
	(1)	**(2)**
Coeficiente	-0,076**	-0.194***
IC (95%)	(0,031)	(0.031)
Hazard ratio	0,927***	0,823***
IC(95%)	(0,867, 0,987)	(0,762, 0,885)
Covariadas		
Fixas	Sim	Sim
Variáveis	Não	Sim
N	17.550	17.436
$\mathbf{R}^2$	0.0003	0.170

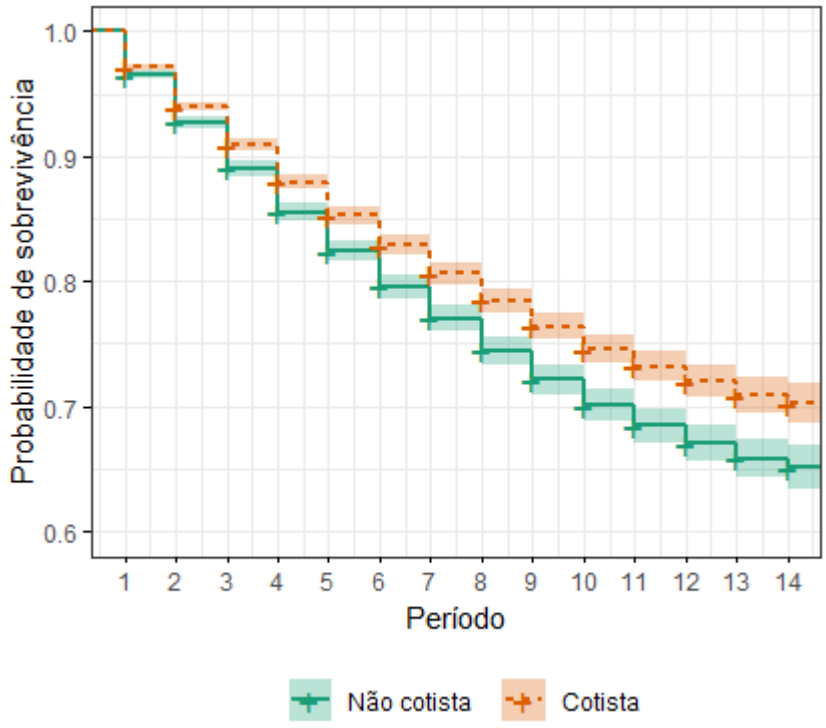
Fonte: Elaboração própria a partir dos microdados do STI/UFPB.
Nota: Níveis de significância: *10%, **5% e ***1%.

Como o modelo (1) desconsidera as mudanças das características dos estudantes que podem vir a afetar o desempenho dos discentes no período de estudo na universidade ao longo do tempo, principalmente no que se refere as variáveis do CRA relativo (que capta o nível de desempenho) como também em relação a variável trancado, consideradas variáveis que sinalizam o comportamento dos alunos e que variam de

semestre a semestre, pode-se afirmar que este superestima o efeito da política de cotas. No entanto, apesar das mudanças nas magnitudes dos coeficientes estimados nos dois cenários, pode-se observar a estabilidade dos resultados, visto que nos diferentes modelos, os coeficientes foram negativos e estatisticamente significantes a pelo menos 5%. Os resultados detalhados em relação as covariadas utilizadas nas estimações em cada um dos modelos de duração de Cox estão nas Tabelas A.3 e A.1 no apêndice.

Por meio do ajuste no modelo de risco proporcional de Cox exposto no modelo (2) na Tabela 3.13, a predição da sobrevivência dos alunos por período letivo na UFPB, entre 2011.1 e 2018.1, e por *status* de tratamento pode ser visualizada também na Figura 3.2. Esta apresenta a curva da sobrevida dos discentes da UFPB, supondo alterações comportamentais médias das variáveis explicativas contínuas e proporcional para as covariadas dicotômicas, relacionadas à mudança no comportamento na variável de tratamento. Ante isso, é presumível avaliar o padrão na ocorrência da "falha" (abandono) de interesse pelos diferentes grupos, controle (não cotista) e tratamento (cotista).

Figura 3.2 – Proporção de Sobrevivência dos alunos da UFPB por período e por status de cotas, 2010-2017.



Fonte: Elaboração própria a partir das estimativas da Tabela 3.13, modelo 3.7.
Nota: Intervalo de confiança representado pelas áreas sombreadas, supondo um nível de 95% de confiança.

Indo ao encontro das evidências iniciais encontradas nas Tabelas 3.6.1 e 3.6.2, a Figura 3.2 evidencia que a probabilidade de sobrevida dos alunos não cotistas é inferior aos alunos cotistas, isto é, o aluno ser cotista reduz o abandono no ensino superior

na UFPB, nos períodos entre 2011.1 e 2018.1. Visto também que não há intercessão entre os intervalos de confiança, pode-se inferir que a diferença entre as probabilidades de sobrevivência dos grupos de controle e tratamento são estatisticamente significativas. Esta evidência ratifica as conclusões encontradas por Mendes Junior, Souza e Waltenberg (2016), os quais afirmam que os alunos ingressantes através do sistema de reserva de vagas possuem maiores estímulos para persistir nos cursos. Em contraponto diverge de Sander (2004), pois este afirma que o abandono é maior entre os cotistas.

A correlação positiva entre o discente ser cotista e a redução do risco de abandono também pode ser visualizada ao analisar o comportamento da inclinação das curvas de sobrevivência nos dois grupos em estudo. A curva de probabilidade de sobrevida do grupo de alunos não cotistas é mais inclinada do que a curva que representa o grupo de alunos cotistas, sinalizando que o declínio da probabilidade de sobrevivência do grupo de tratamento ocorre a uma taxa inferior quando comparada ao decréscimo observado na curva do grupo de controle. Dessa maneira, considerando um dos objetivos iniciais desta pesquisa, verificar se a política de cotas gerou uma externalidade negativa na performance dos discentes, no que se refere a análise de abandono, as estimativas encontradas indicam que há um impacto positivo, reduzindo assim o abandono dos cursos na UFPB, nos períodos em estudo, 2011.1 e 2018.1.

Partindo ainda das predições do modelo de regressão de Cox, modelo (2) na Tabela 3.13, a Tabela 3.15 apresenta o efeito das cotas em termos das diferenças nas probabilidades de sobrevida entre os alunos tratados e não tratados, por período incremental. O que se pode observar é que após o primeiro período letivo, ou seja, após o primeiro semestre cursado na UFPB, a taxa de sobrevivência dos alunos cotistas e não cotistas encontram-se em um nível similar, entretanto os discentes do grupo de controle possuem um menor probabilidade de sobrevida ao serem comparados aos discentes do grupo de tratamento. Observa-se ainda que a diferença da probabilidade ao longo do tempo, por 15 períodos, foi estatisticamente significante a 5%.

Tabela 3.15 – Probabilidade de sobrevivência dos alunos da UFPB por período incremental e por *status* de cotas.

Tempo	Tratamento		Controle		Efeito
(Período)	Probabilidade	Erro-padrão	Probabilidade	Erro-padrão	(%)
1	97,12	0,120	96,52	0,147	0,603*
2	93,94	0,202	92,69	0,250	1,247*
3	90,88	0,278	89,04	0,345	1,840*
4	87,94	0,351	85,56	0,436	2,384*
5	85,27	0,420	82,42	0,523	2,856*
6	82,85	0,485	79,58	0,604	3,268*
7	80,63	0,548	77,00	0,683	3,628*
8	78,45	0,616	74,48	0,767	3,968*
9	76,42	0,687	72,14	0,853	4,272*
10	74,59	0,761	70,05	0,941	4,534*
11	73,16	0,831	68,43	1,023	4,730*
12	71,96	0,909	67,06	1,112	4,891*
13	70,88	1,010	65,85	1,225	5,030*
14	70,25	1,117	65,14	1,345	5,109*

Fonte: Elaboração própria a partir das estimativas da Tabela 3.13, modelo 3.7.
Nota: *p-valor<0,05 para teste de igualdade de diferença das probabilidades.

Por fim, ainda na Tabela 3.15, o referido efeito do programa cresce de forma contínua ao longo dos semestres. A adoção de 15 períodos justifica-se com base no período de 2011.1 a 2018.1 (o período 2011.1 é referente ao período 0, onde nenhum aluno abandonou), ou seja, dois períodos por ano. Uma vez que os dois grupos, tratamento e controle, estão sendo comparados inicialmente em dois períodos diferentes, 2010 onde não existia a política de cotas, e 2011 a qual já existia. No sétimo período, o que equivale a três anos e meio em estudo, o tempo médio de sobrevida dos discentes cotistas é de 3,6% superior aos discentes que não ingressaram na UFPB através do sistema de reserva de vagas. Quando se observa o décimo quarto semestre, a probabilidade de sobrevivência, em média, dos alunos cotistas equivale a 70%, já para os não cotistas é de 65%. Em suma, o programa de reserva de vagas tende a proporcionar oportunidade para os alunos que apresentam maiores chances de sobrevida na UFPB no longo prazo, visto que o efeito da probabilidade de sobrevivência no último período foi de 5%.

3.7 Conclusões

Tendo em vista as características do novo mecanismo de admissão na Universidade Federal da Paraíba (UFPB) implantado em 2011, observa-se que a política de ação afirmativa de cotas selecionou grupos de estudantes com comportamentos bastante heterogêneos, principalmente no que se refere a nota de entrada nesta instituição de ensino superior. Tal fato ilustra a necessidade de se compreender melhor as relações entre os discentes que ingressaram desta instituição por meio do sistema de reserva de vagas e o seu efeito sobre os indicadores educacionais. Dado que o objetivo da

política é a diversificação de alunos tanto com recortes sociais quanto étnico-racial, esta pesquisa culminou ir além desta análise. Partindo com o objetivo de verificar os efeitos diretos e indiretos da política de cotas adotada em uma universidade pública da Paraíba sobre as variáveis de desempenho dos alunos ao longo do curso.

Dessa maneira, este ensaio apresenta direcionamentos empíricos sobre o novo sistema da admissão da universidade, em que busca selecionar grupos que representem as minorias sociais, sobre os fatores de desempenho e abandono dos indivíduos. Como ressaltado, a relevância do estudo faz uma contribuição na literatura nacional brasileira, pois nos trabalhos já existentes não são explorados acerca da relação entre ação afirmativa de reserva de vagas e os fatores determinantes para uma análise de sobrevivência no ensino superior dos aluno cotistas, principalmente com a adoção do PSM integrado ao modelo de risco proporcional de Cox.

Usando-se diferentes instrumentos de pareamento não-experimentais, *Propensity Score Matching* (PSM), *Mahalanobis Distance Matching* (MDM) e *Classification Tree Analysis* (CTA), tendo como variável de resultado o nível de desempenho, medido pelo CRA relativo, o aluno ser cotista na UFPB arca com rendimentos médios significativamente menores em comparação aos rendimentos médios dos alunos ingressantes por ampla concorrência na UFPB. O impacto é maior quando a análise capta as melhores médias da distribuição do CRA relativo.

Complementando a literatura da área, em que há um *gap* de resultados entre os grupos de estudantes cotistas e não cotistas que compõem a universidade, principalmente com foco no comportamento heterogêneo quando a análise parte para as grandes áreas de conhecimento. No que se refere as estimações com controle por área de conhecimento e por cursos, com ênfase no PSM, os maiores impactos negativos se concentram nas áreas de linguística, letras e artes, ciências da saúde, engenharia e ciências humanas.

Tal fato pode justificar-se pela clara relação entre as diferenças de habilidades cognitivas entre os diferentes grupos que compõem o perfil de cada uma destas grandes áreas de conhecimento e as suas respectivas taxas de concorrências. Os cursos das áreas de ciências da saúde, engenharia e ciências humanas detêm as maiores taxas de concorrência, como também possuem as maiores diferenças nas notas de entradas entre os grupos de cotistas e não cotistas. Ou seja, nestas áreas contém os estudantes ingressantes que desempenharam um maior nível de desempenho para ingressar na UFPB.

Por outro lado, a área de linguística, letras e artes se atém a baixa concorrência dos cursos que a compõe, o que pode acarretar consequentemente uma maior concentração de alunos cotistas. Embora os grupos de cotistas e não cotistas não se depararam com uma competição elevada, não demandaram, portanto, um nível tão

elevado de esforço para adentrar no ensino superior na UFPB. Contudo, ainda fica claro a divergência no nível de desempenho quanto ao CRA relativo médio entre os grupos de discentes que compõem os cursos da área supracitada. O que se pode concluir é que nos cursos onde apresentaram os maiores *gap* de habilidades cognitivas entre os alunos, os cotistas apresentam os piores rendimentos no que tange a variável de desempenho acadêmico.

No que se refere ao indicador taxa de abandono, com base no modelo de risco proporcional de Cox, os resultados indicam uma associação positiva entre o sistema de cotas e o acréscimo de sobrevida dos estudantes que ingressaram na UFPB através da política. Embora os resultados deste estudo devam ser vistos cuidadosamente, devido a limitação imposta pela abordagem empírica adotada no que toca ao problema de fatores não observáveis variantes no tempo, observa-se a persistência dos alunos cotistas a partir da estimação da probabilidade de sobrevivência.

De forma geral, a implantação de uma política de ação afirmativa de reserva de vagas pode evidenciar possíveis influências nos resultados educacionais nos discentes da UFPB a partir de 2011. Ao criar relevantes grupos de alunos com características semelhantes entre si, mas diferentes entre os demais ingressantes, o novo sistema de cotas da UFPB afetou tanto o nível de desempenho dos discentes, quanto a taxa de sobrevida, reduzindo a probabilidade de abandono dos alunos beneficiados pela política neste grau de ensino.

Esta correlação inicialmente pode justificar-se pelo fato de que os estudantes que compõem o grupo de cotistas são oriundos de escolas públicas, em que na maioria das vezes suas famílias não tiveram como investir no desenvolvimento cognitivo da educação básica, herdando assim um déficit educacional que o acompanhou até o ensino superior. Assim, ao serem admitidos em uma universidade onde existe os mais diferentes graus de conhecimento, é possível observar um nível de desempenho superior ao do aluno cotista. O mesmo argumento social e econômico justifica a persistência dos alunos para a obtenção do título.

Essa discussão supracitada parte do pressuposto de que o discente conclua seu ensino superior. Por outro lado, o direcionamento da discussão pode seguir também por um outro contexto no qual o estudante cotista persista mais na universidade, mas, não obstante, demande de mais tempo para concluir. Ou seja, o efeito pode se transformar em algo negativo, onde os alunos cotistas têm um nível de rendimento inferior aos do não cotistas e ainda podem ser mais resistentes, entretanto, se essa sobrevivência resultar em uma menor taxa de conclusão e/ou mais tempo para se formar, ou até mesmo abandonar, isso implica em maiores custos para a instituição na formação dessa mão de obra. Em outras palavras, o sistema de cotas na UFPB pode aumentar os custos para a sociedade.

Por fim, ante toda esta exposição, o desempenho relativo auferido pelos cotistas serve como fundamentação para a manutenção da política de cotas no ensino superior, contudo, interligada a outras políticas direcionadas para melhorar a qualidade do ensino fundamental e médio das escolas públicas. Para que possa vir a influenciar de maneira agregada o desempenho dos alunos até que estes ingressem nas IES, evitando o hiato de conhecimento entre os alunos cotistas e não cotistas. Pois o objetivo da ação afirmativa de cotas não deve focar apenas em aumentar a diversidade de raça e socioeconômica entre os grupos de discentes da educação superior, mas também, de maneira conjunta, melhorar a qualidade dos indicadores educacionais dos mesmos.

4 Predição do risco de reprovação no ensino superior usando algoritmos de *Machine Learning*

4.1 Introdução

A retenção e a evasão escolar é uma realidade para muitos dos discentes nos cursos de graduações das universidades públicas no Brasil, considerado um problema complexo nas Instituições de Ensino Superior (IES), o que aumenta os custos de provisão das IES. Várias pesquisas mostram que esses problemas são universais e que devem envolver, para a sua solução, diferentes níveis de intervenção, desde aquelas em nível da família e do indivíduo até as relacionadas com os insumos escolares e diretrizes da educação (GOMES-NETO; HANUSHEK, 1994; LEON; MENEZES-FILHO, 2002; SAMPAIO *et al.*, 2011; DIOGO *et al.*, 2015).

As taxas de reprovação são para muitos autores, um dos maiores problemas no sistema de ensino brasileiro (JÚNIOR; FARIA; LIMA, 2012). De acordo com Souza *et al.* (2012), o discente ao evadir de um determinado curso pode ingressar em um novo, logo em seguida. Por outro lado, a deficiência no aprendizado pode acompanhar o indivíduo e acarretar novas reprovações, superestimando a evasão.

Nos últimos anos, no Brasil, é possível observar uma crescente oferta de cursos, pois de acordo com os dados relativos do Censo da Educação Superior, referente ao ano de 2015, divulgados pelo Instituto Nacional de Estudos e Pesquisa Anísio Teixeira (Inep), 2.368 IES[8] ofertaram 8 milhões de vagas, correspondente a 33 mil cursos de graduação, o que representa um incremento de 11% em relação as vagas ofertadas em 2010 [9]. Contudo, muito embora tenha se observado o referido aumento, boa parte dos alunos ingressantes não chegam a concluir os cursos. Segundo dados do Censo da Educação Superior de 2017, a taxa de conclusão nas instituições públicas gira em torno de 37%, e 36% nas privadas (Inep, 2018).

A evasão é caracterizada como uma interrupção no ciclo de estudos de um estudante, a qual pode gerar prejuízos significativos em vários âmbitos, tais como: acadêmico, sociais e econômicos. No que tange ao setor público, ao evadir dos seus cursos,

[8] Centros Universitários, Faculdades, Universidades, Institutos Federais (IFs) e Centros Federais de Educação Tecnológica (Cefets).

[9] Segundo informações do Censo da Educação Superior em 2010, 29.507 cursos de graduação ofertaram 6.379.299 vagas.

os discentes geram perdas diretas e indiretas, aumentando os custos da diplomação e reduzindo a taxa de retorno dos investimentos em educação. Já no setor privado, a evasão provoca perdas sociais importantes no que tange à redução do estoque de profissionais formados que poderiam aumentar a quantidade de mão-obra-qualificada no mercado e, assim, a produtividade da economia (SILVA-FILHO *et al.*, 2007).

Shirasu, Albuquerque *et al.* (2015), em seu estudo sobre os determinantes da evasão e repetência constataram que a falta de interesse em estudar e repetidas reprovações são os fatores com maiores influências nas taxas de evasão do discente. Para Gomes-Neto e Hanushek (1994), a reprovação, ocasionando repetências, e a evasão estão interligadas de maneira que a combinação entre elas tem sido identificada como uma das falhas do sistema educacional no Brasil, pois esta relação ratifica a ineficiência dos gastos públicos neste sistema. Desse modo, o problema da reprovação dos estudantes pode ser, por muitas vezes, um dos fatores determinantes da evasão.

A reprovação, por sua vez, pode ser considerada como uma consequência de um processo de aprendizado falho que ocorre ao longo de todo um semestre na universidade. É necessário identificar os fatores que influenciam o desempenho do aluno até o mesmo ser reprovado, como por exemplo, sob à ótica do discente: (i) tempo dedicado ao estudo; (ii) preferências do discente e a escolha do curso; (iii) má formação na educação básica; (iv) falta de interesse nas aulas; (v) insatisfação com o ensino. Já na perspectiva da instituição, os fatores que influenciam o status de reprovação podem ser: (i) currículo defasado; (ii) problemas de coordenação e alocação de recursos; (iii) corpo docente com pouca qualificação ou desmotivado, dentre outros.

Dada a abrangência da gama de fatores, do ponto de vista das IES, prever a reprovação acadêmica é necessário pois esta ação pode proporcionar medidas preventivas com o propósito de evitar novos resultados negativos no fim de mais um semestre. Especificamente, é um indicador que não só mede a aprendizagem e desempenho do aluno, mas também o seu impacto em outros desequilíbrios. Mas como prever de forma precisa o risco de reprovação? Quais são os principais fatores que fazem com que os estudantes sejam aprovados ou não? Como adotar medidas preventivas para que essa tendência seja reduzida? Qual é o papel dos gestores educacionais?

A identificação de discentes com maior risco de sofrer reprovação nas disciplinas com os maiores índices de retenção pode estar diretamente relacionada à necessidade de intervenções na educação no ensino superior, as quais buscam reduzir não apenas a retenção, mas, consequentemente, a evasão, para que assim possa evitar custos ocasionados em decorrência desse comportamento nas universidades. Nesse contexto, a análise preditiva pode auxiliar a ponderação entre benefícios e danos, com a finalidade de auxiliar gestores na formulação de políticas públicas direcionadas as intervenções preventivas, como por exemplo reforço escolar, acompanhamento pedagógico, curso

de férias, entre outros.

Dessa maneira, um dos grandes desafios é desenvolver uma estratégia eficiente, passível de operacionalização prática, que consiga prever o resultado dos discentes, de modo a permitir uma intervenção prévia de professores, coordenadores e outros responsáveis institucionais com o escopo de evitar ou minimizar problemas futuros de reprovação e de possível evasão. Uma vez que as IES armazenam dados acadêmicos e socioeconômicos dos seus estudantes, é possível realizar diversas análises na busca por padrões e características relacionadas com a condição de reprovação do alunado. Por sua vez, como o processo exige uma investigação baseada na extração de conhecimento em um extenso volume de dados, as diferentes técnicas de aprendizagem de máquina, *Machine Learning* (ML), apresentam-se como uma opção viável para realizar essa tarefa.

O interesse por modelos preditivos é cada vez mais crescente na sociedade moderna e a área de *data science*, ciência de dados, e de seu conjunto de ferramentas para modelar e compreender conjuntos de dados complexos, de forma autônoma e eficiente, vem ganhando destaque em uma série de aplicações na indústria e no setor de serviços. Essa modelagem abrange os mais diversos algoritmos, como as regressões logística e de *Least Absolute Shrinkage and Selection Operator* (LASSO), árvores de classificação, *boosting* e *support vector machines* (SVM) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013; KUHN; JOHNSON, 2013). Conforme Lantz (2013), o campo de estudo interessado no desenvolvimento de algoritmos de computador para transformar dados complexos em ação inteligente é conhecido como *Machine Learning*.

Assim, esta pesquisa tem como objetivo classificar, de maneira precoce, os discentes com risco de reprovação a partir da aplicação dos algoritmos de *Machine Learning*. Para tanto, foram utilizados registros administrativos e acadêmicos da Universidade Federal da Paraíba (UFPB), de 14 semestres letivos (2010 a 2016) e de mais de 8.500 matrículas em uma disciplina de alta retenção na mencionada instituição, cálculo diferencial e integral I.

Na UFPB, diversas graduações possuem em suas grades curriculares disciplinas da área de matemática, sendo para muitas uma base essencial para a formação do aluno (como nos cursos de matemática, física, ciência da computação, economia e engenharias). Devida à sua relevância, cabe-se prever os possíveis riscos que o fracasso ou insucesso dos discentes nesta disciplina possa vir a contribuir no processo de evasão nesta etapa de ensino[10].

Na literatura de economia da educação, muito se tem debatido acerca da evasão no ensino superior e suas principais causas, principalmente no âmbito da inferência

[10] Uma vez modelado e, posteriormente, testado e implementado no sistema acadêmico das instituições, o modelo de risco de reprovação realizado para a disciplina de cálculo I poderia ser expandido para outras disciplinas.

causal. Contudo, são escassos estudos do ponto de vista preditivo[11]. Costumeiramente, os trabalhos que analisam a evasão adotam uma abordagem estatística mais tradicional, como por exemplo Sampaio *et al.* (2011), preocupando-se nos fatores determinantes da variável de desfecho. Dessa maneira, pensando em contribuir para o desenvolvimento e redução dos problemas de reprovação e, consequentemente, de evasão, este estudo vai usar métodos tradicionais e algoritmos de ML (LASSO, k-vizinhos mais próximos, classificação naïve Bayes, árvore de decisão, entre outros) para fazer uma classificação preditiva mais precisa dos indivíduos com potencial problema de rendimento acadêmico.

Após esta seção introdutória, este ensaio contempla mais quatro Seções. A Seção 4.2 apresenta as variáveis que compõem a base de dados e algumas evidências iniciais das mesmas ao longo do tempo. Por sua vez, a Seção 4.3 descreve as estratégias empíricas que foram aplicadas referentes à todos os algoritmos, sejam eles dos modelos tradicionais de econometria (subseção 4.3.1) como também dos métodos abordados do aprendizado de máquina (subseção 4.3.2) e os critérios de seleção do modelo (subseção 4.3.3). Na Seção 4.4, de resultados, contém as subseções referentes a comparação (subseção 4.4.1), seleção (subseção 4.4.2) e avaliação (subseção 4.4.3) do modelo com melhor previsão, e por fim, a Seção 4.5 trata das conclusões do estudo.

4.2 Base de Dados e Descrição das Variáveis

As informações usadas nesta pesquisa fazem parte dos microdados oriundos da Superintendência de Tecnologia da Informação (STI) da Universidade Federal da Paraíba (UFPB) e contém características sobre os discentes que ingressaram nos cursos de graduação e que demandaram a disciplina de cálculo diferencial e integral I por semestre, no período de 2010 a 2016, como também suas notas do vestibular e características dos respectivos docentes. Logo, a base não está dividida por cursos, e sim por disciplina. Destaca-se que os estudantes tiveram sua identificação preservada.

A base de dados compõe informações de 5.426 discentes (62,7%) que foram reprovados na disciplina de cálculo diferencial e integral I e 3.233 (37,3%) alunos que foram aprovados no mesmo período. A primeira classe compõe os estudantes que não atingiram o nível suficiente de rendimento, ou seja, não tiveram a nota mínima suficiente passa ser aprovado, isto é, não obtiveram a média final igual 7 na disciplina no fim do período letivo, ou foram reprovados por falta. Por sua vez, a segunda classe é composta pelos discentes que foram aprovados com média final igual ou superior a 7.

[11] Modelos preditivos têm sido utilizados na literatura com muita frequência, mas aplicados em outras áreas, como finanças, comércio eletrônico e macroeconomia (KLEINBERG; MULLAINATHAN; RAGHAVAN, 2016; GOEL *et al.*, 2016; BJÖRKEGREN; GRISSEN, 2018).

Logo, para alcançar o objetivo proposto da pesquisa, a variável de resposta, que foi estimada nos métodos tradicionais de econometria e nos algoritmos de classificação de ML, é uma variável binária e assume 1 quando o discente apresenta o status de matricula reprovado em cálculo diferencial e integral I, e 0, caso contrário, ou seja, quando for aprovado. O banco de dados adotado para predizer o desempenho dos estudantes é composto pelas variáveis que detém informações dos discentes da UFPB nos semestres de 2010.1 a 2016.2, como também por outras características em nível dos docentes, da turma, do curso e do centro. A seguir, a Tabela 4.1 reporta as variáveis que compõem cada uma das dimensões adotadas deste estudo.

Tabela 4.1 – Descrição das Variáveis.

Dimensão	Variáveis	Descrição	Fonte
Discente	Nota Vestibular	Nota do vestibular total, variando de 0 a 1000.	STI/UFPB
	Nota Vest. Mat.	Nota do vestibular total na prova de matemática, variando de 0 a 1000.	STI/UFPB
	Casado	*Dummy*: casado assume 1, e 0, caso contrário.	STI/UFPB
	Migrante	*Dummy*: migrante assume 1, e 0, caso contrário.	STI/UFPB
	Raça	*Dummy*: Branco assume 1, e 0, caso contrário.	STI/UFPB
	Sexo	*Dummy*: Feminino assume 1, e 0, caso contrário .	STI/UFPB
	Idade Ingresso	Idade no semestre de ingresso, variando de 15 a 71 anos.	STI/UFPB
	Cotista	*Dummy*: Cotista assume 1, e 0, caso contrário.	STI/UFPB
	Período Ingresso	*Dummy*: 2º semestre assume 1, e 0, caso contrário.	STI/UFPB
	Forma de Ingresso	*Dummy*: Enem assume 1, e 0, caso contrário (PSS).	STI/UFPB
Docente	Tempo de Grad.	Calculada a partir do ano de conclusão da primeira graduação varia de 1 a 40 anos.	CNPq
	Doutorado	*Dummy*: Doutorado assume 1, e 0, caso contrário.	CNPq
	Publicação no Ano	*Dummy*: Publicação no ano assume 1, e 0, caso contrário.	CNPq
	Estrangeiro	*Dummy*: Estrangeiro assume 1, e 0, caso contrário.	CNPq
	Dedic. exclusiva	*Dummy*: Dedicação exclusiva assume 1, e 0, caso contrário.	STI/UFPB
	Sexo	*Dummy*: Feminino assume 1, e 0, caso contrário.	STI/UFPB
Curso	Local do Campus	*Dummy*: João Pessoa assume 1, e 0, caso contrário.	STI/UFPB
	EF do Curso	*Dummies* por curso.	STI/UFPB
	EF do Centro	*Dummies* por centro.	STI/UFPB
Turma	Turno	*Dummy*: Noturno assume 1, e 0, caso contrário.	STI/UFPB
	Carga Horária	*Dummy*: 90 hrs assume 1, e 0, caso contrário .	STI/UFPB
	Média N. Vestibular	Nota do vestibular total média, variando de 0 a 1000.	STI/UFPB
	Média N. Vest. Mat.	Nota do vestibular total média na prova de matemática, variando de 0 a 1000.	STI/UFPB
	Taxa de Cotista	Percentual de discentes cotista na turma.	STI/UFPB

Fonte: Elaboração própria a partir dos microdados do STI/UFPB e Plataforma *Lattes* do CNPq.

- Nível do Discente: foram utilizadas características pré-ingresso ao ensino superior – (i) *Nota Vestibular* e (ii) *Nota Vestibular Matemática*. Como as médias ao final de cada semestre na disciplina são utilizadas como principal e único critério

para a reprovação/aprovação do discente, e como o fracasso na disciplina de cálculo pode estar relacionada também com alguma deficiência agregada no conhecimento geral oriundo do ensino básico, optou-se em selecionar variáveis que apresentam o desempenho no vestibular como um todo, e de maneira mais específica, o desempenho na nota em matemática. Como também essas variáveis podem capturar o desempenho dos estudantes que entram no primeiro semestre do ano, uma vez que quem obtiver as melhores médias é classificado para a primeira chamada e formar a primeira turma do curso.

- Ainda em nível do discente, foram adotados também fatores pessoais como (iii) *Idade* e *dummies* relacionadas as informações se o aluno é (iv) *Casado*, (v) *Migrante*, sua (vi) *Raça* e o (vii) *Sexo*. Como também características ligadas ao ingresso na universidade, como se o aluno é (viii) *Cotista*, captando as informações socioeconômicas; o (ix) *Período de Ingresso*, se foi no 1º ou 2º semestre, e; (x) *Forma de Ingresso*, Enem ou Vestibular (PSS). No que se refere ao período, esta variável *dummy* busca analisar se o discente entrar no ensino superior apenas no segundo semestre poderá impactar negativamente no seu desempenho e provocar a reprovação. Pois, pressupõem-se que ao passar o primeiro semestre de forma ociosa, o discente poderá entrar na universidade e apresentar déficits acarretados devido a interrupção nos estudos.

- Nível do Docente: algumas das variáveis que compõem o conjunto de características dos docentes da UFPB no período em estudo foram construídas a partir das informações disponibilizadas e coletadas na Plataforma *Lattes* do Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq)[12]. Foram construídas as variáveis: (i) *Tempo de Graduação*, a qual busca captar a relação entre o período de experiência em sala de aula como também a idade do docente e a sua influência no desempenho dos seus alunos. Se o docente possui (ii) *Doutorado* ou não, independente da área de formação. Sendo esta uma variável que muda no tempo, para os casos de doutorado em andamento, pois só conta a partir do ano que o discente conclui esta etapa de ensino.

- (iii) *Publicação no Ano* em periódicos, captando a produtividade cientifica, uma vez que quanto mais tempo dedicado a publicação menos tempo dedicado à sala de aula o docente possuirá; se o professor é (iv) *Estrangeiro* ou não, baseado no país onde a universidade a qual foi concluída a sua graduação situa-se. Ao adotar a nacionalidade do docente busca-se identificar se o professor estrangeiro pode ser um problema na transmissão oral do conhecimento da disciplina, visto que está já é considerada um assunto de difícil assimilação entre os discentes; se possui (v) *Dedicação Exclusiva* ao curso, e; por fim, o (vi) *Sexo* do professor,

[12] Ver <http://lattes.cnpq.br/>.

pressupondo que a docente (do sexo feminino) possa vir a ter ou demonstrar metodologias de ensino mais adequadas de acordo com a dificuldade que os alunos se deparam nesta disciplina.

- Nível do Curso e Centro: (i) *Local do Campus*, se é em João Pessoa ou não; (ii) *Efeito Fixo do Curso* e (iii) *Efeito Fixo do Centro*. A disciplina de cálculo diferencial e integral I compõem a grade curricular de 21 cursos de graduações da UFPB, que fazem parte de 6 centros de ensino. Foram construídas 21 *dummies* referentes aos cursos: (1) C. da Computação; (2) C. Atuariais; (3) C. Econômicas; (4) C. Ambiental; (5) Engenharia Civil; (6) E. de Alimentos; (7) E. da Computação; (8) E. de energias renováveis; (9) E. de Materiais; (10) E. de Produção Mecânica; (11) E. de Produção; (12) E. Elétrica; (13) E. Mecânica; (14) E. Química; (15) Estatística; (16) Física; (17) Matemática (Omitida); (18) Matemática (Licenciatura); (19) Matemática Computacional; (20) Química, e; (21) Química Industrial. Como também 6 *dummies* dos centros: (1) C. de Ciências Aplicadas e Educacional (CCAE); (2) C. de C. Exatas e da Natureza (CCEN) (Omitida); (3) C. de C. Sociais e Aplicadas (CCSA); (4) C. de Energias e Alternativas e Renováveis (CEAR); (5) C. de Informática (CI), e; (6) C; de Tecnologia (CT).

- Nível da Turma: (i) *Turno*, é uma *dummy* que assume 1 se o discente cursar a disciplina no período noturno e 0, caso contrário; (ii) *Carga Horária*, se a disciplina possuir 90 hrs a *dummy* assume 1, caso contrário (60 hrs) assume 0. Foram construídas variáveis em nível da turma a partir de variáveis individuais já descritas anteriormente, como: (iii) *Média na Nota do Vestibular*, (iv) *Média na Nota do Vestibular em Matemática* e (v) *Taxa de Cotista*.

É válido destacar que até o ano de 2012, a forma de ingresso na UFPB era por meio do Processo Seletivo Seriado (PSS), onde através deste sistema o estudante era submetido a uma prova dividida por grande área de conhecimento e uma redação. A partir de 2013, a universidade adotou o Exame Nacional do Ensino Médio (Enem) e o Sistema de Seleção Unificada (SiSU) para selecionar os discentes por meio de um cruzamento de informações entre demandantes e ofertantes de vagas no ensino superior. Logo, em meio a transição do sistema de notas dos exames de seleções (PSS e ENEM) no período em análise, 2010 e 2016, foi feita uma padronização das mesmas, como também um cruzamento desta base (que contém as informações dos discentes no processo de seleção e suas notas) com a base que o acompanha ao logo do seu percurso dentro da UFPB.

A justificativa pela adoção da disciplina de cálculo baseia-se por esta ser a disciplina mais demandada em consequência por ser a que apresenta o maior índice de reprovação na UPFB. Nesta universidade, em muitos casos, as disciplinas não são

ofertadas por curso, mas sim, por departamento. Com algumas exceções como no caso do curso de economia, onde o próprio departamento oferta. Os alunos que compõem uma turma podem ser dos mais diversos cursos que têm essa disciplina em sua grade, sendo considerada assim uma turma bem heterogênea, pois o perfil de um estudante do curso de ciências atuariais pode ser diferente de um aluno que ingressou no curso de estatística, por exemplo. Por isso a necessidade também de adotar fatores que possam captar o comportamento em nível da turma nos modelos.

Outra importante justificativa consiste que a reprovação nesta disciplina, por ser presente logo nos primeiros semestres e por ser uma disciplina base dos cursos (pré-requisito de outras disciplinas), principalmente nas graduações em engenharias, pode ter sérias consequências como a retenção (atraso) do discente, como também ser o principal fator que estimule a sua evasão do curso. Dessa maneira, de uma forma mais simples, a aplicação das técnicas de ML busca prever (identificar) os possíveis riscos de reprovações dos discentes para que antes mesmo do início do semestre os coordenadores já tenham o score de risco de reprovação e possam intervir por meio de medidas preventivas.

4.2.1 Comportamento dos discentes por Ano, por Grande Área de Conhecimento e por Curso, na UFPB

Aqui serão expostas algumas evidências iniciais sobre o comportamento da variável de resposta, reprovado ou aprovado, e das variáveis que representam o acumulo de conhecimento adquirido ao longo da vida do discente antes do seu ingresso no ensino superior, média do vestibular geral e média do vestibular em matemática, como também a quantidade de discentes e sua classificação e a média final da disciplina de cálculo diferencial e integral I. Muito embora esta última não seja adotada no modelo, foi exposta apenas a título de comportamento e discussão da média final da referida disciplina entre os grupos. Assim, a Tabela 4.3 apresenta os testes de médias e os intervalos de confiança de cada umas das variáveis anteriormente citadas por status de matrícula separadas por ano, de 2010 a 2016.

A partir da análise anual das matrículas na disciplina de cálculo na Tabela 4.3, pode-se observar que os estudantes foram mais constantes na classe dos reprovados ao longo do período, mesmo apresentando um comportamento heterogêneo na série. De 2010 a 2016 há em torno de 63% dos discentes que cursaram a referida disciplina e obtiveram um resultado negativo ao final do semestre dentro de cada ano, especificamente. Deixando claro portanto que a reprovação na disciplina de cálculo diferencial e integral I é um sério problema que merece atenção na UFPB.

Tabela 4.3 – Evidências iniciais da variável de resultado, por status de matrícula, dos alunos da UFPB, 2010 a 2016 - Testes de médias e intervalo de confiança*.

Ano	Reprovado	Alunos	Média Idade	Média Vestibular	Média Vest. Matemática	Média final Disciplina
2010	Não	546	20,5 20,0 - 20,8	562,3 557,1 - 567,6	588,6 580,8 - 596,4	7,27 7,15 - 7,39
	Sim	929	21,8 21,4 - 22,1	531,4 527,4 - 535,5	547,5 541,6 - 553,5	0,80 0,72 - 0,89
2011	Não	579	20,5 20,1 - 20,8	577,4 571,8 - 582,8	602,6 594,8 - 610,3	7,20 7,09 -07,32
	Sim	1.117	21,8 21,5 - 22,1	533,3 529,1 - 537,5	543,8 538,3 - 549,2	0,86 0,79 - 0,94
2012	Não	650	19,9 19,6 - 20,2	580,6 575,5 - 585,7	617,0 609,1 - 624,8	7,4 7,29 - 7,52
	Sim	1.204	22,2 21,9 - 22,5	535,9 532,0 - 539,8	549,6 544,1 - 555,1	0,84 0,77 - 0,91
2013	Não	670	19,9 19,6 - 20,2	592,4 586,5 - 598,4	635,6 626,8 - 644,4	7,39 7,29 - 7,50
	Sim	1.099	22,3 21,9 - 22,7	546,4 541,4 - 551,3	568,8 561,7 - 575,7	0,80 0,73 - 0,88
2014	Não	632	20,0 19,6 - 20,4	629,3 624,2 - 634,5	675,2 667,8 - 682,6	7,34 7,22 - 7,45
	Sim	999	22,7 22,3 - 23,1	601,3 597,4 - 605,2	630,3 624,5 - 636,1	0,58 0,51 - 0,65
2015	Não	557	20,1 19,8 - 20,5	636,6 631,7 - 641,5	666,8 658,2 - 675,3	7,5 7,41 - 7,65
	Sim	840	23,5 22,9 - 24,0	598,6 594,5 - 602,7	605,8 598,9 - 612,7	1,1 1,00 - 1,20
2016	Não	334	20,5 19,9 - 21,1	599,2 582,3 - 616,0	632,5 607,4 - 657,5	7,61 7,45 - 7,78
	Sim	442	23,0 22,38 - 23,6	553,5 538,6 - 568,4	577,3 554,3 - 600,3	1,19 1,06 - 1,32

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.
Notas: (1) Intervalos com 95% de confiança para as médias calculadas. (2) *Todos os intervalos de confiança apresentam um teste **t** de $0,000$.

Embora esteja decrescendo ao longo dos semestres e, consequentemente, dos anos, a proporção dos reprovados ainda apresenta uma disparidade quando comparada a proporção dos aprovados. É possível observar uma taxa de reprovação acima de 60% em todos os anos até 2015, onde a partir daí a quantidade de alunos com resultados negativos cai em 2016 para 57%, reprovando 442 alunos que cursaram cálculo neste ano, o qual tem também menor volume de discentes nessa classe neste período.

Por sua vez, um dos pressupostos adotados neste ensaio baseia-se que os discentes ingressantes em cursos quem tem a disciplina em estudo em sua grade deveriam apresentar uma maior afinidade em matemática. Por esse motivo uma das variáveis adotadas na análise é a média do vestibular em matemática, como também uma variável que mede o acúmulo de conhecimento geral do discente, a média do vestibular. Antes da análise das médias é importante destacar que as informações

apresentadas podem ser comparadas entre os grupos pois os intervalos afirmam que as médias entre as classes são iguais com sobreposição de 95% de confiança.

Como pode ser visualizado ainda na Tabela 4.3, tanto a média geral como a específica em matemática no vestibular do grupo dos reprovados foram inferiores em comparação aos aprovados em todos os anos. Evidenciando assim, uma possível relação entre essas variáveis, principalmente no que tange o conhecimento específico, onde ocorreu maiores disparidades nas médias entre os grupos. Tornando claro que o déficit no conhecimento adquirido no ensino básico pode-se tornar uma grave herança no ensino superior.

Ante as elevadas taxas de reprovação, Sampaio *et al.* (2011) destaca uma preocupação condizente a esta pesquisa, onde afirma que a evasão acarreta não só vagas ociosas, mas principalmente perdas e elevação nos custos. O que fica claro na discussão já debatida anteriormente é que a reprovação pode ser considerada não só um determinante da retenção, mas também pode vir a provocar estímulo a evasão do discente, visto a dificuldade em ser aprovado em determinadas disciplinas, como por exemplo, em cálculo.

Sendo este um dos principais motivos, e embora ainda muito debatido, torna-se relevante pesquisar a reprovação como principal motivo da evasão dentro da IES (pois a evasão depende de outras intervenções, como por exemplo, familiar e pessoal), onde a partir de uma evidência formal a universidade possa tomar medidas preventivas que tenham como objetivos atenuar não apenas o déficit no ensino, mas também reduzir os custos e tornar o investimento público mais eficiente.

A Tabela 4.5 expõem a evolução da variável de resultado, por status de matrícula, cursos e centros ao longo de 2010 a 2016 na UFPB. Os cursos que tiveram as maiores disparidades entre as classes de matrículas no final de cada período são os do CCEN: física e matemática, onde mais de 70%, em média, da turma foi reprovada em ambos os cursos neste período. O que acaba sendo uma problemática, pois suspeita-se que os discentes que demandam os referidos cursos tenham habilidades em disciplinas fortemente relacionadas a eles, como é o caso de cálculo.

No que concerne os cursos do CT, em sua grande maioria a quantidade dos discentes reprovados superam os aprovados. Por outro lado, embora com uma quantidade elevada de reprovados, os cursos de engenharia civil, engenharia mecânica, engenharia química e alguns anos do curso de engenharia ambiental tiveram um maior índice de aprovações nesse período. Na busca por um análise mais aprofundada do comportamento dessas variáveis, assim como foi apresentado na Tabela 4.3, a Tabela B.1 no apêndice apresenta os testes de médias e intervalos de confiança detalhados dessa vez pelas grandes áreas de conhecimento (ciências exatas e da terra, ciências sociais aplicadas e engenharias) para cada ano em estudo.

Tabela 4.5 – Evolução da variável de resultado, por status de matrícula e cursos, dos alunos da UFPB, 2010 a 2016.

Centro	Curso	Reprovado	Alunos						
			2010	2011	2012	2013	2014	2015	2016
CCAE	Matemática (Licenciatura)	Não	44	21	36	22	41	18	13
		Sim	77	83	68	57	51	19	21
CCEN	Estatística	Não	14	11	-	14	11	5	22
		Sim	34	57	43	33	30	33	17
	Física	Não	29	52	22	34	30	32	27
		Sim	106	97	107	91	106	65	76
	Matemática	Não	56	45	29	46	32	36	21
		Sim	111	131	140	115	93	76	67
	Química	Não	19	12	8	28	26	14	16
		Sim	79	81	60	38	36	29	23
CCSA	C. Atuariais	Não	-	14	35	32	34	29	10
		Sim	-	99	107	99	91	101	72
CEAR	E. de Energias Renováveis	Não	-	-	38	37	38	31	14
		Sim	-	-	64	49	43	45	13
	E. Elétrica	Não	42	36	39	40	41	32	26
		Sim	44	43	32	46	36	30	6
CI	C. da Computação	Não	37	46	32	24	38	48	18
		Sim	89	65	76	76	59	31	19
	E. da Computação	Não	-	41	68	35	33	37	27
		Sim	-	56	28	55	47	46	9
	Matemática Computacional	Não	-	-	24	10	13	5	7
		Sim	-	-	71	85	69	66	24
CT	E. ambiental	Não	35	47	19	54	33	26	10
		Sim	33	35	75	37	43	45	21
	E. Civil	Não	66	78	75	68	48	56	27
		Sim	39	19	24	27	49	31	4
	E. de Alimentos	Não	32	10	17	15	27	19	8
		Sim	53	97	80	73	53	48	20
	E. de Materiais	Não	21	30	26	29	36	16	18
		Sim	77	50	69	59	37	48	6
	E. de Produção Mecânica	Não	21	28	24	25	24	17	12
		Sim	37	44	36	40	28	18	5
	E. de Produção	Não	12	16	12	22	13	14	7
		Sim	18	22	23	20	23	24	8
	E. Mecânica	Não	70	35	56	55	58	62	31
		Sim	40	64	47	45	36	30	6
	E. Química	Não	33	33	70	55	34	46	20
		Sim	56	44	19	29	44	32	7
	Química Industrial	Não	15	24	20	25	22	14	-
		Sim	36	30	35	25	25	23	17

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.
Nota: As linhas do curso de C. Econômicas foram omitidas por ter valores apenas para o status de reprovado em 2016.

Corroborando com algumas dessas evidências iniciais, em sua pesquisa, Diogo *et al.* (2015) afirmam que os cursos com as maiores taxas de reprovações, frequência insuficiente e evasão nas universidades públicas brasileira foram: zootecnia, engenharia de alimentos, ciência e tecnologia alimentar, ciências da computação, química,

engenharia elétrica, engenharia eletrônica, física, engenharia de produção e sistemas e matemática. No que se refere a evasão dos estudantes por disciplinas isoladas, apontam que as pertencentes à grande área de conhecimento das ciências exatas, tais como cálculo, matemática, álgebra linear, física, química e geometria analítica, são as mais frequentes de reprovação.

O que se pode inferir até o momento é que os estudantes recém ingressos no ensino superior se deparam com um nível de exigência diferente do ambiente o qual estavam acostumados no ensino médio. Como também, deparam-se com diversas disciplinas complexas em que seu conhecimento agregado do ensino básico, na maioria dos casos, não se encontra preparado para enfrentar as novas dinâmicas do ensino superior. Dessa maneira, torna-se necessário que os gestores ofereçam o suporte para que as novas demandas a cada início de semestre por essas disciplinas não se configurem em reprovações ao final deles. É preciso então que a IES preocupe-se nas maneiras de reduzir o distanciamento entre o ensino médio e o superior, e não apenas em analisa-lo.

4.3 Estratégia Empírica

Nesta seção serão apresentadas as abordagens empíricas dos modelos tradicionais de econometria, como também os algoritmos do *Machine Learning* (ML), aprendizado de máquina, evidenciando algumas limitações dos modelos microeconométricos convencionais ante aos problemas de previsões, e ressaltando como as técnicas do ML podem contornar essas limitações.

Segundo Wooldridge (2006), econometria é uma ciência que objetiva extrair conhecimento e informações econômicas a partir de uma base de dados. De outro modo, a partir de um conjunto de ferramentas estatísticas, busca entender a relação entre as variáveis econômicas por meio da aplicação de modelos matemáticos. Mais tecnicamente, estimar as relações, testar as teorias, implementar e avaliar o impacto das políticas e predizer resultados.

Por outro lado, o aprendizado de máquina, ML, é a ciência onde os computadores aprendem a realizar as tarefas sem serem programados explicitamente para isso. De maneira mais técnica, de acordo com Mitchell (1997), o ML é quando uma máquina, através de uma experiência E, aprimora sua habilidade em uma tarefa T, seguindo uma métrica de performance P. Para Goodfellow, Bengio e Courville (2016) e Gal (2016), o ML é considerado um avanço na área da inteligência artificial, pois permite que a máquina adquira seu próprio conhecimento ao extrair padrões a partir da base de dados.

Apesar de ser clara a diferença operacional e filosófica entre as duas caixas de

ferramentas, ambas compartilham algumas bases teóricas da modelagem estatística, como por exemplo, o princípio da máxima verossimilhança (LANTZ, 2013; IZBICKI; SANTOS, 2018), como também alguns modelos como a regressão logística e linear (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013). Dessa maneira, essas diferenças são, na verdade, sutis. Em econometria, o foco baseia-se em estabelecer conclusões estatisticamente significantes e válidas, encontrar padrões na base de dados e inferi-los de maneira informativa (WOOLDRIDGE, 2010; GUJARATI; PORTER, 2011). Já no ML, o interesse consiste em gerar boas previsões sob a presença de restrições na máquina (LANTZ, 2013).

Diferentemente dos modelos de econometria, que buscam distribuição normal em torno da média dos estimadores, no aprendizado de máquina o foco ocorre na minimização do erro quadrático médio (ATHEY; IMBENS, 2017). Isto significa que os algoritmos de ML normalmente pagam o custo da presença de viés nas suas estimativas a fim de obter o melhor ajustamento no processo preditivo. James *et al.* (2013) ainda destacam que em econometria outro importante foco na construção do modelo refere-se a inferência, isto é, entender a relação entre as variáveis. Já nos modelos de ML, a preocupação é no poder preditivo, ou seja, prever uma resposta de interesse (variável de saída) a partir das variáveis de entradas.

O ponto central para Mullainathan e Spiess (2017) é que os modelos de aprendizagem de máquina não apenas fornecem novas ferramentas para as análises, mas também resolvem tipos de problemas diferentes. Na verdade, o sucesso do ML em tarefas de inteligência artificial é em grande parte devido à sua capacidade de descobrir estruturas complexas que não foram especificadas com antecedência, isto é, descobrem padrões generalizáveis. São capazes de encaixar formas funcionais complexas e muito flexíveis nos dados, além de encontrar funções que funcionam bem fora da amostra.

Embora que na grande maioria dos estudos em economia, as aplicações empíricas giram em torno da estimação de parâmetros, ou seja, produzir boas estimativas dos parâmetros, β, que fundamentam a relação entre Y e X. É importante destacar que os algoritmos de ML não foram desenvolvidos para essa finalidade. Por exemplo, mesmo quando esses algoritmos produzem coeficientes de regressão, as estimativas raramente são consistentes (MULLAINATHAN; SPIESS, 2017).

Para Athey (2018) a principal vantagem do ML em relação a econometria tradicional é que enquanto nesta abordagem o pesquisador escolhe apenas um modelo baseado em princípios estruturais e estima apenas uma vez, onde o foco está na estimativa de um modelo causal. A análise empírica do ML, por sua vez, consiste, além de um conjunto de algoritmos que constroem um "*tuning*" entre os métodos, na escolha do melhor desempenho, ou seja, o melhor ajuste, permite ainda que os pesquisadores sejam sistemáticos e descrevam todo o processo pelo qual o modelo foi

selecionado.

4.3.1 Modelos Tradicionais

Por a variável de resposta, ou regressando, adotada nesta pesquisa se tratar de uma variável qualitativa, esta subseção trata apenas dos modelos tradicionais de regressão de resposta binária, a saber: Modelos de probabilidade linear, *Logit* e *Probit*. Como um discente reprova ou não uma determinada disciplina, trata-se de uma decisão do tipo *sim* ou *não*, só podendo assumir dois valores, 1 se o estudante reprovou a disciplina e 0 caso contrário. Portanto, sendo considerada uma variável dicotômica ou binária (ALDRICH; NELSON; ADLER, 1984; LIAO, 1994; WOOLDRIDGE, 2010; GUJARATI; PORTER, 2011).

Mas, antes de se iniciar as descrições dos modelos, torna-se importante destacar a fundamental diferença entre os modelos de regressão quando a variável de reposta, Y, é uma variável quantitativa e quando é de natureza qualitativa. De acordo com Gujarati e Porter (2011), quando Y é quantitativo, o objetivo do modelo é estimar o valor esperado, dado o vetor de regressores, $\mathbf{X}$, ou seja, $E(Y_i|X_{1i}, X_{2i}, ..., X_{ki})$. Por outro lado, para os casos em que Y é qualitativo, o modelo busca encontrar a probabilidade de que o evento ocorra, como o aluno reprovar ou não uma determinada disciplina. Por isso, estes últimos são, por muitas vezes, denominados de modelos de probabilidade.

4.3.1.1 Modelo de Probabilidade Linear

Supondo a possibilidade de estimá-lo por Mínimos Quadrados Ordinários (MQO), considere o seguinte modelo de regressão linear, em que a variável de resposta é binária, reprovar ou não. Neste caso, a expectativa condicional de Y_i dado o vetor $\mathbf{X}$, $E(Y_i|\mathbf{X})$, pode ser interpretada como sendo a probabilidade de um discente reprovar uma disciplina cujo o vetor das variáveis explicativas é $\mathbf{X}$. De acordo com Wooldridge (2010), a principal justificativa para a denominação deste modelo ser modelo de probabilidade linear pode ser explicado supondo $E(u_i|\mathbf{X}) = 0$, para obter estimadores não tendenciosos, e obter:

$$E(Y_i|\mathbf{X}) = \beta_1 + \beta_2\mathbf{X} \tag{4.1}$$

ou seja, se P_i = probabilidade de que $Y_i = 1$, a ocorrência do evento, e $(1 - P_i)$ = a probabilidade de que $Y_i = 0$, caso em que o evento não ocorre, a variável Y_i segue uma distribuição de probabilidade de *Bernoulli*.

Pela definição de expectativa matemática:

$$E(Y_i) = 0(1 - P_i) + 1(P_i) = P_i \tag{4.2}$$

Igualando as Equações 4.1 e 4.2, tem-se:

$$E(Y_i|\mathbf{X}) = \beta_1 + \beta_2\mathbf{X} = P_i \tag{4.3}$$

Assim, o modelo de regressão com uma variável dependente binária é denominado de modelo de probabilidade linear (MPL) porque a probabilidade de resposta é linear nos parâmetros β_j. É válido destacar que como Y_i só pode assumir dois valor, β_j não pode ser interpretado como uma variação em Y_i devido ao aumento de uma unidade em X_i, mantendo fixo as demais variáveis, pois Y_i somente muda de um para zero ou o inverso (WOOLDRIDGE, 2010).

Dessa maneira, β_j mensura a mudança na probabilidade de sucesso quando X_i muda, mantendo os outros fatores constantes. De fato, a expectativa condicional do modelo pode ser interpretada como a probabilidade condicional de Y_i (GUJARATI; PORTER, 2011). Como a probabilidade de P_i deve se situar entre 0 e 1, segue a seguinte restrição: $0 \leq E(Y_i|\mathbf{X}) \leq 1$, isto é, a expectativa condicional deve estar nesse intervalo.

Contudo, a aplicação desse modelo pode vir a apresentar vários problemas: (i) Ausência de normalidade dos termos de erro, u_i; (ii) variâncias heterocedásticas dos termos de erro; (iii) Impossibilidade de satisfazer a restrição $0 \leq E(Y_i|\mathbf{X}) \leq 1$, e: (iv) O valor de R^2 como medida de qualidade do ajustamento é questionável. Para um estudo mais aprofundado, ver Wooldridge (2010) e Gujarati e Porter (2011).

4.3.1.2 Modelo *Logit*

De acordo com Wooldridge (2010), modelos *logit* e *probit* compensam os problemas encontrados do MPL, contudo, têm como desvantagem a dificuldade de interpretá-los. A desvantagem fundamental do MPL é que ele pressupõe que $P_i = E(Y = 1|X)$ aumenta linearmente com X, ou seja, o efeito incremental de X permanece constante.

Para Aldrich, Nelson e Adler (1984) e Gujarati e Porter (2011), é necessário um modelo de probabilidade que tenha duas características específicas: (i) Na médida que X_i aumenta, $P_i = E(Y = 1|X)$ aumenta, mas que não saia da faixa 0 e 1, e; (ii) A relação entre P_i e X_i é não linear, ou seja, aproxima-se de zero a taxas cada vez menores na medida que X_i se reduz, e se aproxima de um a taxas cada vez menores quando X_i aumenta muito.

O modelo *logit* tem uma Função de Distribuição Acumulada (FDA) do tipo logística. O MPL adotado foi:

$$P_i = E(Y = 1|\mathbf{X}) = \beta_1 + \beta_2\mathbf{X} \tag{4.4}$$

Segue uma nova representação dessa relação:

$$P_i = E(Y = 1|\mathbf{X}) = \frac{1}{1 + e^{-(\beta_1 + \beta_2 \mathbf{X})}} \tag{4.5}$$

Gujarati e Porter (2011) sugere reescrever a Equação 4.5 como:

$$P_i = \frac{1}{1 + e^{-Z_i}} = \frac{e^Z}{1 + e^Z} \tag{4.6}$$

em que $Z_i = \beta_1 + \beta_2 \mathbf{X}$. Logo, a Equação 4.6 representa a Função de Distribuição Logística (acumulada) (CRAMER, 1991; MADDALA, 1991). Dessa forma, é fácil observar que, como Z_i varia entre $-\infty$ e $+\infty$, P_i varia entre 0 e 1 e também se relaciona de modo não linear com Z_i, ($\mathbf{X}$), a nova forma atende, portanto, as duas exigências citadas anteriormente. Contudo, fica claro um problema de estimação, onde P_i é não linear apenas em $\mathbf{X}$, mas também é não linear nos parâmetros β_j.

A solução apresentada por Wooldridge (2010) e Gujarati e Porter (2011) é linearizar a equação nos parâmetros para que se possa estimar por MQO. Logo, se P_i é a probabilidade do discente reprovar uma disciplina, e é dada pela Equação 4.6, então, $(1 - P_i)$, a probabilidade do estudante não reprovar é:

$$1 - P_i = \frac{1}{1 + e_i^Z} \tag{4.7}$$

Pode ser reescrito por:

$$\frac{P_i}{1 - P_i} = \frac{1 + e_i^Z}{1 + e^{-Z_i}} \tag{4.8}$$

onde $\frac{P_i}{1-P_i}$ é a razão de chances a favor da reprovação do discente, isto é, a razão da probabilidade de que um aluno reprove na disciplina contra a probabilidade de que não reprove. Aplicando o logaritmo natural na Equação 4.8, tem-se:

$$L_i = ln\left(\frac{P_i}{1 - P_i}\right) = Z_i = \beta_1 + \beta_2 \mathbf{X} \tag{4.9}$$

Logo, o logaritmo da razão de chances, L, passa a ser linear não apenas em $\mathbf{X}$, mas também nos parâmetros, sendo denominado de modelo *logit*.

4.3.1.3 Modelo *Probit*

Assim como no modelo *logit*, para explicar o comportamento de uma variável dependente binária precisa-se aplicar uma função de distribuição acumulada de forma adequada. Em alguns casos, a FDS normal é a mais adequada, e o modelo que emerge

dela é o modelo *probit*. O modelo descrito a seguir foi formulado por McFadden *et al.* (1974), com base na perspectiva da escolha racional ou teoria da utilidade.

A decisão do i-ésimo discente reprovar ou não uma disciplina depende de um índice de utilidade observável, conhecido como variável latente I_i, que é determinado por um vetor de variáveis explicativas $\mathbf{X}$, de forma que quanto maior o valor de I_i, maior a probabilidade de que o aluno reprove. O referido índice é expressado por:

$$I_i = \beta_1 + \beta_2 \mathbf{X} \tag{4.10}$$

Gujarati e Porter (2011) destaca que o índice (não observável) se relaciona com a decisão do estudante reprovar ou não de forma que é razoável supor que haja um limiar ou nível crítico dó índice, I_i^*, de modo que se I_i for maior que I_i^*, o aluno reprovará a disciplina, caso contrário, não.

O limiar I_i^* não é observável como o I_i, mas, supondo que aquele se distribui normalmente, tendo a mesma média e variância, se torna possível estimar a Equação 4.10 e obter algumas informações. Dessa maneira, dado o pressuposto de normalidade, a probabilidade de que I_i^* seja menor ou igual a I_i pode ser calculada a partir da FDA normal padronizada (WOOLDRIDGE, 2010; GUJARATI; PORTER, 2011), como:

$$P_i = P(Y = 1|\mathbf{X}) = P(I_i^* \leq I_i) = P(Z_i \leq \beta_1 + \beta_2 \mathbf{X}) = F(\beta_1 + \beta_2 \mathbf{X}) \tag{4.11}$$

em que $P(Y = 1|\mathbf{X})$ é a probabilidade de que um evento ocorra dado o vetor de variáveis independentes, $\mathbf{X}$; e Z_i é a variável normal padronizada, onde $Z\ N(0, \sigma^2)$ [13]

Por fim, para se obter as informações sobre o índice de utilidade, I_i, bem como sobre os β_j, basta tomar o inverso da Equação 4.11, e obtem-se:

$$\begin{aligned} I_i &= F^{-1}(I_i) = F^{-1}(P_i) \\ I_i &= \beta_1 + \beta_2 \mathbf{X} \end{aligned} \tag{4.12}$$

em que F^{-1} é o inverso da FDA normal.

4.3.2 Algoritmos de Machine Learning

Os algoritmos de *Machine Learning* (ML) vem ganhando espaço entre os economistas e nas suas pesquisas nos últimos anos (VARIAN, 2014; BAJARI *et al.*, 2015b;

[13] De acordo com Gujarati e Porter (2011), F é a FDA normal padrão dada por:

$$F(I_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_1 + \beta_2 \mathbf{X}} e^{-z^2/2} dz$$

BAJARI *et al.*, 2015a; ATHEY, 2018). Por se tratar diretamente de problemas que lidam com *big data*, a aplicação do ML tem se tornado eficiente quando o principal objetivo do pesquisador é a previsão do risco de um evento ocorrer. Particularmente, no que se refere no grau de automatização do processo de modelagem, estimação, teste e escolha do melhor modelo de predição.

A discussão sobre ML e o seu desempenho preditivo passa pelo *trade-off* entre viés e variância (*bias-variance trade-off*) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). A relação custo-benefício nesse *trade-off* ocorre quando se torna possível reduzir as incertezas das predições e projeções ao custo de um aumento do viés nos estimadores. Considerada uma das técnicas de *data mining*, o ML avança em relação as abordagens estatísticas mais tradicionais no que tange as melhorias em fazer previsões em conjuntos de dados cada vez maiores e mais complexos, bem como no enfoque para avaliação e seleção dos modelos.

A análise preditiva, por sua vez, baseia-se na aplicação de algoritmos em estruturas de dados existentes na busca por estimar o risco de eventos futuros ocorrerem com base em experiências passadas, e assim, gerar tomadas atuais de decisões (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; KUHN; JOHNSON, 2013). Contudo, a acurácia dessas estimativas é um dos aspectos mais importantes do modelo. Por isso, uma boa parte do debate na literatura em ML se dedica as métricas utilizadas para reduzir a parcela redutível do erro de previsão dos estimadores. Uma vez que a parcela irredutível não pode ser trabalhada, no caso, por exemplo, da omissão de variáveis no modelo.

Os métodos de ML podem ser divididos em aprendizado supervisionado e aprendizado não supervisionado. Nesse estudo foram aplicados métodos de aprendizado supervisionado, onde reúne métodos de estimação em que cada observação do preditor da base de dados mensurado por X_i, $i = 1, 2, ..., n$, há uma variável de interesse (dependente), Y_i. Ou seja, o principal objetivo baseia-se em ajustar o modelo que relacione os preditores, X, a variável de resposta, Y, com a finalidade de prever o evento em observações futuras. Por outro lado, o aprendizado não supervisionado detém os métodos em que para cada observação das covariadas não se tem a variável de resposta correspondente (JAMES *et al.*, 2013; ATHEY, 2018).

O tipo de variável a ser predita pode ser definida em dois subgrupos diferentes no aprendizado supervisionado: regressão, para variáveis quantitativas; e classificação, para variáveis categóricas ou qualitativas. Em ambos os casos, o ajuste dos modelos de ML pode descritos nas seguintes etapas: (i) divisão (aleatória - dependendo do tamanho da base) dos dados em conjuntos de treinamento e teste na etapa de pré-processamento; (ii) na etapa de aprendizado ocorre a seleção do modelo com melhor previsão em dados de treinamento, ante a uma gama de algoritmos; (iii) na terceira

etapa, a predição da resposta de interesse na base de teste, e; por fim, (iv) a avaliação do melhor modelo em novos dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013; RASCHKA, 2017).

Segundo Raschka (2017), a divisão da amostra em conjunto de dados de treinamento e teste é realizada com o intuito de verificar se o modelo apresenta boa predição não apenas nos dados que foram utilizados no ajuste (treinamento), mas também na capacidade de generalização para uma nova amostra (teste). Em geral, dependendo do tamanho da base de dados, as divisões mais adotadas seguem os seguintes padrões: 60:40; 70:30 ou 80:20, logo, quanto maior o número de observações, maior será o conjunto de dados utilizado na etapa de treinamento.

Como o objetivo deste estudo baseia-se na ideia de prever o nível de reprovação dos discentes, o foco será com um problema de classificação, de forma que foi feito uma previsão acurada da variável de interesse, nesse caso: reprovação na disciplina, denotada por $\hat{Y}$, a partir dos valores associados ao vetor de preditores, $\mathbf{X}$, contendo informações sobre o aluno, docentes, turma, curso e centro. O problema de classificação baseia-se na divisão do espaço amostral dos preditores em grupos relacionadas às categorias de resposta de interesse. A fronteira que define a divisão entre esses grupos é denominada de classificador, o qual representa o algoritmo que estima o modelo preditivo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Ante a esse contexto, o objetivo do ML consiste em construir um classificador, $f(X)$, que faça as melhores predições da resposta de interesse com base em observações futuras (IZBICKI; SANTOS, 2018). Para James *et al.* (2013), dentre os diversos algoritmos disponíveis, alguns são considerados como pouco flexíveis, ou menos complexos, como é o caso da regressão logística e da árvore de decisão, em que são interpretáveis. Já os outros algoritmos seguem uma abordagem mais flexível, ou mais complexas, como por exemplo as redes neurais, onde cada preditor está individualmente associado à resposta de interesse, tornando mais difícil a compreensão.

Na etapa de aprendizado, Kuhn e Johnson (2013) destaca que há dois tipos de parâmetros a serem estimados em ML: os parâmetros usuais de um algoritmo, como os pesos de uma regressão logística, e os hiperparâmetros ou parâmetros de ajuste, os quais são relacionados ao controle da flexibilidade de um algoritmo, como por exemplo, os de penalização aplicados nos parâmetros de regressão logística para se obter estimadores viesados, mas com variância mínima. Dessa maneira, controlar a flexibilidade de um algoritmo depende do balanceamento entre viés e variância.

Goldstein, Navar e Carter (2016) discutem que a variância está relacionada à sensibilidade das predições e à variabilidade das observações de treinamento, e o viés refere-se à diferença entre o valor previsto por um modelo, para uma dada observação, e o real valor observado. De maneira geral, um modelo flexível, ou seja, mais complexo,

tem variância alta, pois seu resultado será diferente para base de dados distintos, porém menor viés. Já para os casos dos modelos pouco flexíveis, ou mais simples, têm variância baixa, porém pode apresentar alto viés.

De outra maneira, o ML dá acesso a um leque de possibilidades em que o foco não é apenas o desempenho dos modelos na base de treinamento, mas sim em ter um ótimo desempenho em um conjunto invisível de dados, isto é, na base de teste. Sendo assim, segundo James *et al.* (2013) e Izbicki e Santos (2018), o que não pode acontecer é o *Overfitting*, onde o modelo apresenta alto desempenho no conjunto de treinamento, mas baixo desempenho no conjunto de teste. Logo, o objetivo final é encontrar um algoritmo para o qual ambos (viés e variância) sejam baixos para um determinado problema.

O aprendizado de modelos preditivos é composto por dois principais objetivos: selecionar e avaliar. No que se refere a selecionar, a performance de diferentes modelos é avaliada por meio de critérios de medidas de desempenho, que está descrita na Subseção 4.3.3. Para que a partir de um equilíbrio entre viés-variância, seja selecionado o modelo que resulta em uma melhor acurácia e desempenho no conjunto de treinamento. Já no que se refere o objetivo de avaliar, após a definição da melhor performance, busca-se estimar o modelo em novas observações, na base de teste (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A melhor estratégia para ambos os objetivos supracitados consiste em dividir aleatoriamente a base de dados original em três partes: treinamento, validação e teste. Contudo, em situações em que o conjunto de dados não for grande o suficiente para ser particionados em três partes, as técnicas de reamostragem podem ser adotadas para reaproximar o conjunto de validação através da reutilização das observações do conjunto de treinamento (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Em problemas de ML, a validação cruzada *k-fold* é uma das técnicas de reamostragem mais utilizadas para estimar o desempenho futuro. Consiste na divisão aleatória da base de treinamento em k partes de tamanhos iguais, em que $k-1$ irão compor os dados de treinamento utilizados para o ajuste dos modelos e a outra parte será destinada para a estimação da sua performance. O processo só é encerrado depois que todas as partes tenham participado tanto do conjunto de treinamento como o de validação do modelo, resultado em k estimativas de performance. A precisão dessas estimativas consiste nas repetições desse processo, de modo que, a cada repetição, diferentes divisões do conjunto de treinamento são consideradas para compor cada uma das k partes do processo de validação cruzada (KUHN; JOHNSON, 2013; ATHEY; IMBENS, 2017).

Não há uma regra precisa para a escolha de k, contudo, na literatura é mais comum a divisão dos dados em 5 ou 10 partes (KUHN; JOHNSON, 2013). À medida

que k aumenta, a diferença entre o tamanho da base de treinamento original e dos subconjuntos reamostrados torna-se menor. No entanto, o tempo necessário para chegar ao resultado da validação cruzada torna-se maior. Após definir o valor de k, é preciso escolher uma medida de avaliação para estimar a performance dos modelos que serão ajustados. Importantes tanto na etapa de seleção, quando de avaliação, o cálculo da medida de avaliação objetiva mensurar o quanto o valor previsto para uma observação se aproxima do seu valor observado (JAMES *et al.*, 2013).

Na literatura de ML há um consenso de que não existe um algoritmo que seja capaz de ter uma boa performance em todas as aplicações, logo é importante comparar os diversos métodos com características diferentes entre si para selecionar o modelo com a melhor performance preditiva para o problema abordado (LANTZ, 2013; KUHN; JOHNSON, 2013; IZBICKI; SANTOS, 2018). De um modo geral, na etapa de aprendizado a seguir, estão apresentados inicialmente os algoritmos de ML que podem ser divididos nas seguintes categorias: lineares(regressão linear e regressão logística); não lineares (*K – nearest neighbors*, *naïve bayes classifier*, *neural network* e *support vector machines*) e; modelos baseados em árvores de decisão (*regression trees*, *classification trees*, *bagging*, *random forest* e *gradiente boosting*). Em seguida, a descrição dos critérios que são adotados para avaliar o desempenho e selecionar o modelo com o melhor ajuste.

4.3.2.1 Regressão Linear

O modelo de regressão linear é um método de aprendizado estatístico útil e amplamente utilizado. Além disso, é considerado como um bom ponto de partida para a compreensão das abordagens de ML, uma vez que muitos dos algoritmos mais sofisticadas podem ser vistos como generalizações ou extensões de regressão linear. Logo, justifica-se a importância de se ter uma boa compreensão desse modelo antes mesmo das descrições dos métodos de ML mais complexos (JAMES *et al.*, 2013). Como a presente pesquisa está voltada para um problema de previsão, a natureza inferencial da regressão linear não será discutida mais a fundo nessa subseção [14].

O modelo de regressão linear é uma abordagem utilizada para prever uma resposta de interesse quantitativa Y com base em variáveis preditoras X. Ela assume que há aproximadamente uma relação linear entre X e Y (JAMES *et al.*, 2013). Sob esse contexto, o objetivo do método baseia-se em medir os pesos dos parâmetros da equação que descreva melhor a relação entre a variável resposta e os preditores para utilizá-los na previsão dessa resposta em novas bases de treinamento (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Matematicamente, essa relação linear, dado o vetor

[14] Para mais detalhes, ver Wooldridge (2006) e Gujarati e Porter (2011).

de entrada $X^T = (X_1, X_2, ..., X_p)$, e a saída Y, pode ser escrita como:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j \tag{4.13}$$

O termo $\hat{\beta}_0$ é o intercepto, também conhecido como o viés no *machine learning* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Tornando a denotação mais conveniente, a variável constante 1 será incluída em X, assim como o $\hat{\beta}_0$ será incluído no vetor de parâmetros $\hat{\beta}$, composto por β_j, onde $j = 1, 2, ..., p$, e depois reescrever o modelo linear em forma de vetor:

$$\hat{Y} = X^T \hat{\beta} \tag{4.14}$$

Embora a relação seja linear na função, o modelo permite que sejam aplicadas transformações não lineares nas covariadas, como por exemplo, raiz quadrada, log, representações polinomiais e outras. Dentre os modelos de regressão linear, o método mais comum é o Mínimos Quadrados Ordinários (MQO), considerado um método simples, e, em alguns casos, eficiente. Este modelo pode ser utilizado para estimar os valores de $\hat{\beta}$ para minimizar a Soma dos Quadrados dos Resíduos (SQR), também conhecida como função perda do modelo de regressão linear:

$$SQR(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 \tag{4.15}$$

$SQR(\beta)$ é uma função quadrática nos parâmetros e, portanto, seu mínimo vai sempre existir, mas pode não ser único. A solução é mais fácil de caracterizar em notação matricial. Dessa maneira, a função pode ser reescrita como:

$$SQR(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \tag{4.16}$$

em que $\mathbf{X}$ é uma matriz de dimensão N x p, em que cada linha há um vetor de entrada, preditores, e $\mathbf{y}$ é um vetor de dimensão N para a resposta de interesse, no conjunto de treinamento. Diferenciando $SQR(\beta)$ em relação ao vetor β, obtem-se as equações normais:

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0 \tag{4.17}$$

Se $\mathbf{X}$ tiver posto completo e o termo $\mathbf{X}^T\mathbf{X}$ é positiva definida e não singular, então a solução única é dada por:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{4.18}$$

e o valor ajustado do i-ésimo preditor x_i é $\hat{y}_i = \hat{y}(x_i) = x_i^T\hat{\beta}$.

Apesar de uma série de vantagens, a literatura de *machine learning* destaca algumas características indesejáveis desse método. Hastie, Tibshirani e Friedman (2009) destacam algumas delas: (i) a existência de um elevado número de preditores pode ter como consequência preditores autocorrelacionados, ou ainda problemas relacionados ao número de preditores ser maior do que o de observações, e partir daí o modelo terá infinitas soluções e variância tendendo ao infinito; (ii) embora o método MQO tem como característica estimadores não viesados, podem apresentar também elevada variância. Comprometendo a acurácia das previsões e a interpretação do modelo. Dessa maneira, boa parte da literatura que aplica algoritmos de predição em problemas de regressão utilizam o MQO como uma referência de comparação entre os métodos de estimação (HAN; KAMBER; MINING, 2001; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; KUHN; JOHNSON, 2013; LANTZ, 2013; IZBICKI; SANTOS, 2018).

4.3.2.1.1 *Penalized methods*

Ante ao contexto da elevada variância que pode ocorrer ao aplicar o método de regressão linear para previsões, e como o método de seleção do subconjunto de preditores não sana o referido problema, Hastie, Tibshirani e Friedman (2009) indicam que os métodos de *Shrinkage* são os mais adequados para abordar esse problema, uma vez que são mais contínuos e não sofrem tanto com a elevada variância. Este conjunto de regressores faz parte da classe de estimadores dos métodos com penalidades, os *penalized methods*.

Desse modo, a solução busca penalizar os coeficientes estimados a fim de limitar a variância ao custo de um aumento insignificante de viés (JAMES *et al.*, 2013). O ponto central tratado baseia-se no *trade-off* entre viés e variância, onde é possível obter métodos com menor variância ao acrescentar viés aos estimadores. De acordo com Kuhn e Johnson (2013), através de um pequeno viés nos preditores se torna possível reduzir a variância do modelo, e assim, uma melhora na performance de previsão em novas observações. Os métodos mais conhecidas de penalização são *ridge*, *LASSO* e *elastic net*. Dependendo do tipo de pena, alguns coeficientes podem ser estimados como exatamente zero.

A regressão *ridge* reduz os coeficientes da regressão impondo uma penalidade ao seu tamanho. De outra maneira, as estimativas dos parâmetros são oriundas da

minimização da função perda (SQR) com penalização quadrática (JAMES *et al.*, 2013). É muito semelhante ao MQO, exceto que os parâmetros são estimados minimizando uma quantidade ligeiramente diferente. O algoritmo de *ridge* estimará $\hat{\beta}^{ridge}$ minimizando:

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin}\left\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{4.19}$$

Para $\lambda \geq 0$, λ é um parâmetro de *tunning*, ou ajuste, que controla a quantidade de *shrinkage*, ou seja, quanto maior o valor de λ, maior a penalização, ou tamanho do encolhimento. Os coeficientes são encolhidos em direção a zero. A ideia de penalizar pela soma dos quadrados dos parâmetros também é usada em redes neurais, onde é conhecida como *weight decay* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O intercepto β_0 assumirá o valor médio da resposta de interesse quando todas as covariadas assumirem valor zero, pois este não sofre a penalização, já que o objetivo é penalizar apenas os parâmetros associados aos preditores.

O principal motivo do intercepto β_0 ser deixado de fora da penalização se fundamenta na dependência que método teria da origem escolhida para Y, isto é, adicionar uma constante c a cada um das respostas y_i não resultaria simplesmente em um deslocamento das predições na mesma quantidade c. Assim, a solução para a Equação 4.19 pode ser separada em duas partes: (1) após a reparametrização da matriz de preditores: cada x_{ij} é substituído por $x_{ij} - \bar{x}_j$, e estima-se β_0 por $\bar{y} = \frac{1}{N}\sum_1^N y_i$; (2) os coeficientes restantes são estimados por uma regressão *ridge* sem intercepto, usando o x_{ij} centralizado. Além da reparametrização a qual apresentou média zero para cada um dos preditores, cada um deles poder ser dividido pelo seu desvio padrão, resultando em uma matriz de entrada $\mathbf{X}$ padronizada, com colunas p (em vez de $p+1$) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013). Logo, a SQR penalizada assume:

$$SQR(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta \tag{4.20}$$

e reescrevendo a Equação 4.19 em forma de matriz:

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{4.21}$$

O *LASSO* é um método de *shrinkage* como o *ridge*, com diferenças sutis, mas importantes. Assim como na regressão de *ridge*, pode ser feita a reparametrização da constante β_0 padronizando os preditores, e a solução para $\hat{\beta}_0$ é $\bar{y}$, onde a penalização passa a ser baseado agora no valor absoluto dos parâmetros. A estimativa do *LASSO*

também pode ser escrita no formato lagrangeano:

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin}\left\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{4.22}$$

Observe a semelhança com o problema de regressão *ridge* da Equação 4.19.A penalidade $L2$ da *ridge* $\sum_{j=1}^{p} \beta_j^2$ é substituída pela penalidade $L1$, $\sum_{j=1}^{p} |\beta_j|$, do *LASSO*. Assim, o método *LASSO* também reduz as estimativas dos coeficientes para zero. No entanto, a penalidade em *LASSO* tem o efeito de forçar que algumas das estimativas seja exatamente zero, podendo excluir alguma variável (JAMES *et al.*, 2013). Este tipo de penalização torna as soluções não-lineares em y_i e não apresenta solução analítica para β, ou seja, não há como expressar, de forma fechada, o vetor dos parâmetros estimados, como ocorre na regressão *ridge* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Por fim, o *elastic net* é um algoritmo que envolve as normas de penalidades $L1$ e $L2$ da *shrinkage*, ou seja, é um método de regressão que adota como restrição a combinação linear entre a restrição $L2$ da *ridge* e a $L1$ da *LASSO*. O método contribui tanto para a estimação de soluções esparsas quanto para a restrição das estimativas dos parâmetros (KUHN; JOHNSON, 2013).

$$\hat{\beta}^{Enet} = \underset{\beta}{argmin}\left\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \right\} \tag{4.23}$$

Kuhn e Johnson (2013) ainda destaca que a validação cruzada *k-fold* pode ser adotada no processo de otimização dos parâmetros λ, nos três casos.

4.3.2.2 Regressão Logística

A Regressão logística é um dos modelos de probabilidade linear direcionado para prever respostas qualitativas, é um método linear para classificação. Segundo James *et al.* (2013), prever uma resposta qualitativa para uma classificação de observação pode ser referida como classificando essa observação, uma vez que envolve atribuir a observação a uma categoria ou classe. Por outro lado, muitas vezes os métodos utilizados para classificação primeiro prever a probabilidade de cada um dos as categorias de uma variável qualitativa, como base para fazer a classificação. Nesse sentido, eles também se comportam como métodos de regressão.

Assim como nos modelos direcionados para regressão, nos modelos de classificação, tem-se um conjunto de observações de treinamento, $(x_1, y_1), ..., (x_n, y_n)$, que pode ser usado para construir um classificador. Deve-se ter um bom desempenho não apenas nos dados de treinamento, mas também em observações de teste que não

foram usadas para treinar o classificador. Considere um conjunto de dados padrão (*default*), em que o *default* de resposta cai em uma das duas categorias, *sim* ou *não*. Em vez de modelar a variável de resposta Y diretamente, a regressão logística modela a probabilidade de que Y pertence a uma categoria específica, ou seja, a regressão logística modela a probabilidade de *default*.

De acordo com Kuhn e Johnson (2013), para que essa probabilidade seja estimada, a variável de resposta da base de dados de treinamento é modelada a partir de uma distribuição binomial, que tem como parâmetro (p) a probabilidade da ocorrência de uma categoria específica. É de suma importância ressaltar que a probabilidade estimada deve estar no limiar do intervalo $[0, 1]$ e ao mesmo tempo apresentar uma relação direta com as covariadas (preditores) para cada observação da base de dados (JAMES *et al.*, 2013; KUHN; JOHNSON, 2013). Dessa maneira, para a estimação, utiliza-se a função logística:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{4.24}$$

a qual é responsável por modelar a relação entre o conjunto de preditores, X, e a probabilidade de uma determinada resposta, $p(X) = Pr(Y = k|X = x)$. É valido destacar que no processo de predição não é necessário que o modelo atenda aos pressupostos, para Izbicki e Santos (2018) quando o propósito é a inferência, apenas se espera que um bom classificador seja estimado.

Embora possa ser aplicado mínimos quadrados (não lineares) para ajustar o modelo logístico da Equação 4.24, o método da máxima verossimilhança é o mais comum para estimar o vetor dos parâmetros, β, uma vez que possui as melhores propriedades estatísticas (JAMES *et al.*, 2013).

A intuição básica no ajuste do modelo consiste na busca por estimativas para β_0 e β_1 de tal forma que a probabilidade prevista de *default*, $\hat{p}(x_i)$, para cada observação da base de dados de treinamento, corresponda o mais próximo possível da classe verdadeira observada do indivíduo. De acordo com James *et al.* (2013), a formalização desta intuição pode ser utilizada a partir da equação matemática chamada de função de verossimilhança:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \tag{4.25}$$

Dessa maneira, as estimativas β_0 e β_1 são escolhidas com o fim de maximizar a função de verossimilhança da Equação 4.25. Uma vez que o vetor de parâmetros é estimado, $\hat{\beta}$, este será utilizado para prever a probabilidade da resposta de interesse em novas observações (base de teste) (JAMES *et al.*, 2013; KUHN; JOHNSON, 2013).

Nos casos de respostas dicotômicas, como por exemplo, 0 e 1, se a probabilidade prevista de *default* associada à presença de determinada resposta, a possibilidade $(p/1-p)$ dessa ocorrência será:

$$\frac{Pr(Y=1|X=x)}{Pr(Y=0|X=x)} = \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \tag{4.26}$$

Considerando agora um problema de previsão para resposta binária usando múltiplos preditores, a regressão logística múltipla descreve a relação entre as covariadas e a resposta de interesse. Após a transformação *logit* na Equação 4.26, pode-se observar que o modelo de regressão torna-se linear em X, se tornando generalizado da seguinte maneira:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + ...\beta_p X_p \tag{4.27}$$

em que $X = (X_1, ..., X_p)$ são preditores p. A Equação 4.27 pode ser reescrita como:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+...\beta_p X_p}} \tag{4.28}$$

assim como no modelo simples, é aplicado o método da máxima verossimilhança para estimar $\beta_0, \beta_1, ...\beta_p$.

Por fim, há uma relação entre a fronteira de decisão linear no método de regressão logística e a escolha de um ponto de corte para $p(X)$. Em um exemplo apresentado por James *et al.* (2013), e destacado por Santos (2018), o ponto de corte determinado para $p(X)$ é $0,5$, ou seja, indivíduos com $p(X) > 0,5$ vão ser classificados como um grupo que apresentam a presença de uma resposta específica, assim como aqueles com $p(X) < 5$ representam a ausência de determinada resposta. Portanto, a determinação de um ponto de corte irá definir uma fronteira de decisão linear para o modelo de regressão logística. E assim como na regressão linear, as penalidades *ridge*, *LASSO* e *elastic net* podem também serem aplicados em conjunto com a regressão logística.

4.3.2.3 *K-Nearest Neighbors* (*KNN*)

Considerado um dos algoritmos mais populares do *Machine Learning* (BENEDETTI, 1977; STONE, 1977), o método *K-Nearest Neighbors* (*KNN*), ou o método dos K vizinhos mais próximos, é uma alternativa não paramétrica, tanto para os problemas de classificação como para os de regressão, quando a relação entre a resposta de interesse e os preditores demanda um modelo mais flexível, como por exemplo em casos não lineares (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A abordagem KNN prevê simplesmente uma nova amostra usando as observações mais próximas do conjunto de treinamento. A sua construção consiste somente nas amostras individuais dos dados de treinamento das covariadas e da reposta de interesse. De outra maneira, para prever uma nova amostra para regressão, a KNN identifica os KNNs da amostra no espaço preditivo, e a resposta prevista para a nova amostra é então a média das respostas dos K vizinhos mais próximos. Outras estatísticas, como a mediana, também podem ser usadas no lugar da média para prever a nova amostra (KUHN; JOHNSON, 2013).

Raschka (2017) destaca que esse método possui duas características importantes que precedem a previsão da resposta de interesse para o novo grupo de observações: a determinação do número de K vizinhos mais próximos que serão a vizinhança da nova observação; e a determinação da medida de distância que identificará as K observações da base de dados de treinamento mais próximas à nova observação. A definição do número de K vizinhos mais próximos é tido como um parâmetro que estabelece o quão bem o ajuste do modelo será generalizado para dados futuros.

Existe uma relação direta do limiar entre viés e variância e o processo de escolha de K (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; KUHN; JOHNSON, 2013). Se K=N, onde N é o total de observações da base de treinamento, a resposta prevista será sempre a mesma, o que caracteriza um viés alto e uma baixa variância. Para K=1, o valor previsto representará mais vulnerabilidade a dados discrepantes ou a ruídos presentes na base de treinamento, apresentando assim alta variância, contudo um menor viés. Hastie, Tibshirani e Friedman (2009) e Lantz (2013) ainda destacam que a otimização de K pode ser encontrada a partir de técnicas de reamostragem, como a validação cruzada.

No que se refere a determinação da medida de distância, o método KNN básico depende de como o pesquisador define a distância entre as amostras. A distância euclidiana, a distância em linha reta entre duas amostras, é a métrica mais comumente utilizada e é definida da seguinte maneira (KUHN; JOHNSON, 2013):

$$\left(\sum_{j=1}^{P} (x_{aj} - x_{bj})^2 \right)^{\frac{1}{2}} \tag{4.29}$$

onde x_a e x_b são duas amostras individuais, que correspondem as mensurações das covariadas da base nova de observações e de uma das bases de observações do conjunto de treinamento, respectivamente. Raschka (2017) ainda destaca que para a adoção dessa distância é importante que se estabeleça um padrão na base de dados na etapa de pré-processamento de maneira que a contribuição de cada preditor para a métrica da distância seja igual. Dessa maneira, a otimização de K evidencia que a base de

treinamento por ser adotada para prever a resposta de interesse em novas observações.

Assim, após determinar o número de vizinhos e a medida de distância a serem adotados, no método KNN, em problemas de regressão, a resposta x^* é a média das respostas observadas da sua vizinhança, $N_k(x)$, definidas a partir de K bases de treinamento T com o vetor de covariadas, x_i, dos seus K vizinhos mais próximos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; IZBICKI; SANTOS, 2018) . Especificamente, o KNN busca estimar a distribuição de Y dado x, logo, $\hat{Y}$ é definido da seguinte forma:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{4.30}$$

em que $N_k(x)$ é o total de vizinhos de x definida pelos K pontos mais próximos, x_i, do conjunto de treinamento.

Por sua vez, nos problemas de classificação, a resposta a ser predita deve ser representada pela classe mais comum observada na vizinhança de x^*. Ou seja, para cada classe da resposta de interesse j, calcula-se a probabilidade condicional da nova observação pertencer a *j-ésima* classe através da fração de pontos em $N_k(x)$ cujo valor da resposta é j:

$$P(Y = j|X = x^*) = \frac{1}{K} \sum_{x_i \in N_k(x)} I(y_i = j) \tag{4.31}$$

onde a classe predita j para x^* será representada por a classe que evidenciar a maior probabilidade condicional (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; SANTOS, 2018).

4.3.2.4 *Naïve Bayes Classifier*

Os classificadores *bayesianos*, baseados no teorema de *Bayes*, são classificadores estatísticos, os quais buscam prever a probabilidade de participação na classe, ou seja, a probabilidade de uma determinada "resposta", ou observação, pertencer a uma determinada classe. Ao mesmo tempo, exibiram alta precisão e velocidade quando aplicados a grandes bancos de dados (HAN; KAMBER; MINING, 2001).

Han, Kamber e Mining (2001) e Lantz (2013) afirmam que há um classificador bayesiano simples conhecido como *naïve bayes classifier*, ou classificador *bayesiano* "ingênuo", é comparável em desempenho e performance com os algoritmos dos modelos baseados em árvore de decisão e classificadores de *neural network*. Embora não seja o único método de aprendizado de máquina que utiliza métodos *bayesianos*, é o mais comum.

Os classificadores *naïve bayes* assumem que o efeito de uma característica em uma determinada classe é independente dos valores das outras características. Essa

suposição é chamada de *class conditional independence*, independência condicional de classe, e é imposta para simplificar as análises envolvidas e, por esse motivo, é considerado um classificador "ingênuo".

Para Friedman, Geiger e Goldszmidt (1997), o classificador *naïve bayes* é um tipo de *bayesian network* que, apesar de ser um algoritmo simples, é capaz de criar modelos com alto poder preditivo. Afirma ainda que o método funciona bem tanto com tipos de dados heterogêneos, como com valores omissos, devido ao tratamento independente de cada variável preditora para a construção do modelo. Dessa maneira, esta característica levanta a discussão sobre se um classificador com suposições menos restritivas pode ter um desempenho melhor, quando comparados aos classificadores com suposições mais complexas.

Inicialmente, o teorema de *bayes* é útil na medida em que fornece uma maneira de calcular a probabilidade posterior, $P(y/\mathbf{X})$, de $P(y)$, $P(\mathbf{X}/y)$ e $P(\mathbf{X})$. Assim, assumindo que $\mathbf{X}$ seja um vetor de p covariadas ou parâmetros e y a variável de classe, tem-se o teorema de *bayes* dado por:

$$P(y|\mathbf{X}) = \frac{P(\mathbf{X}|y)P(y)}{P(\mathbf{X})} \tag{4.32}$$

Tendo esta teoria como base, de acordo com Han, Kamber e Mining (2001), o classificador *naïve bayes*, ou classificador *bayesiano* simples, funciona da seguinte maneira:

- Seja D um conjunto de treinamento de preditores e suas classes associadas. Cada preditor é representado por um vetor de atributo, $\mathbf{X} = (x_1, x_2, ..., x_n)$, com dimensão n, representando n medições feitas a partir de n atributos, ou características, respectivamente, $A_1, A_2, ..., A_n$.

- Suponha que existam m classes, $C_1, C_2, ..., C_m$. Dada uma covariada, $\mathbf{X}$, o classificador irá predizer que $\mathbf{X}$ pertence à classe que tem a probabilidade posterior mais alta, condicionada em $\mathbf{X}$. Ou seja, o classificador *naïve bayes* prevê que $\mathbf{X}$ pertence à classe C_i se e somente se:

$$P(C_i|\mathbf{X}) > P(C_j|\mathbf{X}) \quad para \quad 1 \leq j \leq m, \;\; j \neq i \tag{4.33}$$

  A classe C_i para a qual $P(C_i|\mathbf{X})$ é maximizada é chamada de *maximum posteriori hypothesis*. Logo, a Equação 4.34 é reescrita como:

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})} \tag{4.34}$$

- Como $P(X)$ é uma constante para todas as classes, somente $P(\mathbf{X}|C_i)P(C_i)$ precisa ser maximizado. Se as probabilidades anteriores da classe não são conhecidas, então é comumente presumido que as classes são igualmente prováveis, isto é, $P(C_1) = P(C_2) = ... = P(C_m)$, e portanto $P(\mathbf{X}|C_i)$ será maximizado. Caso contrário, $P(\mathbf{X}|C_i)P(C_i)$ é que será maximizado.

- Dado o conjuntos de dados com muitos atributos, seria extremamente caro computar $P(\mathbf{X}|C_i)$. Para reduzir a computação na avaliação de $P(\mathbf{X}|C_i)$, é feita a suposição "ingênua" de independência condicional de classe. Isso pressupõe que os valores dos atributos são condicionalmente independentes um do outro, dado o rótulo de classe do preditor (ou seja, que não há relações de dependência entre as características). Portanto:

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{4.35}$$

$$P(\mathbf{X}|C_i) = P(x_1|C_i)\ x\ P(x_2|C_i)\ x\ ...\ x\ P(x_n|C_i) \tag{4.36}$$

As probabilidades $P(x_1|C_i)$ x $P(x_2|C_i)$ x ... x $P(x_n|C_i)$ das tuplas de treinamento podem serem estimadas. Lembre-se de que x_k se refere ao valor do atributo A_k para o preditor (tupla) $\mathbf{X}$. Para cada atributo, verifica-se se o atributo é categórico ou contínuo. Por exemplo, para calcular $P(\mathbf{X}|C_i)$, considera o seguinte:

(a) Se A_k é categórico, então $P(x_k|C_i)$ é o número de tuplas da classe C_i em D, tendo o valor x_k para A_k, dividido por $|C_{i,D}|$, o número de tuplas da classe C_i em D.

(b) Se A_k é contínuo, o atributo é tipicamente assumido como tendo uma distribuição Gaussiana com uma média μ e desvio padrão σ, definido por:

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\sigma)^2}{2\sigma^2}} \tag{4.37}$$

de modo que,

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \tag{4.38}$$

Dessa maneira, precisa-se calcular μ_{C_i} e σ_{C_i}, que são a média e o desvio padrão, respectivamente, dos valores do atributo A_k para treinar tuplas da classe C_i. Em seguida, conecta-se os resultados na Equação 4.37, juntamente com x_k, para estimar $P(x_k|C_i)$.

- Para prever o rótulo de classe de $\mathbf{X}$, $P(\mathbf{X}|C_i)P(C_i)$ é avaliado para cada classe C_i. O classificador prevê que o rótulo de classe da tupla $\mathbf{X}$ é a classe C_i se e somente se:

$$P(\mathbf{X}|C_i)P(C_i) > P(\mathbf{X}|C_j)P(C_j) \quad \text{para} \quad 1 \leq j \leq m, j \neq i \tag{4.39}$$

  Em outras palavras, o rótulo de classe previsto é a classe C_i para a qual $P(\mathbf{X}|C_i)$ é o máximo.

4.3.2.5 *Neural Network*

Neural Network, ou Redes neurais, são poderosas técnicas de regressão não-linear inspiradas em teorias de como o cérebro funciona (CHENG; TITTERINGTON, 1994; BISHOP *et al.*, 1995; RIPLEY, 1996). De acordo com Hastie, Tibshirani e Friedman (2009), uma rede neural é um algoritmo de aprendizado em máquina, em dois estágios, aplicável tanto para a predição de problemas de regressão como para os de classificação.

Tipicamente representado por um diagrama de rede, Hastie, Tibshirani e Friedman (2009) descreve que para problemas de regressão, em seu caso geral, $K = 1$ e com apenas uma unidade de saída Y_1 no topo. Para os de classificação, com K-classes de respostas definidas, existem K unidades no topo, Y_k, $k = 1, ..., K$, cada uma sendo correspondente com uma variável binária $(0 - 1)$ para as k-classes, em que cada unidade k modela a probabilidade da classe k.

Kuhn e Johnson (2013) afirma que redes neurais modela a variável de resposta, em um primeiro estágio, a partir de um conjunto intermediário de m variáveis, ou como denominado na literatura, *hidden units* (Z_m), resultantes de combinações lineares de preditores originais (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; KUHN; JOHNSON, 2013). Em seguida, o alvo Y_k é modelado como uma função das combinações lineares do Z_m,

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T x), \quad m = 1, ..., M, \tag{4.40}$$

$$T_k = \beta_{0k} + \beta_k^T Z, \quad k = 1, ..., K, \tag{4.41}$$

$$f_k(X) = g_k(T), \quad k = 1, ..., K, \tag{4.42}$$

em que $Z = (Z_1, Z_2, ..., Z_M)$, e $T = (T_1, T_2, ..., T_M)$. Na Equação 4.40, a função de ativação $\sigma(v)$ é geralmente escolhida para ser o sigmoide $\sigma(v) = 1/(1 + e^{-v})$. Às vezes, as funções de base radial gaussiana são usadas para o $\sigma(v)$, produzindo o

que é conhecido como a *Radial Basis Function (RBF) network* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; LANTZ, 2013).

Se $\sigma(v)$ for considerada a função identidade, o resultado será um modelo linear. Portanto, pode-se considerar que uma rede neural representa uma generalização não linear de modelos lineares para problemas de predição (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; LANTZ, 2013). Assim, inserir a transformação não linear θ, acarreta na ampliação considerável da classe de modelos lineares. Por fim, após o número de *hidden units*, Z_m, ter sido definido, cada um é relacionado à resposta de interesse, estabelecendo-se o segundo estágio do modelo, sendo representados nas Equações 4.41 e 4.42.

A função $g_k(T)$ permite uma transformação final do vetor de saídas T. Para problemas de regressão, normalmente utiliza-se a função de identidade $g_k(T) = T_k$. Já nos trabalhos de classificação, uma transformação *softmax*, não linear, é aplicada em T_k com o objetivo de garantir que os valores previstos estejam no intervalo [0,1]:

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}} \tag{4.43}$$

No processo de ajustamento de redes neurais, o algoritmo tem parâmetros desconhecidos, geralmente chamados de *weights*, pesos, e busca-se valores que encaixem bem no conjunto de treinamento. O conjunto completo de pesos é denotado por θ, e consiste em:

$$\{\alpha_{0m}, \alpha_m; \;\; m = 1, 2, ..., M\} \quad M(p+1) \;\; pesos \tag{4.44}$$

$$\{\beta_{0k}, \beta_k; \;\; k = 1, 2, ..., K\} \quad K(M+1) \;\; pesos \tag{4.45}$$

em regressão, a SQR é adotada como medida de ajuste, função perda:

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2 \tag{4.46}$$

já para classificação, utiliza-se o erro quadrado ou a função *cross-entropy* (desvio):

$$R(\theta) = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} log f_k(x_i) \tag{4.47}$$

e o classificador correspondente é $G(x) = argmax_k f_k(x)$. Com a função de ativação *softmax* e a função de erro de *cross-entropy*, o modelo de rede neural é exatamente um modelo de regressão logística linear nas unidades ocultas, e todos os parâmetros são estimados por máxima verossimilhança (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Normalmente, não queremos o minimizador global de $R(\theta)$, pois é provável que seja uma solução de superposição. Em vez disso, é necessária alguma regularização: isso é obtido diretamente por meio de um período de penalidade ou indiretamente pela interrupção antecipada.

Segundo Hastie, Tibshirani e Friedman (2009), a abordagem para minimizar $R(\theta)$ pode representar um desafiador problema de otimização, uma vez que é preciso inicializar os parâmetros, a partir de valores aleatórios, e adotar técnicas de *back-propagation* para a sua solução. Além do mais, por conta do elevado número de parâmetros a serem estimados, a solução de $R(\theta)$ tem tendência ao sobreajuste da base de treinamento, logo, técnicas de regularização devem ser aplicadas.

No que se refere a regularização de redes neurais, a abordagem *weight decay*, decaimento de peso, é a técnica utilizada apresentando como resultado um efeito similar ao da regressão *ridge* no modelos lineares. É adicionado uma penalidade à função perda, que passa a ser representada por $R(\theta) + \lambda J(\theta)$, onde $J(\theta)$ é:

$$J(\theta) = \sum_{km} \beta_{km}^2 + \sum_{ml} \alpha_{ml}^2 \tag{4.48}$$

e $\lambda \geq 0$ é um parâmetro de sintonização, e variam entre 0 e 0,1. Valores maiores de λ tenderão a encolher os pesos em direção a zero e menor o efeito do sobreajuste, em que tipicamente, a validação cruzada é usada para estimar λ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; KUHN; JOHNSON, 2013).

O modelo descrito até aqui refere-se ao método de redes neurais em dois estágios, de outro modo, apenas com uma camada de *hidden units*, representando assim, uma rede neural em sua estrutura mais simples. De acordo com Hastie, Tibshirani e Friedman (2009) e Kuhn e Johnson (2013), o número de *hidden units* depende do número de observações e de preditores base de dados de treinamento, sendo comum adicionar um número grande de observações e treinar o modelo com regularização.

Por outro lado, nos casos em que mais camadas de *hidden units* são utilizadas no processo de predição, este caracteriza-se como redes neurais profundas. Segundo Goodfellow, Bengio e Courville (2016), os algoritmos do *deep learning* são os mais eficientes para treinar as redes neurais nesse contexto. De um modo geral, o *deep learning* destaca-se por alcançar uma alta flexibilidade e performance através dos dados observados como uma ordenação de funções matemáticas, onde cada uma está associada a uma camada diferente do modelo. Dessa forma, a cada aplicação de uma função, é obtida uma nova representação dos dados de entrada, *input*, ou seja, quanto mais profundas forem as camadas, mais abstratos serão os recursos utilizados no processo de extração dos dados para obter uma dada representação.

4.3.2.6 *Support Vector Machines* (*SVM*)

Proposto inicialmente por Cortes e Vapnik (1995), para problemas de classificação, e posteriormente por Smola *et al.* (1996) e Drucker *et al.* (1997), para problemas de regressão, o *Support Vector Machines* (*SVM*) é um algoritmo que modela fronteiras não lineares. Considerado uma generalização do método *Maximal Margin Classifier*, classificador de margem máxima, o SVM difere dos demais algoritmos do *Machine Learning*, direcionados para problemas de classificação, por não estimar probabilidades diretamente, e sim classes da resposta de interesse estimadas em novas observações .

James *et al.* (2013) destacam que embora o *Maximal Margin Classifier*[15] seja um classificador elegante e simples, ele não pode ser aplicado à maioria dos conjuntos de dados, pois requer que as classes sejam separáveis por um limite linear. Dessa maneira, o SVM pode ser aplicado em uma gama maior de casos, a fim de acomodar os limites de classes não lineares. A seguir, serão expostas as características do algoritmo SVM para problemas de classificação.

Em um cenário onde se possa separar perfeitamente as observações da base de treinamento em função da classe em que as mesmas pertencem, o *Maximal Margin Classifier* pode ser adotado para estimar uma fronteira de decisão linear. De outro modo, o classificador estima a equação de um hiperplano [16] cujas regiões de classificação tenham apenas uma das classes da resposta de interesse. No entanto, na grande parte dos problemas reais, não é possível estabelecer o hiperplano, uma vez que as observações das diferentes classes da variável de resposta podem estar sobrepostas, assim destaca James *et al.* (2013).

Nesse caso, James *et al.* (2013) orienta que é necessário considerar os hiperplanos de vetores suporte, visto que este tolera a presença de algumas observações que estejam no lado incorreto da margem da classe à qual pertence. Dessa maneira, o classificador associado a este tipo de hiperplano é o algoritmo *Support Vector Classifier*. O problema de otimização para encontrar a solução deste classificador estar direcionado na identificação dos parâmetros que maximize o tamanho da margem, tendo como restrições à classificação das observações que se localizam na classe errada da base de treinamento.

A solução para o problema do *Support Vector Classifier* envolve apenas os produtos internos das observações (em oposição às próprias observações). O produto interno de dois r-vetores a e b é definido como $\langle a, b \rangle = \sum_{i=1}^{r} a_i b_i$ (JAMES *et al.*, 2013).

[15] Para mais detalhes sobre o algoritmo *Maximal Margin Classifier* e suas principais características ver James *et al.* (2013), o qual apresenta um capítulo destinado a este método.

[16] De acordo com James *et al.* (2013), o hiperplano é um subespaço de um espaço p-dimensional, como por exemplo, em um espaço com duas dimensões, um hiperplano é um subespaço unidimensional plano, em outras palavras, uma linha a qual não precisa passar pela origem.

Assim, o produto interno de duas observações x_i, $x_{i'}$ é dado por:

$$\langle a, b \rangle = \sum_{j=1}^{p} x_{ij}, x_{i'j} \tag{4.49}$$

ante isso, o *Support Vector Classifier* linear pode ser representado por:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \tag{4.50}$$

em que existem n parâmetros α_i, $i = 1, ..., n$, um por observação de treinamento.

Contudo, o algoritmo *Support Vector Classifier*, assim como o *Maximal Margin Classifier*, é uma abordagem em casos em que a fronteira de decisão entre as classes da resposta de interesse é linear. Desta maneira, para os casos em que a fronteira de decisão é não linear, James *et al.* (2013) destacam que o comportamento não linear entre os parâmetros podem ser captados por um classificador que possa, por exemplo, adicionar novas variáveis correspondentes a transformações dos preditores originais, e assim, aumentar o espaço entre eles.

Nessa perspectiva, o SVM é considerado um classificador que representa uma extensão do *Support Vector Classifier*, o qual resulta da ampliação do espaço de recurso de uma maneira específica, utilizando *Polynomial Kernel* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013). A abordagem de *Polynomial Kernel* que está descrita a seguir representa uma das abordagens computacionais mais adotadas[17] em problemas de aprendizado em máquina, os quais detém relações não lineares entre os preditores e a resposta de interesse, ou seja, é uma função que quantifica a similaridade de duas observações. E é representada por:

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij}, x_{i'j}\right)^d \tag{4.51}$$

em que é conhecido como *Polynomial Kernel* de grau d, onde d é um inteiro positivo. Usando $d > 1$, ao invés do kernel linear padrão $K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij}, x_{i'j}$, no algoritmo *Support Vector Classifier* leva a um limite de decisão muito mais flexível. Em outras palavras, equivale a encaixar um *Support Vector Classifier* em um espaço de dimensão superior envolvendo polinômios de grau d, em vez de no espaço de característica original.

[17] Uma outra abordagem bastante comum é o *Radial Kernel*. Representada por:

$$K(x_i, x_{i'}) = exp\left(-\gamma \sum_{j=1}^{p} (x_{ij}, x_{i'j})^2\right)$$

Assim, quando o *Support Vector Classifier* é combinado com um kernel não linear, ou o *Polynomial Kernel*, o classificador resultante é conhecido como *Support Vector Machines*, o (*SVM*). Neste caso, a função (não linear), $f(x)$, tem a forma:

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x_i, x_{i'}) \tag{4.52}$$

verifica-se que α_i é diferente de zero apenas para os vetores de suporte na solução, isto é, se uma observação da base de treinamento não for um vetor de suporte, então seu α_i é igual a zero. Então, S é a coleção de índices desses pontos de suporte (JAMES *et al.*, 2013).

Hastie, Tibshirani e Friedman (2009) apresentam que o algoritmo SVM pode ser adaptado para problemas de regressão, ou seja, problemas com uma resposta quantitativa, de forma a herdar algumas das propriedades do SVM para problmeas de classificação. Inicialmente, considere o modelo de regressão linear $f(x) = x^T\beta + \beta_0$, e, em seguida, lidar com generalizações não lineares. Para estimar β, considere a minimização de H:

$$H(\beta, \beta_0) = \sum_{i=1}^{n} V(y_i - f(x_i)) + \frac{\lambda}{2}||\beta||^2 \tag{4.53}$$

em que:

$$V_\epsilon = \begin{cases} 0, & \text{se} |r| < \epsilon \\ |r| - \epsilon, & c.c \end{cases} \tag{4.54}$$

onde a função V é uma medida de erro "ϵ-intensiva", a qual ignora erros de tamanho menor que ϵ. Nesse caso, o algoritmo SVM visa minimizar a função perda relacionada apenas aos resíduos maiores (em valor absoluto) que determinada constante, condição esta que se estende ao conceito de margem, do *Support Vector Classifier*, para problemas de regressão (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013). Do mesmo modo que os problemas de classificação, a função para problemas de regressão apresenta o mesmo formato exposto na Equação 4.50:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \tag{4.55}$$

Nesse caso, apenas o subconjunto estimado para α_i, referentes aos vetores suporte, serão diferentes de zero. Ao invés de produtos internos das observações propriamente ditas, a solução depende do produto entre os vetores de mensurações dos preditores. Portanto, diferentes *Kernels*, *polinomial* e *radial*, podem ser adotados para os produtos internos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Nesse caso, considere a aproximação da função de regressão em termos de um conjunto de funções de base $h_m(x), m = 1, 2, ..., M$, na forma:

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x) + \beta_0 \tag{4.56}$$

Para alguma medida de erro geral $V(r)$, a solução da Equação 4.56, ou seja, a solução do SVM, tem a forma representada por:

$$\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i K \langle x, x_i \rangle \tag{4.57}$$

onde $K(x, y) = \sum_{m=1}^{M} h_m h_m(y)$.

4.3.2.7 *Decision Tree-Based Methods*

Decision Tree-Based Methods, ou métodos baseados em árvore de decisão, consistem em um conjunto de regras utilizadas para estratificar ou segmentar o espaço do preditor em um número simples de regiões, resumindo em uma árvore (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013). Hastie, Tibshirani e Friedman (2009) afirmam ainda que métodos baseados em árvores de decisão representam boa alternativa de modelo preditivo quando a relação entre as covariadas e a resposta de interesse é complexa e não linear.

A árvore de decisão poder ser aplicada tanto para respostas contínuas, em casos de problemas de regressão, como para respostas categóricas, nos problemas de classificação. Em os ambos os casos, para fazer uma previsão para uma dada observação, normalmente usa-se a média ou a moda das observações de treinamento na região à qual ela pertence. São simples e úteis para interpretações, contudo, não são competitivas com as melhores abordagens de aprendizado supervisionadas, uma vez que apresentam um poder preditivo muito baixo quando comparados aos demais algoritmos (JAMES *et al.*, 2013; IZBICKI; SANTOS, 2018).

4.3.2.7.1 *Regression trees*

O processo de construção de uma árvore de regressão, *regression trees*, pode ser resumido em duas etapas: (i) Divisão do espaço dos preditores, isto é, o conjunto de valores possíveis para $X_1, X_2, ..., X_p$, em J regiões distintas e não sobrepostas, $R_1, R_2, ..., R_j$; (ii) Para cada observação que cai na região R_j, é feita a mesma previsão, em que é calculada a média dos valores de resposta para as observações de treinamento em R_j (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013). Dessa maneira,

a construção de uma árvore de regressão consiste em encontrar $R_1, R_2, ..., R_j$ que minimizem SQR, dado por:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \tag{4.58}$$

em que $\hat{y}_{R_j}$ é a resposta média para as observações de treinamento dentro de cada região R_j. Entretanto, James *et al.* (2013) destaca que é computacionalmente inviável considerar todas as partições possíveis do espaço dos preditores em J regiões. Por essa razão, a sugestão é adotar uma abordagem conhecida como *recursive binary splitting*, divisão binária recursiva. A qual inicia-se no topo da árvore (ponto em que todas as observações pertencem a uma única região) e logo em seguida dividir sucessivamente o espaço do preditor; cada divisão é subdividida por dois novos ramos mais abaixo na árvore.

De acordo com James *et al.* (2013), a divisão binária recursiva consiste em selecionar o preditor X_j e o ponto de corte s de forma que divida o espaço preditivo nas regiões $\{X|X_j < s\}$ e $\{X|X_j \geq s\}$ que resulta no menor SQR. Isto é, considerar todos os preditores $X_1, X_2, ..., X_p$ e todos os valores possíveis do ponto de corte s para cada um dos preditores, e depois escolher o preditor e o ponto de corte de tal forma que a árvore resultante tenha o menor SQR, ou seja, para qualquer j e s, definir um par de regiões:

$$R_1(j,s) = \{X|X_j < s\} \quad e \quad R_2(j,s) = \{X|X_j \geq s\} \tag{4.59}$$

onde busca os valores de j e s que minimizam o SQR da Equação 4.59:

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{4.60}$$

em que $\hat{y}_{R_1}$ e $\hat{y}_{R_2}$ representam a resposta média para a base de treinamento em $R_1(j,s)$ e em $R_2(j,s)$, respectivamente. Em seguida, o processo será repetido com o objetivo de procurar o melhor preditor e o melhor ponto de corte, no entanto, em vez de dividir todo o espaço do preditor, será dividido uma das duas regiões, identificadas anteriormente, agora em três regiões, de modo a minimizar o SQR, e esse processo continua até que um critério de parada seja alcançado (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013). Uma vez que as regiões $R_1, R_2, ..., R_J$ foram criadas, a previsão da resposta para uma dada observação de teste corresponderá à resposta média das observações de treinamento na região na qual a observação de teste pertence.

Por outro lado, Hastie, Tibshirani e Friedman (2009) e James *et al.* (2013) afirmam também que o processo descrito anteriormente pode produzir boas previsões no conjunto de treinamento, no entanto é mais provável que supere os dados, levando a um desempenho ruim no conjunto de testes, resultando assim em uma árvore que pode ser muito complexa. Dessa forma, sugerem que a solução seja uma árvore menor, com menos divisões, ou seja, menos regiões $R_1, R_2, ..., R_J$, em que pode acarretar a uma variação menor e a uma melhor interpretação à custa de um menor viés.

Portanto, a melhor estratégia é "cultivar" uma árvore cheia, T_0, e depois "podá-la" de forma a obter uma subárvore. O objetivo é selecionar uma subárvore T, menos complexa e que leve à menor taxa de erro nas novas observações na base de teste (JAMES *et al.*, 2013). A *Cost complexity pruning*, complexidade do custo de podar, também conhecida como poda de elos mais fracos, evidência uma maneira de fazer isso. Ao invés de considerar todas as subárvores possíveis, considera-se uma sequência de árvores indexadas por um parâmetro de ajuste (*tunning*) não-negativo α.

Em que, de acordo com Kuhn e Johnson (2013), α será responsável por controlar o *trade-off* entre a complexidade do tamanho da subárvore e a qualidade no ajuste do modelo, e a sua otimização pode ser calculada por meio da validação cruzada *k-flod*. Assim, para cada valor de α corresponde uma subárvore $T \subset T_0$, tal que a função perda descrita a seguir:

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T| \tag{4.61}$$

é a menor possível. O termo $|T|$ representa o número de nós terminais (*terminal nodes*) da subárvore T; R_m, por sua vez, é o subconjunto do espaço preditor correspondente ao m nó terminal, e; $\hat{y}_{R_m}$ é a resposta prevista associada a R_m, isto é, a média das observações de treinamento em R_m. Na medida que α aumenta, menor será a subárvore T_α encarregada por minimizar a função perda e, mais interpretável será o modelo ajustado (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013).

4.3.2.7.2 *Classification trees*

Para Hastie, Tibshirani e Friedman (2009), James *et al.* (2013) e Raschka (2017), uma *classification trees*, árvore de classificação, é semelhante a uma árvore de regressão, exceto pelo fato de que aquela é usada para prever uma resposta qualitativa em vez de uma quantitativa. Nesse caso, em uma árvore de classificação, a predição da resposta para cada nova observação está relacionada à classe mais comum de observações da base treinamento na região à qual ela pertence. No que tange as interpretações dos resultados este consiste não apenas na predição da classe correspondente a uma região

de nó terminal em particular, mas também nas proporções das classes em que as observações de treinamento se enquadram.

Assim como no algoritmo de regressão, a divisão binária recursiva também se aplica para criar uma árvore de classificação. No entanto, James *et al.* (2013) destacam que na classificação, o SQR não pode ser aplicado como um critério para fazer as divisões binárias ou o processo de poda. No entanto, outros critérios podem ser utilizados, a saber: a taxa de erro de classificação, índice de Gini e o *cross-entropy* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). No que se refere a taxa de erro de classificação, esta corresponde à fração das observações da base treinamento naquela região que não pertencem à classe mais frequente dessa mesma região:

$$E = 1 - \underset{k}{max}(\hat{p}_{mk}) \tag{4.62}$$

em que $\hat{p}_{mk}$ representa a proporção de observações de treinamento da m-ésima região que são da k-ésima classe. No entanto, Hastie, Tibshirani e Friedman (2009) e James *et al.* (2013) verificaram que o erro de classificação não é suficientemente sensível para o "cultivo" de árvores e, na prática, duas outras medidas são necessárias. O índice de Gini que é definido por:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{4.63}$$

representa uma medida da variância total entre as K classes da reposta de interesse. É possível observar que o índice de Gini assume um valor pequeno se todos os $\hat{p}'_{mk}s$ estiverem próximos de zero ou um (JAMES *et al.*, 2013). Por essa razão, o índice de Gini é referido como uma medida de pureza do nó, de outra forma, pequenos valores indicam que um nó contém, predominantemente, observações de uma única classe.

Uma alternativa ao índice de Gini é a *cross-entropy*, a qual assumirá pequenos valores se o m-ésimo nó for puro. De fato, verifica-se que o índice de Gini e a *cross-entropy* são bastante semelhantes, e esta é definida por:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk}\ log\ \hat{p}_{mk} \tag{4.64}$$

em que $0 \leq \hat{p}_{mk} \leq 1$, segue-se que $0 \leq -\hat{p}_{mk}\ log\ \hat{p}_{mk}$. Assim, pode-se observar que a *cross-entropy* assumirá um valor próximo de zero se os $\hat{p}_{mk}$ estiverem próximos de zero ou próximos de um.

Para Hastie, Tibshirani e Friedman (2009) e James *et al.* (2013), tanto o índice Gini quanto a *cross-entropy* podem ser adotados para avaliar a qualidade da divisão das

regiões, uma vez que ambas são mais sensíveis à pureza do nó do que a taxa de erro de classificação. De modo que, qualquer uma das três abordagens pode ser aplicada para podar a árvore, mas a taxa de erro de classificação é preferível se a precisão (*accuray*) da previsão da árvore podada for a meta final.

Métodos baseados em árvore de decisão simples geram um conjunto de condições fáceis de implementar e interpretar (KUHN; JOHNSON, 2013). Contudo, os algoritmos de *Machine Learning* resultantes de árvores de decisão são considerados instáveis, de modo que pequenas alterações na base de treinamento podem acarretar mudanças estruturais na árvore ou nas suas regras, o que, consequentemente, pode alterar a interpretação do modelo ajustado.

Nesse contexto, foram desenvolvidos diversos métodos a fim de melhorar o desempenho preditivo em árvores de decisão simples. Cada um desses métodos envolve a produção de múltiplas árvores, que são então combinadas para produzir uma única previsão, mais precisa, mais acurada (JAMES *et al.*, 2013; RASCHKA, 2017). Dessa maneira, ante ao exposto, será discutido a seguir que a combinação de um grande número de árvores pode por muitas vezes resultar em melhorias drásticas na acurácia da previsão, às custas de alguma perda na interpretação. Os principais métodos abordados na literatura são *bagging*, *random forests* e *boosting* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013; KUHN; JOHNSON, 2013; LANTZ, 2013; IZBICKI; SANTOS, 2018).

4.3.2.7.3 *Bagging*

O *bootstrap* é o método mais abordado em situações nas quais é difícil, ou mesmo impossível, calcular diretamente o desvio padrão de uma quantidade de interesse. Em *machine learning*, o *bootstrap* pode ser usado em um contexto diferente, a fim de melhorar a performance dos métodos, representando uma técnica de reamostragem. Como também, pode ser adotado em conjunto com outros algoritmos que resultam em modelos instáveis, como por exemplo em árvores de decisão, com o objetivo de uma melhor precisão na performance preditiva, o referido procedimento é conhecido como agregação *bootstrap* ou *bagging* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; KUHN; JOHNSON, 2013).

Hastie, Tibshirani e Friedman (2009) discutem que as árvores de decisão sofrem, costumeiramente, de alta variância. Por exemplo, se dividir os dados da base treinamento em duas partes aleatoriamente e ajustar uma árvore de decisão em ambas as metades, os resultados que obtidos poderão ser bem diferentes. Por outro lado, um procedimento com baixa variância produzirá resultados semelhantes se aplicado repetidamente a conjuntos de dados distintos. Nesse caso, as árvores de decisão é um

dos algoritmos que evidencia uma significativa melhora na sua performance preditiva quando utilizadas em conjunto com o método *bootstrap*, ou *bagging*.

Considere inicialmente um problema de regressão, e suponha o ajuste de um modelo em um conjunto de treinamento $Z = (x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$, obtendo a predição $\hat{f}(x)$ na entrada x. O *bagging*, calcula a média dessa previsão em relação a uma coleção de amostras de *bootstrap*, reduzindo assim sua variância. Para cada amostra de *bootstrap* Z_{*b}, $b = 1, 2, ..., B$, o algoritmo de árvore de decisão é ajustado, resultando em B modelos de previsão, sejam eles de regressão ou de classificação, e, portanto, dando a previsão $\hat{f}^{*b}(x)$ para uma mesma observação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013). A estimativa de *bagging* é definida por:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \tag{4.65}$$

O $\hat{\mathcal{P}}$ será denotado como a distribuição empírica que possui probabilidade de $1/N$ em cada um dos pontos dos dados (x_i, y_i). Na verdade, Hastie, Tibshirani e Friedman (2009) sugere que a estimativa "verdadeira" do *bagging* é definido por $E_{\hat{\mathcal{P}}}\hat{f}^*(x)$, onde $Z^* = (x*_1, y*_1), (x*_2, y*_2), ..., (x*_N, y*_N)$. Dessa maneira, considera que a Equação 4.65 é uma estimativa de Monte Carlo da verdadeira estimativa de *bagging*, aproximando-se como $B \to \infty$.

Suponha agora problemas de classificação, e que a árvore considere um classificador $\hat{G}_b(x)$ para a classe K da resposta de interesse. Nesse caso, se torna aceitável considerar uma função vetorial-indicador subjacente $\hat{f}(x)$, com valor um único e $K-1$ zeros, tal que $\hat{G}(x) = arg\ max_k \hat{f}(x)$. Portanto, a estimativa *bagging* $\hat{f}_{bag}(x)$ da Equação 4.65 é um K-vetor, $[p_1(x), p_2(x), ..., p_K(x)]$, com $p_K(x)$ igual à proporção de árvores que predizem a classe k em x. O classificador *bagging* seleciona a classe com mais "votos" das árvores B, $\hat{G}_{bag}(x) = arg\ max_k \hat{f}_{bag}(x)$ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Logo, a previsão final da resposta de interesse para uma observação será a classe mais votada pelas B árvores que foram agregadas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Contudo, de acordo com Kuhn e Johnson (2013), o algoritmo *bagging* apresenta uma desvantagem no que se refere ao fato das B árvores agregadas evidenciarem alta correlação devido à utilização de todas as covariadas como candidatas em todas as etapas da divisão das B árvores de decisão. Assim, como alternativa para reduzir a correlação supracitada, o algoritmo *random forest* pode ser aplicado ao problema.

4.3.2.7.4 *Random Forest*

A abordagem inicial consistia em construir árvores inteiras com base em subconjuntos aleatórios dos preditores (AMIT; GEMAN, 1997; BARANDIARAN, 1998). Nesse

mesmo sentido, Dietterich (2000) desenvolveu a ideia de seleção de divisão aleatória, em que as árvores são construídas usando um subconjunto aleatório dos principais k-preditores em cada divisão na árvore. Breiman (2000), por sua vez, também tentou adicionar ruído à resposta para perturbar a estrutura da árvore. Por fim, após avaliar cuidadosamente essas generalizações para o algoritmo de *bagging* original, Breiman (2001) construiu um algoritmo unificado chamado *random forest*.

O *Random Forest*, ou florestas aleatórias, proporciona uma melhoria em relação às árvores *bagging*, onde ocorre um pequeno ajuste aleatório a fim de reduzir a correlação existente entre os preditores da base de treinamento das árvores agregadas (JAMES *et al.*, 2013). Assim como no *bagging*, descrito anteriormente, a construção do *random forest* consiste na obtenção de B amostras de tamanho n por amostragem *bootstrap* e com reposição da base de treinamento, em que para cada base estima-se uma árvore de decisão.

Mas ao construir essas árvores, a cada vez que uma divisão binária recursiva é considerada, ou seja, em cada nó da árvore, uma amostra aleatória de preditores, de tamanho m e sem reposição, é escolhida como candidatos do conjunto completo de preditores p. A partir de então, realiza-se, no subgrupo de amostra, a escolha da combinação preditor-ponto de corte responsável por essa segmentação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O referido processo continua até que seja alcançado algum critério de parada.

A divisão binária recursiva tem permissão para usar apenas um desses m preditores. Supondo que para $m = p$, então a predição do *bagging* será igual ao do *random forest*. Por outro lado, Hastie, Tibshirani e Friedman (2009) e James *et al.* (2013) sugerem que para problemas de regressão é recomendado adotar $m \approx p/3$, e em casos de classificação, $m \approx \sqrt{p}$. Onde esses valores representam o hiperparâmetro do algoritmo James *et al.* (2013), e e podem serem encontrador por validação cruzada *k-fold*.

É importante destacar que na construção de uma *random forest*, a cada divisão na árvore, o algoritmo não pode sequer considerar a maioria dos preditores disponíveis. Pois, suponha que haja um preditor muito forte na base de dados de treinamento, junto com vários outros preditores moderadamente fortes. Logo, na coleção de árvores *bagging*, a maioria ou todas as árvores usarão esse forte preditor na divisão superior. Consequentemente, todas as parecerão semelhantes entre si. Portanto, as previsões das árvores *bagging* serão altamente correlacionadas. Infelizmente, calcular a média de muitas grandezas altamente correlacionadas não leva a uma redução tão grande na variância quanto a média de muitas quantidades não correlacionadas. Em suma, isso significa que o *bagging* não levará a uma redução substancial na variação de uma única árvore sob essa configuração (JAMES *et al.*, 2013).

A estimativa da predição da resposta de interesse em *random forest* é similar à descrita no *bagging*. Para problemas de regressão, é denotada por:

$$\hat{f}_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x) \tag{4.66}$$

em que $\hat{f}^{*b}(x)$ é a resposta predita pela B-ésima árvore *random forest*. Do mesmo modo, em problemas de classificação, assim como no *bagging*, considere $\hat{G}_{rf}(x) = arg\ max_k \hat{f}_{bag}(x)$ a predição da B-ésima árvore *random forest*, onde corresponderá a determinada classe da resposta interesse, sendo baseado no voto majoritário (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

4.3.2.7.5 *Boosting*

Os modelos de *boosting* foram originalmente desenvolvidos para problemas de classificação e posteriormente estendidos para problemas de regressão. Surgiram no início dos anos 90, (SCHAPIRE, 1990; FREUND, 1995; FREUND; SCHAPIRE, 1999), sob a influência da teoria da aprendizagem (VALIANT, 1984; KEARNS; VALIANT, 1994), a sua ideia principal consiste em combinar a previsão de um conjunto de classificadores fracos (cuja predição é previamente melhor do que uma classificação aleatória, pois apresenta uma taxa de erro inferior) a fim de construir um classificador superior encarregado pela predição final, caracterizando o método *AdaBoost*.

Segundo Kuhn e Johnson (2013), o algoritmo *AdaBoost* mostrou-se como uma poderosa ferramenta de previsão, geralmente superando qualquer modelo individual. Após a incorporação, através do método *AdaBoost*, dos conceitos da função perda, modelagem aditiva e regressão logística, Friedman *et al.* (2000) apresentaram os resultados nas generalizações para problemas de classificação. Como também, na sua extensão para problemas de regressão, denotando-se como o método *gradient boosting*, que tem como principal objetivo identificar um aditivo direto do modelo que minimize exponencialmente a função perda.

Em sua forma mais simples, o *gradient boosting* baseia-se em dada uma função perda, por exemplo SQR, e um algoritmo fraco, por exemplo árvores de regressão, dessa forma, o *gradient boosting* procura encontrar um modelo aditivo que minimize a função de perda. Em seguida, o algoritmo é inicializado com o melhor palpite da resposta de interesse, por exemplo, a média da resposta na regressão. O gradiente, por exemplo, o residual, é calculado e um modelo é então ajustado aos resíduos para minimizar a função de perda. Assim, o modelo atual é incluído no modelo anterior e o procedimento continua até que um número de interações especificado seja alcançado (KUHN; JOHNSON, 2013).

As árvore de decisão são consideradas excelentes para base na aplicação do *gradient boosting*, devido, principalmente, à possibilidade de adotá-la como um classificador fraco através da construção de uma árvore com poucas divisões, ou seja, com a profundidade restringida (KUHN; JOHNSON, 2013). Sob este cenário, quando a árvore de regressão é usada como base, o *gradient boosting* apresentará dois parâmetros no processo de ajuste do modelo, a saber: profundidade da árvore e o número de interações.

No que se refere a profundidade da árvore, também conhecida como profundidade de interação, está relacionada a quantidade de divisões em cada árvore, por sua vez, o número de interações, m, está relacionada diretamente ao número de árvores do modelo final do *boosting*, e ambos podem otimizados aplicando a validação cruzada *k-fold* (JAMES *et al.*, 2013).

Tanto para os problemas de regressão, como para os problemas de classificação, a árvore de decisão busca subdividir o espaço dos preditores em R_j regiões diferentes, onde $j = 1, 2, ...J$, e prever a resposta de interesse, γ_j, para a região diferente. Dessa maneira, a regra de predição de uma árvore de decisão pode ser apresentada da seguinte forma: $x \in R_j \Rightarrow f(x) = \gamma_j$. Assim, como a árvore completa pode ser representada formalmente como:

$$T(x;\theta) = \sum_{j=1}^{J} \gamma_j \, I(x \in R_j) \tag{4.67}$$

em que $\theta = \{R_j, \gamma_j\}$. As R_j regiões serão estabelecidas a partir de minimização da função perda, $L(.)$:

$$\hat{\theta} = \underset{\theta}{argmin} \sum_{i=1}^{N} L(y_i, T(x;\theta)) \tag{4.68}$$

Logo, a predição da j-ésima região, dada a $\hat{R}_j$, corresponderá à resposta média, para os casos de árvores de regressão, e à classe mais comum, para os casos de árvores de classificação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O algoritmo *gradient boosting*, aplicado em conjunto com as árvores de decisão, tem como objetivo adicionar novas árvores à função de previsão $\hat{f}$, em cada interação m, sem a necessidade de ajustar novamente os coeficientes das árvores recém adicionadas, implicando assim uma aproximação gradual em f (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013). Resultando em um modelo que pode ser apresentado

da seguinte forma:

$$f_M = \sum_{m=1}^{M} T(x; \theta_m) \tag{4.69}$$

onde $\theta_m = \{R_{jm}, \gamma_{jm}\}^{J_m}$. O modelo atual sendo dado por $f_{m-1}(x)$, a função perda é otimizada a partir de:

$$\hat{\theta}_m = \underset{\theta_m}{argmin} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + T(x_i; \theta_m)) \tag{4.70}$$

e a atualização da função de previsão assume:

$$f_m(x) = f_(m-1)(x) + T(x; \theta_m) \tag{4.71}$$

A partir do momento que o número de interações é alcançado, a função de predição denota $\hat{f}(x) = f_M(x)$. Assim, o *gradient boosting* é regularizado a partir da aplicação de penalização, (λ), à árvore de decisão que será adicionada à função de predição:

$$f_m(x) = f_(m-1)(x) + \lambda T(x; \theta_m) \tag{4.72}$$

O termo de penalização, (λ), representa não só a contribuição de cada árvore para a função de previsão final, mas também o parâmetro do *gradient boosting* (JAMES *et al.*, 2013).

4.3.3 Critérios para avaliar o desempenho e selecionar o modelo

Está Subseção parte com o objetivo de apresentar os principais critérios adotados para avaliar e selecionar o modelo com a melhor performance preditiva para o problema abordado. Consistente com a segunda etapa do processo de ajuste dos algoritmos de ML, o aprendizado, conforme foi exposto no início da Subseção 4.3.2. Como o problema adotado neste estudo tem como variável de interesse se o discente reprovou, sim ou não, isto é, uma variável qualitativa, os critérios que estão descritos a seguir tratam-se das medidas de desempenho para os modelos de classificação.

Quando a resposta de interesse é uma variável qualitativa, existem dois tipos de previsões que podem ser obtidos: contínua (p_k^*), que é uma estimativa da probabilidade de a nova observação pertencer a cada uma das classes, k, da resposta de interesse; e outra do tipo categórica, como por exemplo, 0=caso ausente e 1=caso presente, que é uma predição para o valor da resposta de uma nova observação (KUHN; JOHNSON, 2013; MEURER; TOLLES, 2017).

Para James *et al.* (2013), as predições contínuas se tornam interessantes por permitirem a utilização do classificador (modelo ajustado) em diferentes contextos, a partir da imposição de pontos de corte estabelecido pelo pesquisador no conjunto de treinamento. Por exemplo, para o caso de respostas binárias, a nova observação é direcionada à classe $k = 1$ se $(p_k^*) > 0,5$, se $0,5$ for o ponto de corte determinado pelo pesquisador, que pode ser alterado em função do objetivo da análise.

James *et al.* (2013) destaca ainda que na prática, problemas com variáveis binárias podem implicar dois tipos de erros, trazendo para esta análise: atribuir erroneamente um aluno que não reprovou à classe correspondente de reprovação (reprovado = sim); ou ainda, atribuir incorretamente um aluno que reprovou à disciplina de cálculo diferencial e integral I correspondente à classe dos que aprovados (reprovado = não). Logo, a partir da imposição do ponto de corte, uma *confusion matrix*, matriz de confusão, torna-se necessária para observar os valores previstos e reais pertencentes a cada uma das classes.

4.3.3.1 *Confusion Matrix*

De acordo com James *et al.* (2013) e Kuhn e Johnson (2013), a matriz de confusão é uma medida de desempenho para problemas de classificação, onde é apresentada a tabulação cruzada a partir de uma tabela com quatro combinações diferentes de valores reais (observados) e previstos. A Tabela 4.7 a seguir consiste em uma matriz de confusão para um problema contendo duas classes, a classe do evento ocorrer ("*events*", assumindo 1 quando for positivo) e a classe do evento não ocorrer ("*nonevents*", assumindo 0 quando for negativo), para cada tipo de valores, previsto (*predicted values*) e observado (*observed values*) .

Tabela 4.7 – Matriz de Confusão para um problema com duas classes.

		Observed Values	
		Event (1)	**Nonevent (0)**
Predicted Values	**Event (1)**	TP	FP
	Nonevent (0)	FN	TN

Fonte: Adaptação a partir do Livro de Kuhn e Johnson (2013).
Nota: True Positives (TP), False Positives (FP), True Negatives (TN), e False Negatives (FN)

Por sua vez, a Tabela 4.9 adaptada a matriz de confusão da Tabela 4.7 para a problemática abordada nesta pesquisa, com duas possibilidades do evento ocorrer, ou seja, quando o discente é reprovado ou não. As células da tabela indicam o número dos Verdadeiros Positivos (VP), Falsos Positivos (FP), Verdadeiros Negativos (VN) e

Falsos Negativos (FN). Fica claro que as observações que se localizarem nas classes pertencentes a diagonal principal, VP e VN, são as que foram preditas corretamente, isto é, são os estudantes que foram reprovados e aprovados, respectivamente, na disciplina de cálculo diferencial e integral I (JAMES *et al.*, 2013; KUHN; JOHNSON, 2013).

Tabela 4.9 – Matriz de Confusão para prever o risco de reprovação no ensino superior no Brasil.

		Classe Observada	
	Reprovado	**Sim (1)**	**Não (0)**
Classe Prevista	**Sim (1)**	VP	FP
	Não (0)	FN	VN

Fonte: Adaptação a partir do Livro de Kuhn e Johnson (2013).

Entretanto, as observações que situam-se fora da diagonal principal correspondem aos discentes aprovados mas estão classificados como reprovados (FP), e os estudantes reprovados que estão classificados como aprovados (FN) e indicam os erros de classificação (JAMES *et al.*, 2013; KUHN; JOHNSON, 2013). A partir da matriz de confusão é possível determinar os critérios que foram adotados para avaliar a melhor performance preditiva dos algoritmos de ML, a saber: *Accuracy* (acurária), o *Sensitivity* (sensibilidade) e a *Specificity* (especificidade).

4.3.3.2 *Accuracy*, *Sensitivity* e *Specificity*

- *Accuracy*

A *Accuracy* (Acc), ou acurácia, de acordo com Kuhn e Johnson (2013), é a relação mais simples originada da matriz de confusão, e apresenta a concordância entre as classes observadas e previstas, tendo uma interpretação mais direta. No entanto, existem algumas desvantagens: (i) primeiro, as contagens gerais de precisão não fazem distinção entre o tipo de erro cometido, ou seja, em situações em que os custos são diferentes, a precisão pode não medir as características importantes do modelo, como por exemplo, classificar um aluno como reprovado quando o mesmo é aprovado ou o inverso; (ii) em segundo lugar, é preciso considerar as frequências naturais de cada classe, pois está métrica não considera (JAMES *et al.*, 2013; KUHN; JOHNSON, 2013). É dada por:

$$Acc = \frac{VP + VN}{VP + FP + FN + VN} \tag{4.73}$$

- *Sensitivity*

Quando o objetivo é determinar o erro derivado de um classificador, ou seja, conhecer a performance específica de acordo com as classes da resposta de interesse (reprovado ou não), as análises de *Sensitivity* e *Specificity*, sensibilidade e especificidade, respectivamente, podem ser adotadas. A sensibilidade do modelo é a taxa das classes positivas observadas, quando é previsto corretamente. De outra maneira, é a proporção de VP na classe da resposta de interesse [Sim (1)] que foi de fato observada (KUHN; JOHNSON, 2013). Pode ser visualizada em:

$$sensitivity = \frac{VP}{VP + FN} \tag{4.74}$$

Kuhn e Johnson (2013) afirma que a sensibilidade é por vezes considerada a taxa de verdadeiro positivo, visto que mede a precisão na população do evento.

- *Specificity*

Por outro lado, a *Specificity*, ou especificidade, é definida como a proporção de VN na classe da resposta de interesse ausente [Não (0)], que também é observado, assim como na sensibilidade (JAMES *et al.*, 2013; KUHN; JOHNSON, 2013).

$$specificity = \frac{VN}{FP + VN} \tag{4.75}$$

Assumindo um nível fixo de precisão para o modelo, Kuhn e Johnson (2013) destaca a existência de um *trade-off* entre a sensibilidade e a especificidade. Intuitivamente, aumentar a sensibilidade de um modelo é susceptível para uma perda de especificidade, uma vez que mais amostras estão sendo previstas como eventos. Os potenciais *trade-off* podem ser apropriados quando existem penalidades diferentes associadas a cada tipo de erro. A curva ROC (*Receiver Operating Characteristic*) é uma técnica utilizada para avaliar esse *trade-off* e é discutida a seguir.

4.3.3.3 A Curva *Receiver Operating Characteristic* (ROC)

O método gráfico mais comum para combinar sensibilidade e especificidade em um único valor é a curva ROC (PRATI; BATISTA; MONARD, 2008; KUHN; JOHNSON, 2013). Foi projetada como um método geral que, dada um conjunto de dados contínuos, determina um limiar efetivo tal que valores acima do limiar são indicativos de um evento específico. Ressaltando que um algoritmo de classificação gera uma probabilidade de classes, esta ferramenta é apropriada para avaliar a sensibilidade e

especificidade decorrente dos possíveis pontos de corte para p_k^* (ALTMAN; BLAND, 1994; BROWN; DAVIS, 2006; FAWCETT, 2006).

De acordo com Prati, Batista e Monard (2008), a curva ROC também é útil quando se encontra em domínios onde existe forte desproporção entre as classes ou ainda quando deve-se levar em consideração diferentes benefícios/custos para os diferentes acertos/erros dos algoritmos de classificação. Logo, a análise gráfica é utilizada tanto no processo de construção (PRATI; FLACH, 2005) como no ajustamento do modelo (FLACH; WU, 2005).

Ressaltando que a sensibilidade é a taxa de precisão apenas para a população do evento, e a especificidade para os não eventos, ao alterar o limite que maximiza adequadamente o compromisso entre sensibilidade e especificidade, a curva ROC só terá o efeito de tornar as amostras mais positivas (ou negativas, conforme o caso). Na matriz de confusão, ele não pode mover amostras de ambas as células da tabela fora da diagonal. Há quase sempre uma diminuição na sensibilidade ou especificidade quando 1 é aumentado (KUHN; JOHNSON, 2013).

Para James *et al.* (2013), o desempenho geral de um classificador, resumido em todos os possíveis limiares, é dado pela área sob a curva (ROC) (AUC). Uma curva ROC ideal vai abraçar o canto superior esquerdo, então quanto maior a AUC, melhor o classificador. Na tentativa de simplificar a análise da curva ROC, a AUC (*Area Under the ROC Curve*) é a derivada da curva ROC, de maneira que esta busca sintetizar a curva ROC num único valor (KUHN; JOHNSON, 2013), que varia de 0,0 a 1,0, tendo como o limiar de 0,5 entre elas. Isto é, superior a esse limite, o método classifica-se em uma classe e inferior na outra classe. Logo, quanto maior a área AUC, ou seja, mais próxima de 1, melhor a performance do modelo (MEURER; TOLLES, 2017).

As curvas ROC são úteis para comparar diferentes classificadores, pois levam em conta todos os possíveis limiares, pois para cada variação do limiar do classificador são alteradas as taxas de VP e FP. Logo, a cada limiar candidato, a Taxa de Verdadeiro Positivo (TVP), isto é, a sensibilidade, e a Taxa de Falso Positivo, 1 - especificidade, resultantes são representados um contra o outro (KUHN; JOHNSON, 2013). Que nada mais são do que as contagens reais da população em cada classe. Tendo a matriz de confusão como base, as TVP e a TFP são dadas por:

$$TVP = \frac{VP}{VP + FN} \quad \text{e} \quad TFP = \frac{FP}{VN + FP} \tag{4.76}$$

4.4 Resultados

Ante a gama de algoritmos apresentados na Seção de estratégia empírica, tanto no que se refere aos do grupo de modelos mais tradicionais, quanto aos de ML, torna-se necessário o cumprimento de três etapas essenciais: comparação, seleção e avaliação dos modelos. Dessa forma, os resultados desta pesquisa se encontram estruturados em três subseções que objetivam identificar a partir de diferentes abordagens aquela que resulta em uma melhor performance preditiva.

4.4.1 Comparação de performance dos Modelos

A ideia é confrontar a acurácia da previsão aplicando tanto o método em um ambiente de algoritmos estatísticos mais usuais, como também de algoritmos pertencentes ao grupo de ML. Na Tabela 4.11 estão apresentadas as estimações dos modelos tracionais, adequados para variável de resposta binária, com o objetivo de prever o risco de reprovação dos discentes matriculados na disciplina de cálculo diferencial e integral I na UFPB, nos anos de 2010 e 2016.

A base inicialmente estimada é composta por 8.659 observações, sendo 37,34% proporcional a classe dos discentes aprovados e 62,66% a classe dos discentes reprovados. A partir da referida base de dados, no contexto da econometria tradicional, foram estimados os modelo de Probabilidade linear, *Logit* e *Probit*. Tomando para a análise o modelo de regressão logística ou *logit*, este será ajustado para prever a probabilidade do discente reprovar usando as covariadas descritas na Seção 4.2

Tabela 4.11 – Estimações dos modelos tradicionais para prever o risco de reprovação dos discentes, 2010 a 2016.

	Reprovado					
	Prob. Linear		**Logit**		**Probit**	
	Coef.	*Odds Ratio*	Coef.	*Odds Ratio*	Coef.	*Odds Ratio*
Discente						
Nota Vestibular	−0,0003**	0,99	−0,0018**	0,99	−0,0011**	0,99
	(0,0001)		(0,0007)		(0,0004)	
Nota Vest. Matemát.	−0,0008***	0,99	−0,0037***	0,99	−0,0022***	0,99
	(0,0001)		(0,0004)		(0,0003)	
Casado	−0,0119	0,98	−0,0780	0,92	−0,0316	0,96
	(0,0226)		(0,1263)		(0,0729)	
Migrante	−0,0614***	0,94	−0,3014***	0,73	−0,1812***	0,83
	(0,0155)		(0,0758)		(0,0459)	
Raça	−0,00004	0,99	−0,0003	0,99	0,0007	1,00
	(0,0107)		(0,0538)		(0,0322)	
Sexo	−0,0489***	0,95	−0,2356***	0,79	−0,1403***	0,86
	(0,0111)		(0,0555)		(0,0333)	
Idade Ing.	0,0088***	1,00	0,0544***	1,05	0,0294***	1,02
	(0,0010)		(0,0062)		(0,0035)	
Cotista	0,0699***	1,07	0,3373***	1,40	0,2055***	1,22
	(0,0132)		(0,0672)		(0,0401)	
Período Ing.	0,0046	1,00	0,0142	1,01	0,0101	1,01
	(0,0142)		(0,0703)		(0,0424)	
Forma de Ing.	0,0900***	1,09	0,4437***	1,55	0,2608***	1,29
	(0,0167)		(0,0843)		(0,0504)	
Docente						
Tempo de Grad.	−0,0022***	0,99	−0,0108***	0,98	−0,0064***	0,99
	(0,0006)		(0,0031)		(0,0019)	
Doutorado	0,0967***	1,10	0,4938***	1,63	0,2936***	1,34
	(0,0141)		(0,0709)		(0,0424)	
Public. no Ano	−0,0814***	0,92	−0,4215***	0,65	−0,2511***	0,77
	(0,0147)		(0,0744)		(0,0445)	
Estrangeiro	0,0360	1,03	0,2285*	1,25	0,1326*	1,14
	(0,0249)		(0,1322)		(0,0780)	
Ded. exclusiva	0,0524**	1,05	0,2558**	1,29	0,1560**	1,16
	(0,0218)		(0,1078)		(0,0650)	
Sexo	−0,0193	0,98	−0,0871	0,91	−0,0512	0,95
	(0,0118)		(0,0596)		(0,0357)	
Curso						
Local do Campus	0,1179***	1,12	0,5669***	1,76	0,3397***	1,40
	(0,0309)		(0,1552)		(0,0929)	
EF do Curso	X		X		X	
EF do Centro	X		X		X	
Turma						
Turno	−0,0168	0,98	−0,1109	0,89	−0,0660	0,93
	(0,0193)		(0,1032)		(0,0604)	
C. Horária	0,0403	1,04	0,2238	1,25	0,1561	1,16
	(0,0451)		(0,2238)		(0,1350)	
Média N. Vest.	−0,0015***	0,99	−0,0070***	0,99	−0,0040**	0,99
	(0,0005)		(0,0026)		(0,0016)	
Média N. Vest. Mat.	0,0005	1,00	0,0024	1,00	0,0014	1,00
	(0,0004)		(0,0020)		(0,0012)	
Taxa de Cotista	−0,0010**	0,99	−0,0045**	0,99	−0,0026**	0,99
	(0,0004)		(0,0021)		(0,0013)	
Observações	8.659		8.659		8.659	
Log *Likelihood*	-		−5.042,03		−5.046,12	

Fonte: Elaboração própria a partir dos microdados do STI/UFPB e Plataforma *Lattes* do CNPq.
Nota: (1) Níveis de significância: *10%, **5% e ***1%. (2) Os valores entre parenteses são os erros padrões das variáveis.

Assim como adotado por James *et al.* (2013), o ponto de corte para definir a fronteira de decisão linear nesse modelo é de $0,5$, ou seja, para prever se o aluno vai reprovar ou não no ano, deve-se calcular se a probabilidade de reprovação é maior ou menor do que $0,5$. Dadas essas previsões, a Tabela 4.13 expõe a matriz de confusão construída para determinar quantas observações foram classificadas corretamente e incorretamente. Logo, o modelo previu corretamente que 4.611 discentes reprovariam (VP) e que 1.403 seriam aprovados (VN). Neste caso, a regressão logística previu corretamente a reprovação e a aprovação dos estudantes em 69,45% (acurácia).

Tabela 4.13 – Matriz de Confusão do *Logit* - Econometria tradicional.

		Classe Observada	
	Reprovado	**Sim (1)**	**Não (0)**
Classe Prevista	**Sim (1)**	4611	1830
	Não (0)	815	1403

Fonte: Elaboração própria a partir da estimação exposta na Tabela 4.11, Modelo *Logit*.

De início, o modelo apresenta um bom poder de previsão, no entanto, esse resultado é enganoso porque o algoritmo foi treinado e testado no mesmo conjunto de 8.659 observações. Nesse contexto, a taxa de erro de treinamento é de 30,55% (100% - 69,45%). Importante ressaltar que ao aplicar um modelo de econometria em situações de previsão como essa, a taxa de erro de treinamento pode ser, por muitas vezes, excessivamente otimista, tendendo a subestimar a taxa de erro de teste.

Portanto, a fim de obter uma melhor acurácia do modelo de regressão logística, sob a ótica do ML, o ajustamento será feito usando uma parte dos dados, no conjunto de treinamento, e em seguida examinar quão bem sua previsão no restante dos dados, no conjunto de teste. Para James *et al.* (2013), esta estrutura resultará em uma taxa de erro mais realista, de modo que o foco é o desempenho do modelo e o quanto é capaz de prever situações futuras dos discentes onde as quais ainda são desconhecidas.

Uma vez que a base de dados refere-se ao período de 2010 a 2016, optou-se em dividi-la em dois subconjuntos, usando o ano de 2015 como critério de corte. Assim, as observações até 2015 compõem o conjunto de treinamento ($ano < 2015$), e as observações de 2015 e 2016 formam o conjunto de teste ($ano \geq 2015$). A base de treinamento possui 7.127 elementos, enquanto a base de teste é composta por 1.532 observações.

Após o ajuste no primeiro conjunto, a probabilidade será prevista para os próximos alunos que se matricularam na disciplina de cálculo diferencial e integral I e que tem o risco de reprovar no conjunto de teste, isto é, para os anos 2015 e 2016. Deste

modo, a Tabela 4.15 apresenta a regressão logística estimada na base de treinamento, para datas anteriores a 2015, para que em seguida as previsões possam ser calculadas e a respectiva matriz de confusão seja construída.

Tabela 4.15 – Estimação do algoritmo de regressão logística - *Machine Learning* - Base de treinamento

Variável	Coeficiente	Erro-padrão	*Odds Ratio*
Discente			
Nota Vestibular	−0,001	0,001	0,99
Nota Vest. Matemática	−0,004***	0,000	0,99
Casado	−0,062	0,138	0,94
Migrante	−0,453***	0,091	0,63
Raça	0,018	0,057	1,01
Sexo	−0,283***	0,061	0,75
Idade Ingresso	0,054***	0,007	1,05
Cotista	0,431***	0,079	1,53
Período Ingresso	0,155**	0,079	1,16
Forma de Ingresso	0,336***	0,091	1,39
Docente			
Tempo de Graduação	−0,009***	0,003	0,99
Doutorado	0,539***	0,077	1,71
Publicação no Ano	−0,394***	0,083	0,67
Estrangeiro	0,184	0,140	1,20
Dedicação exclusiva	0,256**	0,118	1,29
Sexo	−0,073	0,067	0,92
Curso			
Local do Campus	0,512***	0,167	1,66
EF do Curso	x		
EF do Centro	x		
Turma			
Turno	−0,169	0,117	0,84
Carga Horária	0,300	0,287	1,34
Média N. Vestibular	−0,010**	0,004	0,99
Média N. Vest. Mat.	0,003	0,003	1,00
Taxa de Cotista	−0,003	0,002	0,99
N		7.532	
Log Likelihood		−4.110,11	

Fonte: Elaboração própria a partir dos microdados do STI/UFPB e Plataforma *Lattes* do CNPq.
Nota: (1) Níveis de significância: *10%, **5% e ***1%.

Do mesmo modo, dada às previsões obtidas a partir da fronteira de decisão linear com um limiar de 0,5, agora para uma base de teste contendo 1.532 elementos, a matriz de confusão está exposta na Tabela 4.17. A acurácia calculada previu corretamente 67,16% das situações (reprovado e aprovado) dos discentes matriculados na disciplina de cálculo diferencial e integral I na UFPB nos anos de 2015 e 2016, em que 674 foram reprovados (VP) e 355 aprovados (FP). Assim como previsto, a taxa de erro de teste de 32,83% foi superior a taxa de erro obtida a partir de um processo de previsão treinada e testada na mesma base que foi de 30,55%, visto a partir da matriz

de confusão da Tabela 4.13.

Tabela 4.17 – Matriz de Confusão do *Logit* - ML - Base de teste.

		Classe Observada	
	Reprovado	**Sim (1)**	**Não (0)**
Classe Prevista	**Sim (1)**	674	248
	Não (0)	255	355

Fonte: Elaboração própia a partir da estimação exposta na Tabela 4.15.

4.4.2 Seleção do Modelo - Etapas de Aprendizado, Predição e Avaliação

Após a comparação dos métodos de econometria tradicional e a justificativa para aplicar os métodos de ML, esta Subseção vem abordar algumas etapas fundamentais do processo de ajuste dos modelos de previsão: o aprendizado, a predição e a avaliação dos algoritmos de classificação adotados e selecionados neste estudo. O desenvolvimento do aprendizado de um modelo de previsão é avaliado a partir de um conjunto de base de teste, em que este não foi utilizado no processo de ajuste dos métodos, isto é, não foi utilizado como base de treinamento. Vale lembrar que o conjunto de treinamento busca avaliar a capacidade preditiva de cada variável para a classificação final do estudante, enquanto o conjunto de teste analisa a eficiência do modelo gerado para a previsão em novas bases.

Como já discutido anteriormente e defendido por James *et al.* (2013) e Athey (2018), não há um único algoritmo de ML que domine todos os outros em todas as bases de dados. De outro modo, a aplicação de um determinado algoritmo em um conjunto de dados pode funcionar melhor do que os demais, contudo, o mesmo pode não apresentar o mesmo desempenho em uma outra base com um novo problema.

Com o objetivo de adotar o conjunto de variáveis, dentre as quais estão descritas na Seção 4.2, que irá fornecer a melhor análise preditiva do problema em questão, o critério adotado no processo de seleção se baseou na análise de vários modelos *logit* para fins de previsões. Considerado um modelo clássico entre os algoritmos de ML, foram estimadas nove regressões logísticas com as mais diversas combinações entre os conjuntos de características. Por sua vez, a Tabela 4.19 apresenta todas as combinações entre as covariadas na etapa de aprendizado na base de treinamento, como também os resultados dos indicadores de qualidade (*AUC ROC*, *Accuracy*, *Sensitivity* e *Specifity*) quando os modelos já ajustados são aplicados as novas observações da base de teste na etapa de predição.

Ante aos resultados dos critérios de desempenho expostos na Tabela 4.19, os

modelos 7 e 8 apresentaram os conjuntos de variáveis com as melhores estimativas de previsões em uma nova base de dados: *AUC ROC* (70,36%), *Accuracy* (67,16%) e *Sensitivity* (73,10%). Ambos os modelos detiveram os mesmos resultados, muito embora as *dummies* que mensuram o *EF do Centro* varie entre eles. Sendo assim, optou-se em selecionar a base de dados que contém todas as variáveis que compõem as quatro dimensões: discente, docente, curso e turma. Pois, apesar que no modelo de regressão logística a covariada do *EF do Centro* não tenha impacto na análise de previsão do discente reprovar, uma vez que não houve diferença nos resultados na sua presença ou ausência, a mesma pode ter influência nas estimações dos demais algoritmos de ML.

Tabela 4.19 – Critérios de Seleção das Variáveis - Modelo *Logit*

Dimensão	Variáveis	Modelos								
		1	2	3	4	5	6	7	8	9
Discente	Nota Vestibular	x	x	x	x	x	x	x	x	x
	Nota Vest. Mat.	x	x	x	x	x	x	x	x	x
	Casado	x	x	x	x	x	x	x	x	x
	Migrante	x	x	x	x	x	x	x	x	x
	Raça	x	x	x	x	x	x	x	x	x
	Sexo	x	x	x	x	x	x	x	x	x
	Idade Ing.	x	x	x	x	x	x	x	x	x
	Cotista	x	x	x	x	x	x	x	x	x
	Período Ing.	x	x	x	x	x	x	x	x	x
	Forma de Ing.	x	x	x	x	x	x	x	x	x
Docente	Tempo de Grad.		x	x	x	x	x	x	x	x
	Doutorado		x	x	x	x	x	x	x	x
	Public. no Ano		x	x	x	x	x	x	x	x
	Estrangeiro		x	x	x	x	x	x	x	x
	Dedic. exclusiva		x	x	x	x	x	x	x	x
	Sexo		x	x	x	x	x	x	x	x
Curso	Local do Campus			x	x	x	x	x	x	
	EF do Curso				x		x	x	x	
	EF do Centro					x	x	x		
Turma	Turno							x	x	x
	Carga Horária							x	x	x
	Média N. Vest.							x	x	x
	Média N. V. Mat.							x	x	x
	Taxa de Cotista							x	x	x
	N	1532	1532	1532	1532	1532	1532	1532	1532	1532
Critérios	AUC ROC (%)	70,05	69,32	70,25	70,22	69,54	70,22	70,36	70,36	69,42
	Accuracy (%)	66,77	65,73	66,77	67,55	66,90	67,55	67,16	67,16	66,18
	Sensitivity (%)	69,12	69,27	69,64	70,41	70,17	70,41	73,10	73,10	72,90
	Specificity (%)	60,82	58,05	60,00	61,18	59,87	61,18	58,19	58,19	56,69

Fonte: Elaboração própria a partir dos microdados do STI/UFPB e Plataforma *Lattes* do CNPq.

A seguir, após a determinação do conjunto de variáveis que compõe a presente análise, é necessário comparar os algoritmos e selecionar aquele com a melhor previsão relacionada a etapa de predição do processo dos modelos de ML. A Tabela 4.21 sumariza os resultados dos indicadores de desempenho, descritos na Subseção 4.3.3,

adotados como critérios para selecionar o algoritmo com a melhor análise preditiva: *Accuracy*, *Sensitivity*, *Specifity* e *AUC ROC*. Por se tratar de um problema de classificação, foram estimados doze diferentes métodos, descritos na Subseção 4.3.2, e comparadas as suas respectivas performances na base de teste na etapa de predição.

Em suma, ao analisar a Tabela 4.21 pode-se observar que grande parte dos algoritmos tiveram performances similares, com exceção para os modelos *KNN* e *Naïve Bayes Classifier*, os quais apresentaram as menores *Accuracy* e *AUC ROC*. Como para Prati, Batista e Monard (2008) e Kuhn e Johnson (2013), a curva *ROC* é o método mais comum para combinar a *Sensitivity* e *Specifity*, o desempenho que mensura a área sob a curva *ROC*, a *AUC ROC*, e a *Accuracy* foram os dois critérios, em conjunto, determinantes para selecionar os modelos com as melhores precisões de previsão: Regressão Linear, *Penalized Methods LASSO*, Regressão Logística, SVM e o *Boosting*.

Tabela 4.21 – Estimações dos algoritmos de *Machine Learning* para prever o risco de reprovação dos discentes matriculados na disciplina de cálculo diferencial e integral I na UFPB, nos anos de 2010 e 2016.

Algoritmos		**Critérios para avaliação do desempenho (%)**				
		Accuracy	***Sensitivity***	***Specificity***	***AUC ROC***	***IC 95% (AUC)***
Regressão Linear		66,97	69,42	62,35	69,97	(67,30 - 72,63)
Penalized Methods	***Ridge***[1]	66,64	65,76	64,34	69,45	(66,78 - 72,13)
	Lasso[2]	67,42	72,55	59,53	69,77	(67,10 - 72,44)
	Elastic Net[3]	66,51	65,66	65,33	69,77	(67,10 - 72,44)
Regressão Logística		67,16	72,53	58,87	70,36	(67,70 - 73,01)
K-Nearest Neighbors (KNN)[4]		58,42	50,48	70,14	60,51	(58,07 - 62,94)
Naive Bayes Classifier		57,04	47,72	70,31	59,38	(56,95 - 61,81)
Support Vector Machines (SVM)		67,88	83,10	44,44	63,77	(61,45 - 66,09)
Decision Tree Based Methods	***C. trees***	63,18	67,81	56,05	61,93	(59,45 - 64,42)
	Bagging	64,22	65,44	61,02	67,25	(64,50 - 69,90)
	Random Forest	64,68	68,35	56,88	65,95	(63,16 - 68,73)
	Boosting	66,26	72,55	58,70	69,47	(66,79 - 72,16)

Fonte: Elaboração própria a partir dos microdados do STI/UFPB e Plataforma *Lattes* do CNPq.
Nota: [1]$\lambda_{Ridge} = 0,21$; [2]$\lambda_{Lasso} = 0,01$; [3]$\lambda_{EN} = 0,017$; [4]$K = 300$ (Parâmetros *tunning*).

Ao se deparar com situações como essa, Kuhn e Johnson (2013) sugerem que seja feita inicialmente a comparação dos modelos baseados em termos de performance (como na Tabela 4.21), ponderando alguns benefícios principais: a interpretabilidade do algoritmo, a complexidade computacional e a facilidade de implementação. Em um exemplo de escolha de modelo final, os autores supracitados destacam que os pesquisadores devem avaliar primeiro os modelos mais flexíveis e menos interpretáveis como o SVM, o *boosting* e o *random forest*; em seguida investigar os métodos mais simples como o *LASSO* e o *ridge*. Caso ambos os modelos sejam equivalentes em termos de performance, o pesquisar deve optar pelo algoritmo mais simples que se assemelhe aos modelos mais complexos.

Logo, tendo como base as orientações de Kuhn e Johnson (2013), os modelos que deveriam ser selecionados como modelos finais neste estudo seriam a regressão Linear, o *LASSO* e a Regressão Logística. Porém é interessante destacar algumas desvantagens principalmente da aplicabilidade do modelo de regressão linear vistas na Subseção 4.3.2.1, a qual descreve o referido método. Hastie, Tibshirani e Friedman (2009) fundamentam que as principais características indesejáveis do modelo de regressão linear podem acarretar problemas de variância elevada ou até mesmo tendendo ao infinito. O que é um efeito não desejável no desempenho em um conjunto novo de dados.

Dessa maneira, devido aos problemas debatidos na literatura de ML a respeito da aplicação do método de regressão linear, Hastie, Tibshirani e Friedman (2009) indicam que os métodos lineares com penalidades, mais conhecidos como *penalized methods*, são mais adequados no processo de predição, pois, de acordo com James *et al.* (2013), os algoritmos pertencentes a este grupo buscam penalizar os coeficientes estimados com o objetivo de limitar a variância ao custo de um aumento insignificante no viés. Portanto, os modelos finais selecionados, nas etapas de aprendizado (base de treinamento) e predição (base de teste), para prever o risco de reprovação dos alunos que se matricularão na disciplina de cálculo diferencial e integral I na UFPB nos próximos semestres são: *LASSO* e Regressão Logística.

Por fim, no que se refere a etapa de avaliação, os critérios de *Accuracy* e *AUC ROC* que foram utilizados para a otimização e a seleção dos algoritmos, *LASSO* e Regressão Logística, durante o aprendizado evidenciaram desempenhos superiores a 67,1% e 69,7%, respectivamente, sem grandes diferenças em ambos os métodos. No que se refere a *Accuracy*, os dois algoritmos previram corretamente em torno de 67% das situações dos discentes na UFPB entre 2015 e 2016 (base de teste), situação de reprovado e aprovado. No tocante à interpretação das demais medidas expostas na Tabela 4.21, pode-se afirmar que, com uma Taxa de Falso Positivo (TFP) (1 - *Specifity*) em torno de 41% [1 - 59% (*Specifity*)] para ambos os métodos, é possível prever que 72,5% (*Sensitivity*) dos discentes que se matricularam na disciplina de cálculo diferencial e integral I na UFPB irão reprovar nos anos de 2015 e 2016, ou seja, dos 1.532 estudantes matriculados na base, 1.111 foram corretamente preditos.

4.4.3 Avaliação - Importância das Variáveis

Esta Subseção explicita, especificamente, a importância das variáveis no processo de ajuste dos modelos preditivos. Embora estas informações tenham sido colhidas na etapa de aprendizado, na base de dados de treinamento, torna-se necessário uma análise mais detalhada da relação entre as covariadas dos discentes no âmbito social e econômico do presente problema de pesquisa. Fundamentada, principalmente, na

discussão a qual busca os fatores que mais podem determinar a reprovação dos discentes que se matriculam na disciplina de cálculo diferencial e integral I UFPB entre 2010 e 2016.

Apesar de justificado na Subseção anterior que os modelos finais selecionados foram *LASSO* e Regressão Logística, estão apresentadas também as principais características determinantes dos modelos que obtiveram as melhores métricas de desempenho de previsão: Regressão Linear, SVM e o *Boosting*, a fim de uma maior robustez na discussão. Goldstein, Navar e Carter (2016) frisam que os métodos de ML buscam sintetizar o efeito das covariadas individuais através de métricas específicas, nomeada como a importância das variáveis. Dessa maneira, a Tabela 4.23 exibe o *ranking* de importância dos preditores para os cinco algoritmos supracitados.

Tabela 4.23 – Importância das 10 principais variáveis para a performance dos modelos selecionados - Base de treinamento.

Ranking	Modelos e Variáveis				
	Reg. Linear	***Lasso***	**Reg. Logística**	**SVM**	***Boosting***
1º	N. V. Mat.	L. do Campus	N. V. Mat.	Doutorado	Média N. V.
2º	Idade Ing.	Doutorado	Idade Ing.	Migrante	N. V. Mat.
3º	Doutorado	Cotista	Doutorado	Publ. no Ano	Idade Ing.
4º	Cotista	Migrante	Cotista	Cotista	Média N. V. Mat.
5º	Migrante	Sexo disc.	Migrante	Carga Horária	N. Vest.
6º	Sexo disc.	Forma de Ingr.	Publ. no Ano	Dedic. exclusiva	Doutorado
7º	Publ. no Ano	Publ. no Ano	Sexo disc.	Forma de Ingr.	Tempo de Grad.
8º	Forma de Ingr.	Idade Ingresso	Forma de Ingr.	Sexo disc.	Cotista
9º	L. do Campus	Dedic. exclusiva	*Dum.* Química	L. do Campus	Migrante
10º	*Dum.* Química	Média N. V.	L. do Campus	Estrangeiro	Tx. de Cotista

Fonte: Elaboração própria a partir dos microdados do STI/UFPB e Plataforma *Lattes* do CNPq.

Ainda de acordo com Goldstein, Navar e Carter (2016), embora não tenha efeito inferencial ou causal, a importância das variáveis sinaliza, de forma ordenada, quais são as variáveis que mais contribuíram na performance do modelo. Assim, a Tabela 4.23 contém a contribuição das dez primeiras variáveis de acordo com o ganho de informações no processo de ajuste. Com isso, cada uma delas auxilia, individualmente, a classificação final (reprovados ou aprovados) dos estudantes nos de 2015 e 2016 (etapa de predição). Em suma, o foco é analisar o quanto estas características de fato conseguem prever se aquele aluno vai ser reprovado (retido) ou não, desconsiderando a sua significância

O foco do ML consiste em gerar boas previsões, dada às restrições da máquina (LANTZ, 2013), ao contrário da econometria tradicional, a qual fundamenta-se em estabelecer efeitos de correlações ou causalidades a partir das conclusões estatisticamente significantes (WOOLDRIDGE, 2010; GUJARATI; PORTER, 2011). Logo, este

trabalho não se preocupa na significância do conjunto de características adotadas e estimadas nos modelos, mas sim em saber o quanto os métodos conseguem prever o comportamento observado em uma nova base de dados ao problema de retenção.

Do ponto de vista dos algoritmos, ambos produziram *rankings* com importâncias similares entre as variáveis, porém com sequências diversas. Goldstein, Navar e Carter (2016) justificam tal situação dado que como cada um dos métodos ajusta a sua forma estrutural de modo diferente, deve-se realmente esperar que os *rankings* correspondentes evidenciem diferenças não apenas na ordenação, como também nas variáveis que o compõem, assim como pode ser observado na Tabela 4.23. As variáveis mais constantes nos *rankings* foram: *Nota Vestibular Matemática*, *Nota Vestibular*, se o docente possui *Doutorado*, se o discente é *Cotista*, *Migrante*, *Idade Ingresso*, *Sexo* do aluno, *Publicação no Ano* do docente, *Forma Ingresso* no ensino superior, como também em nível da turma, *Média Nota Vestibular* e *Taxa de Cotista*, e do curso como *Local do Campus*. Logo, a análise do ganho de informação das covariadas adotadas neste estudo mostra que as características que compõem todas as dimensões (discentes, docentes, curso e turma) são importantes na performance de previsão dos modelos.

No que tange aos fatores relacionados à dimensão do aluno, principalmente os que evidenciam o acúmulo de capital humano mensurado pelas nota do vestibular e nota do vestibular em matemática, como também variáveis que sugerem o nível socioeconômico do discente, por exemplo, o aluno ser cotista, apresentaram efeitos mais importantes, sugerindo o que a literatura em economia da educação já debate nos seus estudos (BRUECKNER, 1999; GUIMARÃES; SAMPAIO *et al.*, 2007; CAVALCANTI; GUIMARAES; SAMPAIO, 2010; SAMPAIO *et al.*, 2011).

Os déficits (sucessos) oriundos da formação do ensino básico desse aluno, principalmente na disciplina de matemática, podem influenciar o seu desempenho e adaptação na disciplina de cálculo diferencial e integral I na UFPB. Considerando que esta disciplina é base e faz parte do período inicial dos cursos que a contém em sua grade, assim um desempenho ruim (bom) pode se tornar a principal motivação para a reprovação (aprovação), retenção e até mesmo a evasão dos discentes. Informações em nível de formação dos professores também se destacaram, principalmente no que se refere à qualificação de pós-graduação dos docentes, visto que está variável está presente nas posições iniciais de ambos os métodos.

4.5 Conclusões

A reprovação dos discentes é um sério problema enfrentado pelas instituições de ensino superior, principalmente nas disciplinas que são bases e fazem parte dos cursos nas mais diversas áreas de ensino, como é o caso da disciplina de cálculo

diferencial e integral I na UFPB. Esta compõem a grade curricular de vinte e um cursos de graduações da referida universidade, e varia desde centros que não exige um conhecimento de matemática tão aprofundado como o Centro de Ciências Sociais Aplicadas (CCSA), no curso de Ciências atuarias por exemplo, até os centros que demandam uma base mais consistente como os Centros de Tecnologia (TI), Centro de Informática (CI) e outros já discutidos ao longo desta pesquisa.

A identificação precoce dos alunos com perfil de reprovação pode permitir que os professores, coordenadores e outros atores institucionais planejem ações específicas a fim de evitar futuras reprovações em determinadas disciplinas-chave de cada curso. Evitando até um efeito em cadeia mais a longo prazo, como as reduções dos índices de evasão e, consequentemente, ampliação da taxa de diplomação. Tais fatos podem reduzir os custos da universidade na formação dos discentes, visto que o insucesso em uma disciplina estratégica na grade curricular (que é pré-requisito para várias outras) impacta o tempo de conclusão do curso e, portanto, aumenta o custo de oportunidade dos estudantes e, assim, estimula a evasão.

Nessa temática, esta pesquisa se propôs a aplicar algoritmos de *Machine Learning* como instrumentos para identificar o risco de reprovação dos estudantes que demandam cursar a disciplina de cálculo diferencial e integral I na UFPB no período entre 2010 e 2016, com base em preditores em nível dos discentes, docentes, turma e curso, que podem contribuir na performance dos modelos.

A partir das informações de 8.659 discentes matriculados na disciplina de cálculo diferencial e integral I (62,7% reprovados e 37,3% aprovados), oriundas dos registros administrativos e acadêmicos da UFPB e da Plataforma *Lattes*, foram compostas as duas base de dados necessárias para as etapas de aprendizado e predição do processo de ajuste dos modelos de ML: a base de treinamento e a base de teste. O critério adotado como ponto de corte foi a determinação do ano de 2015, logo a base de treinamento compõe os anos anteriores a 2015, e a base de teste por sua vez contém os anos de 2015 e 2016. Assim, a estratégia do aprendizado em máquina fundamenta-se em traçar um perfil com base em características dos estudantes, na base de treinamento, e a partir de então identificar e acompanhar, de maneira prévia, os estudantes com perfis semelhantes que apresentem o mesmo risco de reprovação em novos dados.

Como medidas de desempenho utilizadas como critérios para avaliar e selecionar os algoritmos com a melhor performance preditiva, para problema de classificação, utilizou-se: o total de previsões classificadas corretamente (*Accuracy*), as taxas de estudantes reprovados classificados (previstos) corretamente (*Sensitivity*) e as taxas de discentes aprovados classificados (previstos) corretamente (*Specificity*), ambas oriundas da *Confusion Matrix*. Como forma de ampliar ainda mais a gama de métricas de avaliação de performance, adotou-se também a área sob as curvas ROC, a AUC ROC.

A qual é adotada para organizar, visualizar e selecionar os algoritmos baseados no seu desempenho, expondo a relação entre o benefício (verdadeiros positivos) e o custo (falso positivos). Logo, os algoritmos que exibirem a melhor performance se sobrepõe as dos métodos menos eficientes.

As performances obtidas pelos 12 métodos, desde os mais tradicionais (regressão linear e logística) aos mais específicos de ML (SVM e Métodos baseados em árvores) foram semelhantes. Contudo, apenas dois deles apresentaram, em conjunto, os melhores desempenhos de previsão: *Penalized Methods LASSO* e Regressão Logística. Dos 1.532 indivíduos que compõem o conjunto de base de teste, a frequência dos alunos com status (reprovados e aprovados) previstos corretamente medida pela *Accuracy* foi de 67%, em ambos os modelos. Por sua vez, 72,5% dos discentes, referentes a 1.111 alunos, foram previstos corretamente como reprovados, medida pela *Sensitivity*, nos anos de 2015 e 2016 na disciplina de cálculo diferencial e integral I na UFPB.

No tocante à importância das variáveis no processo de ajuste dos modelos na etapa de aprendizado e predição, as características que mensuram o conhecimento agregado dos alunos no ensino médio como as notas do vestibular, e nota do vestibular em matemática, influenciam o status do discente no final da disciplina de cálculo. Assim, na busca por suavizar as reprovações ocasionadas principalmente no que se refere aos déficits nos conhecimentos básicos, de maneira específica, em matemática, os professores e coordenadores de cursos de graduações poderiam elaborar estratégias pedagógicas no processo de ensino/aprendizagem, tornando-se fundamentais na obtenção dos resultados positivos.

Tais fatos ilustram que os algoritmos de ML que retornaram previsões para o problema de classificação em estudo indicaram que esta é uma abordagem viável para a identificação dos grupos de riscos de reprovação a partir da aplicação de uma nova base de dados. Como ressaltado, este estudo faz um incremento na literatura nacional apresentando uma nova caixa de ferramentas para problemas de previsão. Tanto no que se refere na descrição dos algoritmos de ML como também na aplicação destes em um problema na área de economia da educação, pois os trabalhos existentes basicamente exploram mais sobre os determinantes da reprovação e evasão usando abordagens empíricas e enfoque mais tradicionais, isto é, não se detêm mais a fundo sobre os mecanismos de mensuração para antever de forma mais eficiente os riscos de reprovação no ensino superior.

Dessa maneira, houve uma preocupação por parte desta pesquisa em expor algumas práticas pedagógicas e estratégias de intervenções que possam vir a auxiliar as ações dos principais agentes transmissores de educação no ensino, são elas: monitorias, tutorias, supervisões, avaliações iniciais aplicadas com o foco de colher indicadores de aprendizagem e correção dos erros cometidos posteriormente; cursos de nivelamento

com conteúdo do ensino básico; incentivos aos professores como cursos de práticas e técnicas pedagógicas, com o objetivo de tornar a disciplina mais atrativa; melhor formação do docente; aulas extras; minicursos e outras.

Por fim, a partir dos resultados encontrados no desenvolvimento da presente pesquisa, e que não estão isentos de limitações, acredita-se que este é um instrumento viável para fornecer um maior suporte as ações dos gestores educacionais que visem a redução dos índices de reprovações em qualquer disciplina dos cursos de graduações de todas as instituições de ensino superior. De maneira que possa auxiliar professores e coordenadores não só para um debate inicial, mas também para identificar as potenciais reprovações com o objetivo de acompanhá-los de maneira mais direta, buscando meios de subsidiá-los nos rendimentos acadêmicos enquanto cursam a disciplina. Consequentemente, reduzir a retenção, a evasão e os custos, por um lado, e aumentar a taxa de diplomação, o estoque de mão-de-obra qualificada e a produtividade do ensino superior público no Brasil.

5 Considerações Finais

Fundamentada nas teorias de financiamento de campanhas políticas (economia da corrupção) e teorias em economia da educação, com foco no ensino superior, a presente tese, estruturada em três ensaios, buscou explorar e investigar novas aplicações de métodos e técnicas que buscam melhores inferências e desempenhos (para o caso dos modelos de previsão) em problemas já existentes em economia aplicada. Dessa maneira, a seguir estão apresentadas as principais evidências referentes a cada um dos capítulos que compõem esta pesquisa.

Diante das evidências encontradas no Capítulo 2, a partir dos dados provenientes do TCE-PB e do TSE para o período entre 2004 e 2016, os resultados encontrados, através do modelo de DD com efeito fixo, validam a hipótese adotada na presente pesquisa, a qual afirma que o financiamento em campanhas eleitorais resulta em benefícios oriundos nos processos de contratos públicos pelas empresas e prestadores de serviços que doaram recursos financeiros aos prefeitos e vereadores eleitos em 2004, 2008 e 2012. Os principais resultados apontam que as doações efetuadas em campanhas políticas por agentes privados geram um retorno, em média, de 42% nos valores contratados para os doadores de candidatos eleitos. Sendo essa taxa de retorno maior para as empresas do que para os prestadores de serviços.

No segundo Capítulo, a partir de diferentes instrumentos de pareamento não-experimentais, *Propensity Score Matching* (PSM), *Mahalanobis Distance Matching* (MDM) e *Classification Tree Analysis* (CTA), tendo como variável de resultado o nível de desempenho, medido pelo CRA relativo, o aluno ser cotista na UFPB arca com rendimentos médios significativamente menores em comparação aos rendimentos médios dos alunos ingressantes por ampla concorrência na UFPB. O impacto é maior quando a análise capta as melhores médias da distribuição do CRA relativo. No que se refere a contribuição do estudo, indicador taxa de abandono, com base no modelo de risco proporcional de Cox, os resultados indicam uma associação positiva entre o sistema de cotas e o acréscimo de sobrevida dos estudantes que ingressaram na UFPB através da política. O desempenho relativo auferido pelos cotistas serve como fundamentação para a manutenção da política de cotas no ensino superior, contudo, interligada a outras políticas direcionadas para melhorar a qualidade do ensino fundamental e médio das escolas públicas.

Por fim, o terceiro parte com o objetivo aplicar algoritmos de *Machine Learning* como instrumentos para identificar o risco de reprovação dos estudantes que demandam cursar a disciplina de cálculo diferencial e integral I na UFPB no período entre

2010 e 2016, os métodos que apresentaram os melhores desempenhos de previsão foram *Penalized Methods Lasso* e Regressão Logística. A partir da modelagem sobre os dados de treinamento (2010 a 2014), os resultados encontrados explicitam que, das 1.532 indivíduos que compõem um novo conjunto de dados (2015 e 2016), a frequência dos alunos com status (reprovados e aprovados) previstos corretamente foi de 67%, em ambos os modelos. Por sua vez, 72,5% dos discentes, referentes a 1.111 alunos, foram previstos corretamente como reprovados. Logo, o presente ensaio evidencia que a identificação precoce dos alunos com perfil de reprovação, por meio de uma nova caixa de ferramentas (ML), pode permitir que os professores, coordenadores e chefes de departamentos planejem ações especificas a fim de evitar futuras reprovações em determinadas disciplinas.

Referências

ALDRICH, J. H.; NELSON, F. D.; ADLER, E. S. *Linear probability, logit, and probit models*. [S.l.]: Sage, 1984.

ALON, S.; MALAMUD, O. The impact of israel's class-based affirmative action policy on admission and academic outcomes. *Economics of Education Review*, Elsevier, v. 40, p. 123–139, 2014.

ALTMAN, D. G.; BLAND, J. M. Diagnostic tests 3: receiver operating characteristic plots. *BMJ: British Medical Journal*, BMJ Publishing Group, v. 309, n. 6948, p. 188, 1994.

AMIT, Y.; GEMAN, D. Shape quantization and recognition with randomized trees. *Neural computation*, MIT Press, v. 9, n. 7, p. 1545–1588, 1997.

ANSOLABEHERE, S.; FIGUEIREDO, J. M. D.; Snyder Jr, J. M. Why is there so little money in us politics? *Journal of Economic perspectives*, v. 17, n. 1, p. 105–130, 2003.

ARCIDIACONO, P.; AUCEJO, E. M.; FANG, H.; SPENNER, K. I. Does affirmative action lead to mismatch? a new test and evidence. *Quantitative Economics*, Wiley Online Library, v. 2, n. 3, p. 303–333, 2011.

ARROW, K. J. Uncertainty and the welfare economics of medical care. In: *Uncertainty in Economics*. [S.l.]: Elsevier, 1978. p. 345–375.

ARVATE, P. R.; BARBOSA, K.; FUZITANI, E. Campaign donation and government contracts in brazilian states. *Working Paper 7*, Center for Applied Microeconomics, São Paulo School of Economics, 2013.

ATHEY, S. The impact of machine learning on economics. In: *The Economics of Artificial Intelligence: An Agenda*. [S.l.]: University of Chicago Press, 2018.

ATHEY, S.; IMBENS, G. W. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, v. 74, n. 2, p. 431–497, 2006.

_____. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, v. 31, n. 2, p. 3–32, 2017.

AYRES, I.; BROOKS, R. Does Affirmative Action Reduce the Number of Black Lawyers? *Stanford Law Review*, v. 57, n. 6, p. 1807–1854, 2005. ISSN 00389765.

BAGDE, S.; EPPLE, D.; TAYLOR, L. Does affirmative action work? caste, gender, college quality, and academic success in india. *American Economic Review*, v. 106, n. 6, p. 1495–1521, 2016.

BAJARI, P.; NEKIPELOV, D.; RYAN, S. P.; YANG, M. *Demand estimation with machine learning and model combination*. [S.l.], 2015.

_____. Machine learning methods for demand estimation. *American Economic Review*, v. 105, n. 5, p. 481–85, 2015.

BARANDIARAN, I. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, v. 20, n. 8, 1998.

BENEDETTI, J. K. On the nonparametric estimation of regression functions. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 39, n. 2, p. 248–253, 1977.

BERTRAND, M.; DUFLO, E.; MULLAINATHAN, S. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, v. 119, n. 1, p. 249–275, 2004.

BERTRAND, M.; HANNA, R.; MULLAINATHAN, S. Affirmative action in education: Evidence from engineering college admissions in India. *Journal of Public Economics*, Elsevier B.V., v. 94, n. 1-2, p. 16–29, 2010. ISSN 00472727. Disponível em: <http://dx.doi.org/10.1016/j.jpubeco.2009.11.003>.

BISHOP, C. M. *et al. Neural networks for pattern recognition*. [S.l.]: Oxford university press, 1995.

BISHOP, J. Drinking from the Fountain of Knowledge: Student Incentive to Study and Learn – Externalities, Information Problems and Peer Pressure. In: HANUSHEK, E.; Finis Welch (Ed.). *Handbook of The Economics of Education*. 1. ed. Amsterdam: North Holland, 2006. v. 2, cap. 15, p. 909–944.

BJÖRKEGREN, D.; GRISSEN, D. Behavior revealed in mobile phone usage predicts loan repayment. *SSRN 2611775*, 2018.

BOAS, T. C.; HIDALGO, F. D.; RICHARDSON, N. P. The spoils of victory: campaign donations and government contracts in brazil. *The Journal of Politics*, Cambridge University Press New York, USA, v. 76, n. 2, p. 415–429, 2014.

BRASIL. *Lei nº 8.713, de 13 de setembro de 1993. Estabelece normas para as eleições de 3 de outubro de 1994.* Brasília-DF: Casa Civil, 1993.

_____. *Lei nº 9.504, de 30 de setembro de 1997. Estabelece normas para as eleições.* Brasília-DF: Casa Civil, 1997.

_____. Lei nº 12.711, de 29 de agosto de 2012. dispõe sobre o ingresso nas universidades federais e nas instituições federais de ensino técnico de nível médio e dá outras providências. *Diário Oficial da União*, Imprensa Nacional Brasília, v. 149, n. 169, 2012.

_____. *Resolução nº 23.463, de 15 de dezembro de 2015. Dispõe sobre a arrecadação e os gastos de recursos por partidos políticos e candidatos e sobre a prestação de contas nas eleições de 2016.* Brasília-DF: Tribunal Superior Eleitoral, 2015.

BREIMAN, L. Randomizing outputs to increase prediction accuracy. *Machine Learning*, Springer, v. 40, n. 3, p. 229–242, 2000.

_____. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BROWN, C. D.; DAVIS, H. T. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, Elsevier, v. 80, n. 1, p. 24–38, 2006.

BRUECKNER, J. *Fiscal decentralization in LDCs: the effects of local corruption and tax evasion*. [S.l.]: Department of Economics, University of Illinois at Urbana-Champaign, 1999.

CALIENDO, M.; KOPEINIG, S. Some Practical Guidance for the Implementation of Propensity Score Matching. *IZA Working Paper n. 1588*, p. 1–32, 2005.

CAMERON, A. C.; TRIVEDI, P. K. *Microeconometrics: methods and applications*. [S.l.]: Cambridge university press, 2005.

CAVALCANTI, T.; GUIMARAES, J.; SAMPAIO, B. Barriers to skill acquisition in brazil: Public and private school students performance in a public university entrance exam. *The Quarterly Review of Economics and Finance*, Elsevier, v. 50, n. 4, p. 395–407, 2010.

CHAN, J.; EYSTER, E. Does Banning Affirmative Action Lower College Student Quality? *The American Economic Review*, v. 93, n. 3, p. 858–872, 2003.

CHENG, B.; TITTERINGTON, D. M. Neural networks: A review from a statistical perspective. *Statistical science*, JSTOR, p. 2–30, 1994.

CLAESSENS, S.; FEIJEN, E.; LAEVEN, L. Political connections and preferential access to finance: The role of campaign contributions. *Journal of financial economics*, Elsevier, v. 88, n. 3, p. 554–580, 2008.

CORTES, C.; VAPNIK, V. Support vector machine. *Machine learning*, v. 20, n. 3, p. 273–297, 1995.

CORTES, K. E. Do bans on affirmative action hurt minority students? evidence from the texas top 10% plan. *Economics of Education Review*, Elsevier, v. 29, n. 6, p. 1110–1124, 2010.

COVIELLO, D.; GAGLIARDUCCI, S. Tenure in office and public procurement. *American Economic Journal: Economic Policy*, v. 9, n. 3, p. 59–105, 2017.

COX, D. R.; OAKES, D. Analysis of survival data. 1984. *Chapman&Hall, London*, 1984.

CRAMER, J. An introduction for economists: The logit model. *London et al.: Edward Arnold*, 1991.

DEE, T. S. Are there civic returns to education? *Journal of public economics*, Elsevier, v. 88, n. 9-10, p. 1697–1720, 2004.

DICKSON, L. M. Does ending affirmative action in college admissions lower the percent of minority students applying to college? *Economics of Education Review*, v. 25, n. 1, p. 109–119, 2006. ISSN 02727757.

DIETTERICH, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, Springer, v. 40, n. 2, p. 139–157, 2000.

DIOGO, M. F.; RAYMUNDO, L. dos S.; WILHELM, F. A.; ANDRADE, S. P. C. de; LORENZO, F. M.; ROST, F. T.; BARDAGI, M. P. Percepções de coordenadores de curso superior sobre evasão, reprovações e estratégias preventivas. *Avaliação: Revista da Avaliação da Educação Superior*, SciELO Brasil, v. 21, n. 1, 2015.

DRUCKER, H.; BURGES, C. J.; KAUFMAN, L.; SMOLA, A. J.; VAPNIK, V. Support vector regression machines. In: *Advances in neural information processing systems*. [S.l.: s.n.], 1997. p. 155–161.

ESTEVAN, F.; GALL, T.; MORIN, L.-P. Redistribution without distortion : Evidence from an affirmative action program at a large Brazilian university . 2016.

FACCIO, M. Politically connected firms. *American economic review*, v. 96, n. 1, p. 369–386, 2006.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FERMAN, B.; ASSUNÇÃO, J. Affirmative action in university admissions and high school students' proficiency. *XXVii Encontro Brasileiro de Econometria. Anais*, 2005.

FISMAN, R. Estimating the value of political connections. *American economic review*, v. 91, n. 4, p. 1095–1102, 2001.

FLACH, P. A.; WU, S. Repairing concavities in roc curves. In: CITESEER. *IJCAI*. [S.l.], 2005. p. 702–707.

FONSECA, T. do N. Doações de campanha implicam em retornos contratuais futuros? uma análise dos valores recebidos por empresas antes e após as eleições. *Revista de Sociologia e Política*, Universidade Federal do Paraná, v. 25, n. 61, p. 31–49, 2017.

FOX, J.; WEISBERG, S. Multivariate linear models in r. *An R Companion to Applied Regression. Los Angeles: Thousand Oaks*, 2011.

FRANCIS, A. M.; TANNURI-PIANTO, M. The redistributive equity of affirmative action: Exploring the role of race, socioeconomic status, and gender in college admissions. *Economics of Education Review*, Elsevier, v. 31, n. 1, p. 45–55, 2012.

FREUND, Y. Boosting a weak learning algorithm by majority. *Information and computation*, Elsevier, v. 121, n. 2, p. 256–285, 1995.

FREUND, Y.; SCHAPIRE, R. E. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, Elsevier, v. 29, n. 1-2, p. 79–103, 1999.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *et al.* Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, Institute of Mathematical Statistics, v. 28, n. 2, p. 337–407, 2000.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. *Machine learning*, Springer, v. 29, n. 2-3, p. 131–163, 1997.

GAL, Y. *Uncertainty in deep learning*. Tese (Doutorado) — PhD thesis, University of Cambridge, 2016.

GALIANI, S.; GERTLER, P.; SCHARGRODSKY, E. Water for life: The impact of the privatization of water services on child mortality. *Journal of political economy*, v. 113, n. 1, p. 83–120, 2005.

GERTLER, P. J.; MARTINEZ, S.; PREMAND, P. Impact evaluation in practice. 2011.

GOEL, S.; RAO, J. M.; SHROFF, R. *et al.* Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. *The Annals of Applied Statistics*, Institute of Mathematical Statistics, v. 10, n. 1, p. 365–394, 2016.

GOLDMAN, E.; ROCHOLL, J.; SO, J. Political connections and the allocation of procurement contracts. *Unpublished paper*, 2008.

GOLDSTEIN, B. A.; NAVAR, A. M.; CARTER, R. E. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal*, Oxford University Press, v. 38, n. 23, p. 1805–1814, 2016.

GOMES-NETO, J. B.; HANUSHEK, E. A. Causes and consequences of grade repetition: Evidence from brazil. *Economic Development and Cultural Change*, University of Chicago Press, v. 43, n. 1, p. 117–148, 1994.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: Cambridge: MIT Press, 2016.

GRAU, N. The impact of college admissions policies on the academic effort of high school students. *Economics of Education Review*, Elsevier, v. 65, p. 58–92, 2018.

GU, X. S.; ROSENBAUM, P. R. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, Taylor & Francis Group, v. 2, n. 4, p. 405–420, 1993.

GUIMARÃES, J.; SAMPAIO, B. *et al.* The influence of family background and individual characteristics on entrance tests scores of brazilian university students. *Anais do XXXV Encontro Nacional de Economia-ANPEC-Associação Nacional dos Centros de Pós-Graduação em Economia, Recife*, 2007.

GUJARATI, D. N.; PORTER, D. C. *Econometria Básica-5*. [S.l.]: Amgh Editora, 2011.

HAN, J.; KAMBER, M.; MINING, D. Data mining: Concepts and techniques. *Morgan Kaufmann*, v. 340, p. 94104–3205, 2001.

HANSEN, B. B. Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, Taylor & Francis, v. 99, n. 467, p. 609–618, 2004.

HANUSHEK, E. A. The economics of schooling: Production and efficiency in public schools. *Journal of economic literature*, JSTOR, v. 24, n. 3, p. 1141–1177, 1986.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference, and prediction, Springer Series in Statistics*. [S.l.]: Springer New York, 2009.

HICKMAN, B. R. *Effort , Race Gaps and Affirmative Action: a structural policy analysis of US college admissions*. Iowa City: University of Chicago, 2009. 50 p.

HINRICHS, P. Affirmative action bans and college graduation rates. *Economics of Education Review*, Elsevier, v. 42, p. 43–52, 2014.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Sinopse Estatística da Educação Superior 2017*. Brasília-DF: Inep, 2018. Disponível em: <http://portal.inep.gov.br/educacao-superior>. Acesso em: 30.07.2019.

IZBICKI, R.; SANTOS, T. M. Machine learning sob a ótica estatística. 2018.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112.

JUNG, J. H.; SUNG, H.-Y.; KIM, H.-S. Affirmative Action in Korea: Its Impact on Women's Employment, Corporate Performance and Economic Growth. In: *2012 Annual Economic Association*. Chicago: [s.n.], 2012. Disponível em: <http://www.aeaweb.org/aea/2012conference/program/retrieve.php?pdfid=188>.

JÚNIOR, F. T.; FARIA, V. B.; LIMA, M. A. de. Indicadores de fluxo escolar e políticas educacionais: avaliação das últimas décadas. *Estudos em Avaliação Educacional*, v. 23, n. 52, p. 48–67, 2012.

KALBFLEISCH, J. D.; ROSS, L. *The statistical analysis of failure time data*. [S.l.]: John Wiley and Sons, 1980.

KANE, T. J. Misconceptions in the debate over affirmative action in college admissions. *Chilling admissions: The affirmative action crisis and the search for alternatives*, Harvard Education Publishing Group Cambridge, MA, p. 17–31, 1998.

KEARNS, M.; VALIANT, L. Cryptographic limitations on learning boolean formulae and finite automata. *Journal of the ACM (JACM)*, ACM, v. 41, n. 1, p. 67–95, 1994.

KING, G.; NIELSEN, R. *Why propensity score should not be used for matching*. Cambridge, MA: Havard, 2016. Disponível em: <http://gking.harvard.edu/publications/why-Propensity-Scores-Should-Not-Be-Used-Formatching>.

KING, G.; NIELSEN, R.; COBERLEY, C.; POPE, J. E. *Comparative Effectiveness of Matching Methods for Causal Inference*. Cambridge, MA: Harvard, 2011. 1–26 p. Disponível em: <http://gking.harvard.edu/publications/comparative-effectiveness-matching-methods-causal-inference>.

KLEINBERG, J.; MULLAINATHAN, S.; RAGHAVAN, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

KRIEGER, G.; RODRIGUES, F.; BONASSA, E. C. *Os donos do congresso: a farsa na CPI do Orçamento*. [S.l.]: Editora Ática, 1994.

KUHN, M.; JOHNSON, K. *Applied predictive modeling*. [S.l.]: Springer, 2013. v. 26.

LANCASTER, T. *The econometric analysis of transition data*. [S.l.]: Cambridge university press, 1992.

LANTZ, B. *Machine learning with R*. [S.l.]: Packt Publishing Ltd, 2013.

LAZZARINI, S. G.; MUSACCHIO, A.; MELLO, R. Bandeira-de; MARCON, R. What do development banks do? evidence from brazil, 2002-2009. *Working Paper 12-047*, Harvard Business School, University of Harvard, 2014.

LEON, F. L. L. d.; MENEZES-FILHO, N. A. Reprovação, avanço e evasão escolar no brasil. Instituto de Pesquisa Econômica Aplicada (Ipea), 2002.

LI, D.; WEISMAN, D. L. Why preferences in college admissions may yield a more-able student body. *Economics of Education Review*, Elsevier, v. 30, n. 4, p. 724–728, 2011.

LIAO, T. F. *Interpreting probability models: Logit, probit, and other generalized linear models*. [S.l.]: Sage, 1994.

LINDEN, A.; YARNOLD, P. R. Identifying causal mechanisms in health care interventions using classification tree analysis. *Journal of evaluation in clinical practice*, Wiley Online Library, v. 24, n. 2, p. 353–361, 2018.

MADDALA, G. S. A perspective on the use of limited-dependent and qualitative variables models in accounting research. *The Accounting Review*, JSTOR, v. 66, n. 4, p. 788–807, 1991.

MCFADDEN, D. *et al.* Conditional logit analysis of qualitative choice behavior. *In P. Zarembka (ed.), Frontiers in Econometrics, Academic Press*, 1974.

Mendes Junior, A. A. F.; SOUZA, A. d. M. e.; WALTENBERG, F. D. Affirmative Action and Access to Higher Education in Brazil: The Significance of Race and Other Social Factors. *Journal of Latin American Studies*, v. 48, n. 2, p. 301–334, 2016. ISSN 1469767X.

MENEZES-FILHO, N. A.; PICCHETTI, P. Os determinantes da duração do desemprego em são paulo. Instituto de Pesquisa Econômica Aplicada (Ipea), 2000.

MEURER, W. J.; TOLLES, J. Logistic regression diagnostics: understanding how well a model predicts outcomes. *Jama*, American Medical Association, v. 317, n. 10, p. 1068–1069, 2017.

MINCER, J. Schooling, experience, and earnings. human behavior & social institutions no. 2. ERIC, 1974.

MIRONOV, M.; ZHURAVSKAYA, E. Corruption in procurement and shadow campaign financing: Evidence from russia. In: *ISNIE Annual Conference*. [S.l.: s.n.], 2012.

MITCHELL, T. M. Does machine learning really work? *AI magazine*, v. 18, n. 3, p. 11–11, 1997.

MOEHLECKE, S. Ação afirmativa: história e debates no brasil. *Cadernos de pesquisa*, SciELO Brasil, n. 117, p. 197–217, 2002.

MULLAINATHAN, S.; SPIESS, J. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, v. 31, n. 2, p. 87–106, 2017.

Nash Jr, J. F. The bargaining problem. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 155–162, 1950.

PEREIRA, J. I. R.; BITTENCOURT, M. V. L.; Silva Junior, W. S. Análise Do Impacto Da Implantação Das Cotas Na Nota Enade 2008. In: *41º Encontro Nacional de Economia (ANPEC)*. Foz do Iguaçu: [s.n.], 2013.

PRATI, R.; BATISTA, G.; MONARD, M. Curvas roc para avaliação de classificadores. *Revista IEEE América Latina*, v. 6, n. 2, p. 215–222, 2008.

PRATI, R. C.; FLACH, P. A. Roccer: An algorithm for rule learning based on roc analysis. In: *IJCAI*. [S.l.: s.n.], 2005. p. 823–828.

RASCHKA, S. *Python machine learning*. [S.l.]: 2 ed. Birmingham: Packt Publishing Ltd, 2017.

RIPLEY, B. D. *Pattern recognition and neural networks*. [S.l.]: Cambridge university press, 1996.

ROSENBAUM, P. R. *Design of Observational Studies*. New York: Springer, 2010. 1–382 p. ISBN 9781441912121.

ROSENBAUM, P. R.; RUBIN, D. B. The central role of the propensity score in observational studies foir causal effects. *Biometrika*, v. 70, n. 1, p. 41–55, 1983.

RUBIN, D. B. Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association*, v. 74, n. 366, p. 318–328, 1979.

_____. Bias Reduction Using Mahalanobis-Metric Matching. *Biometrics,*, v. 36, n. 2, p. 293–298, 1980.

RUMBA, J.; JASČIŠENS, V. Public procurement and political connections: The case of latvia. *Stockholm School of Economics in Riga*, 2009.

SAMPAIO, B.; SAMPAIO, Y.; MELLO, E. P. G. de; MELO, A. S. Desempenho no vestibular, background familiar e evasão: evidências da UFPE. *Economia Aplicada*, v. 15, n. 2, p. 287–309, 2011.

SANDER, R. H. A systemic analysis of affirmative action in american law schools. *Stanford Law Review*, HeinOnline, v. 57, p. 367, 2004.

SANTOS, H. G. d. *Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina*. Tese (Doutorado) — Universidade de São Paulo, 2018.

SCHAPIRE, R. E. The strength of weak learnability. *Machine learning*, Springer, v. 5, n. 2, p. 197–227, 1990.

SCHULTZ, T. W. Investment in human capital. *The American economic review*, JSTOR, p. 1–17, 1961.

SHIRASU, M. R.; ALBUQUERQUE, R. de *et al.* Determinantes da evasão e repetência escolar no ensino médio do ceará. *Revista Econômica do Nordeste*, v. 46, n. 4, p. 117–136, 2015.

SILVA-FILHO, R. L. L.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. C. M. A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, SciELO Brasil, v. 37, n. 132, p. 641–659, 2007.

SMOLA, A. J. *et al. Regression estimation with support vector learning machines*. Tese (Doutorado) — Master's thesis, Technische Universität München, 1996.

Snyder Jr, J. M. Campaign contributions as investments: The us house of representatives, 1980-1986. *Journal of Political Economy*, The University of Chicago Press, v. 98, n. 6, p. 1195–1227, 1990.

SOLOW, R. M. A contribution to the theory of economic growth. *The quarterly journal of economics*, MIT Press, v. 70, n. 1, p. 65–94, 1956.

SOUZA, A. P. d.; PONCZEK, V. P.; OLIVA, B. T.; TAVARES, P. A. Fatores associados ao fluxo escolar no ingresso e ao longo do ensino médio no brasil. *Pesquisa e Planejamento Econômico*, v. 42, n. 1, p. 5–39, 2012.

SOWELL, T. *Affirmative action around the world: An empirical study*. [S.l.]: Yale University Press, 2004.

STONE, C. J. Consistent nonparametric regression. *The annals of statistics*, JSTOR, p. 595–620, 1977.

STUART, E. A. Matching methods for causal inference: A review and a look forward. *Statistical Science*, v. 25, n. 1, p. 1–21, 2010. ISSN 0883-4237. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2943670{&}tool=pmcentrez{&}rendertype=ab>.

STUART, E. a.; RUBIN, D. B. Best practices in quasi-experimental designs: Matching methods for causal inference. In: OSBORNE, J. (Ed.). *Best practices in quantitative methods*. Thousand Oaks, CA: Sage Publications, 2007. cap. 11, p. 155–176. ISBN 9781412940658.

SU, X. Education Hierarchy , Within-Group Competition and Affirmative Action. 2005. Disponível em: <http://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=781>.

Tribunal Superior Eleitoral. *Prestação de Contas Eleitorais 2010*. 2012. Disponível em: <http://http://www.tse.jus.br/eleitor-e-eleicoes/eleicoes/eleicoes-anteriores/eleicoes-2010/prestacao-de-contas/prestacao-de-contas-eleitorais-2010>. Acesso em: 30.04.2018.

VALIANT, L. G. A theory of the learnable. In: ACM. *Proceedings of the sixteenth annual ACM symposium on Theory of computing*. [S.l.], 1984. p. 436–445.

VARIAN, H. R. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, v. 28, n. 2, p. 3–28, 2014.

WINSTON, G.; ZIMMERMAN, D. Peer effects in higher education. In: *College choices: The economics of where to go, when to go, and how to pay for it*. [S.l.]: University of Chicago Press, 2004. p. 395–424.

WOOLDRIDGE, J. M. *Introdução à econometria: uma abordagem moderna*. [S.l.]: Pioneira Thomson Learning, 2006.

_____. *Econometric analysis of cross section and panel data*. [S.l.]: Cambridge, Massachusetts: The MIT press, 2010.

ZHOU, M. Understanding the cox regression models with time-change covariates. *The American Statistician*, Taylor & Francis, v. 55, n. 2, p. 153–155, 2001.

ZYLBERSTAJN, E.; SOUZA, A. P. F. de. Cotas nas universidades e aprendizado escolar : modelo teórico e evidências empíricas. In: *32º Encontro Brasileiro de Econometria*. [S.l.: s.n.], 2010.

Parte I

Apêndice

A Segundo Ensaio

Tabela A.1 – Resultados do modelo de regressão de Cox - *Hazard ratio*.

	Modelo Básico	**Dependência Temporal**
	(1)	**(2)**
Tratamento	0,927***	0,823***
	(0,867,0,987)	(0,762,0,885)
Covariadas		
CRA		0,690***
		(0,677, 0,703)
Sexo		0,855***
		(0,793, 0,917)
Raça		0,958
		(0,895, 1,021)
Média Vestibular		1,001***
		(1,001, 1,002)
Trancado		1,812***
		(1,711, 1,912)
Covariadas		
Fixas	Sim	Sim
Variáveis	Não	Sim
N	17.550	17.436
$\mathbf{R}^2$	0.0003	0.170

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.
Nota: Níveis de significância: *10%, **5% e ***1%.

Tabela A.3 – Resultados do modelo de regressão de Cox.

	Modelo Básico	**Dependência Temporal**
	(1)	**(2)**
Tratamento	-0,076**	-0,194***
	(0,031)	(0.031)
Covariadas		
CRA		-0,371***
		(0,007)
Sexo		-0,157***
		(0,032)
Raça		-0,043
		(0,032)
Média Vestibular		0,001***
		(0,0003)
Trancado		0,594***
		(0,051)
Covariadas		
Fixas	Sim	Sim
Variáveis	Não	Sim
N	17.550	17.436
$\mathbf{R}^2$	0,0003	0,170

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.
Nota: Níveis de significância: *10%, **5% e ***1%.

Tabela A.5 – Evidências iniciais dos indicadores de resultado dos alunos ingressantes por cota por grande área de conhecimento e cursos na UFPB, 2011 - Testes de médias e intervalo de confiança.

	Cota	Ingressantes	Média Idade	Média Vestibular	Taxa de Abandono			CRA Relativo
					Total	1º Ano	2º Ano	
Grande Área de Conhecimento								
C. Sociais Aplicadas	0	1574	22,6	512,3	0,52	0,21	0,34	44,8
			22,3 - 22,9		0,49 - 0,54	0,19 - 0,23	0,31 - 0,36	42,1 - 47,6
			0,000		0,000	0,000	0,000	0,000
	1	409	23,3	470,7	0,53	0,19	0,30	35,7
			22,7 – 23,9		0,48 – 0,57	0,15 – 0,22	0,26 – 0,35	31,4 – 40,0
			0,000		0,000	0,000	0,000	0,000
Ling., Letras e Artes	0	1023	23,9	527,3	0,47	0,22	0,32	31,2
			23,5 - 24,4		0,44 - 0,50	0,20 - 0,25	0,29 - 0,35	29,2 - 33,2
			0,000		0,000	0,000	0,000	0,000
	1	249	23,3	490,7	0,44	0,16	0,30	27,6
			22,5 - 24,1		0,38 - 0,50	0,11 - 0,20	0,24 - 0,36	24,4 - 30,8
			0,000		0,000	0,000	0,000	0,000
Engenharias	0	831	20,6	574,2	0,56	0,18	0,30	47,9
			20,2 - 20,9		0,52 - 0,59	0,16 - 0,21	0,26 - 0,33	43,6 - 52,2
			0,000		0,000	0,000	0,000	0,000
	1	223	21,0	525,5	0,54	0,15	0,26	43,9
			20,6 - 21,5		0,48 - 0,61	0,10 - 0,20	0,21 - 0,32	32,2 - 55,6
			0,000		0,000	0,000	0,000	0,000
C. Humanas	0	978	24,4	530,8	0,41	0,18	0,28	24,5
			23,9 – 24,9		0,38 – 0,44	0,15 – 0,20	0,25 – 0,30	23,2 – 25,9
			0,000		0,000	0,000	0,000	0,000
	1	217	25,7	497,1	0,35	0,08	0,20	20,4
			24,6 - 26,8		0,29 - 0,41	0,04 - 0,12	0,15 - 0,26	17,95 - 22,8
			0,000		0,000	0,000	0,000	0,000

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.

Tabela A.6 – Evidências iniciais dos indicadores de resultado dos alunos ingressantes por cota por grande área de conhecimento e cursos na UFPB, 2011 - Testes de médias e intervalo de confiança. (Continuação)

	Cota	Ingressantes	Média Idade	Média Vestibular	Taxa de Abandono			CRA Relativo
					Total	1º Ano	2º Ano	
Grande Área de Conhecimento								
C. da Saúde	0	821	21,5	583,1	0,33	0,17	0,25	16,5
			21,2 - 21,9		0,30 - 0,36	0,15 - 0,20	0,22 - 0,28	15,6 - 17,4
			0,000		0,000	0,000	0,000	0,000
	1	202	22,1	536,9	0,21	0,10	0,14	12,9
			21,4 - 22,9		0,15 - 0,26	0,06 - 0,15	0,09 - 0,19	11,1 - 14,6
			0,000		0,000	0,000	0,000	0,000
C. Exatas e da Terra	0	796	25,7	525,7	0,69	0,27	0,46	87,7
			25,1 - 26,2		0,66 - 0,73	0,24 - 0,30	0,43 - 0,50	79,0 - 96,3
			0,000		0,000	0,000	0,000	0,000
	1	126	22,4	487,3	0,69	0,31	0,50	72,2
			21,2 - 23,6		0,61 - 0,77	0,23 - 0,39	0,41 - 0,59	48,9 - 109,48
			0,000		0,000	0,000	0,000	0,000
C. Agrárias	0	681	24,8	493,4	0,55	0,26	0,40	48,9
			24,2 - 25,4		0,51 - 0,59	0,22 - 0,29	0,36 - 0,44	45,5 - 52,3
			0,000		0,000	0,000	0,000	0,000
	1	112	21,5	465,3	0,41	0,19	0,31	38,4
			20,4 - 22,6		0,31 - 0,50	0,12 - 0,27	0,22 - 0,39	32,3 - 44,6
			0,000		0,000	0,000	0,000	0,000
C. Biológicas	0	121	26,6	-	0,81	0,47	0,59	50,8
			25,2 - 27,9		0,74 - 0,88	0,38 - 0,56	0,50 - 0,68	40,9 - 60,79
			0,000		0,000	0,000	0,000	0,000
	1	-	-	-	-	-	-	-
	1	26	22,6	472,2	0,23	0,11	0,23	22,8
			19,9 - 25,4		0,05 - 0,40	- 0,01 - 0,24	0,05 - 0,40	16,2 - 29,4
			0,000		0,000	0,000	0,000	0,000
Outras	0	121	23,5	497,4	0,44	0,25	0,37	23,7
			22,5 - 24,5		0,35 - 0,53	0,17 - 0,33	0,28 - 0,45	20,4 - 27,0
			0,000		0,000	0,000	0,000	0,000

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.

Tabela A.7 – Evidências iniciais dos indicadores de resultado dos alunos ingressantes por cota por grande área de conhecimento e cursos na UFPB, 2011 - Testes de médias e intervalo de confiança. (Continuação)

	Cota	Ingressantes	Média Idade	Média Vestibular	Taxa de Abandono			CRA Relativo
					Total	1º Ano	2º Ano	
Cursos com maiores entradas de cotistas								
Direito	0	328	21,4	624,9	0,31	0,09	0,16	16,3
			20,8 - 22,0	620,0 - 629,7	0,26 - 0,36	0,06 - 0,12	0,12 - 0,20	
			0,000	0,000	0,000	0,000	0,000	
	1	97	26,3	551,0	0,26	0,09	0,19	13,6
			24,5 - 28,1	544,4 - 557,6	0,17 - 0,35	0,03 - 0,15	0,11 - 0,27	
			0,000	0,000	0,000	0,000	0,000	
Administração	0	203	21,5	547,6	0,37	0,16	0,26	37,2
			20,8 – 22,2	542,4 - 552,9	0,30 – 0,44	0,11 – 0,21	0,20 – 0,32	
			0,000	0,000	0,000	0,000	0,000	
	1	58	22,5	504,8	0,29	0,06	0,12	31,7
			21,1 – 23,9	492,7 – 516,9	0,17 – 0,41	0,00 – 0,13	0,03 – 0,20	
			0,000	0,000	0,000	0,000	0,000	
Pedagogia	0	315	25,4	475,3	0,43	0,21	0,31	27,2
			24,5 - 26,4	470,7 - 479,9	0,37 - 0,48	0,17 - 0,26	0,26 - 0,36	
			0,000	0,000	0,000	0,000	0,000	
	1	51	25,1	450,9	0,41	0,01	0,19	26,8
			23,3 - 26,9	433,8 - 467,9	0,27 - 0,55	- 0,01 - 0,05	0,08 - 0,30	
			0,000	0,000	0,000	0,000	0,000	

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.

Tabela A.8 – Evidências iniciais dos indicadores de resultado dos alunos ingressantes por cota por grande área de conhecimento e cursos na UFPB, 2011 - Testes de médias e intervalo de confiança. (Continuação)

	Cota	Ingressantes	Média Idade	Média Vestibular	Taxa de Abandono			CRA Relativo
					Total	1º Ano	2º Ano	
Cursos com menores entradas de cotistas								
Química Industrial	0	33	20,2	550,9	0,60	0,18	0,36	79,5
			18,7 - 21,6	531,0 - 570,8	0,43 - 0,78	0,04 - 0,32	0,19 - 0,53	
			0,000	0,000	0,000	0,000	0,000	
	1	11	20,6	496,7	0,63	0,18	0,63	63,4
			17,9 - 23,3	457,5 - 535,8	0,29 - 0,97	- 0,08 - 0,45	0,29 - 0,97	
			0,000	0,000	0,000	0,000	0,000	
Antropologia	0	45	25,1	472,3	0,66	0,26	0,53	82,9
			23,2 - 27,14	458,8 - 485,7	0,52 - 0,80	0,13 - 0,40	0,38 - 0,68	
			0,000	0,000	0,000	0,000	0,000	
	1	10	25,6	454,2	0,90	0,60	0,70	-
			20,2 - 30,9	392,1 - 516,2	0,67 - 1,12	0,23 - 0,96	0,35 - 1,04	
			0,000	0,000	0,000	0,000	0,000	
Sist. de Informações	0	48	21,4	501,9	0,62	0,22	0,39	53,7
			20,3 - 22,5	493,3 - 510,5	0,48 - 0,76	0,10 - 0,35	0,25 - 0,53	
			0,000	0,000	0,000	0,000	0,000	
	1	9	23,7	447,3	0,66	0,22	0,44	44,8
			19,5 - 28,0	423,6 - 471,0	0,28 - 1,05	- 0,11 - 0,56	0,03 - 0,84	
			0,000	0,000	0,000	0,000	0,000	

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.

B Terceiro Ensaio

Tabela B.1 – Evidências iniciais da variável de resultado, por status de matricula e por Grande Área de Conhecimento, dos alunos da UFPB, 2010 a 2016 - Testes de médias e intervalo de confiança*.

Ano	Área	Reprovado	Alunos	Média Idade	Média Vestibular	Média Vest. Matemática	Média final Disciplina
2010	**I**	Não	118	23,7 22,3 - 25,2	534,1 522,4 - 545,8	564,6 547,2 - 582,0	7,18 6,91 - 7,45
		Sim	330	24,0 23,4 - 24,7	497,7 492,1 - 503,3	512,8 504,2 - 521,4	0,52 0,41 - 0,64
	II	Não	44	22,4 20,9 - 23,8	475,3 465,3 - 485,3	496,1 480,7 - 511,6	7,72 7,32 - 8,12
		Sim	77	23,5 22,2 - 24,9	462,5 454,7 - 470,3	474,6 464,1 - 485,2	1,21 0,85 - 1,56
	III	Não	384	19,2 19,0 - 19,4	581,0 575,8 - 586,2	606,6 597,8 - 615,4	7,2 7,11 - 7,40
		Sim	522	20,1 19,7 - 20,4	562,9 558,4 - 567,5	580,3 572,5 - 588,1	0,92 0,81 - 1,04
2011	**I**	Não	120	22,7 21,6 - 23,8	541,2 529,9 - 552,6	572,2 555,2 - 589,2	7,31 7,04 - 7,59
		Sim	366	23,1 22,5 - 23,8	520,2 513,6 - 526,7	534,5 525,1 - 543,8	0,71 0,59 - 0,82
	II	Não	35	26,0 22,9 - 29,2	504,7 486,2 - 523,1	511,8 487,1 - 536,5	6,7 6,26 - 7,21
		Sim	182	23,7 22,7 - 24,7	486,0 476,2 - 495,8	500,4 488,4 - 512,40	0,78 0,60 - 0,96
	III	Não	424	19,4 19,1 - 19,6	595,3 589,6 - 600,9	620,4 611,9 - 628,9	7,21 7,08 - 7,35
		Sim	569	20,4 20,1 - 20,6	558,4 552,9 - 563,8	564,9 557,3 - 572,4	0,99 0,88 - 1,10
2012	**I**	Não	59	21,1 20,2 - 22,0	554,8 535,1 - 574,5	613,6 583,6 - 643,6	7,20 6,76 - 7,65
		Sim	350	24,2 23,5 - 24,9	512,3 505,2 - 519,4	531,5 521,4 - 541,5	0,47 0,37 - 0,57
	II	Não	71	24,64 23,0 - 26,3	511,5 495,6 - 527,5	547,7 520,7 - 574,6	6,99 6,59 - 7,39
		Sim	175	23,2 22,4 - 24,0	482,8 474,7 - 490,9	486,9 475,8 - 498,1	1,01 0,81 - 1,21
	III	Não	520	19,1 18,9 - 19,4	592,9 587,9 - 597,9	626,9 618,7 - 635,0	7,48 7,36 - 7,60
		Sim	679	20,9 20,6 - 21,3	561,9 557,4 - 566,5	575,3 568,2 - 582,4	0,99 0,89 - 1,09

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.
Notas: (1) Intervalos com 95% de confiança para as médias calculadas. (2) *Todos os intervalos de confiança apresentam um teste **t** de $0,000$.
Legenda: Área I = C. Exatas e da Terra, Área II = C. S. Aplicadas, Área III = Engenharias.

Tabela B.2 – Evidências iniciais da variável de resultado, por status de matricula e por Grande Área de Conhecimento, dos alunos da UFPB, 2010 a 2016 - Testes de médias e intervalo de confiança*. (Continuação)

Ano	Área	Reprovado	Alunos	Média Idade	Média Vestibular	Média Vest. Matemática	Média final Disciplina
2013	**I**	Não	122	21,5 20,5 - 22,4	554,6 541,8 - 567,5	613,6 591,7 - 635,5	7,56 7,28 - 7,84
		Sim	277	24,1 23,1 - 25,0	521,2 512,8 - 529,7	545,1 532,0 - 558,1	0,69 0,56 - 0,81
	II	Não	54	24,3 22,2 - 26,4	515,7 494,9 - 536,4	542,8 512,1 - 573,4	7,07 6,63 - 7,51
		Sim	156	22,3 21,5 - 23,1	498,9 486,4 - 511,5	514,9 497,7 - 532,1	0,69 0,51 - 0,87
	III	Não	494	19,1 18,8 - 19,3	609,9 603,6 - 616,2	650,9 641,2 - 660,5	7,39 7,27 - 7,51
		Sim	666	21,6 21,2 - 22,0	567,9 561,7 - 574,2	591,2 582,3 - 600,2	0,88 0,78 - 0,98
2014	**I**	Não	99	22,3 20,8 - 23,7	605,4 594,0 - 616,8	650,7 630,7 - 670,8	7,36 7,04 - 7,67
		Sim	265	23,9 22,9 - 24,8	578,1 571,9 - 584,2	619,2 608,4 - 630,0	0,28 0,19 - 0,38
	II	Não	75	21,9 20,8 - 23,0	537,9 524,2 - 551,7	569,1 549,2 - 589,0	7,12 6,78 - 7,45
		Sim	142	24,4 23,1 - 25,7	548,0 548,0 - 557,9	562,6 548,0 - 577,2	0,79 0,58 - 1,00
	III	Não	458	19,2 18,9 - 19,6	649,5 644,6 - 654,3	697,9 690,5 - 705,3	7,37 7,24 - 7,50
		Sim	592	21,8 21,3 - 22,3	624,5 619,9 - 629,1	651,5 644,4 - 658,6	0,66 0,56 - 0,76
2015	**I**	Não	87	22,7 21,1 - 24,2	603,2 609,4 - 616,5	632,8 589,8 - 656,2	7,64 7,24 - 8,03
		Sim	203	25,5 24,1 - 26,9	570,4 563,6 - 577,2	579,5 566,4 - 592,6	0,61 0,47 - 0,75
	II	Não	47	23,5 21,5 - 25,5	562,7 547,5 - 577,9	558,4 530,9 - 585,90	7,54 7,16 - 7,92
		Sim	120	26,3 24,5 - 28,1	557,2 547,3 - 566,9	546,0 528,9 - 563,1	0,80 0,57 - 1,03
	III	Não	423	19,3 19,0 - 19,5	651,7 647,0 - 956,4	685,8 677,0 - 694,6	7,51 7,38 - 7,64
		Sim	517	22,1 21,6 - 22,6	619,3 614,4 - 624,2	630,0 621,5 - 638,6	1,36 1,23 - 1,49
2016	**I**	Não	86	23,1 21,6 - 24,8	602,4 581,2 - 623,6	631,6 598,5 - 664,8	7,64 7,26 - 8,01
		Sim	183	23,6 22,5 - 24,7	545,6 522,7 - 568,5	579,5 543,9 - 615,0	1,09 0,89 - 1,29
	II	Não	23	26,1 21,1 - 31,1	540,2 490,3 - 590,2	606,2 483,6 - 728,8	7,07 6,44 - 7,70
		Sim	94	24,4 23,0 - 22,5	503,7 474,7 - 607,0	505,4 468,1 - 648,6	1,32 1,02 - 1,45
	III	Não	225	19,0 18,7 - 19,2	610,9 576,4 - 645,40	643,5 592,9 - 694,0	7,66 7,47 - 7,86
		SIm	165	21,6 20,6 - 22,5	584,9 562,9 - 607,0	610,4 572,1 - 648,6	1,24 1,02 - 1,45

Fonte: Elaboração própria a partir dos microdados do STI/UFPB.
Notas: (1) Intervalos com 95% de confiança para as médias calculadas. (2) *Todos os intervalos de confiança apresentam um teste **t** de $0,000$.
Legenda: Área I = C. Exatas e da Terra, Área II = C. S. Aplicadas, Área III = Engenharias.