Predição de risco soberano a partir de indicadores de desenvolvimento do Banco Mundial

Diego Ramon Bezerra da Silva

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, 2018

Diego Ramon Bezerra da Silva

Predição de risco soberano a partir de indicadores de desenvolvimento do Banco Mundial

Monografia apresentada ao curso Engenharia de Computaçãodo Centro de Informática, da Universidade Federal da Paraíba, como requisito para a obtenção do grau de Bacharel em Engenheiro de Computação

Orientador: Thaís Gaudencio do Rêgo

Junho de 2018

Catalogação na publicação
Seção de Catalogação e Classificação

S586p Silva, Diego Ramon Bezerra da.
Predição de risco soberano a partir de indicadores de desenvolvimento do Banco Mundial / Diego Ramon Bezerra da Silva. - João Pessoa, 2018.
53 f.

Orientação: Thaís Gaudêncio do Rego.
Monografia (Graduação) - UFPB/Informática.

1. Rating de Risco Soberano. 2. Aprendizagem de Máquina. 3. Fundamentos Macroeconômicos. 4. Random Forest. 5. Análise de Regressão. I. Thaís Gaudêncio do Rego. II. Título.

UFPB/BC

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Engenharia de Computaçãointitulado ***Predição de risco soberano a partir de indicadores de desenvolvimento do Banco Mundial*** de autoria de Diego Ramon Bezerra da Silva, aprovada pela banca examinadora constituída pelos seguintes professores:

Prof. Dr. Thaís Gaudencio do Rêgo
Universidade Federal da Paraíba

Prof. Dr. Bruno Ferreira Frascaroli
Universidade Federal da Paraíba

Prof. Dr. Tiago Maritan Ugulino de Araújo
Universidade Federal da Paraíba

Coordenador(a) do Departamento Engenharia de Computação
Thaís Gaudêncio do Rêgo
CI/UFPB

João Pessoa, 14 de junho de 2018

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

DEDICATÓRIA

Dedico esse trabalho aos meus familiares.

AGRADECIMENTOS

Agradeço primeiramente ao meu senhor Deus por me guiar neste meu desafio e toda minha vida pessoal e profissional;

Aos meus familiares por me dar a sustentação da família necessária para enfrentar todos os meus desafios;

A Prof. Thaís Gaudêncio do Rêgo, seja pela orientação ao longo deste trabalho, seja por todo apoio dado como coordenadora do Curso de Engenharia da Computação;

Ao Prof. Bruno Ferreira Frascaroli, pela sua co-orientação deste trabalho;

A todos os colegas de graduação, pela convivência, amizade e diversos diálogos abordados a respeito de inúmeros assuntos e temas durante o tempo que passamos juntos

A todos professores do Curso de Engenharia da Computação na Universidade Federal da Paraíba, por todos os conhecimentos compartilhados durante as distintas fases da graduação.

RESUMO

A nota de classificação soberana é um indicador que busca expressar o risco ao que se submetem os investidores estrangeiros ao adquirir títulos de algum país, sendo emitidos por agências de *rating*, empresas independentes de governos ou empresas privadas. As agências de risco estão sendo amplamente criticadas por sua falta de transparência nos processos de classificação. Nesse contexto, a predição de *rating* por meio de modelos de aprendizado de máquina se mostra como uma opção para fins de simulação das notas atribuídas pelas agências. Neste trabalho propõe-se dois estudos, um envolvendo a predição de risco soberano a partir de fundamentos macroeconômicos através do algoritmo *Random Forest*, e outro visando a análise de algumas hipóteses através de um modelo de regressão, mais precisamente responder se a crise financeira de 2008 resultou numa ruptura estrutural nas avaliações de risco soberano, e ainda se as agências de risco possuem diferentes avaliações para países com graus de desenvolvimento econômicos distintos. Os resultados mostram acurácias de até 98,28% no problema de predição, como também através do teste "valor p", sobre o ponto de vista estatístico, verificou-se que houve uma ruptura estrutural nas avaliações após a crise financeira de 2008, e que as agências passaram a avaliar países com economias desenvolvidas de formas diferentes. Esses resultados que indicam que a incapacidade das agências em preverem a crise gerou uma mudança na metodologia das avaliações.

Palavras-chave: *Rating* de Risco Soberano, Aprendizagem de Máquina, Fundamentos Macroeconômicos, *Random Forest*, Análise de Regressão

ABSTRACT

The sovereign classification is an indicator that seeks to express the risk to which foreign investors are subjected when acquiring securities of some country. Being issued by rating agencies, companies independent of governments or private companies. Risk agencies are being widely criticized for their lack of transparency in classification processes. In this context, the prediction of rating through machine learning models is shown as an option for the purpose of simulation of the grades assigned by the agencies. This paper proposes two studies, one involving the prediction of sovereign risk from macroeconomic fundamentals through the algorithm Random Forest, and another aiming to analyze some hypotheses through a regression model, more precisely to answer if the financial crisis of 2008 resulted in a structural break in the sovereign risk assessments, and whether risk agencies have different ratings for countries with distinct degrees of economic development.The results show accuracy up to 98.28% in the prediction problem, but also through the "p-value"test, from a statistical point of view, there was a structural break in assessments after the 2008 financial crisis, and that agencies have come to evaluate countries with developed economies in different ways. These results indicate that the inability of the agencies to predict the crisis has led to a change in the methodology of evaluations.

Key-words: Sovereign Risk Rating, Machine Learning, Macroeconomics Foundations, Random Forest, Regression Analysis

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CART	Árvore de classificação e regressão (*classification and regression tree*)
CSV	Valores separados por vírgula (*comma Separated Values*)
CFS	Seleção de atributos baseada em correlação
FP	Falso positivo (*false positive*)
IA	Inteligência Artificial
LOGIT	Regressão logística (*logistic regression*)
MAE	Média do erro absoluto (*mean absolute error*)
ML	Aprendizado de máquina (*machine learning*)
PCA	Análise de componentes principais (*principal component analysis*)
PPV	Valor preditivo positivo (*positive predictive value*)
RAE	Erro absoluto relativo (*relative absolute error*)
RMSE	Raiz do erro quadrático médio (*root mean square error)*
RRSE	Raiz quadrada do erro relativo *(root relative squared error)*
RF	Floresta aleatória (*random forest*)
ROC	Característica de operação do recepto (*receiver operating characteristic)*
TPR	Taxa positiva verdadeira (*true positive rate*)
TP	Verdadeiro positivo (*true positive*)
UN/DESA	Departamento de Assuntos Econômicos e Sociais das Nações Unidas.
WESP	Situação Econômica Mundial e Perspectivas

Sumário

1 INTRODUÇÃO

1.1 Definição do Problema

A inteligência artificial (IA) é um dos campos mais recentes em ciências e engenharia. O trabalho começou logo após a Segunda Guerra Mundial, primeiramente nomeada por Jonh McCarthly em 1956 na Conferência de Dartmouth. Para a IA ter sucesso, precisamos de inteligência e de um artefato. O computador tem sido o artefato preferido. O computador eletrônico digital moderno foi criado independentemente e quase ao mesmo tempo por cientistas de três países que participavam da segunda Guerra Mundial [2].

O aprendizado de máquina, em inglês *Machine Learning* (ML) é o campo da Inteligência Artificial (IA) responsável pelo desenvolvimento de modelos (hipóteses) gerados a partir de dados, e que automaticamente aperfeiçoam-se com a experiência. Existem diversas aplicações para aprendizado de máquina, tais como processamento de linguagem natural, detecção de fraudes, análise de imagens ou reconhecimento de padrões [9]. Um programa de computador é dito ser capaz de aprender quando este realiza uma certa tarefa, mensurado por certa métrica computável, e melhora com a experiência. De acordo com [17], o ML quando comparado a técnicas de análises manuais, regras de negócios e modelos simplesmente estatísticos, apresenta cinco vantagens: acurácia, automatizado, rápido, customizável e escalável. ML pode ser aplicado em diversas áreas de conhecimento, de problemas financeiros, para recomendação de produtos e clientes em potencial, de monitoramento industrial em tempo real, para análise de sentimento, e diagnóstico médico. O ML é muito aplicado em problemas de economia, sendo as mais comuns desempenho de ações, análise de crédito e previsão de falência [30].

A ciência da economia teve início em 1776, quando o filósofo escocês Adam Smith (1723-1790) publicou *An Inquiry into the Nature and Causes of the Wealth of Nations* [10]. Embora os antigos gregos e outros filósofos tenham contribuído para o pensamento econômico, Smith foi o primeiro a tratá-lo como ciência, usando a ideia de que podemos considerar que as economias consistem em agentes individuais que maximizam seu própria bem-estar econômico [2].

A demanda por sistemas financeiros mais eficientes e confiáveis tem levado a um crescente aprimoramento dos sistemas de análise de dados utilizados pelos agentes financeiros. Essa sofisticação tem levado ao desenvolvimento de sistemas computacionais que sejam capazes de analisar grandes quantidades de dados rapidamente e de gerar relatórios capazes de subsidiar a tomada de decisões. A maioria das principais instituições financeiras, tanto internacionais quanto nacionais, estão utilizando algoritmos de ML para a análise de seus dados. Em várias aplicações na área de finanças, o uso de ML tem possibilitado ganhos financeiros expressivos [5].

Um *rating* de crédito é uma avaliação da qualidade creditícia de cada governo soberano, foca-se nos riscos políticos e econômicos, e é quantitativa e qualitativa [13]. Ele leva em conta a força financeira intrínseca do devedor, incluindo fatores de créditos tradicionais como a qualidade de gestão, posição de mercado e diversidade, flexibilidade financeira, transparência, ambiente regulatório e a capacidade do emissor de cumprir suas obrigações financeiras através dos ciclos de negócios locais, incluindo riscos soberanos, tais como vulnerabilidade a desenvolvimentos políticos, políticas monetárias e fiscais e, em casos raros, risco de conversibilidade e transferência de moeda estrangeira [14]. Quanto maior o risco que investidores assumem em adquirir algum título de um governo soberano, menor a capacidade deste governo em tornar atraente esta aquisição e, portanto, atrair capital estrangeiro. Em consequência, maior é o prêmio remunerado aos investidores para compensá-los por assumir esse risco [16].

Para [31] os *ratings* são importantes não apenas porque alguns dos maiores emissores de dívidas são governos soberanos, mas também porque de acordo com as atribuições de *ratings*, a captação de recursos por governos estaduais, municipais ou empresas privadas localizados nestes países é afetada. A grande quantidade de pesquisas tendo como objetivo fazer a predição dos *ratings* soberanos, a relevância dos resultados obtidos nos últimos anos [16, 24, 25], bem como a necessidade de uma validação independente em função da falta de transparência são motivações para o trabalho desenvolvido, no qual foi realizado dois estudos. O primeiro, visando a predição de *rating* soberano a partir de indicadores de desenvolvimento do Banco Mundial usado o método de ML floresta aleatória (do inglês, *Random Forest* – RF). O segundo, foi a verificação de algumas hipóteses, mais precisamente se a crise de 2008, e o grau de desenvolvimento dos países resultam em avaliações diferentes por parte das agências.

1.2 Premissas e Hipóteses

O trabalho de [16] mostrou a viabilidade de se fazer a predição de *rating* soberano a partir de fundamentos macroeconômicos. No entanto, esse estudo focou-se na análise de países emergentes, que por natureza, cobre uma gama muito pequena dos países do globo. Além disso, os atributos selecionados para serem utilizados no processo de ML foram selecionados por um especialista no domínio do problema.

Nesse trabalho, parte de hipótese que é possível estender esse estudo para os demais países do mundo, restringindo-se a disponibilidade de dados. Isso implicará na perda da possibilidade da seleção de atributos ser feita manualmente, seja em função da grande quantidade de dados ou pelo fato de que num contexto mais geral, torna-se inviável fazer a seleção manual, pois cada país, região ou bloco econômico tem sua peculiaridade. Logo, outra hipótese levantada é que é possível realizar uma seleção de atributos automatizada, mantendo uma acurácia aceitável na predição de *rating* soberano.

1.2.1 Objetivo geral

Prever a classificação de risco soberano de um país a partir dos seus fundamentos macroeconômicos utilizando RF.

1.2.2 Objetivos específicos

1. Obter a classificação de risco soberano a partir de fundamentos macroeconômicos.
2. Obter métricas de acurácia que indiquem a precisão do modelo.
3. Realizar um estudo a partir dos indicadores macroeconômicas mais atuais e abrangido uma maior quantidade de países.
4. Reduzir o número de atributos necessários para uma eficaz predição da classificação de risco soberano.
5. Realizar um estudo sobre o impacto da crise financeira de 2008 nas classificações soberanas, mais precisamente se houve uma ruptura estrutural nas avaliações de risco soberano.
6. Realizar um estudo sobre o impacto do grau de desenvolvimento das economias, mais precisamente se existe avaliações diferentes por parte das agências de classificação.

1.3 Estrutura da monografia

Este trabalho está estruturado da seguinte maneira:

- Capítulo 2: Encontra-se o embasamento, a fundamentação necessária para o desenvolvimento deste estudo.
- Capítulo 3: Descreve-se o método utilizado para realização do trabalho.
- Capítulo 4: Neste capítulo, todos os resultados obtidos são exibidos e discutidos.
- Capítulo 5: Dá lugar às considerações finais, os problemas encontrados, bem como as limitações do trabalho.

2 CONCEITOS GERAIS E REVISÃO DA LITERATURA

2.1 Classificações soberanas

Transações financeiras são intrinsecamente marcadas por assimetrias de informação entre aplicadores e tomadores de recursos. Estes têm necessariamente um maior conhecimento sobre sua própria capacidade de pagamento e sua disposição a pagar do que aqueles que lhes repassam recursos. Portanto, do ponto de vista dos credores, a presença de tal assimetria afetará os prêmios pelos riscos de crédito exigidos em qualquer operação de crédito e aquisição de títulos financeiros [4].

Para [16], o *rating* de risco soberano é um indicador que busca expressar o risco ao que se submetem os investidores estrangeiros ao adquirir títulos de algum país. Os *ratings* de risco soberano são construídos com base em análises das conjunturas econômica, social e política dos países e, por este motivo, podem ser subjetivos porque envolvem julgamento não só das variáveis macroeconômicas internas e externas no presente, mas também das perspectivas das mesmas para o futuro.

As agências de *rating* se apresentam como empresas independentes de quaisquer interesses, quer por parte de governos ou de empresas privadas. Para [11], essa característica lhe permite ter como princípios: independência, objetividade, credibilidade e liberdade de divulgação de avaliações com relação à qualidade de crédito dos emitentes e emissões de dívida. As principais agências de *rating* em nível mundial são a *Standard & Poor's* e a *Moody's Investor Service*, que juntas somam mais de 80% do mercado mundial [12].

Apesar de toda a credibilidade e importância que as agências de risco possuem, é inegável a incapacidade das mesmas de preverem crises econômicas, sendo a mais recente a crise financeira de 2008. A crise financeira de 2008, teve seu epicentro no mercado norte-americano de hipotecas de alto risco (*subprime*). Para [38] começou a tomar forma no ano de 2006, na esteira da inadimplência desse setor de clientes de alto risco. Quando o Sistema de Reserva Federal (do inglês, *Federal Reserve System* - FED) voltou a subir os juros em 2006 e 2007, e a taxa chegou a 5,25 % a.a., veio à tona a incapacidade de grande parte dos devedores saldarem seus débitos. A inadimplência, sendo o estopim do estouro da bolha, teve como consequências as execuções e desvalorizações das hipotecas e dos respectivos derivativos, crise do sistema bancário e encurtamento do crédito.

A classificação *rating* soberano é baseada em dados econômicos, políticos e sociais, sendo necessária uma base de dados que centralize esses indicadores, confiável e com dados atualizados regularmente. Nesse contexto, os indicadores de desenvolvimento mundial tornam-se uma opção tangível.

Os indicadores de desenvolvimento Mundial (do inglês, *World Development Indi-*

cators – WDI) é uma compilação do Banco Mundial de estatísticas internacionalmente comparáveis sobre o desenvolvimento global e a qualidade de vida das pessoas. O WDI é atualizado regularmente e novos dados são adicionados em resposta às necessidades da comunidade. O banco de dados do WDI é resultado da colaboração de inúmeras agências internacionais, mais de 200 escritórios nacionais de estatística e muitos outros [34]. A base é aberta e possui várias formas para obtenção e filtragem de dados.

2.2 Pré-processamento de dados

De posse dos dados de indicadores macroeconômicos e dos *ratings* soberanos fornecidos pelas agências de risco, é possível preparar esses dados para serem usados no processo de aprendizado de máquina. Apesar de algoritmos de ML serem frequentemente adotados para extrair conhecimento de conjuntos de dados, seu desempenho é geralmente afetado pelo estado dos dados. Conjuntos de dados podem apresentar diferentes características, dimensões ou formatos [5]. Antes da aplicação de uma técnica de ML, portanto, pode ser necessário fazer o pré-processamento. Aplicando técnicas de pré-processamento é possível obter um modelo mais robusto, proporcionando assim uma melhor acurácia na predição.

Durante o pré-processamento são eliminadas inconsistências, que podem ocorrem quando dados diferentes são representados pelo mesmo rótulo, ou quando o mesmo dado é representado por rótulos diferentes [20]. Identificação e atenuação de ruídos representados por dados distorcidos também devem ser tratados, já que não representam os valores verdadeiros. Também é feito o tratamento de dados ausentes e a remoção de atributos duplicados e redundantes.

Nem sempre os dados representativos do problema são encontrados de forma direta, sendo os atributos cruciais para o aprendizado do conceito misturados com atributos com pouca ou nenhuma correlação com o problema. Nesses casos é pertinente utilizar técnicas de seleção de atributos. Segundo [20], essas técnicas consistem em encontrar um subconjunto de atributos no qual o algoritmo de AM irá se concentrar. Para [21] as principais razões que justificam o uso de métodos de seleção de atributos são:

- Muitos algoritmos de AM não funcionam bem com uma grande quantidade de atributos, dessa forma, a seleção de atributos pode melhorar o desempenho desses algoritmos;
- O número menor de atributos torna o conhecimento induzido por algoritmos de AM simbólico, frequentemente, mais compreensível;
- Dependendo do domínio, o custo da coleta de dados não é desprezível, métodos de seleção de atributos podem diminuir o custo da aplicação.

2.3 Análise de componentes principais

Além dos problemas atenuados com o pré-processamento de dados, conjunto de dados com alta dimensão ainda podem sofrer com baixa robustez do modelo, bem como tornar os experimentos computacionalmente exaustivos. Portanto, é pertinente realizar uma redução de dimensionalidade ou, em outros termos, realizar a extração dos atributos mais importantes. Na literatura são descritos vários métodos de extração de atributos mais importantes, sendo sua escolha e aplicação baseado na natureza do problema, dentre os quais encontra-se a Análise de Componentes Principais (do inglês, *Principal Component Analysis* – PCA). Também conhecida como transformada Karhunen-Loève, foi descrito primeiramente por [26], sendo uma técnica que consiste em transformar um conjunto de variáveis originais em outro conjunto de variáveis da mesma dimensão denominadas de componentes principais. As variáveis do novo conjunto apresentam as propriedades de serem combinações lineares de todas as variáveis originais, são independentes entre si, e retém o máximo de informação, em termos da variação total contida nos dados.

O PCA é usado principalmente como uma ferramenta na análise de dados exploratórios e para fazer modelos preditivos. Para [27], o PCA pode ser usado como uma maneira de identificar padrões em dados e destacando as semelhanças e diferenças. Uma vez que padrões em dados podem ser difíceis de serem identificados em conjuntos com alta dimensão, onde a representação gráfica torna-se inviável.

Segundo [32], a técnica de análise de componentes principais tem como objetivo geral a redução da dimensionalidade e interpretação de um conjunto de dados. Algebricamente, os componentes principais são combinações lineares de p variáveis aleatórias $X_1, X_2, ..., X_p$. Geometricamente, as componentes principais representam um novo sistema de coordenadas, obtidas por uma rotação do sistema original, que fornece as direções de máxima variabilidade, e proporciona uma descrição mais simples e eficiente da estrutura de covariância dos dados.

Matematicamente [33], a transformação é definida por um conjunto p-dimensional de vetores $w_{(k)} = (w_1, ..., w_p)_{(k)}$ que mapeia cada linha do vetor $X_{(i)}$ para um novo vetor de componentes principais $t_{(i)} = (t_1, ..., t_m)_{(i)}$, dado por $t_{k(i)} = X_{(i)} \cdot W_{(k)}$ para cada $i = 1, ..., n$ e $k = 1, ..., n$ de tal forma que as variáveis individuais t consideradas sobre o conjunto de dados herdam sucessivamente a variância máxima possível de X com cada vetor de pesos w restringindo para ser um vetor unitário.

Primeiro componente

Em ordem para maximizar a variância, o primeiro vetor de pesos w deve satisfazer:

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w} \right)^2 \right\} \tag{1}$$

Equivalentemente, escrevendo na forma de matrizes:

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \{ \|\mathbf{X}\mathbf{w}\|^2 \} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \right\} \tag{2}$$

Desde que $w_{(1)}$ foi definido por um vetor unitário, então satisfaz equivalentemente:

$$\mathbf{w}_{(1)} = \arg\max \left\{ \frac{\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\} \tag{3}$$

Um resultado padrão para uma matriz simétrica tal como $X^T X$ é que esse valor de quociente máximo possível é o maior autovalor da matriz, que ocorre quando w é o correspondente autovetor. Com $w_{(1)}$ obtido, o primeiro componente de um vetor de dados $x_{(i)}$ pode então ser dado como uma pontuação $t_{1(i)} = X_i \cdot w_{(1)}$ nas coordenadas transformadas, ou como o vetor correspondente nas variáveis originais, $\left\{ X_i \cdot w_{(1)} \right\} \cdot w_{(1)}$

Demais componentes

O k-ésimo componente pode ser encontrado subtraindo os primeiros $k-1$ componentes principais a partir de X.

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X} \mathbf{w}_{(s)} \mathbf{w}_{(s)}^{\mathrm{T}} \tag{4}$$

E depois, encontra-se o vetor de pesos que extrai a variância máxima dessa nova matriz de dados:

$$\mathbf{w}_{(k)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \|\hat{\mathbf{X}}_k \mathbf{w}\|^2 \right\} = \arg\max \left\{ \frac{\mathbf{w}^T \hat{\mathbf{X}}_k^T \hat{\mathbf{X}}_k \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\} \tag{5}$$

Acontece que isso dá aos autovetores restantes do $X^T X$, com os valores máximos para a quantidade entre parênteses dados pelos respectivos autovalores correspondentes. Assim, os vetores de pesos são autovetores do $X^T X$. O k-ésimo componente de um vetor de dados $x_{(i)}$ pode, portanto, ser dado como uma pontuação $t_{k(i)} = X_i \cdot w_{(k)}$ nas coordenadas transformadas, ou como o vetor correspondente no espaço do original variáveis $\left\{ X_i \cdot w_{(k)} \right\} \cdot w_{(k)}$, onde $w_{(k)}$ é o k-ésimo autovetor de $X^T X$. A decomposição total dos componentes principais de X pode, portanto, ser dada como:

$$\mathbf{T} = \mathbf{XW} \tag{6}$$

Onde W é uma matriz p x p cujas colunas são os autovetores do $X^T X$. A transposição de W às vezes é chamada de transformação de clareamento ou esfregaço (do inglês, *whitening* ou *transformation*).

Covariâncias

O próprio $X^T X$ pode ser reconhecido como proporcional à matriz de covariância da amostra empírica do conjunto de dados X. A covariância da amostra Q entre dois dos principais componentes principais do conjunto de dados é dada por:

$$\begin{aligned} Q(\mathrm{PC}_{(j)}, \mathrm{PC}_{(k)}) &\propto (\mathbf{X}\mathbf{w}_{(j)})^T (\mathbf{X}\mathbf{w}_{(k)}) \\ &= \mathbf{w}_{(j)}^T \mathbf{X}^T \mathbf{X} \mathbf{w}_{(k)} \\ &= \mathbf{w}_{(j)}^T \lambda_{(k)} \mathbf{w}_{(k)} \\ &= \lambda_{(k)} \mathbf{w}_{(j)}^T \mathbf{w}_{(k)} \end{aligned} \tag{7}$$

Onde a propriedade do autovalor de $w_{(k)}$ foi usada para se mover da linha 2 para a linha 3. No entanto, os autovetores $w_{(i)}$ e $w_{(k)}$ correspondentes aos autovalores de uma matriz simétrica são ortogonais (se os autovalores forem diferentes) ou pode ser ortogonalizado (se os vetores compartilham um valor repetido igual).

O produto na linha final é, portanto, zero; Não há covariância de amostra entre diferentes componentes principais ao longo do conjunto de dados.

Outra maneira de caracterizar a transformação dos componentes principais é, portanto, como a transformação para as coordenadas que diagonalizam a matriz de covariância da amostra empírica. A matriz de covariância empírica para as variáveis originais pode ser escrita como:

$$\mathbf{Q} \propto \mathbf{X}^T\mathbf{X} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T\text{"} \tag{8}$$

A matriz de covariância empírica entre os principais componentes torna-se:

$$\mathbf{W}^T\mathbf{Q}\mathbf{W} \propto \mathbf{W}^T\mathbf{W}\,\mathbf{\Lambda}\,\mathbf{W}^T\mathbf{W} = \mathbf{\Lambda} \tag{9}$$

Onde Λ é a diagonal da matriz de autovalores Λ_k de X^TX.

A transformação $T = Xw$ mapeia um vetor de dados $w_{(i)}$ de um espaço original de variáveis p para um novo espaço de variáveis p que não estão correlacionadas ao longo do conjunto de dados. No entanto, nem todos os componentes principais precisam ser mantidos. Mantendo apenas os primeiros L componentes principais, produzidos usando apenas os primeiros vetores de pesos, fornece a transformação truncada

$$\mathbf{T}_L = \mathbf{X}\mathbf{W}_L \tag{10}$$

Onde T_L e a matriz agora tem n linhas, mas apenas L colunas. Em outras palavras, a PCA aprende uma transformação linear $t = W^Tx, x \in R^p, t \in R^L$, onde as colunas de p x L da matriz w formam uma base ortogonal para os L atributos (os componentes da representação t) que não estão correlacionados [28]. Por construção, de todas as matrizes de dados transformadas com apenas L colunas, esta matriz de pontuação maximiza a variância nos dados originais que foi preservada, ao mesmo tempo em que minimiza o erro total de reconstrução ao quadrado

$$\|\mathbf{T}\mathbf{W}^T - \mathbf{T}_L\mathbf{W}_L^T\|_2^2 \tag{11}$$

Essa redução de dimensionalidade pode ser um passo muito útil para visualizar e processar conjuntos de dados de alta dimensão, enquanto ainda mantém a maior parte possível da variância no conjunto de dados.

2.4 Seleção de atributos baseado em correlação

Embora a análise de componentes principais seja uma técnica robusta e amplamente utilizada em vários tipos de problemas, ela possui uma consequência que pode ser prejudicial em problemas específicos, que é a destruição dos atributos originais, impossibilitando uma análise pós-predição sobre esses atributos e o impacto deles no modelo dado o contexto do domínio do problema estudado.

Uma técnica alternativa para redução de dimensionalidade para esses casos é algoritmo Seleção de atributos baseado em correlação (do inglês, *Correlation-based Feature Selection* - CFS) descrito inicialmente no trabalho de [35]. O conceito base deste algoritmo é a seleção de atributos baseada na correlação deles. Para [35] a hipótese central é que bons conjuntos de atributos são altamentes correlacionados com a classe, e não correlacionados uns com os outros. O funcionamento do algoritmo é dividido em duas etapas: (1) análise de correlação entre os atributos e a classe e (2) a busca por sub-conjuntos compatíveis com a hipótese central.

A análise de correlação segue a hipótese central formalizada através de seguinte heurística equação (12) que calcula o mérito de um subconjunto de atributos S contendo k atributos.

$$\text{Merit}_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}. \tag{12}$$

Onde $k\overline{r_{cf}}$ é o valor médio de todas as correlações de atributos de classificação, e $\overline{r_{ff}}$ é o valor médio de todas as correlações de atributos com atributos. O critério CFS é definido através da fórmula:

$$\text{CFS} = \max_{S_k} \left[\frac{r_{cf_1} + r_{cf_2} + \cdots + r_{cf_k}}{\sqrt{k + 2(r_{f_1 f_2} + \cdots + r_{f_i f_j} + \cdots + r_{f_k f_1})}} \right]. \tag{13}$$

Essa fórmula pode ser reescrita como um problema de programação linear inteira mista que pode resolvida por algoritmos *branch and bound*.

$$\text{CFS} = \max_{x \in \{0,1\}^n} \left[\frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n} x_i + \sum_{i \neq j} 2 b_{ij} x_i x_j} \right]. \tag{14}$$

2.5 Aprendizagem de máquina supervisionada

Os dados resultantes do processo de pré-processamento e extração de atributos são aptos para serem utilizados no processo de aprendizado de máquina (do inglês, *Machine Learning* - ML). O processo de ML é usualmente categorizado em duas vertentes: apren-

dizagem supervisionada e não supervisionada. Na aprendizagem supervisionada, tem-se a figura de um professor externo, possuindo conhecimento do ambiente representado através de um conjunto de entradas e saídas [7]. A máquina aprenderá a partir de um conjunto de exemplos previamente rotulado, chamado de conjunto de treinamento e fará predições, ou seja, rotulará baseado nesse conjunto de treinamento os novos exemplos, ainda não vistos. Cada exemplo do conjunto de treinamento é formado por um conjunto de variáveis independentes chamadas atributos mais uma variável dependente denominada rótulo.

A tarefa de aprendizagem supervisionada pode ser formulada da seguinte maneira: dado um conjunto de treinamento com exemplos de entrada e saída $(x_1, y_1), ..., (x_n, y_n)$, onde cada saída y_i foi gerada por uma função desconhecida $y = f(x)$, obter uma função h que se aproxime da função verdadeira f. Quando a saída y for um conjunto finito de valores, o problema da aprendizagem será chamado de classificação. Quando y for um número, o problema de aprendizagem é chamado de regressão [2].

O processo de ML supervisionada está resumido na Figura 1. O processo inicia-se com a definição do problema, dados representativos do problema são obtidos, tendo o objetivo de entender a relação entre o conjunto de atributos e a sua saída. Em seguida, na fase de modelagem, é realizado o treinamento, onde a máquina irá aprender utilizando o algoritmo selecionado. Também nessa fase, se cabível, será feito o ajuste dos parâmetros do método, sendo esse processo controlado pela avaliação de acurácia do algoritmo. Na fase de predição o modelo para o problema já estará definido e validado, sendo assim, apto para fazer novas predições em novas instâncias do problema não rotuladas.

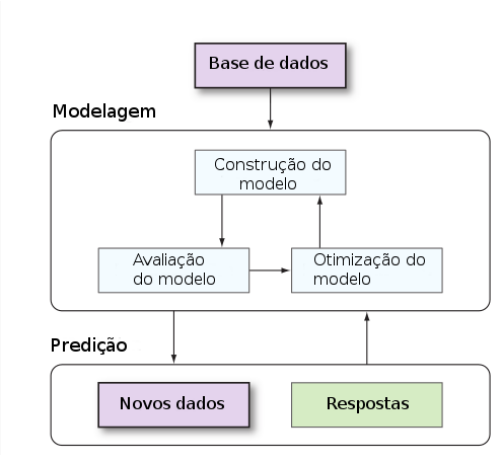

Figura 1: Processo de ML supervisionada. Fonte [17]

A ML supervisionada diferencia-se da aprendizagem não supervisionada pelo fato de que nesta não é fornecido um *feedback* explícito, ou seja, não possuindo um conjunto de variável dependente denominado rótulo. A tarefa mais comum de aprendizagem não supervisionada é o agrupamento: a detecção de grupos de exemplos de entrada potencialmente úteis [2].

2.6 Random forest

Dentre os métodos de ML supervisionada, está a floresta aleatória (do inglês, *Random Forest* - RF). O algoritmo RF foi descrito inicialmente no trabalho de [3]. O algoritmo consiste de uma combinação de árvores de decisão. Uma extensão do algoritmo foi desenvolvida fazendo uso da técnica de *bagging* [6] e da ideia de seleção aleatória de atributos proposta por [3]. Cada uma delas é formada a partir de subconjuntos do conjunto de treino, formados a partir de substituição com amostras do conjunto original. Cada um destes conjuntos é criado por um tipo de amostragem chamado *bootstrapp* [18]. Cada vez que uma amostra é selecionada, é igualmente provável que seja novamente selecionada e adicionada no conjunto de treinamento.

Foi demonstrado por [22] que ganhos substanciais na acurácia no processo de classificação e regressão podem ser obtidos usando conjunto (do inglês, *ensemble*) de árvores, onde cada conjunto cresce de acordo com um parâmetro aleatório. Predições finais são obtidas pela agregação sobre os conjuntos gerados. Como os constituintes básicos dos conjuntos são preditores baseados em árvores, e como cada uma dessas árvores é construída usando uma injeção de aleatoriedade, esses procedimentos são chamados de "florestas aleatórias"[23].

Segundo [6], uma floresta aleatória consiste de uma coleção de classificadores estruturados em árvores $\{h(x, \Theta_k, k = 1...\}$, onde $\{\Theta_k\}$ são vetores aleatórios identicamente distribuídos e cada árvore lança uma unidade de votação para a classe mais popular na entrada x. Supondo que o conjunto de treinamento é representado por T e ainda que existam M atributos, sendo $T = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, Xi é o vetor de entrada $x_{i1}, x_{i2}, ..., xin$ e y_i é o rótulo ou classe da instância. Suponha que a floresta terá S árvores, então serão criados novos S conjuntos de dados com o mesmo tamanho do conjunto original a partir do reescalonamento (do inglês, *Resampling*) aleatório dos dados. Isso resultará em $\{T_1, T_2, ..., T_s\}$ conjunto de dados. Cada um desses conjuntos é chamado de *bootstrap*. Devido ao processo de substituição cada conjunto de dados T_i poderá ter conjunto de dados duplicados ou ausentes quando comparado ao conjunto original. Esse processo é denominado ensacamento (do inglês, *bagging*). Portanto, o RF criará S árvores e usará $m = floor(lnM + 1)$ sub-atributos aleatórios dos M possíveis para criação de cada árvore. Esse processo é chamado de método do subespaço aleatório (do inglês, *Random Subspace Method*).

Como visto, o RF é um algoritmo parametrizado, sendo a sua acurácia altamente dependente da escolha desses parâmetros. Dentre os parâmetros, o tamanho da bolsa (do inglês, *bag*) indica a porcentagem referente ao tamanho do conjunto de dados original criada no reescalonamento aleatório dos dados. A quantidade de atributos indica quantos deles serão utilizados na criação de cada árvore. Por se tratar de um processo iterativo,

também é possível controlar o número máximo de iterações do algoritmo. Por último, pode-se controlar o número de amostras a serem exibidas antes que uma atualização seja realizada através do tamanho do lote (do inglês, *batch size*) ou, de outra forma, o tamanho do lote controla quantas predições serão feitas de cada vez.

Dentre as vantagens do método RF destacam-se seu relativo baixo custo computacional, característica indispensável quando se trabalha com grandes quantidades de dados, além de evitarem sobreajuste (do inglês, *overfitting*) e serem pouco sensíveis a ruídos [6].

2.7 Análise de Regressão Linear

Em diversos problemas das áreas médica, industrial, biológica, econômica entre outras, é de grande interesse verificar se duas ou mais variáveis estão relacionadas de alguma forma. Esse tipo de problema pode ser modelado matematicamente através de um modelo chamado de regressão. Análise de regressão tem como objetivo prever o comportamento de uma variável dependente a partir do comportamento de uma ou mais variáveis independentes (ou regressoras) [36].

Os modelos mais básicos de regressão lineares são Regressão Linear Simples e Regressão Linear Múltipla. Esses modelos diferem basicamente quanto a quantidade de variáveis regressoras presentes. Se a variável dependente está relacionada apenas com uma variável independente, têm-se o caso de Regressão Linear simples. Porém, se a variável dependente está relacionada com mais de uma variável independente, têm-se o caso de Regressão Linear Múltipla. Como o problema abordado neste trabalho se enquadra no modelo de Regressão Linear Múltipla, a mesma terá seu funcionamento explanado.

Modelo

Um modelo de Regressão Linear Múltipla é definido matematicamente através da seguinte equação:

$$Y_1 = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \varepsilon_1 \tag{15}$$

Onde $X_{2i}, X_{3i}, ..., X_{ki}$ são os valores das variáveis independentes (ou explicativas) e $\beta_1, \beta_2, \beta_3, ..., \beta_k$ são os parâmetros ou coeficientes de regressão. O modelo de Regressão Linear Múltipla também possui algumas suposições necessárias para sua existência que são:

- O erro tem média zero e variância σ^2 desconhecida;
- Os erros são não correlacionados

- Os erros têm distribuição normal;
- As variáveis regressoras $X_{2i}, X_{3i}, ..., X_{ki}$ assumem valores fixos.

Estimação de Parâmetros

O método [36, 37] usualmente utilizado para estimar os parâmetros do modelo de regressão múltipla é o método dos mínimos quadrados, que permite encontrar uma reta que minimize a distância entre os pontos observados e a reta, fazendo, em média, a soma dos desvios quadráticos ser igual a zero. Para simplificação será adotada a notação matricial do modelo de Regressão Linear Múltipla:

$$Y = X\beta + \varepsilon \tag{16}$$

Com:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad , \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad , \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{e} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{17}$$

- Y é um vetor $n \times 1$ cujos componentes corresponde às n respostas;
- X é uma matriz de dimensão $n \times (p+1)$ denominada matriz do modelo;
- ε é um vetor de dimensão $n \times 1$ cujos componentes são os erros e
- β é um vetor $(p+1) \times 1$ cujos elementos são os coeficientes de regressão.

O método de mínimos quadrados tem como objetivo encontrar o vetor $\widehat{\beta}$ que minimiza

$$L = \sum_{i=1}^{n} \varepsilon_i^2 \tag{18}$$

Essa soma de quadrados pode ser expandida como:

$$L = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) = Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta = Y'Y - 2\beta'X'Y + \beta'X'X\beta \tag{19}$$

Como as matrizes $Y'X\beta - \beta'X'Y$ são equivalentes, por uma ser a transposta da outra, então:

$$L = Y'Y - 2X'Y + 2X'X\beta. \tag{20}$$

Como é sabido do cálculo, o mínimo da função L pode ser obtida derivando-a e igualando a zero:

$$\frac{\partial L}{\partial \beta} = -2X'Y + 2X'X\beta \tag{21}$$

$$(X'X)\widehat{\beta} = X'Y.$$

Igualando a zero e substituindo o vetor β pelo vetor $\widehat{\beta}$, temos

$$\widehat{\beta} = (X'X)^{-1}X'Y. \tag{22}$$

Portanto, o modelo de regressão linear ajustado e o vetor de resíduos são respectivamente

$$\widehat{Y} = X\widehat{\beta} \quad \text{e} \quad e = Y - \widehat{Y} \tag{23}$$

2.8 Matriz de confusão

Estimar a exatidão de um classificador tem como objetivo validar um modelo, bem como escolher um classificador que melhor se adapte ao modelo, sendo este o objetivo principal do ML, uma predição acurada. Tendo isso em vista é possível elaborar uma matriz de confusão, fornecendo ainda mais informações sobre a acurácia do modelo. A matriz de confusão é uma ferramenta útil para analisar o quão bem um classificador pode reconhecer tuplas de diferentes classes.

Para [18] uma matriz de confusão, conforme visto na Figura 2, pode ser definida da seguinte forma: Dado m classes, uma matriz de confusão é uma tabela de dimensão m por m. Uma entrada, $CM_{i,j}$ nas primeiras m linhas e m colunas indica o número de tuplas da classe i rotuladas pelo classificador como classe j. Para que um classificador tenha uma boa precisão, idealmente a maioria das tuplas seria representada ao longo da diagonal da matriz de confusão, a partir da entrada $CM_{1,1}$ até a entrada $CM_{m,m}$ com o restante das entradas perto de zero. A tabela pode ter linhas ou colunas adicionais para fornecer totais ou taxas de reconhecimento por classe.

		Classe Predita	
		Positivo	Negativo
Classe Verdadeira	Positivo	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 2: Matriz de confusão. Fonte [18]

Através da matriz de confusão é possível extrair as métricas de sensibilidade e especificidade. Sensibilidade também é referida como taxa de verdadeiro positivo, indicando a proporção da tuplas positivas que são corretamente identificadas, enquanto especificidade é a taxa de negativo verdadeiro, indicando a proporção das tuplas negativas que são corretamente identificadas. Em adição, a precisão pode ser definida como a porcentagem das tuplas rotuladas como classe C_i que são corretamente da classe C_i. Essas métricas são definidas nas equações:

$$sensibilidade = \frac{Tpositivo}{positivo} \tag{24}$$

$$especificidade = \frac{Tnegativo}{negativo} \tag{25}$$

Por último, é possível definir acurácia em função de sensibilidade e especificidade:

$$acurácia = sensibilidade\frac{positivo}{positivo + negativo} + especificidade\frac{positivo}{positivo + negativo} \tag{26}$$

A partir dos valores de sensibilidade e especificidade é possível obter duas métricas muito usadas em problemas de classificação: a característica de operação do receptor (do inglês, *Receiver Operating Characteristic* - ROC) e a pontuação F1 (do inglês, *F1-Score*). Uma curva ROC busca estabelecer uma comparação entre modelos de classificação avaliando diferentes pontos de limiar para discriminação; enquanto o eixo vertical do gráfico indica a sensibilidade (taxa de verdadeiros positivos), o eixo horizontal indica a taxa de falsos positivos (1 - especificidade), onde cada ponto no espaço representa os respectivos valores obtidos de uma matriz de confusão [42]. Já a pontuação F1 é uma média harmônica entre precisão e a sensibilidade, sendo especialmente usada em modelos com classes desproporcionais e que não emite probabilidades.

Para garantir que os valores de acurácia de um classificador seja uma estimativa confiável utilizam-se algumas técnicas de avaliação, dentre as quais citam-se: Método *Holdout*, Subamostragem randômica (do inglês, *Random Subsampling*) e Validação cruzada k-grupo (do inglês, *k-Fold Cross-Validation*) [18].

2.9 Validação cruzada k-grupo

Dentre os vários métodos usados para estimar o desempenho de um método ML encontram-se a validação cruzada k-grupo (do inglês, *k-fold cross validation*). A técnica de validação cruzada k-grupo começa dividindo aleatoriamente o conjunto de dados em k subconjuntos disjuntos, sendo normalmente utilizado o valor $k = 10$. Para cada grupo, um modelo é treinado com o conjunto total de dados, exceto os dados desse grupo. Depois de todos os grupos serem percorridos, as previsões para cada grupo são agregadas e comparadas com a variável real a ser predita, avaliando assim sua predição [17]. A Figura 3 ilustra o processo de validação cruzada k-grupo.

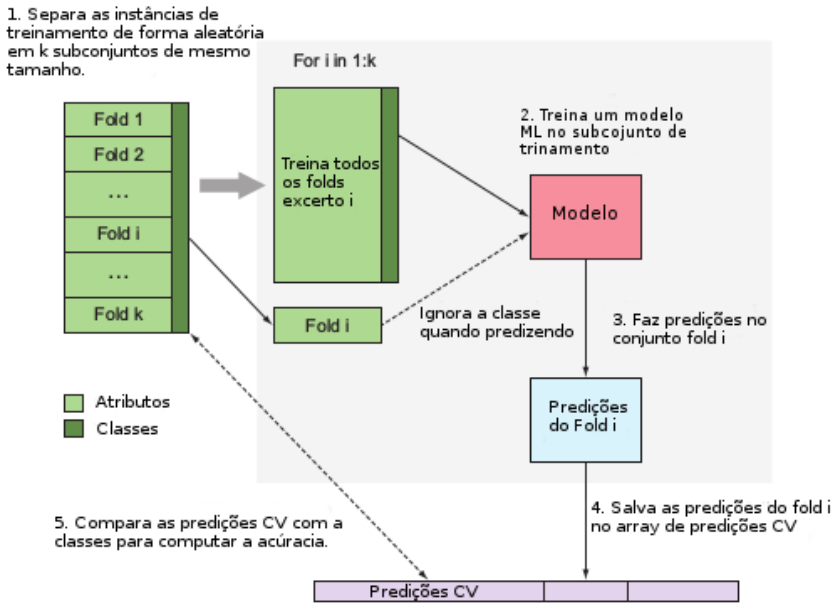

Figura 3: Processo de validação cruzada. Fonte [17]

Validação cruzada k-grupo também é usada para seleção de modelos ou ajuste de parâmetros específicos do método de ML utilizado, onde a melhor parametrização é escolhida [19].

2.10 Trabalhos relacionados

Nesta seção serão apresentados alguns trabalhos relacionados com o problema de previsão da classificação de risco soberano a partir de fundamentos macroeconômicos. Serão citados os trabalhos de [24] e [16] possuem maior semelhança com o trabalho proposto.

A classificação de risco soberano é um tema amplamente pesquisado, sendo estudado por diversos pesquisadores há algum tempo. No trabalho de [24] foi proposto um modelo estimado para classificação de risco soberano. Diferentes modelos estatísticos, tais como regressão múltipla, regressão logística (do inglês, *logistic regression* - LOGIT) e árvore de classificação e regressão (do inglês, *classsification and regression tree* - CART), foram anteriormente propostos na literatura e usados na prática para explicar e predizer a classificação de risco soberano em função de indicadores macroeconômicos. No entanto, esses modelos apresentam algumas falhas importantes. Primeiro, os indicadores são assumidos para se aplicarem em todos os países avaliados, não levando em conta possíveis fatores que podem fazer determinados indicadores possuírem mais peso em determinado país. Segundo, o nível de ajustamento de dados é alcançado apenas no contexto das técnicas de otimização inerentes aos modelos estatísticos escolhidos *a priori*.

Nesse contexto, foi proposto no trabalho um procedimento que aplica uma generalização da regressão logística para vincular a classificação de risco soberano de um país e os indicadores político-econômicos. As estimativas para os parâmetros do modelo são obtidas através de um modelo de programação matemática desenvolvido de forma independente, em vez de se basear em técnicas clássicas de otimização com a maioria dos modelos estatísticos. Os resultados do

procedimento proposto foram comparados com dois modelos estatísticos amplamente utilizados: LOGIT e CART. Os resultados indicaram que o método proposto é superior aos métodos estatísticos, no que diz respeito a erros de estimação e validação.

Outro trabalho de grande relevância foi realizado por [16], onde foi realizado um estudo sobre classificação de risco soberano dos diversos países emergentes a partir de fundamentos macroeconômicos utilizando Redes Neurais Artificiais. Para realização do estudo foi considerado um conjunto de dados formado por indicadores macroeconômicos, contendo 467 amostras, referentes a faixa de 1989 até 2012, obtidos da base de dados do Banco Mundial denominada Indicadores do desenvolvimento Mundial (do inglês, *World Development Indicators* – WDI) e de *ratings* obtidos a partir das principais agências de classificação no âmbito mundial.

Variáveis (em %)	Observações 2003	*Cenário* 1	*Cenário* 2	*Cenário* 3	*Cenário* 4
Balança externa / Produto	6	6,4	7,6	7,6	7,6
Dívida externa total / Exportações	265	235	235	235	265
Crescimento do produto *per capita*	-1	2	1	2	2
Nível de reservas totais / Produto	9	15,75	15,75	13,75	15,75

Figura 4: Simulações de cenários macroeconômicos. Fonte [16]

Nesse estudo observou-se o grau de homogeneidade entre as atribuições de *ratings* e os fundamentos macroeconômicos dos países da amostra, no qual quatro fundamentos mostraram-se intimamente ligados às atribuições. A Figura 4 mostra os resultados de um exercício de estática comparativa. Utilizou-se o modelo para fazer simulações de cenários das condições externas de crédito para o Brasil ao modificar os fundamentos macroeconômicos. Estes cenários indicam que as agências esperam por crescimento mais acentuado do produto per capita e pela diminuição da dívida pública.

O trabalho proposto diferencia-se principalmente pela ausência de uma pré-seleção de atributos por um especialista no domínio do problema, optando-se por usar todos os atributos presentes na base de dados do Banco Mundial, com 1374 atributos, e fazer uso da técnica de análise de componentes principais para realizar a redução de dimensionalidade. Outra diferença fundamental desse trabalho é que o modelo pode ser usado principalmente como uma caixa preta (do inglês, *Black Box*), já que o uso do PCA torna uma análise posterior a predição inviável, pelo fato do mesmo criar variáveis a partir das originais, que deixam de ser consideradas.

3 METODOLOGIA

Neste trabalho, é proposta a realização da predição do *rating* soberano a partir de fundamentos macroeconômicos. Utilizado o método de ML denominado RF. A escolha desse método é motivada, principalmente, por se tratar de um método menos sensível à ruídos e seu custo computacional ser moderado, quando comparado com outros métodos de ML supervisionado. Sendo assim, uma boa opção para ser utilizando com grandes conjuntos de dados, garantindo assim maior rapidez na realização do experimento. No entanto, esse trabalho pode ser reproduzido utilizado qualquer método de ML supervisionada presente na literatura. A metodologia do experimento é esquematizada pelo fluxograma apresentado na Figura 5.

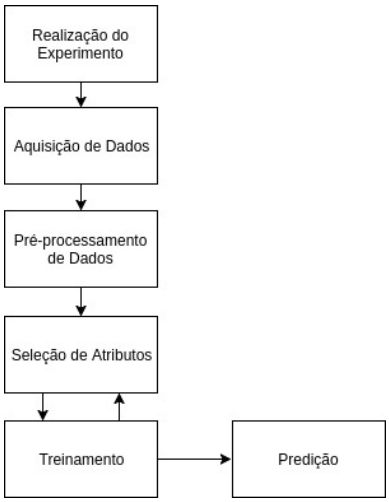

Figura 5: Diagrama de fluxo da metodologia. Fonte: Elaborada pelo autor.

3.1 Aquisição de dados

Para a realização do experimento foi necessária à obtenção de três bases de dados diferentes e a partir de fontes diferentes, uma contendo os indicadores macroeconômicos, uma contendo os *ratings* soberanos das principais agências de *rating* no âmbito internacional e, por fim, uma contendo o grau de desenvolvimento econômico de cada país.

A base de dados com indicadores macroeconômicos provindas do Banco Mundial, denominada Indicadores do desenvolvimento Mundial (do inglês, *World Development Indicators* – WDI), é a base principal de indicadores de desenvolvimento do Banco Mundial, compilada a partir de fontes internacionais oficialmente reconhecidas. Ela apresenta os dados de desenvolvimento global mais atuais e precisos disponíveis, e inclui estimativas nacionais, regionais e globais. A base é pública e apresenta diversas formas de obtenção e exploração de dados.

A base de dados contendo os *ratings* soberanos, ao contrário da base do Banco Mundial não é completamente pública, sendo disponibilizado apenas os *ratings* mais atuais para cada país. Para obter uma base com uma série histórica dos *ratings* foi feito uso de um script de *web scraping* que fez a extração e agrupamento desses dados automaticamente a partir de páginas da web.

A base de dados contendo o grau de desenvolvimento econômico foi obtida através do relatório da Situação Econômica Mundial e Perspectivas (do inglês, *World Economic Situation and Prospects* - WESP). Provinda do esforço de vários órgãos internacionais, encabeçada pelo departamento de Assuntos Econômicos e Sociais das Nações Unidas (UN/DESA), fornece um prognóstico da situação econômica mundial, em especial, uma classificação do nível de desenvolvimento de cada economia.

Finalmente, a base usada no experimento será obtida através do interseção das três bases citadas anteriormente.

3.2 Weka

Para realizar o pré-processamento e treinamento do modelo foi utilizado o software Weka. O Weka (*Waikato Environment for Knowledge Analysis*) surgiu na Nova Zelândia na Universidade de Waikato, tendo seu desenvolvimento iniciado em 1993, na linguagem de programação Java, disponibilizado sobre a licença de código fonte aberto GPL (do inglês, *General Public License*), sendo possível estudar e alterar o respectivo código fonte. O Weka fornece diversos algoritmos já implementados para realização de tarefas de pré-processamento de dados e aprendizagem de máquina.

3.3 Pré-processamento de dados

Antes de partir para a aplicação de uma técnica de ML, é necessário fazer o pré-processamento da base de dados bruta. Aplicando técnicas de pré-processamento é possível obter um modelo mais robusto, proporcionando, assim, uma melhor acurácia na predição.

A base resultante do processo de aquisição de dados possuía ausência de um grande número de valores para os atributos considerados para cada instância. Na literatura essa perda de dados é chamada de *missing data* ou dados faltantes, fazendo-se necessária à aplicação de uma técnica de imputação de dados faltantes. A imputação de dados ausentes consiste na técnica de preencher os dados faltantes com valores plausíveis. Nesse experimento, a técnica de imputação utilizada foi a técnica de imputação por mediana e imputação por moda.

Também observou-se na base o fato da mesma possuir uma grande quantidade de atributos em escalas diferentes, situação que pode afetar negativamente a robustez do modelo, pois métodos baseados em instâncias são mais eficazes se os atributos de entrada estiverem na mesma escala. Para contornar esse problema, recorre-se a técnicas de normalização e centralização dos atributos da base de dados. A aplicação dessas duas técnicas em conjunto resultará em uma base padronizada, ou seja, com valor médio de 0 e um desvio padrão de 1. A metodologia do pré-processamento é esquematizado pelo fluxograma apresentado na Figura 6.

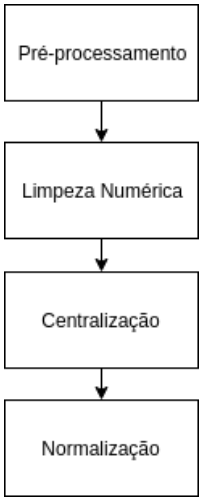

Figura 6: Diagrama de fluxo do pré-processamento. Fonte: Elaborada pelo autor.

Após a aplicação desse fluxo de procedimentos de pré-processamento as bases adquiriram os seguintes aspectos:

Classificação de Desenvolvimento

A base com classificação de desenvolvimento dos países, cujo os dados foram extraídos do relatório WESP (2017), resultante do pré-processamento de dados que será usada no experimento é sumarizada na Tabela 1 e conta com 107 países distintos, sendo cerca 57,94% dos países classificados como economias em desenvolvimento (categoria C), 31,77% como países com economias desenvolvidas (categoria A) e 10,28% como economias em transição (categoria B).

	País	**A**	**B**	**C**
	Andorra	0	0	0
	United Arab Emirates	0	0	1
	Albania	0	1	0
	...	...	...	...
	Vietnam	0	0	1
	South Africa	0	0	1
Total	107	34	11	62

Tabela 1: Descrição geral dos dados da base (UN/DESA). Fonte: Elaborada pelo autor.

Ratings soberanos

A base com classificação de desenvolvimento dos países que será de fato usada no experimento é sumarizada na Tabela 2 e conta com 3596 instâncias, de 137 países distintos, com dados históricos de 38 anos distintos, na faixa de anos de 1958 até 2017, com 22 classificações soberanas distintas.

	País	Ano	Nota
	AD	2016	9
	AO	2017	15
	...	...	...
	VE	2011	14
	ZM	2015	15
Total	137	38	22

Tabela 2: Descrição geral dos dados de *ratings*. Fonte: Elaborada pelo autor.

3.4 Seleção de atributos

A base de dados bruta possui uma grande quantidade de atributos (do inglês, *features*) que não necessariamente representam bem o problema a ser resolvido. Esses atributos redundantes podem afetar negativamente a robustez do modelo diminuindo assim sua acurácia, e causa o aumento do custo computacional necessário para realização do experimento. Portanto, se faz necessário a aplicação de uma técnica de seleção dos atributos que tem mais correlação com a estimativa do *rating* soberano. A dimensão da base antes e depois da aplicação do PCA pode ser observado na Figura 7.

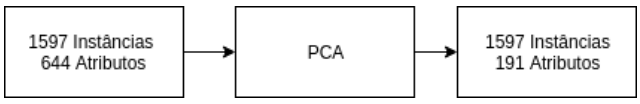

Figura 7: Extração de atributos com o PCA. Fonte: Elaborada pelo autor.

3.5 Divisão da base de dados

Tendo a base de dados resultante dos processos de pré-processamento e seleção de atributos, têm-se que dividir essa base em conjuntos de treinamento, conjunto de teste e conjunto de validação. Para esse trabalho, utilizou-se as seguintes proporções: 80% dos dados para fase treinamento e 20% para fase testes.

3.6 Treinamento

Após a aplicação de técnicas de pré-processamento para adequação dos dados para aplicação de métodos de ML, e a aplicação do PCA para redução de dimensão dos atributos, a base original com dimensão de 3489 instâncias, com cada instância contendo 644 atributos, diminuiu para uma base com 3489 instâncias, com cada instância contendo 294 atributos. Essa base resultante está pronta para aplicação de qualquer método de ML supervisionada. Como o método RF é parametrizado, e as escolhas desses parâmetros afetam fortemente a acurácia de predição, foi necessário realizar uma bateria de testes empíricos para realizar a escolha dos

parâmetros que resultam na maior acurácia possível. Para esse experimento, os parâmetros ótimos para o algoritmo RF são mostrados na Tabela 3.

bagSizePercent	batchSize	maxDepth	numFeatures	numIterations
100	100	Máx	Máx	150

Tabela 3: Parâmetros que obtiveram melhor acurácia. Fonte: Elaborada pelo autor.

Onde *bagSizePercent* é o tamanho de cada saco, como porcentagem do tamanho do conjunto de treinamento, *batchSize* é o tamanho do lote desejado para previsão em lote, *maxDepth* a profundidade máxima da árvore, e *numFeatures* é o número de recursos usados na seleção aleatória.

3.7 Agrupamento de classes

Além da redução de dimensão proporcionada pelo uso da análise de componentes principais, também foi feito um estudo aplicando a técnica de agrupamento (do inglês, *clustering*) nas classes do problema. Com isso, espera-se a diminuição do número de classes através do agrupamento de classes ou *ratings* semelhantes nos mesmos *clusteres* e um eventual aumento de acurácia. Na Figura 8 é possível observar onde esse processo entra, no fluxograma da metodologia já exposta.

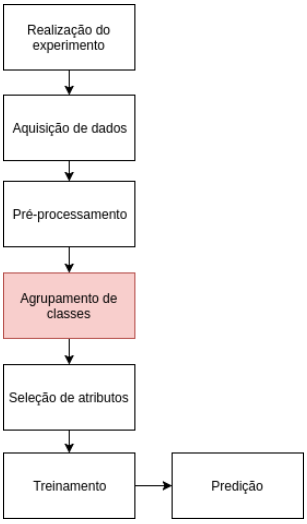

Figura 8: Diagrama de fluxo da metodologia com agrupamento de classes. Fonte: Elaborada pelo autor.

3.8 Análise de Regressão

O fluxograma geral visto na Figura 5 ainda é válido para realizar a análise de regressão e testar se as hipóteses levantadas são verdadeiras ou não. Porém, a etapa de seleção de atributos teve que ser alterada. Isso deve-se a natureza do método de análise de componentes principais,

que constrói novos atributos, transformados a partir dos originais, que deixam de existir impedindo eventual análise do impacto destes atributos no modelo. No seu lugar foi utilizado o método CFS para fazer a redução da quantidade de atributos da base.

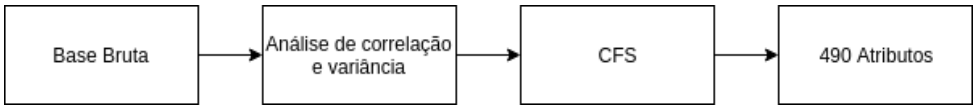

Figura 9: Diagrama de fluxo da metodologia. Fonte: Elaborada pelo autor.

Como visto na Figura 9 antes da aplicação do CFS foi realizada uma análise de correlação e variância manual, isso foi motivado pelo fato da base ter muitos atributos com muita falta de dados, e se simplesmente aplicássemos imputação geraríamos uma variância muito baixa, não sendo útil para o modelo.

Regressão sem variáveis binárias

A regressão sem variáveis binárias, esquematizada na Figura 10, será feita para ter um comparativo em relação ao poder preditivo do modelo, ou seja, se de fato a variável dependente (*rating soberano*) pode ser explicada pelas variáveis regressoras (indicadores de desenvolvimento do banco mundial).

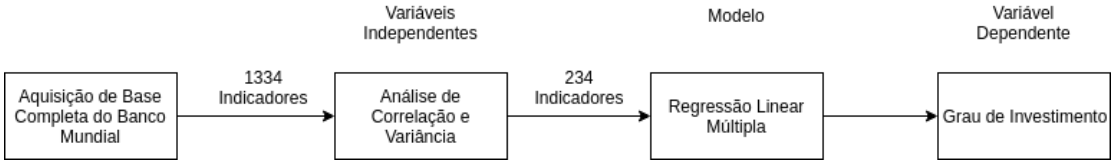

Figura 10: Fluxo da metodologia para regressão sem variáveis binárias. Fonte: Elaborada pelo autor.

O modelo de regressão múltipla linear utilizado é apresentado a seguir:

$$Y_1 = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + \varepsilon_1$$

Onde Y_1 é a variável dependente, ou seja, a classificação soberana, $X_{2i}, X_{3i}, ..., X_{ki}$ são os valores das variáveis independentes (ou explicativas), ε_1 é o erro de cada observação e $\beta_1, \beta_2, \beta_3, ..., \beta_k$ são os parâmetros ou coeficientes de regressão linear com $k = 1, 2, 3, ..., 420$.

Regressão com variável binária período

A regressão com variável binária período, esquematizada na Figura 11, terá como objetivo verificar se a hipótese de que houve uma ruptura estrutural na forma que as agências de risco avaliam os países ocasionada pela crise financeira americana de 2008. Para isso, conforme consta

na literatura, será adicionado ao modelo uma variável *dummy* (variável binária) responsável por dividir o conjunto de dados em dois períodos distintos: anterior a crise financeira de 2008 e posterior a crise financeira de 2008. Finalmente, será analisada a significância estatística dessa variável para o modelo, se a variável for estatisticamente significante significa que há evidências de que a hipótese é verdadeira.

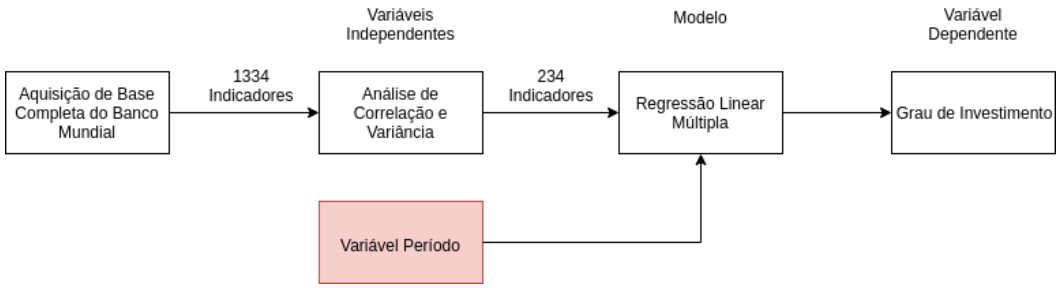

Figura 11: Fluxo da metodologia para regressão com variável binária período. Fonte: Elaborada pelo autor.

O modelo de regressão múltipla linear utilizado é apresentado a seguir:

$$Y_1 = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + z + az + \varepsilon_1$$

Onde Y_1 é a variável dependente, ou seja, a classificação soberana, $X_{2i}, X_{3i}, ..., X_{ki}$ são os valores das variáveis independentes (ou explicativas), ε_1 é o erro de cada observação e $\beta_1, \beta_2, \beta_3, ..., \beta_k$ são os parâmetros ou coeficientes de regressão linear com $k = 1, 2, 3, ..., 421$. A variável z é a variável binária introduzida e a é o ano.

Regressão com variável binária grau de desenvolvimento

A regressão com variável binária grau de desenvolvimento, esquematizada na Figura 12, terá como objetivo verificar se a hipótese de que existe uma avaliação diferente por parte das agências de risco para países com diferentes graus de desenvolvimento econômico usado a classificação de desenvolvimento econômica provinda do Departamento de Assuntos Econômicos e Sociais das Nações Unidas (UN/DESA). Para isso, novamente, será adicionada ao modelo uma variável *dummy* (variável binária) responsável por dividir o conjunto de dados conforme as classificações para grau de desenvolvimento do WESP. Finalmente, será analisada a significância estatística dessa variável para o modelo, se a variável for estatisticamente significante significa que há evidências de que a hipótese é verdadeira.

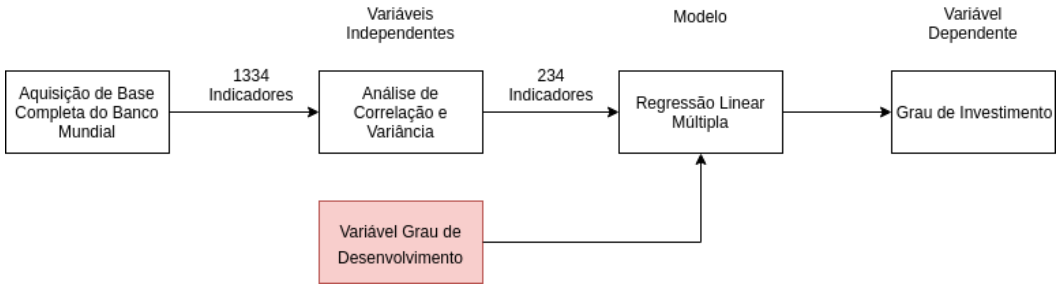

Figura 12: Fluxo da metodologia para regressão com variável binária grau de desenvolvimento. Fonte: Elaborada pelo autor.

O modelo de regressão múltipla linear utilizado é apresentado a seguir:

$$Y_1 = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + g + \varepsilon_1$$

Onde Y_1 é a variável dependente, ou seja, a classificação soberana, $X_{2i}, X_{3i}, ..., X_{ki}$ são os valores das variáveis independentes (ou explicativas), ε_1 é o erro de cada observação e $\beta_1, \beta_2, \beta_3, ..., \beta_k$ são os parâmetros ou coeficientes de regressão linear com $k = 1, 2, 3, ..., 422$. A variável g é a variável binária introduzida

4 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Os resultados obtidos na execução da metodologia apresentada neste trabalho são apresentados nesta seção sobre a forma de tabelas e gráficos com métricas de verificação de acurácias, sendo considerado dois cenários. No primeiro, centrado em predição de risco soberano são apresentado os resultados para dois experimentos, o primeira com aplicação da técnica de análise de componentes principais, a segunda com agrupamento das classes, por meio de técnicas de clusterização. No segundo, a análise de regressão visando verificar a veracidade das hipóteses levantadas.

4.1 Predição de risco soberano

O classificador construído a partir de 1597 instâncias, contendo 191 atributos artificiais gerados pelo algoritmo PCA obteve acurácia na predição de *rating* soberano de 78,52 % com o uso de validação cruzada com k = 10. Observa-se a dificuldade em encontrar uma relação direta entre os indicadores macroeconômicos e os *ratings*, característica intrínseca do problema [16]. As demais métricas são mostradas na Tabela 4.

KAPPA	MAE	RMSE	RAE	RRSE
0,7705	0,033	0,1191	38,545 %	57,7464 %

Tabela 4: Sumário da validação cruzada classificação sem agrupamento. Fonte: Elaborada pelo autor.

4.1.1 Seleção de atributos com agrupamento de classes

Tentando contornar esse problema foi realizado a diminuição da quantidade de classes, sendo em dois cenários distintos. No primeiro, foi realizado uma fusão de classes de forma empírica, fundamentada em uma análise manual de semelhança entre as classes. No segundo, visando fazer essa fusão de forma mais formal, foi adotado o uso de uma algoritmo de clusterização, motivando-se pelo que instâncias semelhantes tenderão a ser agrupadas no mesmo cluster.

4.1.2 Seleção manual

No cenário de redução de classes por semelhança entre classes, procedimento empírico, resultando na diminuição de 24 classes para apenas 4. O classificador conseguiu obter a acurácia de 90,91 % na predição de (*ratings*). As demais métricas do modelo são mostradas nas Tabelas a seguir.

KAPPA	MAE	RMSE	RAE	RRSE
0,8654	0,0838	0,1945	24,795 %	47,3009 %

Tabela 5: Sumário da validação cruzada seleção manual. Fonte: Elaborada pelo autor.

A Tabela 5 apresenta o coeficiente de concordância (KAPPA), o erro médio absoluto (MAE), a raiz do valor quadrático médio (RMSE), o erro absoluto relativo (RAE) e raiz do erro quadrático relativo (RRSE) que são diferentes métricas de erros e com diferentes interpretações, mas que no geral apresentam bons valores, compatıveis com a acurácia de 90,91%.

	TP	FP	PPV	TPR	F1	ROC	Classe
	0,943	0,027	0,953	0,943	0,948	0,989	A
	0,869	0,042	0,869	0,869	0,869	0,971	B
	0,923	0,054	0,903	0,923	0,913	0,976	C
	0,651	0,007	0,75	0,651	0,697	0,935	D
$\bar{x}$	0,909	0,04	0,909	0,909	0,909	0,978	

Tabela 6: Detalhamento da acurácia por classe seleção manual. Fonte: Elaborada pelo autor.

Na Tabela 7 encontra-se a matriz de confusão para o modelo com seleção manual, é possível observar que a quantidade de erros aumenta para clusters com classificações mais baixas.

a	b	c	d	classe
1229	53	21	0	a = A
47	741	65	0	b = B
13	58	1133	23	c = C
0	1	36	69	d = D

Tabela 7: Matriz de confusão para seleção manual. Fonte: Elaborada pelo autor.

As métricas visualizadas acima mostram que o classificador apresenta um bom desempenho quando reduzidas a quantidade de classes. Porém, o impacto dessa ação deverá ser discutida posteriormente.

4.1.3 Seleção por clusterização

No cenário de redução de classes através da aplicação do algoritmo de clusterização *SimpleKMeans*, resultando na diminuição de 24 classes para apenas 4. Com isso houve um aumento na acurácia do classificador para 98,28 %. As demais métricas do modelo são mostradas nas Tabelas a seguir.

KAPPA	MAE	RMSE	RAE	RRSE
0,9764	0,0309	0,1014	8,47 %	23,7712 %

Tabela 8: Sumário da validação cruzada seleção por clusterização. Fonte: Elaborada pelo autor.

A Tabela 8 apresenta as métricas de erro para o modelo com seleção por clusterização, todas as métricas apontam valores melhores do que no modelo com seleção manual, indicando assim que as notas de classificação estão melhores agrupadas é, por conseguinte, melhores preditas.

	TP	FP	PPV	TPR	F1	ROC	Classe
	0,995	0,003	0,99	0,995	0,992	1	cluster1
	0,975	0,006	0,968	0,975	0,971	0,997	cluster2
	0,981	0,005	0,987	0,981	0,984	0,998	cluster3
	0,98	0,01	0,982	0,98	0,981	0,998	cluster4
$\bar{x}$	0,983	0,006	0,983	0,983	0,983	0,998	

Tabela 9: Detalhamento da acurácia por classe seleção por clusterização. Fonte: Elaborada pelo autor.

Na Tabela 10 encontra-se a matriz de confusão para o modelo com seleção por clusterização, é possıvel observar que a quantidade de erros aumenta para clusters com classificações mais baixas.

a	b	c	d	classe
760	0	0	4	a = cluster1
0	509	5	8	b = cluster2
0	9	978	10	c = cluster3
8	8	8	1182	d = cluster4

Tabela 10: Matriz de confusão para seleção por clusterização. Fonte: Elaborada pelo autor.

Analisando os dois cenários é possível observar que a aplicação de técnicas de clusterização nas classes, resulta no aumento de acurácia na classificação, de no mínimo 12,36 %. No entanto, podemos observar a perda de sensibilidade.

4.2 Análise de Regressão

4.2.1 Regressão sem variáveis binárias

As métricas R^2 e R^2 ajustado são 0,914 e 0,912, respectivamente. O que indica que aproximadamente cerca de 91% da variável dependente pode ser explicada pelos regressores presentes no modelo. Esses resultados são sumarizados na tabela a seguir.

R^2	R^2 **ajustado**	**Teste F**
$0,914$	0,912	334,3

Tabela 11: Métricas qualitativas do modelo de regressão. Fonte: Elaborada pelo autor.

4.2.2 Regressão com variável binária período

As métricas R^2 e R^2 ajustado são 0,915 e 0,913, respectivamente. O que indica que aproximadamente cerca de 91% da variável dependente pode ser explicada pelos regressores presentes no modelo. Houve um aumento de 0,001 no poder explicativo do modelo. Esses resultados são sumarizados na tabela a seguir.

R^2	R^2 **ajustado**	**Teste F**
$0,915$	0,913	338,7

Tabela 12: Métricas qualitativas do modelo de regressão. Fonte: Elaborada pelo autor.

A métrica valor-p mostra que a variável período é estatisticamente significante (valor-p $< 0,05$). Portanto, há evidências estatísticas de que a avaliação do *rating* soberano antes e depois da crise de 2008 são diferentes. Portanto, a hipótese é verdadeira. A Tabela 13 sumariza as métricas da regressão efetuada, porém com termos omitidos para fins de visualização.

X_i	β	σ	**t**	**P-valor**
X_1	0,0131	0,001	10,252	0,000
X_2	0,0012	0,006	0,219	0,827
X_3	-0,0250	0,011	-2,261	0,024
X_4	-0,0157	0,008	-2,029	0,042
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
X_{490}	0,0017	0,008	0,222	0,825
período	0,0384	0,011	3,611	0,000

Tabela 13: Coeficientes estimados e métricas para verificação de hipóteses. Fonte: Elaborada pelo autor.

Esse resultado já era esperado, e coerente com outros resultados na literatura [39, 40, 41] tendo em vista a incapacidade das agências de risco em prever a crise financeira do setor imobiliário americano. Portanto, sugere-se que a metodologia usada pelas agências de risco foi ajustada após a deflagração da crise.

4.2.3 Regressão com variável binária grau de desenvolvimento

As métricas R^2 e R^2 ajustado são 0,916 e 0,913, respectivamente. O que indica que aproximadamente cerca de 91% da variável dependente pode ser explicada pelos regressores presentes no modelo. Houve um aumento 0,001 no poder explicativo do modelo. Esses resultados são sumarizados na tabela a seguir.

R^2	R^2 **ajustado**	**Teste F**
$0,915$	0,913	338,7

Tabela 14: Métricas qualitativas do modelo de regressão. Fonte: Elaborada pelo autor.

A métrica valor-p mostra que a categoria C (Economias em desenvolvimento) não é estatisticamente significante. Porém, as categorias B (Economias em transição) e A (Economias desenvolvidas) são estatisticamente significantes. Logo, há evidências estatísticas de que a avaliação do *rating* soberano para países com economia em desenvolvimento são semelhantes. No entanto, para economias desenvolvidas e economias em transição o mesmo não se verifica. A Tabela 15 sumariza as métricas da regressão efetuada, porém com termos omitidos para fins de visualização.

X_i	β	σ	**t**	**valor P**
X_1	0,0131	0,001	10,252	0,000
X_2	0,0012	0,006	0,219	0,827
X_3	-0,0250	0,011	-2,261	0,024
X_4	-0,0157	0,008	-2,029	0,042
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
X_{490}	0,0008	0,008	0,101	0,919
A	-0,0236	0,006	-3,871	0,000
B	0,0088	0,002	5,028	0,000
C	-0,0093	0,006	-1,474	0,140

Tabela 15: Coeficientes estimados e métricas para verificação de hipóteses. Fonte: Elaborada pelo autor.

Os resultados sugerem que as agências de risco avaliam países com economias em transição e economias desenvolvidas de formas diferentes. No entanto será necessário fazer algumas observações sobre esses resultados.

No caso de economias em transição, é necessário ressaltar que a maioria desses países são da organização Comunidade dos Estados Independentes, grupo que envolve 11 países que pertenciam à antiga União Soviética. A maioria desses países possuem contextos próprios que podem invalidar esses resultados. Por exemplo, a Rússia está sofrendo sanções por parte do

ocidente; a Ucrânia passou por um princípio de guerra civil, que levou a separação da Crimeia, ou seja, fatores geopolíticos que podem ser a causa do tratamento diferenciado por parte das agências de risco.

No caso de economias desenvolvidas, uma possível explicação para esse tratamento diferenciado por parte das agências de risco também é a crise financeira de 2008, que forçou as agências de risco a mudaram sua metodologia, mais precisamente, conforme [39] aumentarem os pesos de entradas quantitativas nos modelos. Essas adequações levaram a uma diminuição das notas de economias desenvolvidas. Isso pode ser verificado também através do dados apresentados na Tabela 16, a grande diferença entre a média e o desvio padrão indica volatilidade para economias desenvolvidas.

A	**Soma**	**média**	**std**
0	2086	10,52	4,514
1	1403	6,374	4,431

Tabela 16: Métricas estatísticas para países desenvolvidos (categoria A). Fonte: Elaborada pelo autor.

5 CONCLUSÕES E TRABALHOS FUTUROS

A nota de classificação soberana é um indicador que busca expressar o risco ao que se submetem os investidores estrangeiros ao adquirir títulos de algum país, sendo emitidos por agências de *rating*, empresas independentes de governos ou empresas privadas. A predição da nota de classificação soberana é importante para diversos tipos de análises econômicas. Nesse trabalho, propôs-se um modelo para predição dos *ratings*, utilizando-se o algoritmo de classificação *Random Forest*.

A eficiência do modelo proposto foi comprovada através dos testes de predição, por meio de métricas, conforme descrito na literatura. Os resultados obtidos mostram que o método se configura como uma abordagem possível para predição de classificações soberanas. Esse *rating* previsto pode ser usado principalmente como uma caixa preta (do inglês, *Black Box*), já que o uso do PCA torna uma análise posterior a predição inviável, pelo fato do mesmo criar variáveis a partir das originais, que deixam de ser consideradas.

O modelo de regressão sugere uma ruptura estrutural na forma com que as agências de risco avaliam os países motivada pela incapacidade das agências de preverem a crise financeira de 2008, resultando em uma mudança de metodologia por parte das agências, e uma diminuição das notas de países com economias desenvolvidas. No entanto, a análise para economias em transição não foi conclusiva, em função das particularidades geopolíticas de muitos dos países desse grupo, que podem ser a causa da avaliação diferenciada pelas agências de risco.

Apesar dos resultados obtidos, este trabalho possui limitações que motivam trabalhos futuros. Uma delas é a grande quantidade de atributos, mesmo com a utilização de inúmeras técnicas de redução de dimensionalidade, a quantidade de atributos ainda é muito grande, principalmente no contexto da análise de regressão, essa quantidade torna extremamente massante manipulá-los e fazer análises sobre esses atributos. A segunda seria a relativa baixa acurácia no contexto de predição, apesar de ser compatível com resultados encontrados na literatura, novas investigações podem ser feitas visando melhorá-la.

Dito isso, algumas propostas de trabalhos futuros são:

1. Reduzir ainda mais a dimensão da base de dados, sem perder informações.
2. Automatizar a geração de novas bases de dados, eliminando o trabalho manual de aquisição e pré-processamento de dados.
3. Realizar novas análises sobre a base de dados, por exemplo usar outras classificações de desenvolvimento de economias.

REFERÊNCIAS

[1] LOESCH, Claudio. **Métodos estatísticos multivariados - 1ª** Edição. Saraiva, 2012.

[2] RUSSEL, S.; NORVIG, P.; **Artificial Inteligence: A Modern Approach**. Prentice-Hall, Second Edition, 2003.

[3] HO, Tin Kam. Random decision forests. **In Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1)**,Volume 1, ICDAR '95, pages 278–, Washington, DC, USA, 1995. IEEE Computer Society.

[4] CANUTO, O; SANTOS, P.F. dos. Risco-soberano e prêmios de risco em economias emergentes.**Ministério da Fazenda, Secretária de Assuntos Internacionais, Temas de economia internacional**, Jan, 2003.

[5] CARVALHO, André Carlos Ponce de Leon de, FACELI, Katti, LORENA, Ana Carolina, GAMA, João. **Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina**. LTC, 08/2011.

[6] BREIMAN, Leo. **Random forests**. Machine Learning, Statistics Departament, University of California, Berkeley, CA, USA, Jan. 2001

[7] HAYKIN S. **Neural Networks and Learning Machines**. Prentice-Hall, Nova Jersey, terceira edição, 2009.

[8] BARANAUSKAS, J. A. e MONARD, M. C. Reviewing some **Machine Learning Concepts and Methods**. Relatório Técnico 102, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2000. Disponível em: <ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/rel_tec/RT_102.ps.zip>.

[9] MITCHELL, T. M. **Machine learning**. New York: McGraw-Hill. 1997.

[10] SMITH, Adman. **A riqueza das nações: investigação sobre sua natureza e suas causas** . São Paulo: Abril Cultural, 1983.

[11] STANDARD & POOR'S,**Corporate Ratings Criteria**, 3 p. 2002, Disponível em: <http://www.standardandpoors.com>. Acesso em: 5 de Maio de 2017

[12] AFONSO, ANTÓNIO. **Understanding the Determinants of Government Debt Ratings: Evidence for the Two Leading Agencies**. CISEP working paper, Fev 2002. Disponível em: <http://pascal.iseg.utl.pt> Acesso em: 5 de Maio de 2017.

[13] Standard & Poor's (2008). **Critérios - comentários: ratings de crédito soberano: principais conceitos** . Standard & Poor's. 29 de Maio de 2008, Disponível em: <http://www.standardandpoors.com>. Acesso em: 5 Maio 2017.

[14] Moody's Investors Service (2009). **Moody's rating symbols & definitions**. Moody's Investors Service. Disponível em: <http://www.moodys.com> Acesso em: 5 set 2017.

[15] Moody's Investors Service (2016). **Moody's rating symbols & definitions** . Moody's Investors Service. Disponível em: <http://www.moodys.com>. Acesso em: 5 set 2017.

[16] FRASCAROLI, B. F.; SILVA, L. C. ; SILVA FILHO, O. C. **OS RATINGS DE RISCO SOBERANO E OS FUNDAMENTOS MACROECONÔMICOS DOS PAÍSES: UM ESTUDO UTILIZANDO REDES NEURAIS ARTIFICIAIS**. Revista Brasileira de Finanças (Impresso), v. 7, p. 73-106, 2009.

[17] BRINK, H.; RICHARDS, J. **Real-Word Machine Learning**. Manning Publications Co., Shelter Island, 2017.

[18] HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques**. 3ed. San Francisco: Morgan Kaufmann Publishers, 2011.

[19] SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding Machine Learning: From Theory to Algorithms**. Cambridge University Press, 2014.

[20] BATISTA, G. E. A. P. A. **Pré-processamento de Dados em Aprendizado de Máquina Supervisionado**. ICMC-USP, 2003.

[21] LEE, H. D.; MONARD, M. C. **Applying Knowledge-Driven Constructive Induction: Some Experimental Results**. Technical Report 101, Department of Computer Science. University of São Paulo, São Carlos, SP, 2000.

[22] BREIMAN, Leo. Bagging predictors. **Machine Learning**, Statistics Departament, University of California, Berkeley, CA, USA. 2001

[23] BIAU, Gérard. **Analysis of a Random Forests Model**. LSTA & LPMA, Universite Pierre et Marie Curie, Paris, 2012.

[24] ORAL, M.; KETTANA, O.; COSSET, J.C; DAOUAS, M. **An estimation model for country risk rating**. International Journal of Forecasting, 1992

[25] YIM, J.; MITCHELL, H. **Comparison of country risk models: hybrid neural networks, logit models, discriminant analysis and cluster techniques** . School of Economics and Finance, RMIT University, Melbourne, Australia, 2005.

[26] PEARSON, K. **On Lines and Planes of Closest Fit to Systems of Points in Space** . Philosophical Magazine. 2(11), 1901.

[27] SMITH, L. I. **A tutorial on Principal Components Analysis** . Disponível em: <http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf>. Acesso em: 5 set 2017.

[28] BENGIO, Y; et al. **Representation Learning: A Review and New Perspectives Pattern Analysis and Machine Intelligence** . 35 (8): 1798–1828. 2013

[29] **Principal Components Analysis** . WIKIPEDIA. Disponível em: <https://en.wikipedia.org/wiki/Principal_component_analysis>. Acesso em: 12 jun 2017.

[30] WUERGS, A. F. E; BORDA. J. A. **Redes Neurais, Lógica Nebulosa e Algoritmos Genéticos: Aplicações e Possibilidade em Finanças e Contabilidade** . Revista de Gestão da Tecnologia e Sistemas de Informação. Vol. 7. No. 1, p.163-182, 2010.

[31] CANTOR, Richard; PACKER, Frank. **Determinants and Impact of Sovereign Credit Ratings**. New York: Federal Reserve, 1996.

[32] JOHNSON, R.A. & WICHERN, D.W.**Applied multivariate statistical analysis**. New Jersey: Prentice Hall, 2002.

[33] I.T. JOLLIFFE, I.T.**Principal Component Analysis**. New Jersey: Springer, 2002.

[34] **WDI HIGHLIGHTS** Disponível em: <http://databank.worldbank.org/data/download/site-content/wdi-2016-highlights-featuring-sdgs-booklet.pdf>. Acesso em: 13 jun 2017.

[35] KAUFMANN, M. Correlation-based feature selection for discrete and numeric class machine learning. In Proc. of the 17th Int. Conf. on Machine Learning, San Francisco, CA, pp. 359–366. 2, 11.

[36] **Portal Action**. Disponível em: <http://www.portalaction.com.br/analise-de-regressao/22-estimacao-dos-parametros-do-modelo>. Acesso em: 13 jun 2017.

[37] MORAIS, N. F. **Análise de regressão linear com estudo de caso em acidentes de trânsito**. Monografia de TCC. Universidade Estadual da Paraíba: Campina Grande-PB, 2010.

[38] PAGOT, R, JARDIM, E. B. **Os BRICs frente aos estados unidos após a crise financeira de 2008: alternativa a uma hegemonia declinante?** Textos de Economia, Florianópolis, v.17, n.2, p.128-150, jul./dez.2014

[39] AMSTAD, M., PACKER, F. **Sovereign ratings of advanced and emerging economies after the crisis**. BIS Quarterly Review, December, 2015

[40] BASU, K., DE, S., RATHA, D., TIMMER, H. **Sovereign Ratings in the Post-Crisis World: An Analysis of Actual, Shadow and Relative Risk Ratings.** Outubro 2013.

[41] REUSENSA, P., CROUXA, C., **Sovereign credit rating determinants: the impact of the European debt crisis**,Faculty of Economics and Business, KU Leuven, Belgium.

[42] DEL GROSSI, A. A., **COMPARAÇÃO E AVALIAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA INDICAÇÃO DE BIÓPSIA PARA O CÂNCER DE PRÓSTATA**, Universidade Estatual de Londrina, Londrina, 2013.

ANEXO A – Tabela com grupos de países resultante de clusterização

Os grupos resultantes da clusterização podem ser visualizados na Tabela 17. Os grupos estão rotulados em ordem alfabética, portanto espera-se que países com maiores notas estejam em clusters mais próximos de A e países com as menores notas estejam em clusters mais próximos de B.

Grupo	Países
A	Austria, Australia, Belgium, Canada, Switzerland, Germany, Denmark, Spain, Finland, France, United Kingdom, Ireland, Italy, Japan, Korea, Rep., Luxembourg, Netherlands, Sweden, United States
B	Albania, Bosnia and Herzegovina, Brazil, Grenada, Greece, Lebanon, Philippines, Romania, Serbia, El Salvador, Venezuela
C	Armenia, Angola, Argentina, Barbados, Bangladesh, Burkina Faso, Bulgaria, Benin, Bolivia, Bahamas, The, Botswana, Belarus, Belize, Congo, Dem. Rep., Cote dIvoire, Chile, Cameroon, Costa Rica, Cabo Verde, Dominican Republic, Ecuador, Egypt, Arab Rep., Ethiopia, Fiji, Gabon, Ghana, Gambia, The, Guatemala, Honduras, Indonesia, India, Iraq, Jamaica, Jordan, Kenya, Cambodia, Kazakhstan, Lesotho, Morocco, Moldova, Montenegro, Macedonia, FYR, Mali, Mongolia, Malawi, Mozambique, Nigeria, Nicaragua, Peru, Papua New Guinea, Pakistan, Paraguay, Russian Federation, Senegal, Turkmenistan, Tunisia, Trinidad and Tobago, Ukraine, Uganda, Uruguay, Vietnam, Zambia
D	Andorra, United Arab Emirates, Azerbaijan, Bahrain, Chile, China, Colombia, Cyprus, Czech Republic, Estonia, Georgia, Hong Kong SAR, China, Croatia, Hungary, Israel, Iran, Islamic Rep., Iceland, Kuwait, Liechtenstein, Sri Lanka, Lithuania, Latvia, Libya, Malta, Mexico, Malaysia, New Caledonia, Norway, Oman, Panama, Poland, Qatar, Rwanda, Saudi Arabia, Seychelles, Singapore, Slovenia, Slovak Republic, San Marino, Suriname, Thailand, Turkey, South Africa

Tabela 17: Agrupamento de países feito pelo algoritmo de clusterização