UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO
ENGENHARIA DE PRODUÇÃO MECÂNICA

JOÃO VICTOR SOARES DE CARVALHO

DESENVOLVIMENTO DE UM MODELO DE PREDIÇÃO DE DESLIGAMENTO VOLUNTÁRIO DE COLABORADORES: UMA ABORDAGEM MULTIMODAL

JOÃO PESSOA
2024

JOÃO VICTOR SOARES DE CARVALHO

DESENVOLVIMENTO DE UM MODELO DE PREDIÇÃO DE DESLIGAMENTO VOLUNTÁRIO DE COLABORADORES: UMA ABORDAGEM MULTIMODAL

Trabalho de Conclusão de Curso apresentado à Banca Examinadora do Curso de Graduação em Engenharia de Produção Mecânica da Universidade Federal da Paraíba como requisito parcial para a obtenção do grau de bacharel em Engenharia de Produção Mecânica.

Orientador: Profª. Dra. Renata de Oliveira Mota

JOÃO PESSOA
2024

Catalogação na publicação
Seção de Catalogação e Classificação

C331d Carvalho, Joao Victor Soares de.
Desenvolvimento de um modelo de predição de desligamento voluntário de colaboradores: uma abordagem multimodal / Joao Victor Soares de Carvalho. - João Pessoa, 2024.
62 f. : il.

Orientação: Renata de Oliveira Mota.
TCC (Graduação) - UFPB/CT.

1. Predição. 2. Análise de Risco. 3. Gestão de Recursos Humanos. 4. Machine Learning. I. Mota, Renata de Oliveira. II. Título.

UFPB/BSCT CDU 658.5(043.2)

Elaborado por ROSANGELA GONCALVES PALMEIRA - CRB-216

JOÃO VICTOR SOARES DE CARVALHO

DESENVOLVIMENTO DE UM MODELO DE PREDIÇÃO DE DESLIGAMENTO VOLUNTÁRIO DE COLABORADORES: UMA ABORDAGEM MULTIMODAL

Trabalho de Conclusão de Curso apresentado à Banca Examinadora do Curso de Graduação em Engenharia de Produção Mecânica da Universidade Federal da Paraíba como requisito parcial para a obtenção do grau de bacharel em Engenharia de Produção Mecânica.

Orientador: Profª. Dra. Renata de Oliveira Mota

BANCA EXAMINADORA:

Dra. Renata de Oliveira Mota
Universidade Federal da Paraíba

Membros:

Mª. Alessandra Berenguer de Moraes
Universidade Federal da Paraíba

Dr. Paulo Rotella Junior
Universidade Federal da Paraíba

João Pessoa, 20 de Dezembro de 2024

JOÃO VICTOR SOARES DE CARVALHO

DESENVOLVIMENTO DE UM MODELO DE PREDIÇÃO DE DESLIGAMENTO VOLUNTÁRIO DE COLABORADORES: UMA ABORDAGEM MULTIMODAL

Trabalho de Conclusão de Curso submetido à **Coordenação de Graduação do Curso de Engenharia de Produção Mecânica** da UFPB, apresentado em sessão de defasa pública realizada em 18/12/2024, obtendo o conceito APROVADO, nota 9.5, sob avaliação da banca examinadora a seguir:

gov.br Documento assinado digitalmente
RENATA DE OLIVEIRA MOTA
Data: 30/12/2024 22:06:05-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Renata de Oliveira Mota - Orientadora - DEP/CT/UFPB

gov.br Documento assinado digitalmente
ALESSANDRA BERENGUER DE MORAES
Data: 30/12/2024 22:09:11-0300
Verifique em https://validar.iti.gov.br

Profª. Me. Alessandra Berenguer de Moraes - Membro - DEP/CT/UFPB

Profº. Dr. Paulo Rotella Junior - Membro - DEP/CT/UFPB

João Pessoa (PB)

DEZEMBRO/2024

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

Coordenação de Graduação em ENGENHARIA DE PRODUÇÃO MECÂNICA

TRABALHO DE CONCLUSÃO DE CURSO - TCC

Prezado(a) Professor (a),

Pedimos a gentileza preencher este formulário com a sua avaliação sobre o TCC que lhe foi encaminhado. Preencha todos os seus campos e devolva-o à Coordenação de Curso de **Engenharia de Produção Mecânica**, responsável pela consolidação da ATIVIDADE **TCC** no SIGAA.

FORMULÁRIO DE AVALIAÇÃO DE TCC

Discente: João Victor Soares de Carvalho Matrícula: 20180040716

Título do trabalho: Desenvolvimento de um modelo de predição de desligamento voluntário de colaboradores: uma abordagem multimodal

Data da defesa: 18/12/2024 Local: Google Meet Início: 10:00h Término: 11:45h

AVALIAÇÃO QUANTITATIVA							
TRABALHO ESCRITO		Item	Atribua notas de 0 a 10 para os quesitos a seguir:	Máx.	Orient.	Aval. 1	Aval. 2
	Introdução	1	**Clareza e precisão na delimitação do tema e do problema**	2	2	2	1,7
		2	Clareza e coerência dos objetivos geral e específicos				
	Fundamentação teórica	3	**Robustez e sincronismo (escopo) da fundamentação teórica**	2	1,5	2	1,5
		4	Abrangência e assertividade da fundamentação teórica				
	Procedimentos metodológicos	5	**Delineamento e coerência dos procedimentos metodológicos**	2	2	1,7	2
		6	Detalhamento da coleta e do tratamento dos dados				
	Resultados	7	**Análise dos dados e correlação com a fundamentação teórica**	2	2	1,5	2
		8	Clareza e consistência dos resultados alcançados				
	Conclusão	9	**Clareza e sustentação das conclusões**	2	1,7	2	2
	&	10	Qualidade gramatical e ortográfica/formatação (ABNT 6023 e 14724) do texto				
	Geral	11	**Certeza da autoria do trabalho (originalidade)**				
	Nota da Avaliação do Trabalho Escrito			10	**9,2**	**9,2**	**9,2**
EXPOSIÇÃO/DEFESA ORAL		Item	Atribua notas de 0 a 10 para os quesitos a seguir:	Máx.	Orient.	Aval. 1	Aval. 2
	Apresentação	1	**Qualidade dos slides e sequência dos itens**	6	6	6	6
		2	Abrangência do trabalho na apresentação				
		3	**Segurança e domínio dos conteúdos apresentados**				
		4	Clareza e objetividade na exposição				
		5	**Coerência e Postura (equilíbrio e naturalidade)**				
		6	Apresentação no tempo determinado				
	Arguição	1	**Entendimento das perguntas**	4	4	4	4
		2	Segurança nas respostas				
		3	**Assertividade e qualidade das respostas**				
	Nota da Avaliação da Exposição/Defesa Oral			10	**10**	**10**	**10**

CONSOLIDAÇÃO DA AVALIAÇÃO QUANTITATIVA **Calcula-se as médias** de avaliação do trabalho escrito e da apresentação		
Média de Avaliação do **Trabalho Escrito (60%)**	Média de Avaliação da **Apresentação (40%)**	**Média Final de Avaliação do TCC**
9,2	10	9,5

Caso a média geral seja maior ou igual a nove (9), ou menor ou igual a cinco (5), **explicite** os **motivos** desta **avaliação** fora da média:

O trabalho apresenta uma contribuição relevante e inovadora para a área de Gestão de Pessoas e Engenharia de Produção ao aplicar técnicas avançadas de análise de dados em um tema de grande importância para organizações: a retenção de talentos. O uso de dados multimodais (quantitativos, qualitativos ou provenientes de diversas fontes) demonstra uma compreensão profunda da complexidade do problema, ampliando a precisão e aplicabilidade do modelo desenvolvido. O trabalho aborda um desafio real enfrentado por empresas, conectando teorias acadêmicas a soluções práticas, o que amplia sua aplicabilidade no mercado.

AVALIAÇÃO QUALITATIVA

Registre os Pontos Fortes e Oportunidade de Melhoria do texto analisado:

Pontos Fortes	Oportunidades de Melhoria
• A escolha das técnicas de modelagem e predição é fundamentada em literatura robusta e implementada de maneira rigorosa, demonstrando domínio técnico e analítico. • A apresentação de resultados claros, com métricas de validação que comprovam a eficiência do modelo, evidencia a qualidade do estudo. • Além de prever o desligamento, o trabalho oferece importantes implicações para a formulação de estratégias de retenção, ampliando seu impacto.	• Embora a abordagem multimodal seja um ponto forte, a análise pode ser enriquecida com uma discussão mais aprofundada sobre fatores qualitativos, como cultura organizacional, clima de trabalho e aspectos emocionais que influenciam a decisão de desligamento. • Poderia haver uma seção mais detalhada sobre as limitações do modelo, como viés nos dados, restrições nas fontes de dados ou desafios na aplicação em organizações com características muito diferentes.

Aponte as correções / melhorias a serem introduzidas no exemplar definitivo do TCC:

- Considerar o risco de prejuízos ao processo de contratação devido ao excesso de análises de dados.
- Incluir nas considerações finais uma reflexão sobre possíveis implicações éticas, como a discriminação gerada pelo uso inadequado ou enviesado dos dados.
- Explorar como esses métodos podem impactar os critérios de seleção em processos futuros, abordando possíveis mudanças na forma como decisões de contratação são tomadas.
- Detalhar os métodos de aprendizado supervisionado, trazendo exemplos práticos para cada categoria de aprendizado.
- Ampliar a explicação sobre o conceito e aplicações de machine learning no contexto do trabalho.
- Explicar melhor o que é o CRISP-DM, incluindo no referencial teórico uma definição clara e detalhada sobre sua aplicação e relevância.
- Garantir uma transição mais fluida entre os capítulos 2 e 3, contextualizando o uso do CRISP-DM no desenvolvimento da pesquisa.
- Comparar os métodos utilizados com outros disponíveis, destacando os benefícios específicos da abordagem adotada no estudo.

Banca Examinadora
Nome completo **e** matrícula **SIAPE** da equipe de avaliação
Profa. **Orientadora:** 1277271 – Renata de Oliveira Mota
Profa. **Avaliadora 1:** 1287839 - Alessandra Berenguer de Moraes
Prof. **Avaliador 2:** 23171983 **-** Paulo Rotella Junior

Dedico este trabalho a todos que ao longo da minha trajetória, me ofereceram apoio, confiança e incentivo. Àqueles que acreditaram em mim nos momentos de desafio, me impulsionaram nos momentos de dúvida e celebraram comigo cada conquista. O sucesso desse trabalho é, em grande parte, graças a vocês.

AGRADECIMENTOS

A todos que, de alguma forma, fizeram parte desta jornada. Em especial:

Minha orientadora Renata, pelo empenho, paciência e apoio em todos os momentos, fundamentais para o meu crescimento. Muito obrigado por tudo que fez e faz pela minha vida acadêmica.

Aos meus pais, Ladjane e Salmineor, pelo amor incondicional, pelo suporte constante e por todos os sacrifícios que fizeram ao longo de toda a minha jornada. Obrigado por sempre me lembrarem de onde eu vim, das minhas prioridades e das realizações que almejo. Sem vocês, nada disso seria possível, muito obrigado.

À minha família, que esteve sempre presente em todos os momentos. Pelo apoio, compreensão, por cada acolhida nos momentos difíceis e que tanto fizeram por mim. Cada gesto de carinho foi fundamental para a manutenção da minha motivação, muito obrigado.

Às minhas tias, Lindinalva e Lucimar, com todo o meu carinho e gratidão, em especial pelo apoio financeiro nos momentos em que as dificuldades financeiras quase me fizeram desistir do curso. O suporte de vocês foi fundamental para que eu seguisse em frente, e sou eternamente grato por tudo.

Minha amada namorada Nathalia, presente em todos os momentos da minha trajetória acadêmica, dividindo comigo o fardo desta jornada, minha fonte de força, paciência e inspiração. Sempre ao meu lado, no bom e no mau.

Aos professores do curso, pela importante contribuição na minha formação, monetária desta caminhada.

Sei que meu sucesso também é o sucesso de vocês.

À todos os meus mais sinceros agradecimentos.

“E eu posso mudar meu destino
Não seja cego jovem negro
Não sinta medo jovem negro
Sem desespero jovem negro
Você pode escolher seu caminho“
(Leall, 2021)

RESUMO

Este trabalho consiste no desenvolvimento de um modelo preditivo para a previsão do desligamento voluntário de colaboradores de uma plataforma digital que atua no setor de comércio eletrônico dentro do período de um ano. Utilizando a metodologia Processo Padrão Inter-Indústrias para Mineração de Dados, *(Cross-Industry Standard Process for Data Mining (CRISP-DM))*. A partir de uma análise aprofundada dos dados, foram utilizadas variáveis demográficas, organizacionais e de desempenho para treinar e testar diversos algoritmos de aprendizado de máquinas. O processo de modelagem incluiu a combinação de resultados de três modelos para definir níveis de risco, priorizando a redução de erros do tipo II, que são críticos nesse contexto. Os resultados obtidos evidenciam a eficácia dos modelos preditivos utilizados. Com *F-beta Scores* de 95,3%, 84,7% e 82,2% para *Random Forest*, *LightGBM* e *CatBoost*, respectivamente, os modelos foram capazes de estimar com precisão as probabilidades de desligamento, permitindo a criação de uma estratégia de risco integrada às decisões de remuneração e retenção. A abordagem adotada demonstrou grande valor ao permitir antecipar possíveis desligamentos e auxiliar a gestão de recursos humanos na implementação de ações mais assertivas e focadas na retenção de talentos. O estudo evidencia como a aplicação de técnicas de aprendizado de máquina pode agregar valor estratégico à gestão de pessoas e à tomada de decisões organizacionais.

Palavras-Chave: Predição; Análise de Risco; Gestão de Recursos Humanos, Machine Learning

ABSTRACT

This research developed a predictive model for forecasting voluntary employee turnover within one year in a digital platform operating in the e-commerce sector, within a one-year period, using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. Through a comprehensive analysis of the data, demographic, organizational, and performance variables were used to train and test several machine learning algorithms. The modeling process included combining the results of three models to define risk levels, prioritizing the reduction of type II errors, which are critical in this context. The results highlight the effectiveness of the predictive models used. With F-beta Scores of 95.3%, 84.7%, and 82.2% for Random Forest, LightGBM, and CatBoost, respectively, the models were able to accurately estimate turnover probabilities, enabling the creation of a risk strategy integrated with compensation and retention decisions. The approach proved valuable for anticipating turnover and assisting HR management in implementing more targeted and effective retention actions. The study demonstrates how the application of machine learning techniques can add strategic value to people management and organizational decision-making.

Keywords: Prediction; Risk Analysis; Human Resources Management, Machine Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A evolução histórica dos estudos da rotatividade de colaboradores revela o crescente interesse no tema e a busca por entender suas causas, consequências e características em contextos diversos. Inicialmente tratada como um fenômeno individual, a pesquisa evoluiu para abordar a rotatividade coletiva em níveis de grupo, unidade e organização. Essa mudança de perspectiva permitiu uma compreensão mais abrangente dos fatores que influenciam a saída dos colaboradores das organizações(Akiba et al., 2019).

Ferreira e Freire (2001), em seu trabalho sobre a rotatividade na função de frentista, aponta que o fenômeno da rotatividade apresenta-se como uma preocupação relevante para as empresas em um contexto de competição global. Nesse ambiente, a necessidade de competir implica na oferta de produtos e serviços de qualidade, evidenciando a importância de uma política de gestão de recursos humanos que promova a retenção de talentos dentro da organização. Isso possibilita que os profissionais exerçam suas funções com eficiência e eficácia.

Moreira e Nantes (2024) destaca que a administração de recursos humanos é crucial para diminuir a rotatividade de funcionários dentro de uma organização. Quando os empregados estão motivados e satisfeitos, eles tendem a se comprometer mais com a empresa e têm menos chances de deixar suas funções, o que ajuda a manter uma força de trabalho estável e produtiva, contribuindo assim para a redução da rotatividade.

Segundo Liu, Kwong e Mohammadi (2024), a intenção rotatividade de colaboradores diminui a eficácia operacional de um negócio, aumentando os custos e reduzindo a lucratividade. Além disso, pode ter implicações generalizadas, frequentemente causando ocorrências de rotatividade coletiva entre os funcionários. Liu, Kwong e Mohammadi (2024) ainda aponta que para as organizações gerenciarem corretamente a rotatividade, é preciso que elas entendam a fundo o fenômeno e implementem diferentes abordagens de retenção.

Das e Behera (2017) aponta o aprendizado de máquinas como um paradigma capaz de melhorar performances futuras por meio de dados históricos. Devido à habilidade de reconhecer padrões de comportamento que precedem ações, como desligamentos voluntários, e eventos, o aprendizado de máquina se apresenta como uma estratégia promissora para melhorar a gestão de talentos e otimizar os recursos humanos.

A retenção de colaboradores envolve a identificação proativa e a compreensão de quais colaboradores podem estar inclinados a deixar a empresa, assim como o momento e as razões por trás de sua possível partida. A utilização de análises permite a integração de dados dos colaboradores, informações organizacionais e tendências de mercado para prever e analisar os comportamentos de colaboradores de alto desempenho, aprimorando as estratégias de retenção (Isson; Harriott, 2016).

Utilizando três modelos, o estudo procurou combiná-los para obter uma melhor precisão na previsão do nível de risco de saída dos colaboradores de uma plataforma digital

que atua no setor de comércio eletrônico.

A importância desta pesquisa reside na sua capacidade de fornecer ferramentas analíticas que ajudam as organizações a enfrentar o problema da rotatividade de maneira mais estratégica e fundamentada.

1.1 OBJETIVOS

1.1.1 Objetivos Gerais

O objetivo deste trabalho é desenvolver um modelo de predição de desligamento voluntário de colaboradores, utilizando uma abordagem multimodal que integre diferentes tipos de dados, incluindo informações quantitativas e qualitativas, com foco em técnicas de aprendizado de máquina, a fim de minimizar os custos diretos e indiretos associados à rotatividade e melhorar a gestão de recursos humanos.

1.1.2 Objetivos Específicos

Diante do objetivo geral proposto, almeja-se alcançar com este estudo tais objetivos específicos:

- Comparar o desempenho de diferentes algoritmos de predição e configuração de base de dados para a predição do desligamento voluntário.
- Utilizar métricas de avaliação, para medir o desempenho dos modelos preditivos e identificar os modelos com as melhores capacidades preditivas.
- Testar sua aplicabilidade e eficácia na identificação de colaboradores com alta probabilidade de desligamento.
- Elaborar estratégias e intervenções baseadas nas previsões do modelo para melhorar a retenção de colaboradores e reduzir a rotatividade dentro da organização.
- Estabelecer uma classificação de risco que categorize os colaboradores com base na probabilidade estimada de desligamento voluntário resultante dos 3 melhores modelos.

2 REFERENCIAL TEÓRICO

2.1 *PEOPLE ANALYTICS*

De acordo com Marler e Boudreau (2017), *People Analytics* é uma abordagem de recursos humanos suportada por tecnologia da informação que emprega análises descritivas, representações visuais e estatísticas de dados vinculados a processos de RH, gestão do capital humano, desempenho organizacional e referências econômicas externas para determinar o impacto nos negócios e facilitar decisões fundamentadas em dados.

Com o avanço das tecnologias de *big data* e a crescente disponibilidade de dados comportamentais e operacionais dos funcionários, as empresas têm se voltado cada vez mais para essas abordagens quantitativas para tomar decisões estratégicas mais informadas sobre gestão de pessoas (Isson; Harriott, 2016) .

Esse campo emergente tem se tornado cada vez mais necessário em um ambiente corporativo cada vez mais competitivo, pois permite que as organizações tomem decisões baseadas em dados para não apenas melhorar a produtividade e o bem-estar dos colaboradores, mas também aprimorar processos de recrutamento, retenção e desenvolvimento. Essa abordagem transformou a forma como as empresas gerenciam suas operações de pessoas. Isso inclui a análise de dados como tempo de permanência, desempenho, satisfação no trabalho, rotatividade, dados de recrutamento e outros fatores que influenciam diretamente os resultados organizacionais. As aplicações de *People Analytics* são amplas e podem trazer benefícios significativos para as organizações. Estas são algumas das principais áreas onde a análise de dados tem sido empregada:

- Recrutamento e seleção: o time de *People Analytics*s facilita a otimização dos processos de recrutamento e seleção ao identificar padrões de sucesso entre os colaboradores atuais, que podem ser aplicados nas contratações. A análise de currículos, entrevistas e dados de desempenho ajuda a minimizar o viés humano e a aprimorar a qualidade das contratações (Isson; Harriott, 2016).

- Engajamento e retenção de funcionários: por meio da análise de dados relacionados à satisfação, **feedback** dos colaboradores, absenteísmo e outras métricas, as organizações conseguem identificar os fatores que contribuem para a rotatividade e desenvolver estratégias específicas para aumentar o engajamento (Isson; Harriott, 2016).

- Desempenho e desenvolvimento: a avaliação do desempenho individual e em equipe pode ser realizada com o uso da análise de dados, permitindo identificar lacunas de habilidades, oportunidades para treinamento e estratégias de desenvolvimento. Além disso, isso possibilita uma gestão mais personalizada do crescimento profissional dos colaboradores com base em suas necessidades particulares (Isson; Harriott, 2016).

Embora os benefícios do *People Analytics* sejam amplamente reconhecidos, também é importante considerar os desafios e as questões éticas que surgem. A coleta e o uso de informações pessoais dos funcionários geram preocupações em relação à privacidade e à transparência. O uso inadequado ou a má interpretação de dados podem levar a situações de discriminação, preconceito ou desconfiança entre os colaboradores.

Conforme apontado por Charlwood et al. (2016), é fundamental que as instituições estabeleçam diretrizes transparentes relativas à coleta, armazenamento e uso de dados, garantindo que os colaboradores estejam cientes do procedimento e tenham fornecido seu consentimento. Além disso, a análise dos dados deve ser realizada de forma ética, visando apoiar os colaboradores e aprimorar o ambiente organizacional, em vez de puni-los ou manipulá-los negativamente.

2.1.1 Desligamento voluntário e rotatividade

Segundo Chiavenato (2004) a rotatividade de pessoal é o resultado da saída de alguns colaboradores e a entrada de outros para substitui-los no trabalho. Price e Mueller (1981) dividem a rotatividade em 2 tipos: voluntária e involuntária. A rotatividade involuntária ocorre quando o empregador opta pelo desligamento do funcionário. Também ocorre em casos de aposentadoria ou falecimento. Já a rotatividade voluntária, ou evitável, ocorre quando a iniciativa do desligamento parte do trabalhador e não da organização. A rotatividade voluntária é também conhecida como evitável pois a organização é capaz de realizar manobras ou traçar estratégias para mitiga-la.

Rotatividade é um conceito analisado por meio de três principais perspectivas. A primeira perspectiva foca nos modelos de rotatividade, que a consideram um resultado da satisfação no trabalho dos colaboradores e seu comprometimento com a organização. A segunda perspectiva é baseada na literatura sobre liderança e na teoria da troca entre líder e membro. A terceira perspectiva está fundamentada na teoria do suporte organizacional (Mathieu et al., 2016).

O desligamento voluntário é um fenômeno amplamente estudado no campo da gestão de pessoas, dado seu impacto significativo tanto para os colaboradores quanto para as organizações (Hom; Griffeth, 1995). O desligamento voluntário, especificamente, ocorre quando o colaborador opta por deixar a organização por diversos motivos, que podem estar relacionados tanto a questões internas da empresa quanto a fatores externos (Mobley, 1977).

Segundo Hom et al. (2017) as primeiras pesquisas sobre o tema de rotatividade datam de 1920 com Bills (1925), publicando o primeiro estudo empírico sobre rotatividade. Bills (1925) aponta a influência do trabalho dos seus pais na frequência em que os colaboradores pedem ou não demissão. Hom et al. (2017) no seu trabalho *One Hundred Years of Employee Turnover Theory and Research* desenvolveu a linha do tempo das

pesquisas sobre rotatividade (Quadro 1).

Quadro 1 – Linha do tempo das pesquisas sobre rotatividade

Período	Eventos
1970 Modelos Fundamentais	• 1950: March e Simon publicam o primeiro modelo formal de rotatividade. • Estudos continuam sobre os antecedentes da rotatividade, como satisfação no trabalho e dados demográficos. • 1973: Porter e Steers revisam a literatura e propõem a teoria de atendimento às expectativas. • 1977: Price desenvolve a taxonomia dos determinantes de rotatividade.
1980 Teste de Teoria	• Price e Mueller propõem modelo causal de determinantes e fatores. • 1983: Rusbult e Farrell propõem o teste da teoria de investimento. • 1985: Hulin aborda o papel das oportunidades de trabalho.
1990 Modelo de Desdobramento	• 1992: Lee e Mitchell propõem o modelo de desdobramento. • 1996: Teorias passam a focar em nível organizacional e desdobramento.
2000 Pesquisa do século 21	• 2001: Trevor apresenta "movimento de capital" e analisa impactos de rotatividade no desempenho. • 2005: Shaw propõe novos frameworks de rotatividade organizacional. • 2010–2014: Revisão de domínios conceituais e teorias. • 2015: Comemoração dos 100 anos do JAP (2017).

Fonte: Adaptado de (Hom et al., 2017)

Isson e Harriott (2016) aponta 3 tipos principais de desligamento voluntário, aposentadoria, promoção e mudança de empregador. O desligamento por aposentadoria acontece quando o funcionário decide se retirar das atividades laborais ao atingir a idade ou os critérios exigidos para a aposentadoria. O desligamento por promoção refere-se à saída do colaborador em razão de sua ascensão a uma nova função. Por fim, o desligamento por troca de empregador ocorre quando o profissional opta por mudar de empresa em busca

de novas oportunidades ou melhores condições de trabalho no mesmo cargo.

2.2 CAUSAS E FATORES ASSOCIADOS

O desligamento voluntário é frequentemente influenciado por uma combinação de fatores pessoais, organizacionais e externos. As causas mais comuns incluem a insatisfação no trabalho, a falta de reconhecimento, o desalinhamento de valores entre o colaborador e a organização, a ausência de oportunidades para crescimento profissional, remuneração inadequada e condições de trabalho desfavoráveis (Mobley, 1977).

Kreitner e Kinicki (2013) sugere que quanto maior a insatisfação de um colaborador em relação a aspectos como o trabalho em si, os relacionamentos no ambiente de trabalho e os benefícios, maior será a propensão ao desligamento voluntário. Conforme Herzberg (1966), na sua teoria dos dois fatores, nos apresenta os conceitos de Fatores Motivacionais e Fatores de Higiene. Onde os fatores motivacionais são os que são diretamente ligados à satisfação e motivação no trabalho, como reconhecimento, responsabilidade e o próprio prazer no trabalho, por exemplo. Já os Fatores de Higiene, são os fatores que a sua ausência causa insatisfação, mas a sua presença não necessariamente gera motivação por si só. Como salário, benefícios e segurança no trabalho.

Os Fatores de Higiene de Herzberg (1966) influenciam diretamente nas vontades de sair ou no desejo de continuar na mesma organização. Hom e Griffeth (1995) em seu estudo também mostram que salários baixos, benefícios inadequados e falta de segurança, por exemplo, são causas comuns para a tomada de decisão de pedir desligamento do seu emprego atual.

A Teoria da Expectativa, proposta por Vroom (1964), também se aplica ao desligamento voluntário, pois sugere que os colaboradores avaliam as recompensas de suas ações, comparando seus esforços ao retorno que recebem da organização. Se essa expectativa de recompensa não for atendida ou superada, o colaborador pode se sentir desmotivado e descontente com a organização em que se encontra e buscar alternativas fora da empresa, resultando no desligamento voluntário.

Além disso, fatores externos também podem desempenhar um papel importante. A disponibilidade de melhores oportunidades de emprego em outros locais ou mudanças no mercado de trabalho podem tornar mais atraente para o colaborador deixar a organização, especialmente se ele perceber que outras empresas oferecem melhores condições de trabalho ou crescimento profissional (Hom et al., 2017).

2.2.1 Custos e Consequências

A rotatividade por si só traz diversos impactos para a organização e para os funcionários que deixam a empresa. Os desligamentos voluntários são ainda mais impactantes para empresa pois surge de maneira inesperada. No quadro Quadro 2, Mobley (1982) traz

algumas consequências da rotatividade, abordando a visão da organização e do funcionário que deixa está deixando a empresa.

Quadro 2 – Consequências da rotatividade

Consequências	**Organização**	**Funcionários que Deixam a Empresa**
Negativas	Custos econômicos com a demissão, reposição e treinamento; Perda de produtividade; Deterioração da qualidade do serviço; Perda de oportunidades de negócios; Aumento da carga administrativa; Desmoralização dos funcionários que permanecem.	Perda de antiguidade e benefícios; Estresse na transição para um novo emprego; Custos de realocação; Término de redes sociais pessoal e familiar; Perda de serviços da comunidade valorizada; Dificuldade na carreira do cônjuge.
Positivas	Substituição de funcionários com baixo desempenho e burnout; Inclusão de novos conhecimentos e tecnologias; Novos empreendimentos; Redução de custos trabalhistas; Melhores oportunidades de promoção para os funcionários que permanecem; Empoderamento dos funcionários que permanecem.	Obtenção de um emprego melhor em outro lugar; Evasão de um antigo emprego estressante; Renovação do compromisso com o trabalho; Busca de novos empreendimentos; Mudança para uma comunidade mais desejável; Melhora na carreira do cônjuge.

Fonte: Adaptado de Mobley (1982)

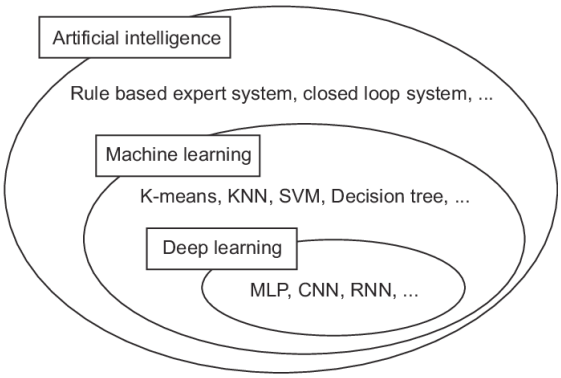
Financeiramente falando, os custos com o desligamento podem ser segmentados em 3 componentes principais: Custos demissionais, custos com substituição e custos com treinamento (Boudreau; Berger, 1985; Flamholtz; Das; Tsui, 1985; Cascio, 1991). Os custos demissionais referem-se às despesas diretamente associadas ao processo de desligamento de funcionários, como as entrevistas de desligamento e os custos com a emissão de documentos comprobatórios, entre outros. Os custos de substituição, por sua vez, incluem os gastos com recrutamento e seleção, divulgação de vagas e contratação de consultorias especializadas para o preenchimento das posições abertas. Por fim, os custos com treinamento, como o próprio nome sugere, referem-se aos gastos relacionados à orientação e capacitação do novo colaborador.

Além dos impactos financeiros, os desligamentos voluntários impactam negativamente em outros aspectos da organização. A qualidade do serviço pode ser prejudicada pela perda de colaboradores experientes, o que compromete a eficiência das operações. Além disso, resulta em uma queda na moral e motivação da equipe restante, gerando um ambiente de trabalho mais desmotivador. Por fim, a imagem da empresa no mercado de trabalho também é afetada, pois um alto índice de desligamentos voluntários prejudica a reputação da organização, dificultando a atração de novos talentos.

2.3 APRENDIZADO DE MÁQUINA

Segundo Mitchell e Mitchell (1997) "O aprendizado de máquina é o estudo de algoritmos computacionais que melhoram automaticamente por meio da experiência."O aprendizado de máquinas, do inglês *Machine Learning*, é uma subárea da inteligência artificial que se concentra na criação de algoritmos capazes de aprender e fazer previsões a partir de dados. O principal objetivo do aprendizado de máquinas é permitir que sistemas computacionais melhorem seu desempenho em uma tarefa específica sem a necessidade de programação explícita para cada possibilidade, utilizando dados para aprender e generalizar novos padrões. A figura 1, ilustra a organização das áreas da inteligência artificial.

Figura 1 – Diagrama de Venn da inteligência artificial



Fonte: Lee e Jung (2018).

O aprendizado de máquinas é geralmente classificado em três principais categorias, dependendo do tipo de dados disponíveis e da natureza do problema a ser resolvido: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço.

2.3.1 Aprendizado supervisionado

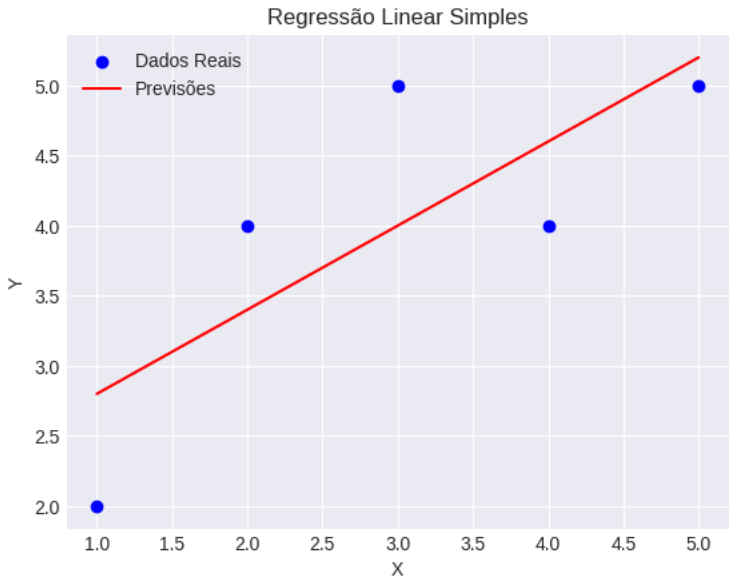
O aprendizado supervisionado ocorre quando um modelo é treinado com dados rotulados, ou seja, cada entrada de dados está associada a uma resposta conhecida. O modelo aprende a partir desses exemplos e torna-se capaz de prever as respostas para novos dados. Algoritmos desse tipo operam compreendendo a relação entre os dados de entrada e as respostas associadas, e, a partir dessa compreensão, geram a saída para novos dados de entrada. Em problemas de classificação, as saídas são categorias, como, por exemplo: se vai chover ou não, se um colaborador tende ou não a pedir demissão, se uma compra é ou não fraudulenta, entre outras. Essas categorias estão associadas à probabilidade de pertencimento àquela classe. Em problemas de regressão, as saídas são valores numéricos, como, por exemplo: o preço do aluguel de um apartamento em uma determinada região, o consumo de energia elétrica de uma fábrica na próxima hora, entre outros. Algoritmos

como Regressão Linear, Máquinas de Vetores de Suporte (SVM) e Árvores de Decisão são utilizados em problemas desse tipo.

Mujumdar e Vaidehi (2019) utilizou o aprendizado supervisionado, aplicando diversos algoritmos para classificar os pacientes e apontar os que tem diabetes. No mesmo estudo, Mujumdar e Vaidehi (2019) ainda aponta que no futuro esse estudo pode apontar a probabilidade de pessoas, que não tem diabetes, desenvolver a doença dentro de alguns anos. Feng et al. (2024) desenvolveu um *framework* para aplicar algoritmos de aprendizado de máquina para prever a gravidade de acidentes marítimos.

Na figura 2, temos um exemplo da plotagem do resultado de um problema de regressão simples.

Figura 2 – Exemplo de problema de aprendizado supervisionado, um problema de regressão



Fonte: Depieri (2023).

2.3.2 Aprendizado não supervisionado

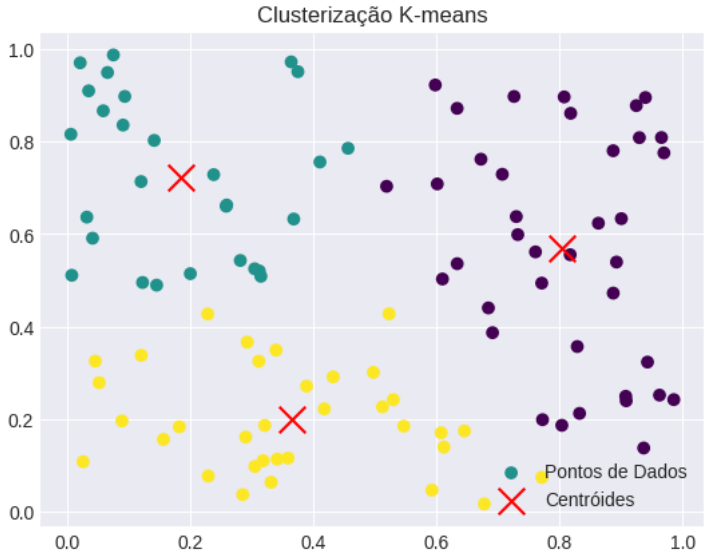
No aprendizado não supervisionado, os dados de treinamento não possuem rótulos, e o objetivo do modelo é identificar padrões e estruturas subjacentes nos dados. Nesse tipo de abordagem, o algoritmo busca padrões nos dados de entrada e tenta agrupar os itens em subgrupos ou categorias dentro desse conjunto de dados. Exemplos desse tipo de problema incluem segmentação de clientes e detecção de anomalias. Técnicas como clusterização e redução de dimensionalidade são amplamente utilizadas nessa abordagem.

Mozumder et al. (2024) aplica e compara 3 algoritmos de aprendizado não supervisionado para segmentar os clientes do setor bancário. Essa segmentação permite que os

bancos conheçam melhor o comportamento e características dos seus clientes e ofereçam produtos e serviços de forma mais eficiente. Chen et al. (2024) desenvolve um modelo de aprendizado não supervisionado para segmentar os clientes do setor elétrico com base em seus padrões de consumo de energia.

Na figura 3, temos um exemplo da plotagem do resultado de um problema de clusterização.

Figura 3 – Exemplo de problema de aprendizado não supervisionado, um problema de clusterização



Fonte: Depieri (2023).

2.3.3 Aprendizado por reforço

O aprendizado por reforço é um tipo de aprendizado no qual um agente aprende a tomar decisões sequenciais visando maximizar uma recompensa acumulada ao longo do tempo. Um paralelo comum com a vida real pode ser feito com o treinamento de um cachorro: ao obedecer ao comando, ele é recompensado, enquanto a desobediência não resulta em recompensa. Dessa forma, o comportamento do agente vai sendo condicionado ao sucesso, com o erro sendo progressivamente reduzido ao longo do tempo. Sistemas de recomendação são exemplos típicos desse tipo de problema. Na figura 4, tem-se um exemplo da plotagem do resultado de um problema de aprendizado por reforço.

2.3.4 Aprendizado de máquinas aplicado à área de recursos humanos

Mendonça et al. (2017) destaca que o RH se transformou em um departamento de grande relevância para as empresas, e a Inteligência Artificial desempenhou um papel

Figura 4 – Exemplo de problema de aprendizado por reforço

Fonte: (Depieri, 2023).

importantíssimo nesse processo. O emprego dessa tecnologia ofereceu ao departamento de recursos humanos a chance de operar de forma mais estratégica, visando um crescimento mais sólido dentro da empresa (Mendonça et al., 2017).

A aplicação de modelos de aprendizado de máquina em Recursos Humanos (RH), embora tenha levado algum tempo para se consolidar, tem se intensificado significativamente nos últimos anos (Faggella, 2019). Com o objetivo de tornar os processos mais ágeis e eficientes, o uso de aprendizado de máquina em Sistemas de Rastreamento de Candidatos (ATS), leitores biométricos, assistentes virtuais e *chatbots* se tornou cada vez mais comum.

Garg et al. (2022) aponta que a gestão dos recursos humanos tem cada vez mais adotado práticas e técnicas de aprendizado de máquinas, principalmente nas aplicações que dispõem de uma grande quantidade de dados.

Nas aplicações sobre engajamento dos colaboradores, Anitha (2014) utiliza do aprendizado de máquina para determinar o nível de engajamento e as variáveis que mais impactavam nesse fenômeno. Além de como o engajamento afetava a performance do colaborador.

Para predizer a performance do colaborador, Li et al. (2016) utilizou K-Vizinhos mais próximos *(K-Nearest Neighbor (KNN))*. Nesse estudo, foram comparados outros algorítimos de aprendizado de máquina e o *KNN* foi o mais performático.

Bianchi (2023) utilizou regressão logística e análise de sobrevivência para prever o desligamento voluntário de trabalhadores de uma empresa de educação e tecnologia. Já Guedes (2024) utilizou floresta aleatória para prever o mesmo evento em uma fábrica de

Quadro 3 – Funções de RH e Objetivos para Aplicações de ML

Função de RH	**Objetivos para Aplicação de ML**
Recrutamento	Avaliar a adequação de candidatos a posições; extrair informações de currículos e analisar perfis de candidatos.
Seleção	Identificar atributos decisivos para seleção e desenvolver modelos de seleção.
Engajamento dos funcionários	Compreender o engajamento dos funcionários com a marca; analisar o sentimento atual dos funcionários e fatores que aumentam o estresse no trabalho.
Treinamento e desenvolvimento	Identificar necessidades de treinamento; recomendar cursos relevantes e medir a eficácia do treinamento.
Gestão de desempenho	Avaliação de desempenho; previsão de desempenho; detecção de viés no processo de avaliação; estimativa do nível de especialização dos funcionários e desenvolvimento de incentivos personalizados.
Rotatividade de funcionários	Prever a rotatividade dos funcionários com base em fatores pessoais e relacionados ao trabalho.
Dinâmica de equipe	Recomendar membros para equipes; avaliar o desempenho de equipes; compreender sentimentos das equipes e os padrões de interação entre elas.
Alocação de recursos humanos	Alocar pessoas em diferentes categorias.

Fonte: Garg et al. (2022)

celulose.

3 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa foi desenvolvida seguindo a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*). CRISP-DM é amplamente utilizada em projetos de ciência de dados por oferecer uma abordagem estruturada e iterativa, dividida em seis etapas principais (Chapman et al., 2000). Desenvolvida na década de 1990 por um consórcio de empresas, o CRISP-DM se tornou um modelo de processo padrão para a realização de projetos de mineração de dados. A metodologia é composta por uma série de etapas estruturadas que permitem organizar o processo de análise de dados de maneira sistemática e eficaz. Sua flexibilidade, robustez e aplicabilidade em diferentes setores fazem dela uma escolha popular tanto para iniciantes quanto para profissionais experientes da área (Chapman et al., 2000). As etapas da metodologia CRISP-DM são:

- Entendimento do Negócio: Fase inicial, onde o objetivo do projeto é compreendido e os requisitos são definidos. A principal tarefa é entender o problema de negócio e os objetivos da análise, para que a solução seja relevante.
- Entendimento dos Dados: Após entender o negócio, os dados são coletados e analisados para explorar sua qualidade, consistência e relevância para o projeto. Nessa etapa, são realizados testes iniciais e é verificado o que pode ser usado para responder às perguntas de negócio.
- Preparação dos Dados: Envolve a limpeza e transformação dos dados para que possam ser utilizados nas etapas seguintes. Isso inclui o tratamento de dados ausentes, eliminação de inconsistências e a criação de variáveis relevantes.
- Modelagem: Aqui, diferentes algoritmos de modelagem são aplicados aos dados. São testados modelos diversos, ajustando parâmetros e utilizando técnicas e abordagens apropriadas para o tipo de problema (classificação, regressão, clusterização, etc.).
- Avaliação: Após a construção dos modelos, é necessário avaliá-los, verificando se atendem aos objetivos de negócio. Esta etapa também envolve a interpretação dos resultados e a decisão sobre a implementação do modelo final.
- Implantação: A última etapa consiste em implementar a solução proposta, o que pode envolver a criação de relatórios, integração com sistemas existentes ou a adoção do modelo em operações práticas. Também é importante planejar como o modelo será mantido e monitorado.

A escolha do CRISP-DM como a metodologia para a realização de um projeto de aprendizado de máquinas é justificada por sua abordagem estruturada e iterativa, que permite uma adaptação contínua e refinamento do processo. Além disso, sua independência de

domínio, ou seja, sua capacidade de ser aplicada em qualquer tipo de negócio, setor ou indústria, torna o CRISP-DM uma solução universal para problemas de análise de dados (Shearer, 2000).

Por ser uma abordagem estruturada, o CRISP-DM pode ser aplicado em diversos tipos de problemas. PÁDUA et al. (2017) utiliza a metodologia para estruturar o desenvolvimento de um modelo de predição do risco de evasão dos cursos técnicos EAD do Instituto Federal do Piauí (IFPI). Silva e Timo (2022) utilizou a metodologia para o desenvolvimento de um modelo aplicado à atenção domiciliar e predição de condição de óbito por meio da identificação de pacientes de alto risco.

No contexto deste trabalho, ela orienta o desenvolvimento do modelo preditivo de desligamentos voluntários de colaboradores, garantindo uma integração lógica entre o entendimento do problema, preparação de dados, construção do modelo e a implementação de soluções práticas.

3.1 COMPREENSÃO DO NEGÓCIO

A primeira etapa do CRISP-DM, visa compreender o contexto do negócio e os objetivos da organização para assegurar que as soluções de aprendizado de máquina atendam às suas necessidades. Este projeto envolve uma plataforma digital de grande porte que atua no setor de comércio eletrônico, oferecendo um serviço de classificados online. A empresa conecta compradores e vendedores de uma ampla gama de produtos e serviços, permitindo que os usuários publiquem anúncios de itens novos e usados.

Como toda grande organização digital, a empresa busca constantemente aprimorar seus processos internos, com foco na gestão de recursos humanos, especialmente no que tange aos desligamentos voluntários. Este trabalho visa compreender os fatores que influenciam a tomada de decisão relacionada aos desligamentos voluntários e reduzir a taxa de saída por meio da predição da probabilidade de desligamento dos colaboradores.

3.2 ENTENDIMENTO DOS DADOS

Na etapa de Entendimento dos Dados envolve o processo de coleta, exploração e análise inicial dos dados disponíveis. Esse processo visa não apenas a compreensão das características dos dados, mas também a identificação de padrões, *outliers*, lacunas e relacionamentos entre as variáveis, que serão necessários para o sucesso da modelagem preditiva. O entendimento adequado dos dados proporciona a base necessária para o desenvolvimento de um modelo robusto e eficiente.

Os dados apresentados neste estudo foram alterados para garantir a privacidade dos colaboradores e da empresa. Modificações incluíram a adição de ruído estatístico, remoção de identificadores diretos, e generalização de categorias.

3.2.1 Coleta de dados

No contexto deste trabalho, os dados necessários para a predição de desligamento voluntário de colaboradores são coletados de diversas fontes e sistemas, tanto internos quanto externos à organização. Essas fontes de dados abrangem informações organizacionais, demográficas e de desempenho dos colaboradores, proporcionando uma visão ampla do perfil de cada indivíduo dentro da empresa.

Os dados são coletados de fontes diversas. Uma das principais fontes é o sistema de gestão de recursos humanos da empresa. Desse sistema, são extraídos relatórios manuais que contêm informações organizacionais sobre os colaboradores. Esses dados incluem, entre outros, o salário, cargo ocupado, tempo de serviço na empresa, departamento em que atuam, entre outros dados relacionados à estrutura organizacional e perfil do colaborador.Essas informações são muito importantes para compreender o contexto de cada colaborador dentro da organização e como essas variáveis podem influenciar o desligamento voluntário. Por exemplo, o cargo ocupado e o tempo de serviço são variáveis que podem ter uma forte correlação com a propensão ao desligamento, já que colaboradores com maior tempo de casa ou em cargos de maior responsabilidade podem ter comportamentos distintos em relação ao desligamento quando comparados a colaboradores recém-contratados ou em cargos operacionais.

Além dos dados organizacionais extraídos do sistema de RH, também são coletadas informações sobre o desempenho dos colaboradores, por meio de APIs (*Application Programming Interface*) de plataforma gestão de desempenho. Por meio dessas APIs são obtidos dados como notas de desempenho, *feedbacks* de líderes e colegas de trabalho, e outros indicadores qualitativos e quantitativos do comportamento, desempenho e produtividade do colaborador. Esses dados de desempenho oferecem uma dimensão adicional à análise, pois podem revelar padrões de insatisfação ou desmotivação que não seriam capturados apenas pelos dados organizacionais. Dessa forma, para a composição da base de dados que servirá como entradas para o treinamento dos modelos, foram utilizados principalmente 3 tipos de variáveis:

- Variáveis Demográficas: incluem dados como idade, gênero, tempo de serviço e formação educacional, região, se reside em capitais, se reside nos estados em que a empresa tem escritório dentre outras.
- Variáveis Organizacionais: englobam dados sobre o cargo ocupado, departamento em que o colaborador está alocado, salário, posicionamento na faixa salarial, se ocupam cargos de gestão e etc.
- Variáveis de Desempenho: avaliações de desempenho, *feedbacks* de gestores e colegas, produtividade, e outros indicadores qualitativos e quantitativos.

Por fim, os dados das diversas fontes são unidos em uma tabela *Excel* que será utilizada. Essa ferramenta permitiu uma rápida conferência e maior agilidade à unificação das bases.

Quadro 4 – Dicionário parcial de dados

Coluna	**Descrição**	**Exemplo de Valor**	**Tipo**
Chave	Chave Unica	CPF-Nome	Categórico
cargo	Cargo do empregado	Gerente	Categórico
Voluntário?	Empregado Ativo ou Desligado voluntariamente	ATIVO	Categórico
CC	Centro de Custo	56230	Categórico
Diretoria	Diretoria do colaborador	GENTE	Categórico
Área Tech	Se o colaborador atua em uma área de Tecnologia	SIM	Binário
Posicionamento	% da Faixa Salarial	80%	Numérico
Fases da Vida	Fase da vida do colaborador de acordo com a idade	Até os 30 anos	Categórico
Estado Civil	Estado civil do empregado	Casado	Categórico
Filhos	Se o colaborador possui filhos	1	Binário
Afastamento	Se o colaborador pediu afastamento no ultimo 1ano	0	Binário
Virou Gestor	Se o colaborador se tornou gestor no ultimo 1ano	1	Binário
Escolaridade	Nível macro da escolaridade do colaborador	Superior Completo	Categórico
Tempo de Casa	Tempo de casa do colaborador	Até 1 ano	Categórico

3.2.2 Análise Exploratória de Dados (EAD)

Para realizar essa análise, foi escolhida a linguagem *Python*, amplamente utilizada em projetos de ciência de dados devido à sua sintaxe simples e à vasta quantidade de bibliotecas específicas para análise de dados, como *Pandas, NumPy, Matplotlib e Seaborn*. A análise foi realizada no ambiente de notebooks do *Visual Studio Code (VSCode)*, que oferece uma excelente plataforma para experimentação, permitindo a execução de código interativo e visualização imediata dos resultados. Além disso, o *VSCode* facilita o versionamento de código por meio de integração com o *GitHub*, garantindo o controle de versões e a colaboração eficiente durante o desenvolvimento.

O primeiro passo da análise exploratória foi investigar a presença de dados faltantes nas variáveis. Caso algum erro tenha acontecido durante a unificação das bases ou até mesmo da extração dos dados, seria apontado aqui.Para isso, utilizou-se a função *isnull()* da biblioteca Pandas, que permite verificar a presença de valores ausentes em cada coluna do *dataset*.

Em seguida, foi realizada uma análise estatística descritiva das variáveis numéricas do conjunto de dados. O objetivo dessa análise foi entender a distribuição e o comportamento das variáveis contínuas, como salário, tempo de serviço e avaliação de desempenho, entre outras. Para isso, foram calculadas métricas como a média, mediana, desvio padrão, mínimo, máximo e quartis. A análise incluiu a visualização das distribuições dessas variáveis por meio de gráficos como histogramas e *boxplots*, que permitem observar a forma da distribuição e a presença de *outliers* e possíveis anomalias nos dados, respectivamente.

Outro aspecto importante da análise exploratória foi a verificação das variáveis categóricas. Para isso, foi realizada a contagem da frequência de cada categoria nas variáveis, como cargo, departamento e sexo, entre outras. Com essa análise, busca-se entender a distribuição das categorias e verificar a possibilidade de desequilíbrios nas classes, o que pode afetar a performance do modelo.

Foi utilizado o método *value_ counts()* do *Pandas* para contar o número de ocorrências de cada categoria, e as distribuições foram visualizadas por meio de gráficos de barras. Em alguns casos, categorias com baixa frequência foram combinadas, a fim de evitar que categorias muito pequenas gerassem distorções nos modelos desenvolvidos.

3.3 PREPARAÇÃO DOS DADOS

A etapa de preparação dos dados é a etapa no desenvolvimento de modelos preditivos que visa garantir que os dados estejam no formato correto e prontos para alimentar os algoritmos de aprendizado de máquina.

3.3.1 Tratamento de dados faltantes

Durante a etapa de análise exploratória, é verificada a presença de valores faltantes no conjunto de dados. A forma de lidar com essa ausência varia de acordo com o tipo da variável, da distribuição e da proporção de valores faltantes.

Em variáveis numéricas, pode-se substituir os valores faltantes por:

- Média: Adotar a média caso o conjunto da variável siga a distribuição normal
- Mediana: Caso o conjunto não siga a distribuição normal
- Regressão: Há casos em que é possível, utilizando as outras variáveis, predizer o valor faltante da variável.

Já em variáveis categóricas, para substituir os valores ausentes, é comum o uso da categoria mais frequente, moda, do conjunto. Em variáveis com uma % de dados faltantes relativamente alta, o comum é excluir todo o conjunto da variável, caso seja possível.

3.3.2 Codificação de variáveis categóricas

Foi necessário tratar as variáveis categóricas, uma vez que a maioria dos algoritmos de aprendizado de máquina não consegue processar dados em formato de texto ou categorias diretamente. Para isso, utilizou-se 2 processos de codificação para transformar variáveis categóricas em variáveis numéricas:

- Uma codificação manual: Nas variáveis categóricas ordinais, como fases da vida, escolaridade dentre outras. Essa abordagem não afeta a dimensionalidade da variável.
- *Binary Enconding*: Nas variáveis categóricas nominais foi utilizado o *Binary Enconding*, onde cada categoria é convertida em um número único e esse número é convertido para o formato binário, resultado em colunas de 0 ou 1. Uma variável com d valores únicos, resulta em *log2(d)* novas colunas (Seger, 2018).

3.4 MODELAGEM

Durante esta fase, foram realizados testes com diversos algoritmos de aprendizado de máquina, e as melhores combinações de algoritmos e configurações de base foram selecionadas para a predição do desligamento voluntário dos colaboradores.

3.4.1 Análise de correlação e Configurações de base

Antes de realizar o treinamento, realizou-se uma análise de correlação para identificar quais variáveis poderiam ter mais influência no desligamento voluntário dos colaboradores. A partir dessa análise, foi criado três configurações de base com diferentes combinações de variáveis, que foram usadas para treinar os modelos de aprendizado de máquina:

- Variáveis com Correlação Positiva: Foi selecionado variáveis que possuem uma correlação positiva com o desligamento voluntário, ou seja, variáveis que aumentam a probabilidade de desligamento à medida que seus valores aumentam.
- Features com Correlação Negativa: Neste caso, foram escolhidas variáveis que apresentam correlação negativa com o desligamento voluntário, ou seja, valores maiores dessas variáveis indicam uma redução na probabilidade de desligamento.

Além disso, um outro método de seleção de variáveis foi utilizado, o *SelectKBest* do pacote *Scikit-learn* (Scikit-learn, 2024a). Essa técnica avalia a importância de cada variável com base em uma métrica estatística, como o teste de qui-quadrado, ANOVA ou correlação, e seleciona as K variáveis mais relevantes para o modelo (Guyon et al., 2002).

- Seleção de variáveis com *SelectKBest*: Utilizado para selecionar as variáveis mais relevantes com base na pontuação de cada uma delas, usando o teste de qui-quadrado para classificar a importância das características.

3.4.2 Treinamento dos modelos e Pipelines

Para realizar o treinamento dos modelos, foram escolhidos seis algoritmos de aprendizado de máquina amplamente utilizados e com bom desempenho geral em problemas com características como o do problema abordado nesse trabalho. Os algoritmos selecionados foram:

- *Random Forest*: um modelo baseado em múltiplas árvores de decisão, que utiliza o voto da maioria das árvores para fazer a previsão. Esse algoritmo é robusto, eficiente e ajuda a reduzir o *overfitting*, sendo uma excelente escolha para dados com alta dimensionalidade (Breiman, 2001).

- *LightGBM* (LGBM): um algoritmo de *boosting*baseado em gradiente, conhecido pela sua alta velocidade de treinamento e pela capacidade de lidar eficientemente com grandes volumes de dados (Ke et al., 2017).

- *CatBoost*: outro algoritmo de *boosting*, que se destaca na manipulação de variáveis categóricas e também apresenta alta performance e baixa propensão ao *overfitting*. O *CatBoost* é especialmente eficaz em conjunto de dados complexos numa alta variedade de desafios práticos(Dorogush; Ershov; Gulin, 2018).

- *Support Vector Machine* (SVM): um modelo de classificação que tenta encontrar um hiperplano ótimo para separar as classes, mesmo em dados de alta dimensionalidade e volume (Lorena; Carvalho, 2007).

- Árvore Binária : algoritmo de aprendizado de máquina baseado em árvore de decisão, que realiza a divisão dos dados de acordo com perguntas binárias até chegar a uma previsão(Quinlan, 1986).

- Regressão Logística (Logistic Regression): um modelo estatístico utilizado para prever a probabilidade de ocorrência de um evento binário. A regressão logística é uma técnica simples, mas eficaz, especialmente quando a relação entre as variáveis independentes e a variável dependente é linear (Jr; Lemeshow; Sturdivant, 2013).

Todos os modelos foram treinados usando *pipelines* do *Scikit-learn* . Os *pipelines* permitem organizar o fluxo de trabalho de forma eficiente e reproduzível, desde o pré-processamento dos dados até o treinamento do modelo. A utilização de *pipelines* também garante que o processo seja realizado de maneira consistente em todas as iterações, facilitando a automação, o versionamento do código e prevenindo vazamento de dados.

3.5 AVALIAÇÃO DOS MODELOS

A quinta etapa do CRISP-DM é dedicada à avaliação do modelo. Essa etapa tem como principal objetivo analisar e validar os modelos treinados, garantindo que o modelo

escolhido para produção seja capaz de fornecer previsões confiáveis e robustas para o problema em questão. A avaliação não se limita apenas a verificar a precisão do modelo, mas envolve a análise das métricas de desempenho, identificação de possíveis melhorias e a compreensão das limitações do modelo. A avaliação visa garantir que o modelo escolhido esteja alinhado com os objetivos do negócio e seja capaz de generalizar bem para dados não vistos, ou seja, tenha bom desempenho tanto nos dados de treino quanto nos dados de teste e reproduza esse desempenho com os novos dados em produção. Além disso, a etapa de avaliação permite comparar o desempenho de diferentes modelos, possibilitando a escolha do melhor modelo para a aplicação.

3.5.1 Matriz de confusão

A matriz de confusão é uma ferramenta amplamente utilizada na avaliação de modelos de classificação. Ela mostra o desempenho do modelo em termos de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos.

- Verdadeiro Positivo (TP): Quando o modelo corretamente classifica a classe positiva.
- Falso Positivo (FP): Quando o modelo erroneamente classifica a classe negativa como positiva (erro tipo I).
- Verdadeiro Negativo (TN): Quando o modelo corretamente classifica a classe negativa.
- Falso Negativo (FN): Quando o modelo erroneamente classifica a classe positiva como negativa (erro tipo II).

Figura 5 – Matriz de confusão

		Predito	
		Positivo	**Negativa**
Real	**Positivo**	Verdadeiro Positivo	Falso Negativo
	Negativo	Falso Positivo	Verdadeiro Negativo

A matriz de confusão é fundamental para calcular várias métricas de desempenho, que ajudam a entender melhor como o modelo se comporta em relação a diferentes tipos de erro. Ela oferece uma visão detalhada das predições feitas pelo modelo, mostrando não apenas a quantidade de acertos e erros, mas também o tipo de erro que o modelo está cometendo.

3.5.2 Tipos de erro em modelos de classificação

Existem dois tipos principais de erro em problemas de classificação:

- Erro Tipo I (Falso Positivo): Esse erro ocorre quando o modelo classifica erroneamente uma instância negativa como positiva. Ele é particularmente crítico em situações onde o custo de uma falsa detecção é elevado, como em modelos de identificação de fraudes ou diagnóstico de doenças, onde uma falsa alarme pode gerar consequências significativas.
- Erro Tipo II (Falso Negativo): Aqui, o modelo classifica incorretamente uma instância positiva como negativa. Esse erro é crucial em contextos em que perder uma instância positiva acarreta grandes custos, como em sistemas de recomendação.

3.5.3 Métricas de Avaliação Para Modelos de Classificação

Para medir o desempenho de modelos de classificação, existem várias métricas que podem ser utilizadas, dependendo do contexto e dos objetivos do problema. As principais métricas incluem:

A acurácia é uma das métricas mais simples e amplamente utilizadas em problemas de classificação. Ela mede a proporção de classificações corretas (verdadeiros positivos e verdadeiros negativos) em relação ao total de instâncias.

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN}$$

No entanto, a acurácia pode ser enganosa em cenários de desequilíbrio entre as classes, ou seja, quando uma classe é muito mais representativa do que a outra, ou seja quando há desbalanceamento na variável alvo.

A precisão mede a capacidade do modelo de classificar corretamente as instâncias positivas.

$$\text{Precisão} = \frac{TP}{TP + FP}$$

A precisão é útil em cenários em que o custo de falsos positivos é elevado, como em sistemas de detecção de fraude. Quando é preciso minimizar o erro de tipo I.

O *recall* (ou sensibilidade) mede a capacidade do modelo de identificar todas as instâncias positivas.

$$\textit{Recall} = \frac{TP}{TP + FN}$$

O *recall* é importante quando o custo de falsos negativos é significativo, como na detecção de doenças graves, onde é crucial identificar todos os casos positivos. Quando é preciso minimizar o erro de tipo II.

O *F1-score* é a média harmônica entre precisão e recall. Ele é particularmente útil quando se busca um equilíbrio entre essas duas métricas.

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \textit{Recall}}{\text{Precisão} + \textit{Recall}}$$

Essa métrica é recomendada quando se lida com dados desbalanceados, pois leva em consideração tanto os falsos positivos quanto os falsos negativos.

A Área sob a Curva ROC (AUC) é uma métrica baseada na curva ROC (*Receiver Operating Characteristic*), que é um gráfico que mostra a taxa de verdadeiros positivos versus a taxa de falsos positivos. A AUC fornece uma avaliação da capacidade do modelo em distinguir entre as classes.

$$\text{AUC} = \int_0^1 \text{TPR}(f)\, d\text{FPR}(f)$$

Uma AUC de 1 indica um modelo perfeito, enquanto uma AUC de 0,5 indica que o modelo não é melhor do que uma classificação aleatória, como jogar uma moeda.

O *F-Beta Score* é uma métrica de avaliação utilizada em problemas de classificação, especialmente em cenários onde há um desequilíbrio entre as classes. Ele é uma extensão do *F1-Score* e tem como objetivo balancear a importância entre a *precisão* e o *recall* (sensibilidade) de um modelo.

O F-Beta Score é uma média ponderada entre a *precisão* e o *recall*, onde o parâmetro β controla a ponderação relativa de cada uma dessas métricas. A fórmula geral do F-Beta Score é dada por:

$$\textit{F-Beta Score} = \frac{(1 + \beta^2) \cdot \text{Precisão} \cdot \textit{Recall}}{\beta^2 \cdot \text{Precisão} + \textit{Recall}}$$

O parâmetro β no *F-Beta Score* define o peso relativo dado à precisão e ao *recall*. Ele permite ajustar a ênfase entre essas duas métricas, dependendo da importância de cada uma no problema específico:

- Quando $\beta = 1$, o *F-Beta Score* é igual ao *F1-Score*, ou seja, F-Beta = F1-Score, equilibrando igualmente a precisão e o *recall*.
- Quando $\beta > 1$, o *recall* recebe um peso maior, priorizando a captura de mais instâncias positivas, o que é útil em cenários onde é mais importante evitar falsos negativos, como em tarefas de diagnóstico médico.

- Quando $\beta < 1$, a precisão é priorizada em relação ao *recall*, sendo adequado para problemas onde os falsos positivos têm um custo maior do que os falsos negativos, como em sistemas de detecção de fraude.

Em outras palavras, o valor de β permite que o *F-Beta Score*se adapte à natureza do problema em questão, ajustando a relevância das duas métricas. É comum utilizar o β sendo $1/2$, 1 ou 2, para se ajustar à natureza do problema abordado.

3.5.4 Melhoria de Hiperparâmetros

A melhoria dos hiperparâmetros é uma etapa crucial na otimização de um modelo de aprendizado de máquina. Hiperparâmetros são parâmetros externos ao modelo que controlam seu funcionamento, como a profundidade da árvore de decisão, a taxa de aprendizado e o número de estimadores no caso de métodos de *ensemble*.

Existem várias técnicas para a melhoria de hiperparâmetros, sendo as mais comuns:

- *Grid Search*: testa todas as combinações possíveis de hiperparâmetros em um espaço definido. Embora seja uma abordagem simples, pode ser muito demorada, especialmente quando o espaço de busca é grande. A busca em grade pode ser eficiente para um número limitado de parâmetros e valores discretos, mas pode se tornar impraticável à medida que o número de hiperparâmetros e as opções aumentam (Bergstra et al., 2012).

- *Random Search*: testa combinações aleatórias de hiperparâmetros dentro de um intervalo. Essa abordagem é geralmente mais eficiente do que a busca em grade, pois explora mais rapidamente o espaço de busca, e pode encontrar boas soluções em menos tempo (Bergstra et al., 2012).

- *Otimização Bayesiana*: método probabilístico que busca os melhores parâmetros de forma mais eficiente, com base em testes anteriores. A otimização Bayesiana ajusta os parâmetros de forma iterativa, visando maximizar a função objetivo. Essa técnica é frequentemente usada quando o tempo de execução de cada experimento é alto (Snoek; Larochelle; Adams, 2012).

- *Optuna*: o *Optuna* é uma biblioteca moderna e poderosa para otimização de hiperparâmetros, baseada em um algoritmo de otimização de tipo *Tree-structured Parzen Estimator (TPE)*. Essa técnica é eficiente e escalável, podendo ser aplicada a uma ampla variedade de modelos e conjuntos de dados. O *Optuna* aprende com as tentativas anteriores e ajusta a busca para encontrar a melhor combinação de hiperparâmetros de maneira mais eficaz, reduzindo o tempo necessário para essa otimização em comparação com as abordagens tradicionais (Akiba et al., 2019).

Ao ajustar os hiperparâmetros de um modelo, é fundamental ter cuidado para não comprometer os resultados. Um dos riscos é o *overfitting*, que ocorre quando o modelo se adapta excessivamente aos dados de treinamento, prejudicando sua capacidade de generalizar para dados novos. Além disso,deve-se levar em consideração o tempo de computação, a busca por hiperparâmetros ideais, dependendo do método escolhido, pode ser muito custosa em termos de recursos computacionais.

3.5.5 Importância das Variáveis

A análise da importância das variáveis tem como objetivo entender como o modelo está tomando suas decisões e para fornecer informações e entendimento sobre o problema. A maioria dos algoritmos de aprendizado de máquina aplicados em problemas de classificação já possuem um bom nível de explicabilidade e interpretabilidade dos seus resultados.

Ferramentas como o *feature importance* ou *permutation importance* são amplamente utilizadas para determinar a contribuição de cada variável.

A importância das variáveis refere-se a técnicas usadas para identificar quais características dos dados têm maior impacto nas previsões feitas pelo modelo. Existem vários métodos para calcular a importância das variáveis, sendo os mais comuns:

- *Random Forest*: modelos baseados em árvores, como Random Forest e Árvores de Decisão, fornecem uma medida de importância das variáveis de forma natural. A importância é calculada com base na redução do impureza ao longo das divisões feitas em cada nó da árvore. Variáveis que contribuem significativamente para a redução da impureza são consideradas mais importantes (Liaw; Wiener, 2002).

- Permutação: esse método calcula a importância das variáveis através da avaliação do impacto na performance do modelo quando as variáveis são permutadas aleatoriamente. A técnica baseia-se em medir quanto a performance do modelo é degradada ao embaralhar os valores de uma variável específica (Altmann et al., 2010).

- *SHAP (SHapley Additive exPlanations)*: é um método baseado em teoria dos jogos que explica o valor de uma previsão com base na contribuição individual de cada variável. Ele calcula uma contribuição de cada variável para a previsão de uma instância específica, utilizando a ideia de valores de Shapley. O SHAP é uma das abordagens mais populares por sua interpretação intuitiva e capacidade de oferecer explicações localmente e globalmente (Lundberg, 2017).

- *LIME (Local Interpretable Model-agnostic Explanations)*: é uma técnica de explicabilidade de modelo que cria um modelo interpretável localmente para cada instância de teste. Ele perturba os dados de entrada e observa como o modelo reage, criando uma aproximação local linear que ajuda a entender o comportamento do modelo

para uma instância específica. A partir daí, é possível verificar quais variáveis influenciam a previsão (Ribeiro; Singh; Guestrin, 2016).

Compreender quais variáveis os modelos consideram mais relevantes ajuda a:

- Aprimorar a interpretabilidade do modelo, facilitando a compreensão das decisões e previsões feitas pelo algoritmo.
- Realizar a seleção de variáveis, descartando aquelas que têm pouco impacto, o que pode melhorar a eficiência do modelo.
- Detectar correlações ou dependências entre variáveis, trazendo a tona padrões ocultos nos dados e proporcionando uma visão mais aprofundada dessas relações.

3.6 IMPLEMENTAÇÃO

A sexta etapa do CRISP-DM, Implementação, refere-se à fase na qual o modelo preditivo, validado e testado nas etapas anteriores, é colocado em produção para atender aos objetivos de negócio. É nesta fase que as soluções desenvolvidas se tornam utilizáveis, e os modelos preditivos começam a gerar valor real para a organização. O modelo, uma vez em produção, pode ser utilizado para automatizar decisões ou auxiliar na tomada de decisões, baseando-se nas previsões realizadas. A implementação pode ser feita de diversas formas, de acordo com as necessidades do negócio, a infraestrutura disponível e o contexto específico.

Durante a implementação, algumas decisões sobre a forma como o problema vai ser abordado e sobre a infraestrutura da solução ser tomadas, como:

- Onde o modelo será executado: será em um servidor interno, em nuvem ou localmente?
- Qual a frequência das atualizações dos dados e das previsões: o modelo será executado em tempo real ou de forma periódica?
- Monitoramento e Manutenção: como o modelo será monitorado para garantir que ele continue performando bem ao longo do tempo, especialmente com dados novos.

A implementação também envolve a produtização do modelo em um ambiente de produção, onde ele pode começar a fazer previsões em tempo real ou em *batch*, dependendo do fluxo de trabalho desejado. Além disso, é importante garantir que as previsões sejam apresentadas de maneira compreensível e útil para os tomadores de decisão.

Existem diversas formas de colocar um modelo em produção, dentre elas :

- Implementação em Servidor Interno: processo de execução do modelo em servidores próprios da organização, ao invés de utilizar plataformas de nuvem ou externas.

- Implementação em Nuvem: processo de execução do modelo em servidores e infraestrutura externa.
- Implementação em *Batch*: processamento de grandes volumes de dados em intervalos de tempo pré-determinados, como uma vez por dia ou por semana, para gerar previsões baseadas nos dados mais recentes

3.7 ESTRATÉGIA DE DEFINIÇÃO DE RISCO

A definição de risco será baseada nas probabilidades atribuídas pelos três melhores modelos, sendo que a decisão final sobre o risco de desligamento será feita através da combinação das probabilidades de cada modelo. O processo de definição de risco segue os seguintes critérios:

- **Risco 1: se nenhum modelo apontar probabilidade maior que 50%**
  Quando todos os modelos indicarem probabilidades abaixo de 50%, isso será interpretado como um baixo risco de desligamento. Ou seja, a probabilidade de o colaborador deixar a organização voluntariamente é considerada mínima.
- **Risco 2: se apenas um modelo apontar probabilidade maior que 50% e menor que 60%**
  Caso apenas um modelo forneça uma probabilidade superior a 50% e inferior a 60%, isso indica um risco moderado baixo, pois o modelo está sugerindo uma chance significativa de desligamento, mas a incerteza é razoável, dado que apenas um modelo está apresentando esse valor.
- **Risco 3: se mais de um modelo apontar probabilidade maior que 50% e menor que 60%**
  Neste caso, a probabilidade de desligamento aumenta, já que mais de um modelo está sugerindo uma chance moderada, embora não muito alta, de o colaborador se desligar da organização. Isso representa um risco moderado, com alguma chance de ação preventiva ser necessária.
- **Risco 4: se mais de um modelo apontar probabilidade entre 60% e 70%**
  Aqui, a situação se torna mais grave. Com mais de um modelo indicando probabilidades acima de 60%, há uma forte indicação de que o colaborador possui um risco considerável de se desligar da empresa. O risco é classificado como alto, e estratégias de retenção podem ser necessárias.
- **Risco 5: se mais de um modelo apontar probabilidade maior que 70%**
  Quando mais de um modelo sugere uma probabilidade maior que 70%, o risco é classificado como muito alto. A alta concordância entre os modelos indica que

as chances de o colaborador deixar a empresa são substanciais, o que exige ações imediatas para tentar mitigar esse risco, como intervenções no processo de gestão de pessoas.

A estratégia de combinar as probabilidades dos modelos foi escolhida com base na ideia de que diferentes algoritmos podem capturar diferentes aspectos dos dados e, portanto, podem fornecer perspectivas valiosas para a previsão de desligamentos. Esse tipo de ensemble pode ajudar a melhorar a robustez da previsão, minimizando a influência de possíveis vieses ou falhas de um modelo individual. Essa abordagem leva em consideração tanto a probabilidade do modelo como as diferentes perspectivas oferecidas por múltiplos algoritmos, melhorando a precisão e a confiabilidade das decisões tomadas.

4 RESULTADOS E DISCUSSÕES

Esta seção apresenta os resultados obtidos ao longo do desenvolvimento deste projeto, com base na aplicação da metodologia CRISP-DM. Esta abordagem, amplamente reconhecida na ciência de dados, foi escolhida para estruturar o processo desde a compreensão do problema até a implementação do modelo preditivo. Os resultados aqui descritos refletem o impacto das decisões tomadas durante o desenvolvimento e oferecem uma análise detalhada do desempenho dos modelos testados, além das soluções adotadas para garantir a eficácia do sistema preditivo dentro do contexto organizacional.

Foi considerado um histórico de 2 anos de desligamentos voluntários para a construção da base de treino. Esse intervalo foi adotado para garantir uma maior confiabilidade nos dados e para afirmar que todas as variáveis utilizadas já estavam sendo coletadas à época.

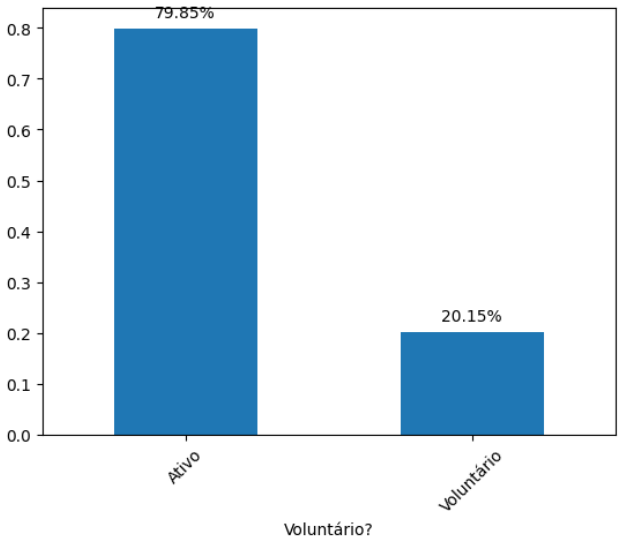
Assim, a base de treino, incluindo colaboradores ativos e desligamentos voluntários, contém cerca de 2000 amostras e pouco menos de 40 variáveis.

4.1 ANÁLISE DOS DADOS

A análise dos dados iniciou-se com a exploração das variáveis presentes no *dataset*. A distribuição das variáveis categóricas e da variável alvo foi analisada para entender o comportamento das variáveis e a estrutura do *dataset*.

A figura 6 apresenta a distribuição da variável alvo. Nela, nota-se desbalanceamento entre as categorias da variável.

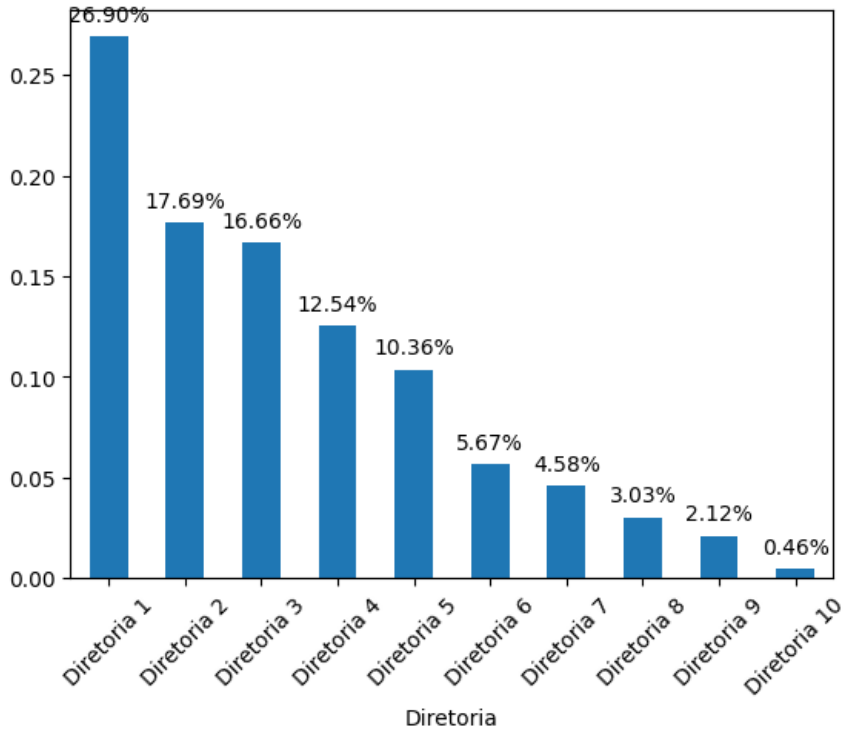
Figura 6 – Distribuição da variável alvo



Na Figura 7, tem-se a distribuição da variável diretoria. Os nomes das diretorias presentes na variável Diretoria foram anonimizados para manter a confidencialidade da

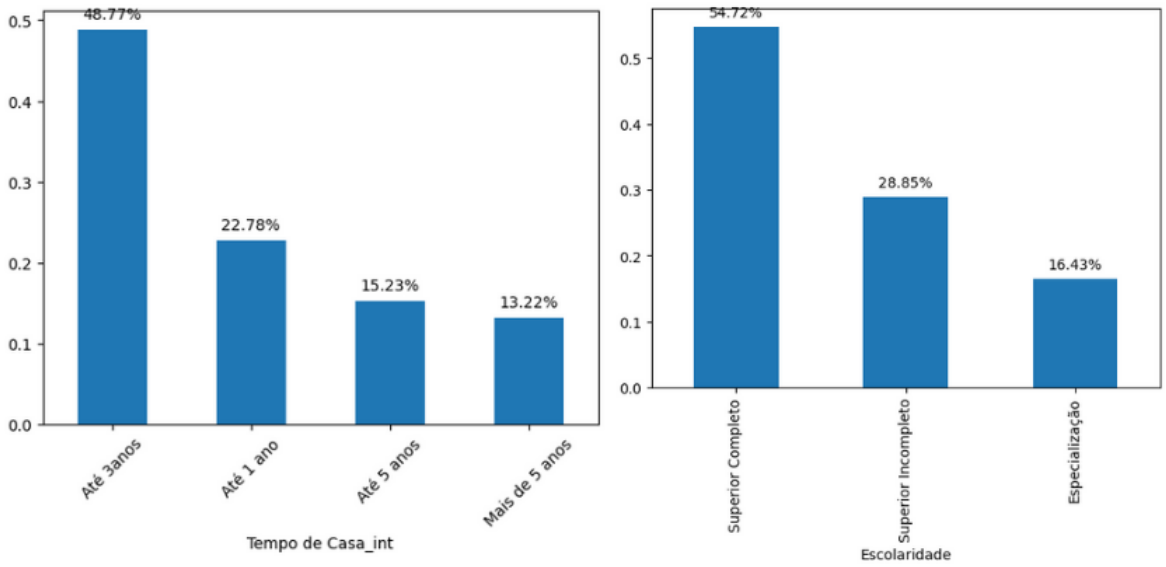
organização onde o projeto deste trabalho foi desenvolvido.

Figura 7 – Distribuição da variável de diretoria



A Figura 8 retrata a distribuição das variáveis de nível de escolaridade e de tempo de casa.

Figura 8 – Distribuição das variáveis de tempo de casa e escolaridade



Os valores da variável tempo de casa são calculados da seguinte forma:

- Para colaboradores ativos: a diferença entre a data de referência do conjunto de dados e da data de admissão.

- Para colaboradores Desligados: a diferença entre a data de desligamento e data de admissão

  Os valores da variável escolaridade são definidos da seguinte forma:

- Superior Incompleto: colaboradores que possuem superior incompleto ou ensino médio completo.
- Superior Completo: colaboradores que possuem ensino superior completo
- Especialização: colaboradores que possuem: Mestrado, Doutorado, especializações.

4.2 PREPARAÇÃO DOS DADOS

A preparação dos dados incluiu etapas para garantir que as informações estivessem prontas para o treinamento do modelo. A primeira etapa foi a limpeza dos dados, com a verificação da presença de valores faltantes. Foi atestado que não há a presença de dados faltantes, mostrando o sucesso na gestão dos dados e no processo de unificação e criação da base de treinamento. Tornando os modelos criados ainda mais robustos.

Além disso, foi realizada a codificação das variáveis categóricas utilizando o método manual, como exemplificado na Figura 9, e o método *Binary Enconding* para garantir que as variáveis fossem tratadas corretamente pelos algoritmos de aprendizado de máquina, que geralmente exigem entradas numéricas.

Figura 9 – Exemplo da função de codificação da variável fases da vida

```python
def fases_vida(fase:str) -> int():
    if fase == 'Até 30 anos':
        return 0
    elif fase == 'Dos 30 aos 40 anos':
        return 1
    elif fase == 'Dos 40 aos 50 anos':
        return 2
    else: return 3
```

Em seguida, criou-se o *pipeline* de pré-processamento das variáveis de entrada do modelo. Para que quando os modelos forem colocados em produção e caso haja dados faltantes nos dados novos, os modelos funcionem corretamente, utilizou-se a estratégia de preencher os valores faltantes com a média, para dados numéricos, e com o valor mais frequente em variáveis categóricas.

Figura 10 – Exemplo do *pipeline* de pré-processamento

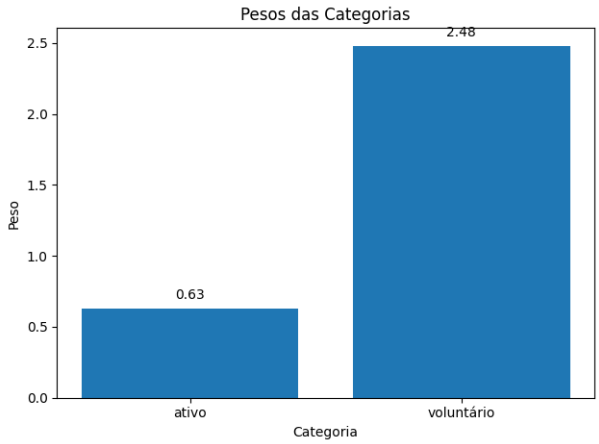
```python
colunas_cat = X.select_dtypes(include=['object']).columns
colunas_num = X.select_dtypes(exclude=['object']).columns


preprocessor = ColumnTransformer(
    transformers=[
        ('num', Pipeline([
            ('imputer', SimpleImputer(strategy='mean')),
            ('scaler', StandardScaler())
        ]), colunas_num),

        ('cat', Pipeline([
            ('imputer', SimpleImputer(strategy='most_frequent', fill_value='missing')),
            ('encoder', BinaryEncoder(drop_invariant=True))
        ]), colunas_cat)
    ])
```

4.3 MODELAGEM

4.3.1 Desbalanceamento da Variável Alvo

Como observado na figura 6, a variável alvo da modelagem apresenta um nível considerável de desbalanceamento. Como o objetivo é prever um evento binário, é comum que o conjunto de dados seja desbalanceado, no caso do problema deste trabalho, mais exemplos de colaboradores ativos do que aqueles que se desligaram voluntariamente. Combinada à relativamente baixa quantidade de amostras da base de dados, faz-se necessário tratar esse desbalanceamento. Foi utilizado o método *compute_class_weight* do *Scikit-learn* para calcular os pesos relativos de cada categoria (Scikit-learn, 2024c). Sendo eles: 0,63 para a classe majoritária, ativo. E 2,48 para a classe dos desligamentos voluntários.

Figura 11 – Pesos relativos das categorias da variável alvo



4.3.2 Configuração de bases e separação em treino e teste

Foi realizada uma análise de correlação entre as variáveis para identificar quais delas eram mais relevantes para a predição do desligamento voluntário. A partir dessa análise, selecionou-se as variáveis mais correlacionadas e construímos três configurações de base para o treinamento: uma com variáveis de correlação positiva, outra com variáveis de correlação negativa, e uma terceira utilizando o método *SelectKBest* para selecionar as variáveis mais significativas.

Em seguida, utilizou-se a função *train_test_split* do Scikit-learn (2024b) para fazer a separação do conjunto dos dados em treino e teste. Foi adotada a proporção de 70/30 dos dados em treino e teste. Nessa separação, foi utilizado como parâmetros os pesos apresentados na figura 11 e o parâmetro *Stratify* com a variável alvo como referência. Esse parâmetro faz com que ambos os conjuntos, treino e teste, mantenham a mesma proporção na variável alvo.

4.3.3 Treinamento dos Modelos

Os modelos foram treinados utilizando os seguintes algoritmos: *Random Forest, LightGBM* (LGBM), *CatBoost*, *Support Vector Machines*, Árvores Binárias e Regressão Logística.

Através de um *pipeline* de treinamento, garantimos que todas as etapas de pré-processamento, Figura 10, e treinamento aplicada de forma segura.

O *pipeline* de treinamento exemplificado na Figura 12 foi utilizado na configuração de base sem seleção prévia de variáveis. E será utilizado as 20 variáveis que apresentarem melhor resultado.

Figura 12 – Exemplo do *pipeline* de treinamento com *SelectKBest*

```
pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('select_kbest', SelectKBest(score_func= chi2, k=20)),
    ('classifier', RandomForestClassifier())
])
```

O *pipeline* da Figura 13 foi utilizado na configuração de base de variáveis de correlação positiva e na configuração de variáveis de correlação negativa.

Figura 13 – Exemplo do *pipeline* de treinamento

```
pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier())
])
```

4.4 AVALIAÇÃO DO MODELO

4.4.1 Métrica de Avaliação e Tipo de Erro

Dada a natureza do problema abordado neste trabalho, optou-se por priorizar a minimização do erro Tipo II. O custo de perder um bom colaborador por classificá-lo com baixo risco de desligamento é mais elevado do que o custo de mantê-lo, mesmo quando ele não apresenta propensão a sair. Para isso, foi utilizada a métrica *F-beta score* com $\beta = 2$.

A tabela a seguir apresenta os resultados da métrica *F-beta score* para os 6 modelos treinados, considerando as 3 configurações de base diferentes:

Tabela 1 – Resultados da métrica *F-beta score* para os modelos treinados

Modelo	**Variáveis de correlação positiva**	**Variáveis de correlação negativa**	***SelectKBest***
Random Forest	0.783	0.755	0.821
LightGBM	0.804	0.777	0.812
CatBoost	0.791	0.749	0.802
SVM	0.726	0.701	0.755
Árvore Binária	0.698	0.682	0.716
Regressão Logística	0.701	0.712	0.743

Foi observado que os modelos *Random Forest, LGBM e CatBoost* foram os mais eficazes, com *F-beta scores* mais altos, especialmente na configuração utilizando o *SelectKBest*.

4.4.2 Melhoria de Hiperparâmetros

Para melhorar ainda mais esses modelos, utilizou-se o *Optuna*, para otimizar os parâmetros dos três melhores modelos: *Random Forest, LGBM e CatBoost*. O processo de *tuning* foi realizado de forma automatizada, o que resultou em uma melhoria significativa no desempenho dos modelos.

Os hiperparâmetros ajustados para cada modelo são apresentados abaixo:

- **Random Forest:** número de estimadores = 350, profundidade máxima = 3, Número mínimo de amostras necessárias em uma folha = 2, Número mínimo de

Figura 14 – Exemplo da função de otimização do *Optuna*

```python
def objective(trial):

    # Definir os hiperparâmetros a serem otimizados
    n_estimators = trial.suggest_int('n_estimators', 200, 2000, step=50)
    max_depth = trial.suggest_int('max_depth', 3, 10)
    min_samples_split = trial.suggest_int('min_samples_split', 2, 10)
    criterion = 'gini'
    min_samples_leaf = trial.suggest_int('min_samples_leaf', 2, 10)
    cv = StratifiedKFold(n_splits = 5, random_state = SEED, shuffle = True)
    # Criar o modelo de Random Forest com os hiperparâmetros sugeridos pelo Optuna
    rfc = RandomForestClassifier(n_estimators=n_estimators,
                                 max_depth=max_depth,
                                 min_samples_split=min_samples_split,
                                 criterion='gini',
                                 random_state=SEED)

    # Treinar o modelo
    rfc.fit(X_train_enc, y_train_ros)

    # Treinando o modelo
    y_pred_train = rfc.predict(X_train_enc)


    # Calculando a performance
    fbeta = fbeta_score(y_train, y_pred_train, beta=2)


    return  fbeta
```

amostras necessárias para dividir um nó interno = 9, critério = "gini"

- **LGBM:** número de árvores = 280, taxa de aprendizado = 0.011374418646603454, profundidade máxima = 2, número de folhas = 35
- **CatBoost:** número de iterações = 573, taxa de aprendizado = 0.006506569982300885

Após a otimização dos hiperparâmetros, as novas métricas de teste foram:

Tabela 2 – Resultados da métrica *F-beta score* para os modelos otimizados

Modelo	***F-beta score***
Random Forest	0.853
LGBM	0.847
CatBoost	0.822

4.4.3 Importância das Variáveis

Analisou-se as as variáveis mais importantes de cada um dos modelos utilizando o método *Feature_ Importance*. Os variáveis mais importantes para os modelos são:

Importância das variáveis do modelo *Random Forest*:

- Posicionamento: % do salário do colaborador na faixa salarial do cargo
- Afastamento: variável binária de se o colaborador solicitou afastamento nos ultimos 12 meses
- Desempenho: nota de desempenho organizacional
- Cargo gestao: variável binária de se o colaborador ocupa um cargo de gestão ou liderança
- SMs: salário do colaborador em salário mínimo (R$ 1412,00)

Importância das variáveis do modelo *CatBoost*:

- Tempo de Casa: tempo de serviço do colaborador na empresa.
- Filhos: variável binária de se o colaborador possui filhos.
- SMs: salário do colaborador em salário mínimo (R$ 1412,00)
- Afastamento: variável binária de se o colaborador ocupa um cargo de gestão ou liderança
- Desempenho: nota de desempenho organizacional

Importância das variáveis do modelo *LihgtGBM*:

- Posicionamento: % do salário do colaborador na faixa salarial do cargo
- Afastamento: variável binária de se o colaborador solicitou afastamento nos ultimos 12 meses
- Ultima Promoção: tempo desde a ultima promoção
- Desempenho: nota de desempenho organizacional
- Tempo de Casa: tempo de serviço do colaborador na empresa.

4.5 IMPLEMENTAÇÃO

O modelo final foi implementado utilizando uma abordagem de batch, sendo executado a cada dois meses, com dados atualizados, garantindo que as previsões reflitam as informações mais recentes disponíveis. A execução do modelo será realizada na infraestrutura de servidores internos da organização, proporcionando controle e segurança no processamento e armazenamento, dado a alta sensibilidade dos dados.

4.6 PREDIÇÃO E ESTRATÉGIA DE RISCO

A estratégia de risco foi definida com base na combinação das probabilidades dos três melhores modelos. A classificação de risco foi dividida em cinco níveis, conforme descrito anteriormente. A Figura 15 exemplifica criação dos níveis de risco de acordo com as probabilidades de saída dos colaboradores, resultantes dos 3 modelos.

Figura 15 – *Loop for* de criação da categorização dos riscos

```python
for index, row in df_real.iterrows():
    probas = row['Predict_ctb_PROBA'], row['Predict_rf_PROBA'], row['Predict_lgbm_PROBA']
    menor_60 = [x for x in probas if x >= 0.5 and x <0.6]
    menor_70 = [x for x in probas if x < 0.7 and x >= 0.6]
    maior_70 = [x for x in probas if x >= 0.7]
    # 2-3 modelos com Proba X >= 70%
    if len(maior_70) > 1:
        df_real.at[index, 'Risco'] = 5
        continue
    # 2-3 modelos com 60% <= X < 70%
    if len(menor_70) > 1:
        df_real.at[index, 'Risco'] = 4
        continue
    # 2-3 modelos com 50% <= X < 60%
    if len(menor_60) > 1:
        df_real.at[index, 'Risco'] = 3
        continue
    # Somente um modelo 50% <= X < 60%
    if len(menor_60) == 1:
        df_real.at[index, 'Risco'] = 2
        continue
    df_real.at[index, 'Risco'] = 1
```

A seguir, a distribuição de risco para a predição dos colaboradores ativos, com base nas probabilidades obtidas pelos modelos, para o mês de referência, que ocorre seis meses após o término do intervalo de dados utilizados no treinamento:

- **Risco 1:** 91% dos colaboradores
- **Risco 2:** 7,2% dos colaboradores
- **Risco 3:** 1,2% dos colaboradores
- **Risco 4:** 0,51% dos colaboradores
- **Risco 5:** 0,09% dos colaboradores

Essa abordagem de combinar os resultados dos modelos tem como objetivo reduzir a incerteza associada à predição e aumentar a robustez da classificação do risco de desligamento. Em cenários de risco incerto, o consenso entre múltiplos modelos pode fornecer uma visão mais confiável da probabilidade de saída do colaborador.

A análise da distribuição dos níveis de risco indicou que a maior parte dos colaboradores foi classificada no Nível 1, o que sinaliza uma baixa probabilidade de desligamento para esses casos. Esse resultado estava alinhado com as expectativas, já que o histórico da organização mostrava que a maioria dos colaboradores tende a permanecer na empresa. Assim, o Nível 1 reflete a tendência observada nos dados anteriores, em consonância com a taxa de desligamento voluntário, que é inferior a 10%.

4.7 SÍNTESE DA ANALISE

A análise dos resultados dos modelos e da classificação dos níveis de risco, demonstram que o conjunto de dados coletados é capaz de explicar o fenômeno do desligamento voluntário e se mostra capaz de prever e antecipar as saídas. As abordagens escolhidas e decisões técnicas adotadas se mostraram eficazes na potencialização do uso dos dados e solucionador do problema.

Além de ser uma ferramenta importante para a gestão de recursos humanos e para antecipar a necessidade de ações preventivas para retenção de talentos, o nível de risco de saída também desempenha um papel crucial nas estratégias de remuneração da empresa. Este indicador é usado para tomar decisões sobre ajustes salariais e bonificações.

Por exemplo, colaboradores classificados nos níveis de risco mais altos e que performam bem podem ser alvo de ações de retenção mais imediatas e incisivas, como aumentos salariais, bônus ou ofertas de benefícios especiais. Estas ações são voltadas para diminuir a probabilidade de saída desses colaboradores, aumentando seu engajamento e, consequentemente, sua permanência na empresa.

Assim, a definição do risco de saída se torna um forte critério na construção de pacotes de remuneração personalizados, alinhados aos objetivos de retenção de talento da empresa, além de garantir que os recursos da organização sejam direcionados de forma estratégica para os colaboradores com maior risco de saída.

5 CONSIDERAÇÕES FINAIS

Esta pesquisa teve como objetivo desenvolver um modelo preditivo para prever o desligamento voluntário de colaboradores em uma organização, utilizando a metodologia CRISP-DM *(Cross-Industry Standard Process for Data Mining)*. Ao seguir as etapas dessa metodologia, o estudo foi conduzido de maneira estruturada, desde o entendimento do negócio e a coleta dos dados até a implementação e avaliação dos modelos. Os resultados obtidos oferecem informações para as práticas de gestão de recursos humanos, contribuindo para a retenção de talentos e decisões estratégicas.

A primeira etapa, o entendimento do negócio, foi fundamental para compreender as necessidades da organização e como a previsão de desligamento pode ser crucial para o planejamento e gestão do capital humano. A definição clara dos objetivos do negócio orientou a coleta de dados e a construção dos modelos preditivos. Na preparação dos dados, realizou-se a limpeza e a análise das variáveis mais relevantes, como características demográficas, de desempenho e organizacionais, que influenciam diretamente o risco de desligamento.

Na fase de modelagem, diferentes algoritmos de aprendizado de máquinas, como *Random Forest, LightGBM e CatBoost*, foram treinados. Técnicas como o *SelectKBest* foram empregadas para selecionar as variáveis mais significativas. A combinação de modelos (ensemble) se revelou eficaz, utilizando a probabilidade predita pelos três melhores modelos para determinar o risco de desligamento. A definição de uma estratégia de risco, com base em diferentes limiares de probabilidade, proporcionou uma visão mais robusta e confiável sobre o comportamento dos colaboradores.

Durante a avaliação do modelo, a prioridade foi dada aos erros do tipo II, que são mais críticos no contexto de desligamento, pois um falso negativo pode resultar na negligência de um colaborador com alto risco de saída. A escolha do *F-beta score*, com beta = 2, permitiu ajustar o modelo para minimizar esse erro. A utilização do *Optuna* para tunar os hiperparâmetros dos modelos selecionados foi fundamental para otimizar o desempenho e garantir a eficiência na aplicação do modelo em um ambiente de produção.

A implementação do modelo, adotando uma estratégia *batch*, mostrou-se adequada para a organização, permitindo atualizações periódicas a cada dois meses. A aplicação do modelo à base de dados de colaboradores ativos para prever desligamentos em agosto foi eficaz, com a distribuição dos níveis de risco alinhada aos padrões históricos de desligamento, validando a eficácia do modelo.

5.1 IMPLICAÇÕES TEÓRICAS

Uma das contribuições teóricas mais significativas deste estudo é a sugestão de uma abordagem multimodal para enfrentar o problema dos desligamentos voluntários. Historicamente, a rotatividade de funcionários tem sido examinada sob uma única perspectiva

ou modelo. Contudo, este estudo propõe que a união de diferentes abordagens, como a teoria do suporte organizacional, a troca líder-membro e as teorias sobre satisfação no trabalho, pode proporcionar uma visão mais completa dos fatores que afetam as decisões de saída. Dessa forma, combinando dados que integram essas abordagens. Essa perspectiva integrada permite considerar tanto os elementos individuais quanto os organizacionais, resultando em um modelo teórico mais sólido e dinâmico. Ao combinar múltiplos fatores, torna-se possível prever o comportamento dos colaboradores com maior precisão e oferecer soluções mais eficazes para reduzir a rotatividade.

Este trabalho reforça a importância da integração entre a área de dados e a gestão de pessoas, evidenciando o papel crescente da análise preditiva na transformação das práticas de recursos humanos e na criação de ambientes de trabalho mais eficientes e satisfatórios para os colaboradores. Além da combinação das probabilidades resultantes de diferentes modelos para classificação do risco de perda do colaborador.

5.2 IMPLICAÇÕES GERENCIAIS

No contexto gerencial, este estudo apresenta relevantes implicações para a tomada de decisões estratégicas dentro da organização. A implementação do modelo desenvolvido aumenta a precisão nas decisões relacionadas à gestão de pessoas, possibilitando que os gestores detectem precocemente sinais de desengajamento. Com isso, é viável criar estratégias de retenção mais personalizadas e eficazes, favorecendo um ambiente que atenda melhor às necessidades e expectativas dos funcionários.

Adicionalmente, a capacidade de prever saídas representa uma vantagem considerável que a empresa pode conquistar ao utilizar o modelo. Ter essa informação de forma antecipada, permite que os gestores se preparem para a sucessão de cargos e elaborem planos de ação que assegurem a continuidade das operações. Essa abordagem não só reduz os efeitos adversos da rotatividade, mas também fortalece a cultura organizacional e o capital humano da empresa.

A classificação de cada colaborador em um nível de risco torna mais fácil a tomada de decisão sobre qual estratégia deverá ser adotada para a retenção.

5.3 LIMITAÇÕES DA PESQUISA E PROPOSTAS DE TRABALHOS FUTUROS

Embora a pesquisa tenha demonstrado resultados significativos, é crucial mencionar algumas restrições. Primeiramente, fatores externos que não podem ser controlados, como alterações econômicas, inovações tecnológicas ou variações no mercado de trabalho, afetam as taxas de desligamento de forma direta, mas são muito difíceis de serem capturados em dados para contribuir na base de treino.

Outra limitação da pesquisa reside no tamanho relativamente pequeno do histórico de dados, o que limita a plena generalização dos resultados. Além disso, toda e qualquer

implementação de variável no conjunto de treinamento, necessita que já venha sendo coletado no mesmo período do histórico considerado ou essa nova variável limitará o intervalo considerado para o conjunto de treino.

É importante ressaltar que, embora não seja esse o objetivo, os resultados das pesquisas sobre esse tipo de problema podem acabar sendo utilizados como um critério ou barreira na seleção de novos colaboradores. Por isso, é extremamente necessário um controle rigoroso nas pessoas que tem acesso aos resultados desse tipo de pesquisa e a conscientização de como a explicação desse fenômeno deve ser utilizada para retenção e entendimento.

Para mitigar tais limitações, sugere-se a estruturação de um plano de coleta de dados externos à organização. Onde possa capturar as tendências de mercado e da sociedade e que esses dados sejam incluídos no conjunto de treinamento.

Além disso, o uso de técnicas de aprendizado profundo pode ser uma boa solução para contornar a pequena amostra de dados. Essas técnicas são capazes de encontrar padrões em pequenos volumes de dados e entregar bons resultados.

Outra proposta de continuação dessa pesquisa é a predição do motivo do desligamento com base nas respostas coletadas nas entrevistas de desligamento. Dessa forma, ao invés de ter uma predição binária, sai ou não sai, associando a um nível de risco, se predirá o motivador da saída. Tornando possível otimizar a estratégia de retenção.

REFERÊNCIAS

AKIBA, T. et al. Optuna: A next-generation hyperparameter optimization framework. In: **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**. [S.l.: s.n.], 2019. p. 2623–2631. Citado 2 vezes nas páginas 17 e 39.

ALTMANN, André et al. Permutation importance: a corrected feature importance measure. **Bioinformatics**, Oxford University Press, v. 26, n. 10, p. 1340–1347, 2010. Citado na página 40.

ANITHA, Jagannathan. Determinants of employee engagement and their impact on employee performance. **International journal of productivity and performance management**, Emerald Group Publishing Limited, v. 63, n. 3, p. 308–323, 2014. Citado na página 27.

BERGSTRA, J. et al. Random search for hyper-parameter optimization. **Journal of Machine Learning Research**, v. 13, p. 281–305, 2012. Citado na página 39.

BIANCHI, Fernanda Buffon. People analytics: previsão de desligamentos por meio das técnicas de regressão logística e análise de sobrevivência. 2023. Citado na página 27.

BILLS, M. A. Social status of the clerical worker and his permanence on the job. **Journal of Applied Psychology**, v. 9, n. 4, p. 424–427, 1925. Citado na página 20.

BOUDREAU, John W; BERGER, Chris J. Decision-theoretic utility analysis applied to employee separations and acquisitions. **Journal of Applied Psychology**, American Psychological Association, v. 70, n. 3, p. 581, 1985. Citado na página 23.

BREIMAN, Leo. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001. Citado na página 35.

CASCIO, Wayne F. **Costing human resources**. [S.l.]: South-Western Educational Publishing Boston, MA, 1991. v. 21. Citado na página 23.

CHAPMAN, Pete et al. Crisp-dm 1.0: Step-by-step data mining guide. **SPSS inc**, v. 9, n. 13, p. 1–73, 2000. Citado na página 29.

CHARLWOOD, Andy et al. Why hr is set to fail the big data challenge. **LSE Business Review**, London School of Economics and Political Science, 2016. Citado na página 20.

CHEN, Yafeng et al. K-means clustering method based on nearest-neighbor density matrix for customer electricity behavior analysis. **International Journal of Electrical Power & Energy Systems**, Elsevier, v. 161, p. 110165, 2024. Citado na página 26.

CHIAVENATO, I. **Gestão de Pessoas: E o Novo Papel dos Recursos Humanos nas Organizações**. 9ª reimpressão. ed. Rio de Janeiro: Elsevier, 2004. Citado na página 20.

DAS, Kajaree; BEHERA, Rabi Narayan. A survey on machine learning: concept, algorithms and applications. **International Journal of Innovative Research in Computer and Communication Engineering**, v. 5, n. 2, p. 1301–1309, 2017. Citado na página 17.

DEPIERI, Odemir. **Tipos de aprendizado em machine learning**. 2023. Acesso em: 06 Dezembro 2024. Disponível em: https://www.dataviking.com.br/post/tipos-de-aprendizado-em-machine-learning. Citado 3 vezes nas páginas 25, 26 e 27.

DOROGUSH, Anna Veronika; ERSHOV, Vasily; GULIN, Andrey. Catboost: gradient boosting with categorical features support. **arXiv preprint arXiv:1810.11363**, 2018. Citado na página 35.

FAGGELLA, D. **Machine Learning in Human Resources – Applications and Trends**. 2019. https://emerj.com/ai-sector-overviews/machine-learning-in-human-resources/. Acesso em: 2024-12-05. Citado na página 27.

FENG, Yinwei et al. Prediction of the severity of marine accidents using improved machine learning. **Transportation Research Part E: Logistics and Transportation Review**, Elsevier, v. 188, p. 103647, 2024. Citado na página 25.

FERREIRA, Mário César; FREIRE, Odaléa Novais. Carga de trabalho e rotatividade na função de frentista. **Revista de administração contemporânea**, SciELO Brasil, v. 5, p. 175–200, 2001. Citado na página 17.

FLAMHOLTZ, Eric G; DAS, Tapan Kumar; TSUI, Anne S. Toward an integrative framework of organizational control. **Accounting, organizations and society**, Elsevier, v. 10, n. 1, p. 35–50, 1985. Citado na página 23.

GARG, Swati et al. A review of machine learning applications in human resource management. **International Journal of Productivity and Performance Management**, Emerald Publishing Limited, v. 71, n. 5, p. 1590–1610, 2022. Citado 2 vezes nas páginas 27 e 28.

GUEDES, Rodolfo Lemos. Análise preditiva de recursos humanos em uma indústria de celulose: identificando riscos de saída voluntária de colaboradores. 2024. Citado na página 27.

GUYON, Isabelle et al. Gene selection for cancer classification using support vector machines. **Machine learning**, Springer, v. 46, p. 389–422, 2002. Citado na página 34.

HERZBERG, Frederick. Work and the nature of man. **World**, 1966. Citado na página 22.

HOM, P. W.; GRIFFETH, R. W. **Employee turnover**. Cincinnati, OH: South-Western College Publishing, 1995. Citado 2 vezes nas páginas 20 e 22.

HOM, Peter W et al. One hundred years of employee turnover theory and research. **Journal of applied psychology**, American Psychological Association, v. 102, n. 3, p. 530, 2017. Citado 3 vezes nas páginas 20, 21 e 22.

ISSON, Jean Paul; HARRIOTT, Jesse S. **People analytics in the era of big data: Changing the way you attract, acquire, develop, and retain talent**. [S.l.]: John Wiley & Sons, 2016. Citado 3 vezes nas páginas 17, 19 e 21.

JR, David W Hosmer; LEMESHOW, Stanley; STURDIVANT, Rodney X. **Applied logistic regression**. [S.l.]: John Wiley & Sons, 2013. Citado na página 35.

KE, Guolin et al. Lightgbm: A highly efficient gradient boosting decision tree. **Advances in neural information processing systems**, v. 30, 2017. Citado na página 35.

KREITNER, R.; KINICKI, A. **Comportamento Organizacional**. 9ª edição. ed. São Paulo: McGraw-Hill, 2013. Citado na página 22.

LEE, Hyung-Chul; JUNG, Chul-Woo. Anesthesia research in the artificial intelligence era. **Anesthesia and Pain Medicine**, v. 13, p. 248–255, 07 2018. Citado na página 24.

LI, Ni et al. Human performance modeling for manufacturing based on an improved knn algorithm. **The International Journal of Advanced Manufacturing Technology**, Springer, v. 84, p. 473–483, 2016. Citado na página 27.

LIAW, A.; WIENER, M. Classification and regression by randomforest. **R News**, v. 2, p. 18–22, 2002. Citado na página 40.

LIU, Nanxi; KWONG, SCM; MOHAMMADI, Alireza. The impact of colleague departures on employee turnover intentions: A study on chinese enterprises. **Journal of International Business and Management**, v. 7, n. 3, p. 1–15, 2024. Citado na página 17.

LORENA, Ana Carolina; CARVALHO, André CPLF De. Uma introdução às support vector machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007. Citado na página 35.

LUNDBERG, Scott. A unified approach to interpreting model predictions. **arXiv preprint arXiv:1705.07874**, 2017. Citado na página 40.

MARLER, Janet H; BOUDREAU, John W. An evidence-based review of hr analytics. **The International Journal of Human Resource Management**, Taylor & Francis, v. 28, n. 1, p. 3–26, 2017. Citado na página 19.

MATHIEU, Cynthia et al. The role of supervisory behavior, job satisfaction and organizational commitment on employee turnover. **Journal of Management & Organization**, Cambridge University Press, v. 22, n. 1, p. 113–129, 2016. Citado na página 20.

MENDONÇA, Afonso Paulo Albuquerque de et al. A tecnologia atrelada ao resultado-recursos humanos frente as novas posturas e atribuições. **Razão Contábil e Finanças**, v. 8, n. 2, 2017. Citado 2 vezes nas páginas 26 e 27.

MITCHELL, Tom M; MITCHELL, Tom M. **Machine learning**. [S.l.]: McGraw-hill New York, 1997. v. 1. Citado na página 24.

MOBLEY, William H. Intermediate linkages in the relationship between job satisfaction and employee turnover. **Journal of applied psychology**, American Psychological Association, v. 62, n. 2, p. 237, 1977. Citado 2 vezes nas páginas 20 e 22.

MOBLEY, W. H. **Employee Turnover: Causes, Consequences and Control**. Reading, MA: Addison-Wesley, 1982. Citado 2 vezes nas páginas 22 e 23.

MOREIRA, Fabiano Greter; NANTES, Luana da Silva. Fatores que influenciam na rotatividade de pessoal nas organizações:: Um estudo bibliográfico. **Revista Estudos e Pesquisas em Administração**, v. 8, n. 1, 2024. Citado na página 17.

MOZUMDER, Md Abu Sufian et al. Optimizing customer segmentation in the banking sector: A comparative analysis of machine learning algorithms. **Journal of Computer Science and Technology Studies**, v. 6, n. 4, p. 01–07, 2024. Citado na página 25.

MUJUMDAR, Aishwarya; VAIDEHI, Vb. Diabetes prediction using machine learning algorithms. **Procedia Computer Science**, Elsevier, v. 165, p. 292–299, 2019. Citado na página 25.

PÁDUA, Antonio Francisco Lima de Oliveira et al. Predição do risco de evasão dos cursos técnicos a distância do ifpi: uma aplicação da metodologia crisp-dm em dados educacionais. Universidade Federal Rural de Pernambuco, 2017. Citado na página 30.

PRICE, James L; MUELLER, Charles W. A causal model of turnover for nurses. **Academy of management journal**, Academy of Management Briarcliff Manor, NY 10510, v. 24, n. 3, p. 543–565, 1981. Citado na página 20.

QUINLAN, J. Ross. Induction of decision trees. **Machine learning**, Springer, v. 1, p. 81–106, 1986. Citado na página 35.

RIBEIRO, M.; SINGH, S.; GUESTRIN, C. Why should i trust you? explaining the predictions of any classifier. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2016. p. 1135–1144. Citado na página 41.

SCIKIT-LEARN. **sklearn.feature_selection.SelectKBest**. 2024. Acesso em: 2024-12-06. Disponível em: https://scikit-learn.org/dev/modules/generated/sklearn.feature_selection.SelectKBest.html. Citado na página 34.

SCIKIT-LEARN. **sklearn.model_selection.train_test_split**. 2024. Acesso em: 2024-12-06. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html. Citado na página 48.

SCIKIT-LEARN. **sklearn.utils.class_weight.compute_class_weight**. 2024. Acesso em: 2024-12-06. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html. Citado na página 47.

SEGER, Cedric. **An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing**. 2018. Citado na página 34.

SHEARER, Colin. The crisp-dm model: the new blueprint for data mining. **Journal of data warehousing**, THE DATA WAREHOUSE INSTITUTE, v. 5, n. 4, p. 13–22, 2000. Citado na página 30.

SILVA, Daniel Henrique Cordeiro; TIMO, Elisa Maria do Nascimento. Machine learning aplicado à atenção domiciliar para predição de condição de óbito. **Research, Society and Development**, v. 11, n. 14, p. e230111436078–e230111436078, 2022. Citado na página 30.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. Practical bayesian optimization of machine learning algorithms. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2012. v. 25. Citado na página 39.

VROOM, V. H. **Work and Motivation**. [S.l.]: Wiley, 1964. Citado na página 22.