MINISTÉRIO DA EDUCAÇÃO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Especialização em Ciência de Dados

Análise preditiva de recursos humanos em uma indústria de celulose: identificando riscos de saída voluntária de colaboradores

Rodolfo Lemos Guedes

João Monlevade, MG
2024

Rodolfo Lemos Guedes

Análise preditiva de recursos humanos em uma indústria de celulose: identificando riscos de saída voluntária de colaboradores

Trabalho de conclusão de curso apresentado ao curso de Ciência de Dados do Instituto de Ciências Exatas e Aplicadas da Universidade Federal de Ouro Preto, como parte dos requisitos necessários para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Janniele Aparecida Soares Araujo

João Monlevade, MG

2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

G924a Guedes, Rodolfo Lemos.
Análise preditiva de recursos humanos em uma indústria de celulose [manuscrito]: identificando riscos de saída voluntária de colaboradores. / Rodolfo Lemos Guedes. - 2024.
46 f.: il.: color., gráf., tab..

Orientadora: Profa. Dra. Janniele Aparecida Soares Araujo.
Produção Científica (Especialização). Universidade Federal de Ouro Preto. Departamento de Engenharia de Produção.

1. Algoritmos. 2. Análise de regressão. 3. Aprendizado do computador. 4. Controle preditivo. 5. Industria de celulose. 6. Logística. 7. Mineração de dados (Computação). 8. Recursos humanos. I. Araujo, Janniele Aparecida Soares. II. Universidade Federal de Ouro Preto. III. Título.

CDU 519.2:004.85

Bibliotecário(a) Responsável: Flavia Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Rodolfo Lemos Guedes

Análise preditiva de recursos humanos em uma indústria de celulose: identificando riscos de saída voluntária de colaboradores

Monografia apresentada ao Curso de Especialização em Ciência de Dados da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Especialista em Ciência dos Dados

Aprovada em 05 de Julho de 2024

Membros da banca

Dra. Janniele Aparecida Soares Araujo - Orientador(a) Universidade Federal de Ouro Preto
Me. Alexandre Magno de Sousa - Universidade Federal de Ouro Preto
Dr. Guilherme Luiz de Jesus - CENIBRA

Janniele Aparecida Soares Araujo, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 16 de agosto de 2024

Documento assinado eletronicamente por **Janniele Aparecida Soares Araujo**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 16/08/2024, às 19:01, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0761387** e o código CRC **036AE686**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0819 - www.ufop.br

Agradecimentos

Gostaria de expressar minha profunda gratidão, primeiramente a Deus, pela força e sabedoria concedidas ao longo desta jornada. À minha querida esposa Luiza, pelo apoio constante, paciência e encorajamento inabalável que tornaram possível a realização deste projeto. Ao meu amado filho Antônio, cuja presença e amor me motivaram a continuar e alcançar meus objetivos. À minha professora orientadora Janniele, pelo valioso conhecimento compartilhado, orientação e dedicação ao desenvolvimento deste projeto. Ao Vander Saldanha, por me proporcionar a oportunidade de realizar o curso de Ciência de Dados e por acreditar no meu potencial. Aos meus professores, por todo o ensinamento e suporte ao longo desta caminhada. Ao Douglas Oliveira, meu mentor, e ao Felipe Lannes, sponsor deste projeto. E ao meu pai Antônio Guedes, que não está mais presente, cujo legado me inspira e motiva a ser cada dia melhor. A todos, meu sincero muito obrigado.

Resumo

Este trabalho propõe uma abordagem preditiva para identificar riscos de desligamento voluntário de colaboradores em uma indústria de celulose, utilizando técnicas de aprendizado de máquina. Foram aplicados algoritmos de Regressão Logística, Floresta Aleatória e Árvore de Decisão para modelar os dados históricos de empregados. A análise incluiu a aplicação de validação cruzada para garantir a robustez dos modelos. Os resultados indicaram que os modelos de Regressão Logística e Floresta Aleatória apresentaram os melhores desempenhos, destacando-se pela combinação equilibrada entre precisão e *recall*. A análise das variáveis mais influentes sugere que fatores como tempo na empresa, idade e última movimentação são cruciais na previsão de desligamentos.

Palavras-chaves: Mineração de Dados, Aprendizado de Máquina, Previsão de Desligamento. Recursos Humanos. Validação Cruzada. Floresta Aleatória. Regressão Logística. Árvore de Decisão.

Abstract

This work proposes a predictive approach to identifying voluntary employee attrition risks in a pulp industry using machine learning techniques. Logistic Regression, Random Forest, and Decision Tree algorithms were applied to model the historical employee data. The analysis included cross-validation to ensure model robustness. Results indicated that the Logistic Regression and Random Forest models exhibited the best performance, notably due to their balanced combination of precision and recall. Analysis of the most influential variables suggests that factors such as time at the company, age, and last job change are crucial in predicting employee attrition.

Keywords: Data Mining. Machine Learning. Employee Attrition Prediction. Human Resources. Cross-Validation. Random Forest. Logistic Regression. Decision Tree.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ERP *Enterprise Resource Planning*

FPR *False Positives Rate*

GI *Gini Index*

KDD *Knowledge Discovery in Databases*

ML *Machine Learning*

PCA *Principal Component Analysis*

S4/HANA *SAP ERP over HANA*

SVM *Support Vector Machine*

TPR *True Positives Rate*

AED Análise Exploratória de Dados

CAGED Cadastro Geral de Empregados e Desempregados

GP Gestão de Pessoas

IA Inteligência Artificial

LGPD Lei Geral de Proteção de Dados Pessoais

RH Recursos Humanos

Sumário

1 Introdução

O contexto econômico contemporâneo é marcado por dinâmicas complexas que impactam diretamente o mercado de trabalho (FRANCO; DRUCK; SELIGMANN-SILVA, 2010). O crescimento alarmante no número de pedidos de demissão, conforme indicado pelos dados do Cadastro Geral de Empregados e Desempregados (CAGED), entre janeiro e agosto de 2023, apresenta-se como um desafio significativo a ser compreendido e abordado. Com um aumento de 45% em relação ao mesmo período de 2021, os 1,6 milhão de pedidos de demissão registrados representam uma parcela expressiva, atingindo 28,75% de todas as saídas formais do mercado de trabalho nesse intervalo (Ministério do Trabalho e Emprego, 2023). Segundo Santos e Cruz (2019) este fenômeno revela uma tendência preocupante e sugere a necessidade urgente de investigação para compreender os fatores subjacentes a esse aumento e desenvolver estratégias eficazes para mitigar suas consequências.

O problema em questão suscita uma série de questionamentos acerca das razões que levam os profissionais a solicitarem demissão em um período relativamente curto. Segundo Queiroz (2020) é fundamental analisar a fundo os motivos que impulsionam uma crescente onda de desligamentos, considerando fatores econômicos, sociais e individuais que possam estar interagindo de maneira complexa. Além disso, é imperativo compreender como esses pedidos de demissão afetam não apenas os indivíduos diretamente envolvidos, mas também as empresas e a economia como um todo.

Nesse sentido, o presente trabalho propõe a desvendar os elementos que contribuem para o aumento dos pedidos de demissão, buscando identificar padrões, tendências e características específicas desse fenômeno. Pretende-se, assim, construir um entendimento mais profundo sobre as variáveis envolvidas nesse processo e fornecer subsídios para o desenvolvimento de estratégias eficazes de gestão de recursos humanos, tanto para as organizações quanto para os próprios profissionais.

Especificamente na indústria que será objeto deste estudo, os pedidos de demissão têm se mostrado uma variável crítica. Em meados de 2021, a empresa observou um aumento significativo nos pedidos de desligamento voluntário, seguindo a tendência nacional. O índice de pedidos de demissão em relação ao total de desligamentos estava em ascensão , conforme ilustrado na Figura 1. Embora tenha havido uma redução nos índices mais recentemente, compreender as causas desses desligamentos continua a ser uma prioridade para a gestão de recursos humanos.

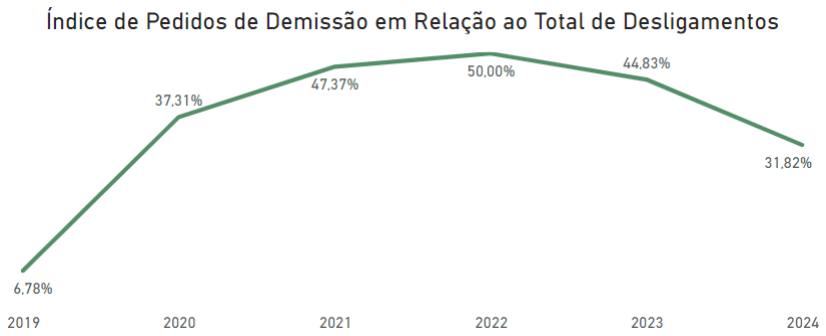


Figura 1 – Índice de Pedidos de Demissão em Relação ao Total de Desligamentos (2019-2024).

Fonte: Elaborado pelo autor

Ao compreender as raízes desse aumento significativo nos pedidos de demissão, como menciona Queiroz (2020), será possível desenvolver intervenções e políticas que visem a minimizar os impactos negativos, promovendo um ambiente de trabalho mais saudável e resiliente. Portanto, este trabalho não apenas aborda um problema atual e relevante no cenário laboral, mas também se propõe a contribuir para a construção de soluções concretas que beneficiem tanto empregadores quanto empregados. De acordo com Cintra (2018) a complexidade desse fenômeno demanda uma abordagem multidisciplinar, incorporando elementos da psicologia organizacional, economia do trabalho, sociologia e gestão estratégica.

Ao longo dos próximos capítulos, serão apresentados os fundamentos teóricos que norteiam este trabalho, a metodologia adotada para investigação, os resultados obtidos e suas implicações. Este estudo visa, portanto, preencher uma lacuna no entendimento do fenômeno em questão, além de fornecer uma base sólida para futuros trabalhos e intervenções práticas no âmbito do mercado de trabalho contemporâneo.

1.1 Objetivo geral

O objetivo geral deste trabalho é implementar e avaliar diferentes modelos de Aprendizado de Máquina (*Machine Learning* (*ML*)) para predizer a probabilidade de saída voluntária de colaboradores e fornecer *insights* que auxiliam nas estratégias de retenção de talentos da empresa.

1.1.1 Objetivos específicos

Para cumprimento do objetivo geral pretende-se alcançar os seguintes objetivos específicos:

- desenvolver e treinar modelos de *ML* utilizando algoritmos como regressão logística, árvores de decisão, floresta aleatória, utilizando conjuntos de dados históricos de saída voluntária de colaboradores;

- avaliar a eficácia dos modelos implementados por meio de métricas de desempenho, como precisão, acurácia, *recall* e *F1-score*, utilizando conjuntos de dados e de teste independentes;
- identificar os principais fatores e variáveis influentes na predição da probabilidade de saída de colaboradores, por meio de análises de importância de recursos e interpretabilidade dos modelos;
- ranquear os empregados, identificando a probabilidade de deixarem a empresa.

1.2 Contribuições

Este trabalho representa uma contribuição significativa para a literatura em gestão de recursos humanos, especialmente no contexto da indústria de celulose, por meio de sua abordagem na análise preditiva da saída voluntária de colaboradores. Primeiramente, destaca-se o avanço na literatura ao adotar uma metodologia que incorpora diferentes modelos de *ML*, permitindo uma análise aprofundada e integrada. Esta abordagem enriquece o entendimento das estratégias de retenção de talentos, e também amplia o espectro de técnicas utilizadas em estudos similares, estabelecendo um novo marco no campo.

Um dos aspectos centrais deste estudo é a identificação precisa de padrões que influenciam a saída de colaboradores. Ao investigar as variáveis mais relevantes e sua correlação com a decisão dos colaboradores de deixar a empresa, o trabalho fornece *insights* valiosos para a gestão de recursos humanos. Esta análise detalhada facilita a compreensão dos motivos por trás da rotatividade de pessoal, permitindo que gestores desenvolvam estratégias mais eficazes para antecipar e mitigar tais ocorrências.

Além disso, as conclusões derivadas deste estudo oferecem uma base sólida para a proposição de estratégias inovadoras de retenção de talentos. A compreensão aprofundada dos fatores-chave que influenciam a decisão dos colaboradores possibilita a criação de políticas internas mais assertivas e eficientes, visando a redução dos riscos associados à perda de colaboradores estratégicos.

1.3 Organização do Trabalho

Neste capítulo, é apresentada uma introdução ao tema do trabalho, fornecendo contexto e justificativa para o mesmo. O objetivo geral e os objetivos específicos são delineados, dando uma visão geral do que será abordado. Além disso, são destacadas as contribuições esperadas do trabalho para a literatura e para a prática no campo de estudo.

O Capítulo 2 refere-se a revisão da literatura que abrange uma análise detalhada das pesquisas anteriores relacionadas ao tema. São discutidos os principais conceitos, teorias e modelos relevantes para a predição de saída de colaboradores e estratégias de retenção de talentos. Esta seção busca contextualizar o trabalho dentro do panorama existente e identificar lacunas de conhecimento que motivam o trabalho.

No Capítulo 3, são detalhados os métodos e procedimentos utilizados para atingir os objetivos propostos. A descrição da abordagem de pesquisa, a coleta e preparação de dados, a escolha dos algoritmos de *ML*, os critérios de avaliação de desempenho e a análise estatística são abordados de maneira clara e sistemática. A metodologia fornece uma base para a replicação do estudo e a compreensão de suas limitações.

O Capítulo 4 apresenta e discute os resultados obtidos a partir da implementação e avaliação dos modelos de *ML*. Gráficos, tabelas e métricas de desempenho são utilizados para ilustrar a eficácia dos modelos na predição da saída voluntária de colaboradores. Além disso, são destacados *insights* sobre os fatores mais influentes na tomada de decisão dos modelos.

No Capítulo 5, por fim, são recapitulados os objetivos do estudo e as principais descobertas apresentadas nos capítulos anteriores. São discutidas as implicações práticas e teóricas dos resultados, bem como suas limitações. O trabalho é encerrado com sugestões para futuros trabalhos e recomendações para a aplicação prática dos *insights* obtidos, consolidando assim o impacto do estudo na área de gestão de recursos humanos.

2 Revisão da Literatura

A evolução constante da Gestão de Pessoas (GP) no cenário empresarial antecipa um futuro onde a predição de desligamentos desempenhará um papel central nas estratégias organizacionais. À medida que as dinâmicas do mercado exigem respostas ágeis e adaptativas, a gestão eficaz do capital humano se torna um diferencial competitivo essencial. A análise de dados emergirá como uma prática importante, fornecendo dados valiosos para a tomada de decisões.

A análise preditiva de desligamentos, como parte integrante das práticas de GP, não apenas mitigará custos associados à rotatividade, mas também contribuirá para a construção de ambientes de trabalho mais saudáveis e produtivos. O entendimento aprofundado dos motivos que levam ao desligamento será uma ferramenta essencial para moldar políticas internas e aprimorar as condições de trabalho, proporcionando às empresas uma vantagem sustentável no mercado competitivo.

2.1 *Turnover*

A rotatividade de empregados, conhecida como *turnover*, constitui um processo dinâmico e interativo entre a empresa e o mercado de trabalho (CHIAVENATO, 2003). Embora a movimentação de colaboradores seja considerada saudável em certa medida, a recorrência de desligamentos pode estar diretamente ligada à gestão ineficiente dos recursos humanos, resultante de políticas e processos internos inadequados (BOHLANDER; SNELL, 2016). Assim, a compreensão dos fatores-chave que influenciam o turnover e seus efeitos torna-se um objeto de estudo crucial.

A categorização dos desligamentos em voluntários e involuntários, conforme Chiavenato (2003), destaca a importância de discernir entre saídas por iniciativa do empregado e aquelas por iniciativa do empregador. Os desligamentos voluntários ocorrem quando as oportunidades fora do atual emprego superam os benefícios de permanecer nele. Já os desligamentos por iniciativa da empresa visam à substituição de empregados com desempenho aquém das expectativas, redução de custos internos ou readequação do quadro de colaboradores (CHIAVENATO, 2014).

A motivação por trás do abandono de empregos tem sido objeto de estudo relevante, dada sua significativa influência na sobrevivência das empresas e nas análises preditivas (RUBENSTEIN *et al.*, 2015). Ferramentas como pesquisas de clima e entrevistas de desligamentos têm se mostrado cruciais para identificar motivos que impactam as taxas de desligamento (CHIAVENATO, 2014).

Carless (2005) propõem uma análise multifacetada das variáveis associadas ao *turnover*, categorizando-as em diferentes perspectivas. Desde atributos individuais do trabalhador até contextos organizacionais e fatores externos, as variáveis abrangem aspectos como envolvimento no trabalho, condições pessoais e comportamento do trabalhador.

A rotatividade de pessoal não apenas representa um desafio organizacional, mas também implica custos substanciais para as empresas, exigindo uma monitorização constante. Os custos associados à reposição de empregados abrangem recrutamento, seleção, treinamento e os próprios processos de desligamento, enfatizando a necessidade de um gerenciamento proativo desse indicador (CHIAVENATO, 2003).

Diante do exposto, observa-se que os desligamentos e a redução da rotatividade de empregados emergem como imperativos estratégicos para as organizações, visando a eficiência operacional e a sustentabilidade financeira, e também a construção de ambientes de trabalho saudáveis, motivadores e propícios ao desenvolvimento profissional.

2.2 *People Analytics*

A área de gestão de pessoas é responsável pela execução de todas as atividades ligadas ao capital humano de uma organização, incluindo processos de seleção e contratação, definição de salários e benefícios, avaliações de ambiente organizacional e términos de contrato, entre outros. Segundo (BERSIN, 2016), o capital humano é considerado o recurso mais crucial para as empresas, sendo um elemento chave para a competitividade no mercado. Diante dessa realidade, tornou-se essencial para a área de Recursos Humanos (RH) evoluir de uma função meramente administrativa para um papel mais estratégico dentro da organização. É neste cenário que o conceito de *People Analytics* ganha destaque, representando um agrupamento de técnicas e processos, apoiados por tecnologias avançadas, que utilizam análises descritivas, visuais e estatísticas para analisar informações relacionadas aos empregados e aos procedimentos de RH (MARLER; BOUDREAU, 2016). Este movimento em direção a uma cultura baseada em dados permite sua aplicação em diversos procedimentos do setor, como na criação de perfis ideais para vagas disponíveis, em modelagens para definir políticas salariais, na geração de visualizações de métricas e indicadores para comparar o desempenho organizacional com o do mercado, além de modelagens preditivas que visam reduzir o *turnover* de colaboradores.

De acordo com (BERSIN, 2016), existem quatro estágios de desenvolvimento em *People Analytics* dentro das organizações, ilustrados na Figura 2.

Figura 2 – Maturidade em People Analytics

Fonte: Adaptado e traduzido de (BERSIN, 2016).

Bersin (2016) observa que a maioria das empresas está nos estágios iniciais 1 e 2. No entanto, enfrentam-se desafios significativos para progredir aos níveis 3 e 4, que estão mais associados a uma contribuição estratégica ampliada por parte dos departamentos de RH, por meio do uso de análises e modelagens preditivas. Angrave *et al.* (2016) sugerem que um dos obstáculos para essa evolução é a lacuna de conhecimento: muitos profissionais de RH carecem de competências em análise de dados, e, de forma similar, equipes de análise muitas vezes não possuem familiaridade com os aspectos de RH. Portanto, torna-se claro o papel crucial de especialistas em *People Analytics* no cenário atual de RH, visando não apenas operacionalizar, mas também participar ativamente na formulação de estratégias empresariais, contribuindo assim para um valor agregado significativo à organização. Além disso, este trabalho pode auxiliar na transição da organização para os níveis 3 e 4 de maturidade em *People Analytics*, empregando modelagens preditivas para mover o RH de uma atuação reativa para uma mais proativa.

Deste modo, uma das questões mais relevantes no contexto de RH é a previsão da saída de colaboradores. Visto que a contratação de um novo colaborador pode levar a uma queda temporária na produtividade e envolve custos significativos com demissões e contratações (LUCENA, 2007). Estudos indicam que o custo médio da rotatividade é de 21% do salário anual do funcionário (BOUSHEY; GLYNN, 2012), sugerindo que, salvo quando a troca de colaboradores é estrategicamente vantajosa a longo prazo, os custos associados à rotatividade podem ter um impacto financeiro considerável na empresa.

2.3 *Machine Learning*

O *ML*, uma subárea da Inteligência Artificial (IA), capacita computadores a aprenderem e melhorarem a partir dos dados para resolver problemas (MITCHELL, 2017), sem serem explicitamente programados para tarefas específicas. Esta capacidade de extrair conhecimento e aprimorar algoritmos através da experiência com os dados permite fazer predições precisas ou melhorar o desempenho em uma vasta gama de aplicações (MITCHELL, 2017). O *ML* destaca-se especialmente em contextos onde problemas complexos requerem soluções que adaptam-se continuamente a novas informações, sem a necessidade de ajustes manuais extensivos.

2.3.1 Categorias de Sistemas de *Machine Learning*

Os sistemas de *ML* classificam-se em quatro categorias principais, baseadas na presença ou ausência de supervisão humana durante o treinamento: aprendizagem supervisionada, não supervisionada, semissupervisionada e por reforço.

Aprendizagem supervisionada é um método que envolve o treinamento do algoritmo com dados previamente etiquetados, ou seja, cada exemplo de treinamento é composto por um par de entrada (como características de um objeto) e a saída desejada (a etiqueta ou resultado). O objetivo é que o algoritmo aprenda a mapear as entradas para as saídas. Assim que o modelo é treinado, ele pode prever a saída para novos dados que nunca viu antes, baseando-se no conhecimento adquirido durante o treinamento (GéRON, 2019). Isso é particularmente útil em cenários onde a relação entre a entrada e a saída é bem definida, mas complexa demais para ser modelada explicitamente através de programação convencional. As tarefas comuns incluem:

- modelagem preditiva de classificação: Neste caso, o algoritmo aprende a classificar as entradas em categorias distintas, como diferenciar e-mails legítimos de spam ou identificar qual espécie animal uma imagem representa;
- modelagem preditiva de regressão: Aqui, o objetivo é prever um valor contínuo para novas entradas, como o preço de uma casa baseado em seu tamanho, localização e características, ou a temperatura de um dia futuro a partir de dados meteorológicos históricos;

Aprendizagem não supervisionada, diferentemente da aprendizagem supervisionada, lida com dados que não são etiquetados. Ou seja, o algoritmo tenta identificar padrões diretamente dos dados de entrada sem referência a saídas conhecidas. O desafio é descobrir a estrutura subjacente nos dados – agrupar objetos semelhantes, identificar padrões comuns ou reduzir a complexidade dos dados para análises futuras (GéRON, 2019). Aplicações típicas incluem:

- agrupamento (*clustering*): Identificação de conjuntos de exemplos similares dentro dos dados, útil para segmentação de mercado, organização de grandes bibliotecas de mídia, ou detecção de padrões de uso em grandes conjuntos de dados;

- detecção de anomalias: Localização de casos atípicos que desviam do padrão normal, importante para detecção de fraudes, falhas em sistemas ou doenças raras em diagnósticos médicos;
- redução de dimensionalidade: Simplificação dos dados ao reduzir o número de variáveis a serem consideradas, preservando tanto quanto possível a estrutura original dos dados, facilitando a visualização e análises subsequentes.

Aprendizagem semissupervisionada combina elementos das aprendizagens supervisionada e não supervisionada. Utiliza-se um pequeno conjunto de dados etiquetados em conjunto com um grande volume de dados não etiquetados. O objetivo é melhorar a eficiência e precisão do aprendizado, aproveitando ao máximo os dados disponíveis, mesmo quando as etiquetas são escassas ou caras de obter (GéRON, 2019).

Na aprendizagem por Reforço, o algoritmo (chamado de agente) aprende a tomar decisões através da experimentação e recebimento de *feedback* em forma de recompensas ou penalidades. O foco está em aprender a melhor política de ação, isto é, uma estratégia para escolher ações que maximizem a recompensa acumulada ao longo do tempo. Este método é amplamente utilizado em jogos, robótica, e para otimizar sistemas de recomendação (SUTTON; BARTO, 2018).

2.3.2 Técnicas

No processo da ciência de dados, a fase de modelagem é crucial para a seleção das metodologias de aprendizado de máquina mais eficientes na resolução do problema em questão. Neste segmento, discutiremos os modelos aplicados neste estudo e o funcionamento do mecanismo de aprendizado.

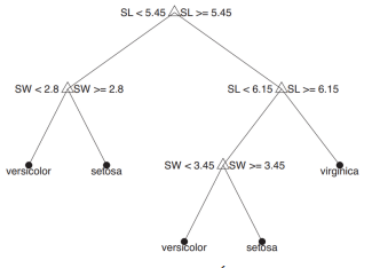
2.3.2.1 Árvores de Decisão

Empregando estruturas de árvores binárias, as árvore de decisão iniciam por um nó raiz, progredindo através de nós intermédios até atingir um nó terminal ou folha, que representa o resultado. Exceto pelas folhas, cada nó testa uma condição, onde as ramificações delineiam os possíveis desfechos das decisões, resultando em uma classificação ou um valor de regressão (MURPHY, 2012).

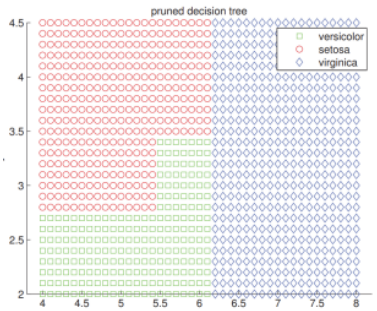
Para a escolha das condições nos nós intermédios, estratégias como ganho de informação, entropia e, predominantemente, o índice de impureza de Gini (*Gini Index* (*GI*)) são utilizadas (MURPHY, 2012). O *GI* de cada atributo é avaliado de maneira ponderada, considerando o impacto da condição sobre o conjunto de dados. A condição com o menor *GI* é selecionada, evidenciando a capacidade de segmentar os dados em grupos com maior homogeneidade.

A Figura 3a exemplifica uma árvore de decisão, evidenciando o processo de avaliação das condições a partir do nó raiz até a classificação no nó folha. A Figura 3b apresenta a disposição das classificações de uma árvore de decisão num espaço bidimensional, onde a especificidade do algoritmo resulta em divisões retangulares, limitando sua eficácia em conjuntos de dados de distribuições complexas.

Figura 3 – Exemplificação do funcionamento de árvores de decisão



(a) Representação de uma árvore de decisão



(b) Representação gráfica da predição em árvores de decisão

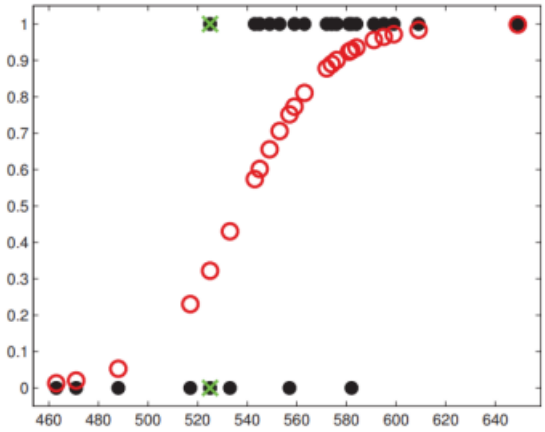
Fonte: Murphy (2012).

2.3.2.2 Regressão Logística

Diferentemente da regressão linear, que busca estabelecer uma linha que melhor se ajusta aos dados em um gráfico de dispersão, a regressão logística desenvolve uma curva em forma de S, conhecida como função logística, indicativa da probabilidade de um dado ser classificado em uma categoria específica (RUSSELL; NORVIG, 2002). A Figura 4 demonstra como a regressão logística modela essas probabilidades.

A figura apresenta um gráfico de dispersão com duas classes de dados: uma classe é representada por círculos vermelhos vazados e a outra por pontos pretos. A variável independente está no eixo horizontal, enquanto a variável dependente, que indica a probabilidade de pertencer a uma classe específica, está no eixo vertical. A curva em forma de S, característica da regressão logística, é evidenciada pela disposição dos círculos vermelhos que começam na base inferior do gráfico, aumentando a probabilidade conforme a variável independente aumenta, até atingir o platô próximo a 1.0 no eixo vertical. Além disso, há pontos pretos agrupados na base e no topo do gráfico, indicando classificações binárias distintas. Essa visualização ilustra claramente como a regressão logística pode separar duas classes diferentes com base nas probabilidades calculadas.

Figura 4 – Predição com Regressão Logística



Fonte: Murphy (2012).

2.3.2.3 Floresta Aleatória

A técnica da floresta aleatória para aprendizado de máquina é avançada e combina várias árvores de decisão para criar um modelo mais acurado e eficiente. Ela utiliza um método de reamostragem para criar subconjuntos do conjunto de dados original, que são usados para treinar diferentes árvores de decisão. Cada árvore é desenvolvida usando um subconjunto aleatório de características, ao invés de todas as características disponíveis. As previsões de cada árvore são então combinadas, geralmente calculando-se a média das previsões para problemas de regressão ou usando a votação majoritária para problemas de classificação (LIAW; WIENER, 2002).

Esta abordagem é eficaz porque ao construir árvores com amostras e subconjuntos de características diferentes, ela aumenta a diversidade entre as árvores no modelo, o que geralmente leva a um desempenho geral melhor e maior robustez, especialmente em cenários com grande variabilidade nos dados. Ajustar o número de árvores e o número de características consideradas em cada divisão são parâmetros importantes que podem ser ajustados para otimizar o desempenho do modelo.

A floresta aleatória é uma das técnicas mais poderosas para lidar com grandes conjuntos de dados e é altamente eficiente para lidar com variáveis de entrada de tipos diferentes (numéricas e categóricas) e muitos dados faltantes. É amplamente utilizada em diversos campos, como na biologia para classificação de espécies, no setor bancário para detecção de fraudes, e no comércio eletrônico para recomendação de produtos (LIAW; WIENER, 2002).

2.3.3 Análise de Componentes Principais

A *Principal Component Analysis (PCA)*, é uma das técnicas estatísticas multivariadas mais utilizadas para análise de dados em diversas áreas do conhecimento, devido à sua capacidade de sintetizar dados com base na correlação entre várias variáveis medidas. Esta técnica permite a análise de dados de forma reduzida, possibilitando a demonstração de resultados similares ou diferentes entre amostras em um determinado conjunto de dados. Além disso, a *PCA* elimina sobreposições e proporciona meios mais representativos desses dados, a partir de combinações de eixos das variáveis originais (LATTIN; CARROLL; GREEN, 2011).

A *PCA* facilita a compreensão de dados multivariados, permitindo a reorientação dos mesmos de forma que as primeiras dimensões concentrem a maior parte da variação existente, especialmente útil em presença de redundâncias nos dados (LATTIN; CARROLL; GREEN, 2011). A *PCA* envolve a determinação de autovalores e autovetores da matriz de correlação, simplificando a expressão dos dados através dos escores dos componentes principais.

O objetivo principal da *PCA* é explicar a estrutura da variância e covariância de um vetor aleatório, composto de *p*-variáveis aleatórias, por meio de combinações lineares das variáveis originais. Essas combinações lineares são chamadas de Componentes Principais e não são correlacionadas entre si (HONGYU; SANDANIELO; JUNIOR, 2016).

2.4 Ambiente Computacional

Os ambientes computacionais que serão utilizados nesta pesquisa para coleta, tratamento e modelagem dos dados, são:

• *SAP ERP over HANA* (*S4/HANA*) (versão 1909): sistema *Enterprise Resource Planning* (*ERP*) padrão da empresa onde estão armazenados os dados dos empregados em questão desta pesquisa;

• Google Colab: software de aplicação em nuvem da empresa Google disponibilizado em versão gratuita para desenvolvimento de algoritimos em Python (versão 3) com memória RAM do sistema de 12.7GB e 107.7GB de disco.

2.5 Trabalhos Relacionados

Chourey, Phulre e Mishra (2019), em sua pesquisa sobre previsão de rotatividade de empregados usando técnicas de *ML*, afirmaram que a rotatividade de empregados é um problema significativo destacado em todas as organizações. Há uma necessidade por métodos e algoritmos para prever a rotatividade de empregados usando técnicas de mineração de dados.

Norrman (2020) estudou a previsão de rotatividade de empregados com algoritmos de *ML* supervisionados em nível individual e seus efeitos em uma organização. No estudo, Norrman avaliou o modelo floresta aleatória, *Support Vector Machine* (*SVM*) e regressão logística múltipla e concluiu que nenhum dos métodos poderia prever a rotatividade de empregados com os conjuntos de dados testados. O estudo realizado também recomenda investigar modelos mais precisos para uso no mundo real.

Kane-Sellers (2007) usou análise de mineração de dados para investigar o impacto de inúmeras variáveis pessoais e profissionais na rotatividade voluntária de empregados. No estudo, o autor usou regressão logística binomial e recomendou seu uso futuro, mas deve ser testado rigorosamente para garantir a robustez do modelo.

Nos resultados apresentados no estudo de Sayah (2021), observa-se uma clara distinção na performance dos modelos de árvores de decisão e florestas aleatórias quando aplicados à predição de rotatividade de empregados.

A árvore de decisão alcançou uma precisão de 100% nos dados de treinamento, refletindo sua capacidade de capturar perfeitamente as relações nos dados de treino. No entanto, a precisão caiu para 77,78% nos dados de teste, com uma precisão de classificação para empregados que saíram de apenas 25,97%. Este resultado sugere um *overfitting* significativo, onde o modelo aprende os detalhes e ruídos nos dados de treinamento a ponto de prejudicar sua performance em novos dados.

Por outro lado, a floresta aleatória, que agrega múltiplas árvores de decisão, também mostrou uma precisão perfeita nos dados de treino mas apresentou um desempenho superior nos dados de teste, com uma precisão geral de 86,85%. A precisão para prever empregados que saíram foi de 63,64%, demonstrando uma melhoria significativa em relação à árvore de decisão. Este aumento na precisão pode ser atribuído à habilidade da floresta aleatória de reduzir o *overfitting* através da média das previsões de múltiplas árvores, proporcionando um modelo mais generalizável.

Esses resultados destacam a floresta aleatória como uma abordagem mais robusta para a previsão de rotatividade de empregados em comparação com a árvore de decisão isolada. A capacidade de mitigar o *overfitting* e fornecer previsões mais confiáveis em dados não vistos justifica a preferência pela floresta aleatória em contextos organizacionais, onde a precisão na previsão de rotatividade é crucial para a tomada de decisão estratégica.

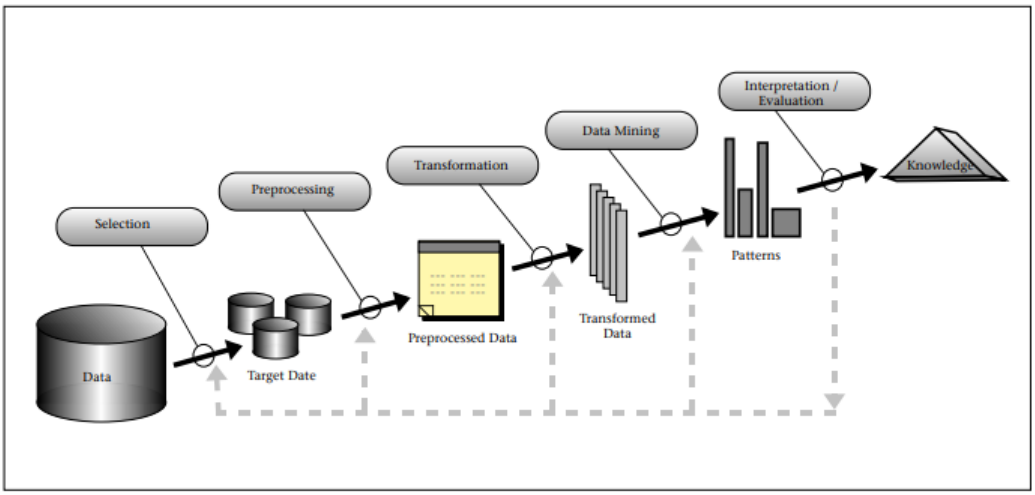
Na revisão da literatura deste estudo, identificou-se uma variedade de abordagens aplicadas na previsão de rotatividade de empregados utilizando técnicas de *ML*, conforme descrito por Chourey, Phulre e Mishra (2019). A seleção de algoritmos específicos, como regressão logística, árvores de decisão e floresta aleatória, foi informada e orientada pelos resultados e discussões destes trabalhos anteriores. Cada um desses algoritmos foi escolhido com o objetivo de abordar e mitigar desafios específicos destacados na literatura, tais como a precisão na classificação e o risco de *overfitting*. Assim, esta escolha estratégica visa não apenas replicar as metodologias bem-sucedidas, mas também ampliar a robustez e a aplicabilidade dos modelos em cenários organizacionais reais, conforme a necessidade apontada por Kane-Sellers (2007) para testes rigorosos que garantam a robustez dos modelos.

3 Metodologia

Este estudo adota uma abordagem quantitativa, exploratória e preditiva, visando analisar e predizer a probabilidade de saída voluntária de colaboradores em uma indústria de celulose utilizando algoritmos de *ML*. A empresa, sediada no estado de Minas Gerais, com atuação multinacional possui mais de 4 mil colaboradores distribuídos nas áreas de fabricação, manutenção, sustentabilidade, recursos humanos, tecnologia, financeira, comercial, jurídico, silvicultura, colheita, planejamento florestal, suprimentos e auditoria. Atualmente a empresa apresenta um *turnover* mensal de aproximadamente 1,73%, e nesse contexto ter a capacidade de antever quais são os colaboradores que estão mais propensos a deixar a empresa pode resultar em reter talentos, evitar custos associados a desligamentos, e alterações indesejadas no clima organizacional.

Neste Capítulo são apresentadas as etapas da metodologia *Knowledge Discovery in Databases* (*KDD*), que define um processo de descoberta de conhecimento em bases de dados, proposta por Fayyad, Piatetsky-Shapiro e Smyth (1996) e desenvolvidas neste trabalho. Conforme as etapas da metodologia apresentadas na Figura 5, a etapa de seleção é descrita na Seção 3.1, o pré-processamento e transformação na Seção 3.2, e a mineração de dados na Seção 3.3.

Figura 5 – Etapas do KDD



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996)

3.1 Seleção

A coleta de dados é o primeiro passo significativo tanto em aprendizado de máquina quanto no processo de *KDD*. Este passo é crítico porque a qualidade e disponibilidade dos dados coletados predeterminarão o quão preciso será o modelo preditivo e as descobertas que podem ser feitas a partir dos dados.

No contexto deste trabalho, a etapa de Seleção do *KDD* foi realizada principalmente a partir dos dados dos empregados mantidos no sistema de gestão interno SAP, sendo esta a principal fonte de dados. Tais registros foram obtidos por meio de consultas a bases de dados. Os dados levantados, apesar de estarem na mesma base de dados, estavam em tabelas diferentes e, após sua extração, necessitaram de integração para formar um conjunto de dados unificado para o trabalho. O arco temporal para a extração de dados abrangeu janeiro de 2014 a fevereiro de 2023, abarcando todos os empregados ativos neste período, incluindo os que foram desligados por iniciativa própria. A coleta resultou em um total de 5.293 empregados, sendo que 657 destes foram desligados voluntariamente.

A etapa de Seleção no *KDD* é essencial para garantir que os dados relevantes sejam escolhidos a partir de diversas fontes de dados. Esta fase envolve identificar e selecionar os dados apropriados que podem ser utilizados para a análise subsequente, assegurando que as variáveis relevantes estejam disponíveis para construção de um modelo robusto de aprendizado de máquina.

A partir da revisão da literatura e precedentes na área, especificamente os estudos de Khera (2019), Yedida *et al.* (2018), foram escolhidas 26 variáveis, descritas na Tabela 1.

3.2 Pré-Processamento e Transformação

Com a definição clara da fonte de dados, procedeu-se à fase de pré-processamento e transformação, preparando-a para o processamento pelo modelo. Nesta etapa foram utilizadas as ferramentas *Excel* da *Microsoft* e *Google Colab*. Inicia-se com a anonimização ou exclusão de dados sensíveis seguindo diretrizes da Lei Geral de Proteção de Dados Pessoais (LGPD). Os dados considerados sensíveis por poder identificar os empregados são: Matrícula, Nome e CPF.

Os dados de matrícula foram anonimizados, de tal forma que, para cada matrícula existente foi gerado um identificador. Em seguida, houve a exclusão do campo Nome e CPF, por se tratar de informações irrelevantes e desnecessárias. Com esta etapa concluída, foi possível ter uma fonte de dados sem a presença de dados que venham infringir a LGPD. Dando sequência no pré-processamento, as colunas irrelevantes para a análise foram excluídas para evitar distorções nos resultados da modelagem.

Tabela 1 – Dicionário de Dados

Coluna	Descrição	Exemplo de Valor	Tipo
ID2	Matrícula do empregado	1175	Inteiro
cargo	Cargo do empregado	Auxiliar Líder	Categórico
status	Empregado Ativo ou Desligado	Ativo	Categórico
Grupo_AV	Grupo de avaliação	Supervisores	Categórico
Desligamento_ou_Hoje	Data demissão ou data atual para ativos	10/06/2015	Data/Hora
Convenio_nao6_meses	Utilização pl. saúde nos últimos 6 meses	sim	Categórico
CC	Código de custo	56230	Categórico
UO	Unidade Organizacional	60000016	Inteiro
grupo	Grupo em que o empregado se encontra	Empregados	Categórico
area	Área de atuação do empregado	DEFAB-S	Categórico
Grade	Nível e salarial dentro da empresa	10	Número Real
Posicionamento	% da Curva Salarial	80%	Número Real
area_rh	Região em que o empregado está lotado	Nova Era	Categórico
SubareaRH	Sindicato do empregado	Sintcel Oper.	Categórico
Idade	Idade do empregado	35	Inteiro
Estado_Civil	Estado civil do empregado	Casado	Categórico
Tempo_Empresa	Tempo de trabalho na empresa em anos	5	Inteiro
tempo_cargo	Tempo no cargo atual em anos	3	Inteiro
ultima_movimentacao	Tempo última mov. salarial em anos	1	Inteiro
Cidade	Cidade de residência do empregado	Ipatinga	Categórico
Qtde_Filhos	Quantidade de filhos do empregado	2	Inteiro
Filhos<3Anos	Possui filhos menores de três anos	Sim	Categórico
sexo	Sexo do empregado	masculino	Categórico
raça	Raça do empregado	Parda	Categórico
Insalubridade	Se o emprego tem condições insalubres	sim	Categórico
Periculosidade	Se o emprego tem condições perigosas	não	Categórico

Fonte: Elaborado pelo autor

Ainda nesta fase, as variáveis categóricas, especificamente dicotômicas do tipo 'sim' ou 'não', foram convertidas em formatos binários. Esta conversão foi realizada atribuindo-se o valor 1 para 'sim' e 0 para 'não'. Esta transformação facilita a aplicação de algoritmos de aprendizado de máquina que requerem entradas numéricas, além de simplificar a análise estatística dos dados.

As colunas com datas foram convertidas em idade, pois em se tratando de modelos preditivos, isso oferece vantagens significativas, incluindo maior relevância direta da variável para o modelo, simplificação e redução da dimensionalidade dos dados, facilitando a normalização e a interpretação. A idade, sendo um valor numérico contínuo, integra-se mais facilmente com outras variáveis demográficas, melhorando a generalização do modelo em novos dados e minimizando potenciais vieses relacionados a datas específicas. Ademais, essa transformação pode agilizar o treinamento e a execução do modelo, além de ser crucial para manter a precisão sem necessidade de recálculos frequentes devido à passagem do tempo.

Para garantir a comparabilidade e melhorar a performance dos modelos preditivos, foi realizada a codificação de variáveis categóricas e posteriormente a normalização de variáveis numéricas. Primeiramente, as variáveis categóricas foram codificadas utilizando o método de codificação ordinal, transformando-as em variáveis numéricas. Em seguida, aplicou-se a técnica de normalização *Min-Max Scaling* para padronizar todas as variáveis numéricas, incluindo tanto as colunas inicialmente numéricas (UO, Idade, Tempo Empresa, tempo cargo, Posicionamento, Grade, Qtde Filhos, ultima movimentacao) quanto as colunas categóricas que foram transformadas em numéricas através da codificação ordinal. Essa padronização foi feita para garantir que todas essas variáveis estivessem no intervalo entre 0 e 1, minimizando a influência de diferentes escalas de variáveis. Essa etapa permitiu calcular com precisão a matriz de correlação e contribuir para a construção de modelos de classificação mais robustos e eficientes.

3.2.1 Análise Exploratória de Dados AED

Com um conjunto de dados unificado e preparado, prosseguiu-se para a etapa de análise exploratória, onde foram examinadas estatísticas descritivas, distribuições e possíveis correlações entre as variáveis, especialmente focando em características que podem influenciar a decisão de um empregado em deixar a empresa.

A Tabela 2 representa a análise descritiva realizada sobre a população de colaboradores, a mesma trouxe informações relevantes sobre as características demográficas e profissionais. Estabeleceu-se, portanto, que a faixa etária dos empregados dentro desta organização varia de 19 a 75 anos, com uma média em torno de 40 anos. Isso mostra literalmente que há uma mistura de diferentes gerações nesta organização. Em termos de vinculação com a empresa, a média é de 9 anos – um indicativo de um nível alto de retenção e acumulação de experiência organizacional. Quanto ao tempo na posição atual, somando tudo, a média chega perto de quase 5,7 anos, o que pode significar que a maioria deles está estável em sua função. Além disso, o *escore* médio mostrou que cada empregado tem cerca de 1,3 filhos, o que mostra a gama nas diferentes fases da vida familiar entre eles.

Tabela 2 – Resultado parcial do método describe().

index	status	Grade	Pos.	Idade	Tempo_Empresa	tempo_cargo	ultima_mov.	Qtde_Filhos
count	5293	5293	5293	5293	5293	5293	5293	5293
mean	0.8758	4,8	95,65	39,84	9	6	1	1
std	0.3297	4,9	12,18	11,53	8	5	2	1
min	0	1	13,93	19,00	0	0	0	0
25%	1	1	90,96	30,00	3	2	0	0
50%	1	1	100,00	40,00	8	4	0	1
75%	1	10	100,00	49,00	11	10	2	2
max	1	28	180,11	75,00	48	48	19	10

Fonte: Elaborado pelo autor

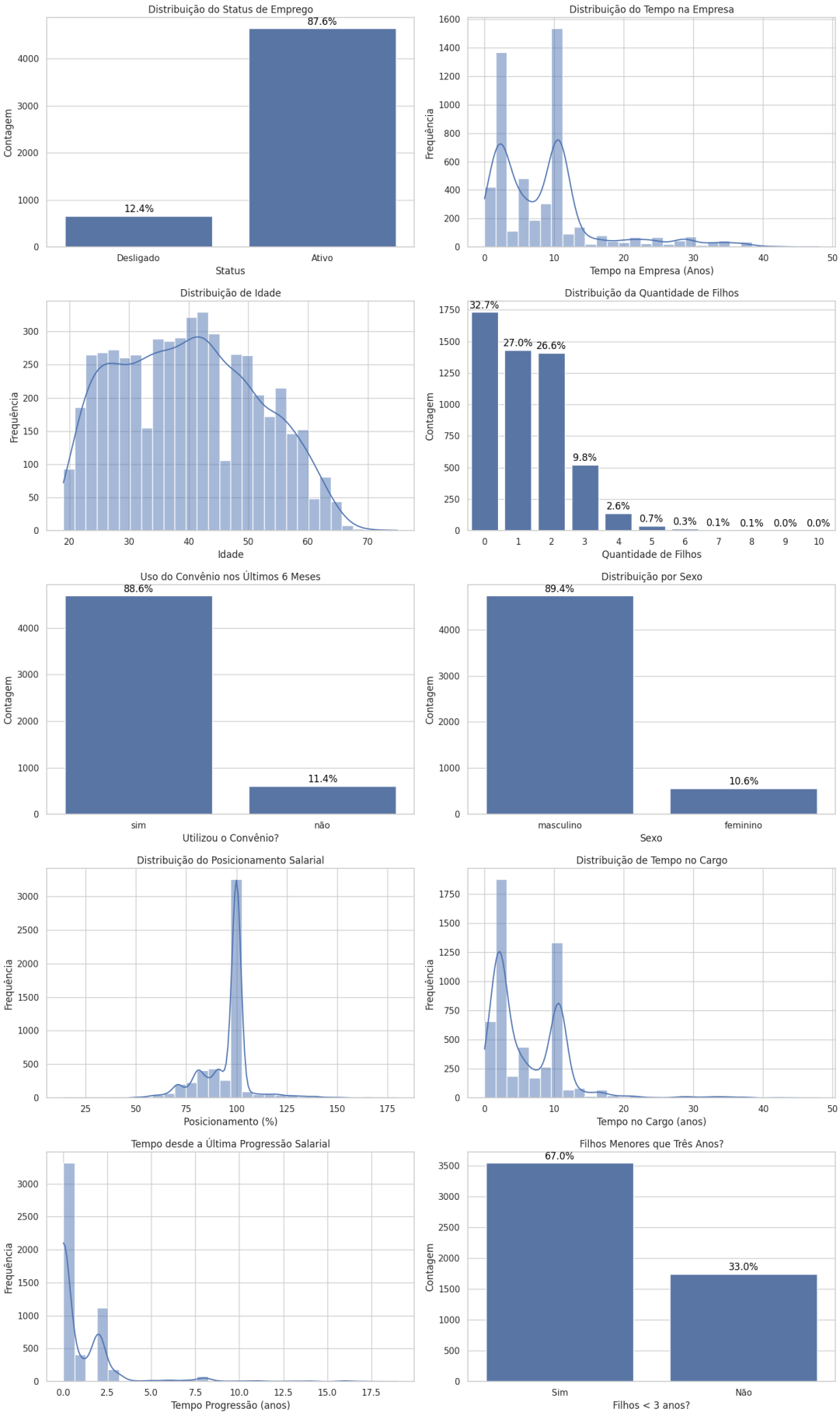
A Figura 6 fornece uma análise abrangente das variáveis categóricas e contínuas presentes na base de dados. Observa-se que a maioria dos colaboradores são ativos (87.6%), enquanto uma menor proporção está desligada (12.4%). Este desequilíbrio significativo na classe de *status* de emprego pode influenciar modelos de aprendizado de máquina, tendendo a prever a maioria dos casos como ativos. A distribuição do tempo na empresa mostra que a maioria dos colaboradores possui até 10 anos de serviço, com picos notáveis entre 0-5 e 5-10 anos. Esse padrão é em grande parte resultado do programa de primarização iniciado pela empresa em 2013, que levou à contratação direta de cerca de 3 mil trabalhadores que anteriormente estavam vinculados a prestadoras de serviços na área florestal.

A distribuição de idade revela que a maior parte dos colaboradores está na faixa dos 30 aos 50 anos, com uma queda acentuada após os 50 anos, possivelmente devido a políticas de aposentadoria ou menor retenção de trabalhadores mais velhos. Em relação à quantidade de filhos, a maioria dos colaboradores possui de 0 a 2 filhos, com poucos tendo mais de 3 filhos. A utilização do convênio médico nos últimos 6 meses é alta (88.6%), destacando a importância deste benefício para os trabalhadores.

A análise da distribuição por sexo revela uma predominância masculina (89.4%), o que pode refletir uma concentração de funções operacionais predominantemente ocupadas por homens, sugerindo a necessidade de iniciativas de diversidade e inclusão para equilibrar a composição de gênero. Apesar do baixo número de mulheres, a empresa já tem trabalhado para aumentar a presença feminina em funções antes predominantemente ocupadas por homens, promovendo ações de recrutamento e treinamento voltadas para a inclusão de mulheres em diversos níveis da organização. A distribuição do posicionamento salarial mostra uma concentração na faixa de 75% a 125%, com um pico significativo em 100%, refletindo que muitos trabalhadores florestais recebem o piso salarial, justificando esse alto volume de trabalhadores posicionados exatamente a 100%. O tempo no cargo tem uma distribuição similar ao tempo na empresa, com a maioria dos colaboradores tendo menos de 10 anos em suas posições atuais.

O tempo desde a última progressão salarial é predominantemente menor que 5 anos, indicando uma política de progressão relativamente frequente. Finalmente, a análise da variável "Filhos Menores que 3 Anos" mostra que 67% dos colaboradores possuem filhos nesta faixa etária. Este aspecto pode ser um fator relevante na análise de retenção de empregados, já que colaboradores com filhos pequenos podem ter uma maior necessidade de estabilidade e benefícios familiares robustos. Avaliar como esses benefícios impactam a decisão dos empregados de permanecer na empresa pode fornecer informações importantes sobre a eficácia das políticas de retenção atualmente implementadas. Em conjunto, esses dados fornecem uma visão detalhada da força de trabalho, permitindo direcionar estratégias de gestão e políticas internas de forma mais eficaz.

Figura 6 – Distribuição dos dados



Fonte: Elaborado pelo autor

A Figura 7 apresenta uma análise comparativa detalhada entre empregados ativos e desligados em diversas variáveis demográficas e profissionais, destacando padrões que podem influenciar decisões estratégicas organizacionais. Os histogramas fornecem uma visão clara das diferenças nos perfis dos empregados ativos e desligados em termos de tempo de empresa, idade, tempo no cargo, última movimentação, estado civil e quantidade de filhos.

Os dados mostram que os empregados com maior tempo de empresa tendem a permanecer na organização, evidenciado pelo pico significativo de empregados ativos com até 10 anos de serviço. Esta tendência é reforçada pela análise do tempo no cargo e da última movimentação salarial, onde vemos que a maioria dos desligados possui menor tempo nessas variáveis, indicando que mudanças recentes ou falta de progressão salarial podem estar associadas ao desligamento.

A distribuição de idade revela que empregados mais jovens apresentam uma maior taxa de desligamento, enquanto a faixa etária de 30 a 50 anos possui maior estabilidade, possivelmente devido a maiores responsabilidades e comprometimento com a carreira. A análise do estado civil indica que empregados casados são mais propensos a permanecer na organização, sugerindo que estabilidade familiar pode estar correlacionada com estabilidade profissional.

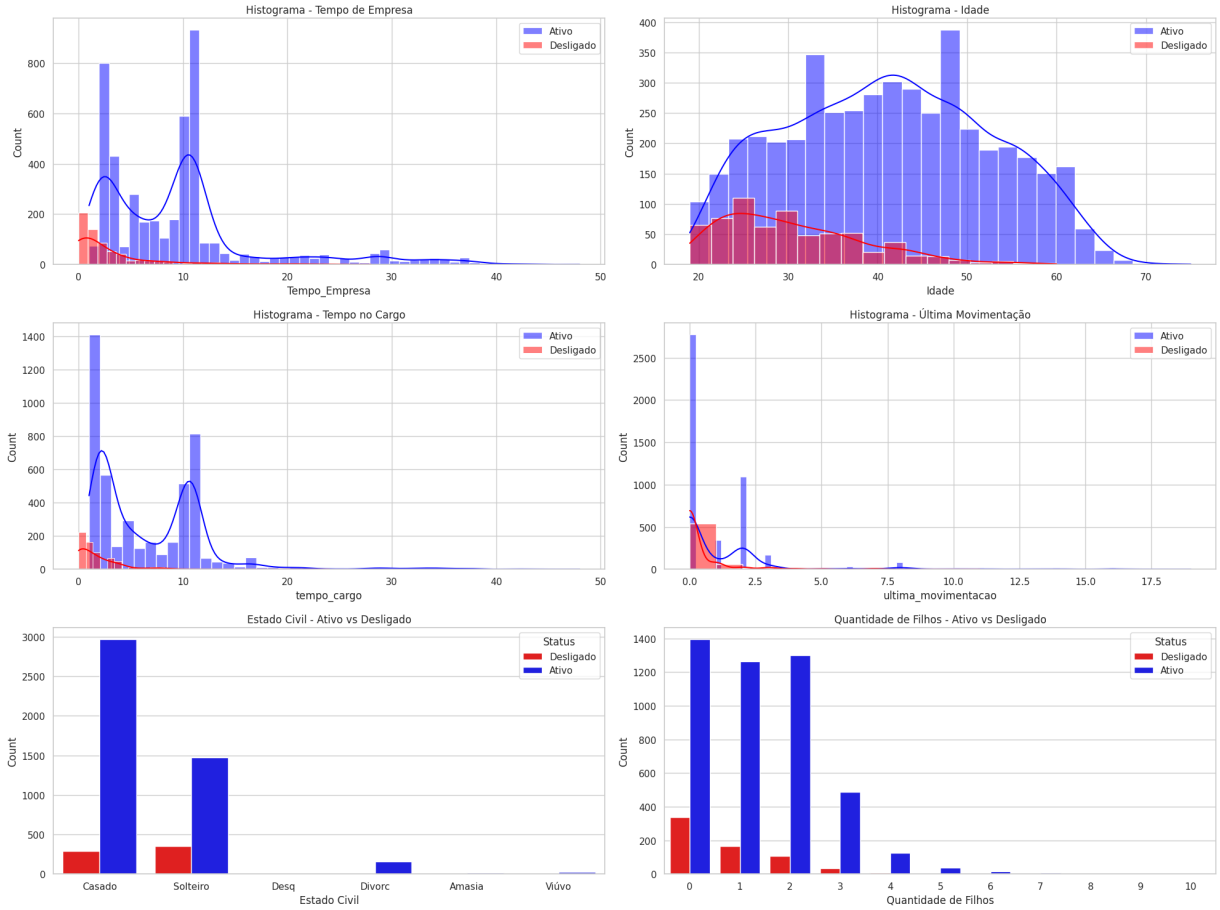
Adicionalmente, a quantidade de filhos parece influenciar a retenção, com empregados que possuem filhos apresentando maior taxa de permanência. Este dado pode refletir a importância do plano de assistência à saúde e suporte familiar oferecidos pela empresa como fator de retenção. Por outro lado, empregados solteiros ou com menos filhos mostram uma propensão maior ao desligamento, possivelmente devido a maior mobilidade e menos responsabilidades familiares.

Estes *insights* são cruciais para a compreensão das dinâmicas de retenção dentro da empresa. A identificação desses padrões permite a implementação de políticas de recursos humanos mais eficazes, focadas na retenção de talentos, oferecendo suporte adequado e oportunidades de progressão que atendam às necessidades dos empregados em diferentes estágios de suas vidas e carreiras.

3.2.2 Correlações e Redução de Dimensionalidade

Para compreender as relações entre as variáveis no conjunto de dados pré-processados e transformados, foi calculada a matriz de correlação utilizando o coeficiente de correlação de Pearson. As colunas 'ID2' e 'Desligamento ou Hoje' foram excluídas antes do cálculo, garantindo que a análise fosse focada nas variáveis mais relevantes. A visualização da matriz de correlação foi realizada com a biblioteca *Seaborn*, por meio de um mapa de calor que ilustra graficamente a força e a direção das associações entre as variáveis.

Figura 7 – Histograma Ativos x Desligados



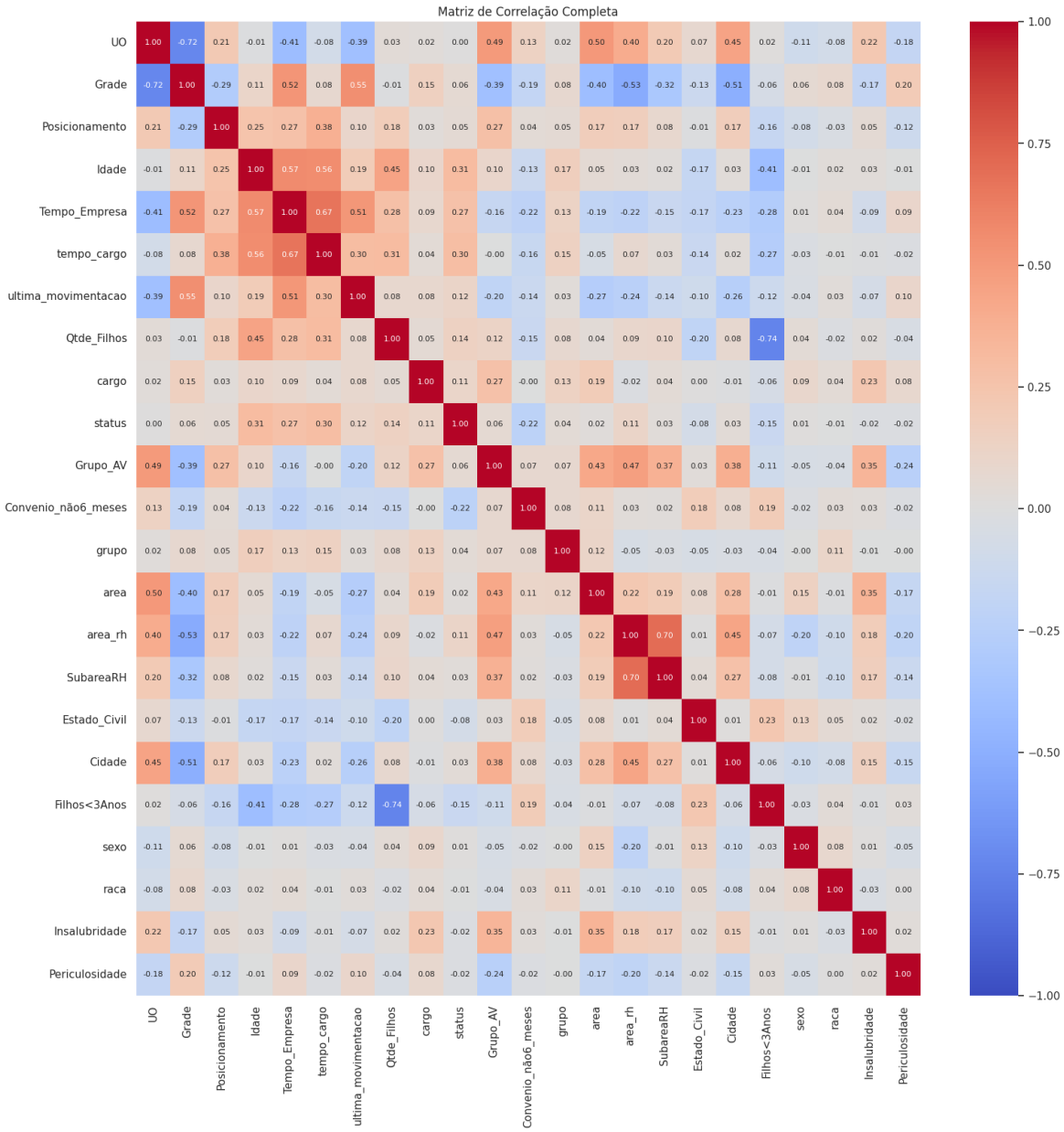
Fonte: Elaborado pelo autor

A Figura 8 apresenta a matriz de correlação. A análise revelou várias relações significativas entre as variáveis, conforme os critérios de força de correlação definidos por Cohen (1988). Primeiramente, observou-se uma correlação forte entre "Grade" e "Idade" (0.565), e entre "Grade" e "Tempo_Empresa" (0.520), indicando que, conforme a idade e o tempo de empresa aumentam, o "Grade" também tende a aumentar (por grade, entenda-se agrupamento de cargos por níveis salarias). Adicionalmente, uma correlação forte foi encontrada entre "Tempo_Empresa" e "tempo_cargo" (0.667), sugerindo que empregados com mais tempo na empresa também possuem mais tempo no cargo atual. Além disso, uma correlação moderada foi identificada entre "Estado_Civil" e "Filhos<3Anos" (0.234), sugerindo que o estado civil pode estar associado à quantidade de filhos pequenos.

No que se refere a correlações negativas fortes, foi identificado que "UO" e "Grade"possuem uma correlação de (-0.722), indicando que certas unidades organizacionais (UO) tendem a ter grades mais baixos. Similarmente, "Grade" e "Grupo_AV" (-0.388), e "Grade" e "area" (-0.401) também apresentaram correlações negativas fortes, sugerindo variações nas grades com esses grupos e áreas específicas.

Por fim, correlações moderadas foram encontradas entre "Posicionamento" e "Idade" (0.251), indicando que o posicionamento pode aumentar ligeiramente com a idade, e entre "Posicionamento" e "Tempo_Empresa" (0.268), sugerindo que um melhor posicionamento pode estar associado a mais tempo de empresa.

Figura 8 – Matriz de Correlação

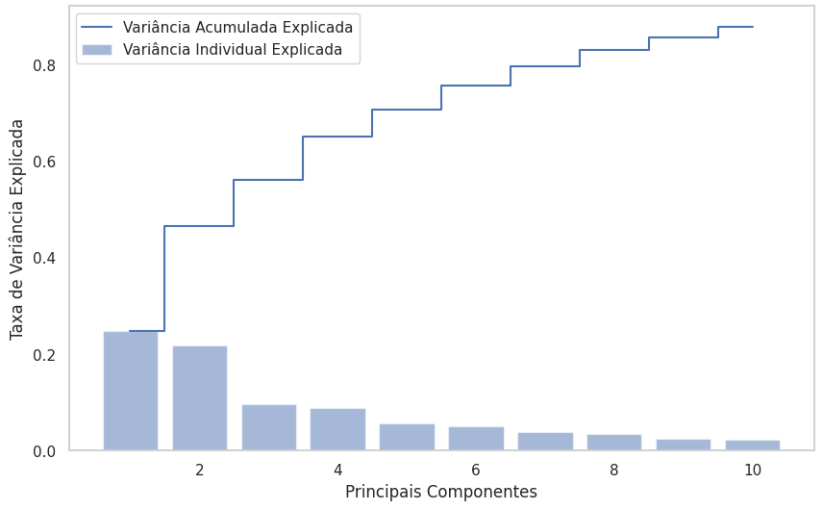


Fonte: Elaborado pelo autor

Em seguida, para reduzir a dimensionalidade dos dados, foi aplicada a *PCA*, selecionando 10 componentes principais. Isso permitiu capturar uma fração significativa da variância total presente nos dados. A variância explicada acumulada por cada componente foi visualizada em um gráfico, fornecendo uma representação clara da contribuição de cada um dos 10 componentes principais para a variabilidade total. Por fim, os componentes principais foram integrados em um novo *DataFrame*, juntamente com a variável alvo status, para continuar as análises e modelagem subsequentes.

O gráfico apresentado na Figura 9 mostra tanto a variância individual explicada por cada componente principal quanto a variância acumulada explicada pelos componentes. As barras representam a variância individual explicada por cada componente principal, enquanto a linha escalonada representa a variância acumulada explicada. Este gráfico é útil para entender o quanto cada componente principal individualmente contribui para a explicação da variância total nos dados e como essa contribuição se acumula à medida que mais componentes são considerados. Ele permite identificar os componentes que mais contribuem para a variância explicada e entender a distribuição dessa variância entre os componentes.

Figura 9 – Variância Individual e Acumulada Explicada pelos Componentes Principais

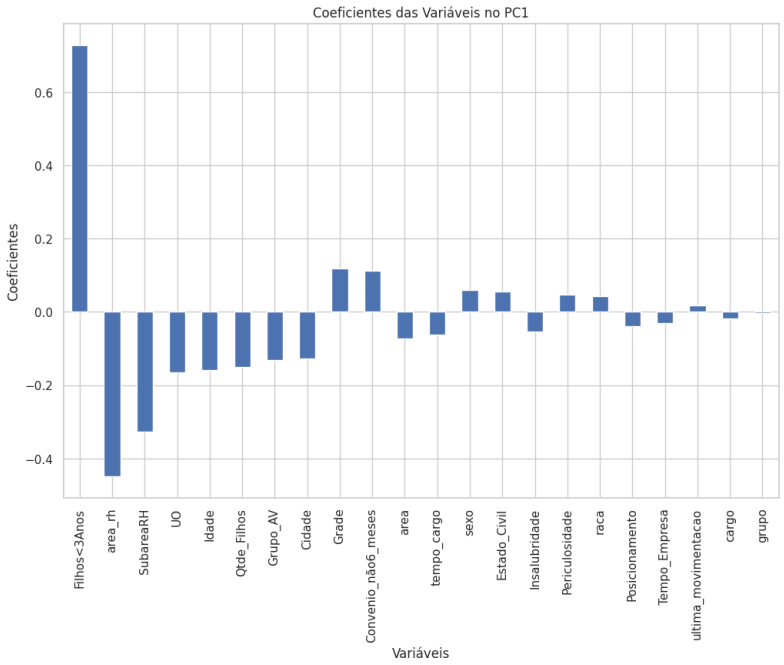


Fonte: Elaborado pelo autor

Observa-se que os primeiros componentes principais têm uma contribuição significativa para a variância total, com uma diminuição gradual na contribuição dos componentes subsequentes. O gráfico evidencia que com 10 componentes, quase a totalidade da variância dos dados originais é explicada, aproximando-se de 90%, o que demonstra a eficácia da PCA em reduzir a dimensionalidade mantendo a maior parte das informações relevantes do conjunto de dados.

A Figura 10 apresenta os coeficientes das variáveis no primeiro componente principal (PC1).

Figura 10 – Coeficientes das Variáveis no PC1



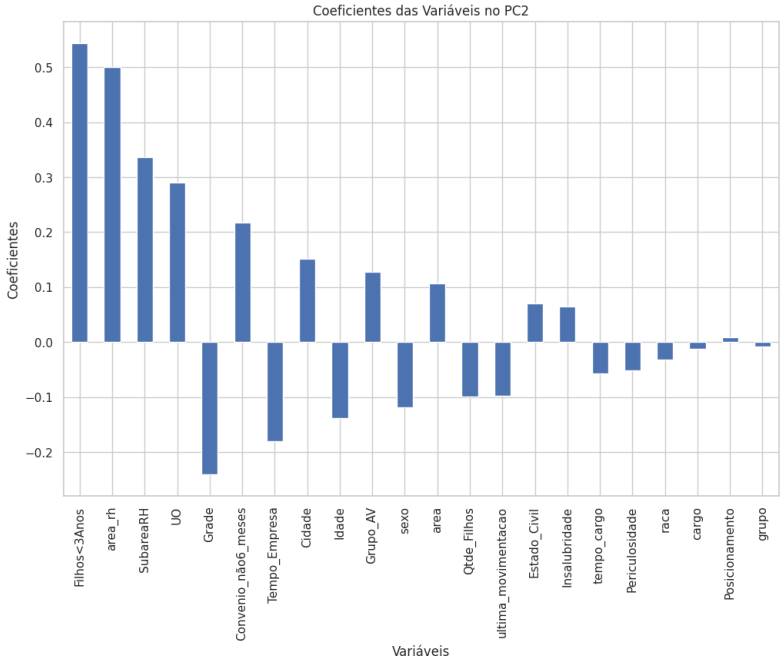
Fonte: Elaborado pelo autor

Esses coeficientes indicam a contribuição de cada variável original na formação do PC1. Observa-se que a variável "Filhos < 3 Anos"tem a maior contribuição positiva, enquanto "area_rh"possui a maior contribuição negativa. Essa análise é fundamental para entender quais variáveis têm maior impacto na variância capturada pelo primeiro componente principal. Os coeficientes, também conhecidos como autovetores, representam a influência das variáveis na composição linear que define cada componente principal.

A Figura 11 ilustra os coeficientes das variáveis no segundo componente principal (PC2). Assim como no PC1, esses coeficientes mostram a importância relativa de cada variável na composição do PC2. Notamos que "Filhos < 3 Anos" e "area_rh" novamente possuem contribuições significativas, mas em magnitudes diferentes comparadas ao PC1, indicando que diferentes padrões de variância são capturados pelo segundo componente principal.

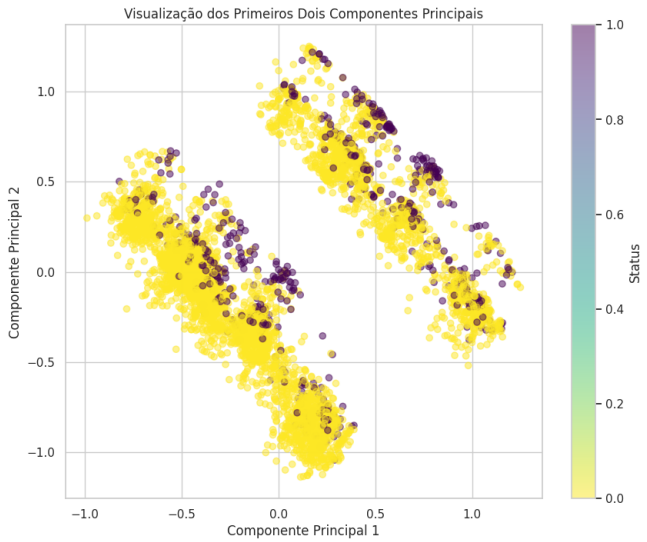
O gráfico apresentado na Figura 12 ilustra a visualização dos primeiros dois componentes principais (PC1 e PC2) obtidos através da PCA aplicada ao conjunto de dados. Cada ponto no gráfico representa um registro individual de colaborador, onde a coloração varia de roxo a amarelo, indicando o *status* de desligamento, com roxo representando colaboradores desligados e amarelo representando colaboradores ativos.

Figura 11 – Coeficientes das Variáveis no PC2



Fonte: Elaborado pelo autor

Figura 12 – Coeficientes das Variáveis PC1 e PC2



Fonte: Elaborado pelo autor

Os eixos X e Y correspondem aos componentes principais que capturam a maior e a segunda maior variância nos dados, respectivamente. Observa-se que os pontos estão dispersos em diferentes regiões do gráfico, sugerindo que os componentes principais identificam padrões variados nos dados. Áreas com maior concentração de um *status* em relação ao outro são visíveis, mas não há uma separação clara e distinta entre as classes. Esta dispersão indica a complexidade inerente ao conjunto de dados e a possibilidade de sobreposição nas características dos colaboradores desligados e ativos.

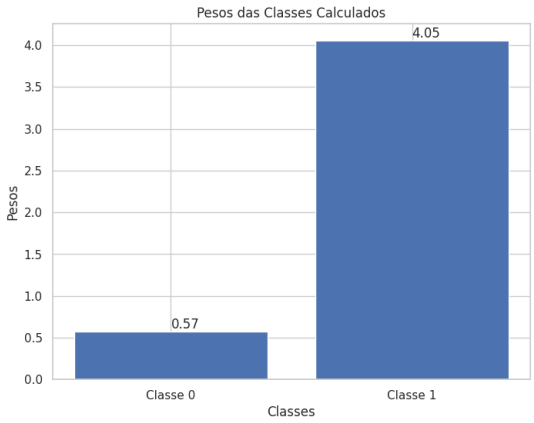
A visualização dos componentes principais facilita a compreensão da distribuição dos dados e da variabilidade capturada por esses componentes. Ela destaca a importância de utilizar técnicas adicionais ou modelos mais avançados para melhorar a classificação dos registros de desligamento, dado que a PCA, apesar de útil para redução de dimensionalidade e identificação de padrões gerais, pode não ser suficiente para separar claramente as classes de colaboradores desligados e ativos.

3.3 Mineração de Dados

Na etapa de mineração de dados do método *KDD*, é essencial preparar os conjuntos de treino e teste de maneira equilibrada para garantir a eficácia dos modelos preditivos. Para isso, o conjunto de dados transformado foi dividido em variáveis independentes (*features*) e a variável dependente (*target*) '*status*', que indica se o empregado está ativo ou desligado.

Foi utilizada a função *train_test_split* da biblioteca *sklearn* para dividir os dados em 70% de treinamento e 30% de teste. Como a variável alvo é desbalanceada, ou seja, há uma predominância do *status* 'Ativo', os pesos das classes foram calculados usando a função *compute_class_weight*, atribuindo pesos maiores às classes minoritárias para compensar o desbalanceamento. O cálculo resultou nos seguintes pesos: 4.0536 para a classe 0 (desligado) e 0.5704 para a classe 1 (ativo).

Figura 13 – Balanceamento de Classes



Fonte: Elaborado pelo autor

Essa estratégia de balanceamento, ilustrada na Figura 13, visa corrigir o desequilíbrio nas classes durante o processo de treinamento, fornecendo mais importância aos exemplos minoritários, o que potencialmente melhora a capacidade preditiva dos modelos apresentados nas próximas subseções.

3.3.1 Construção de Modelos Preditivos

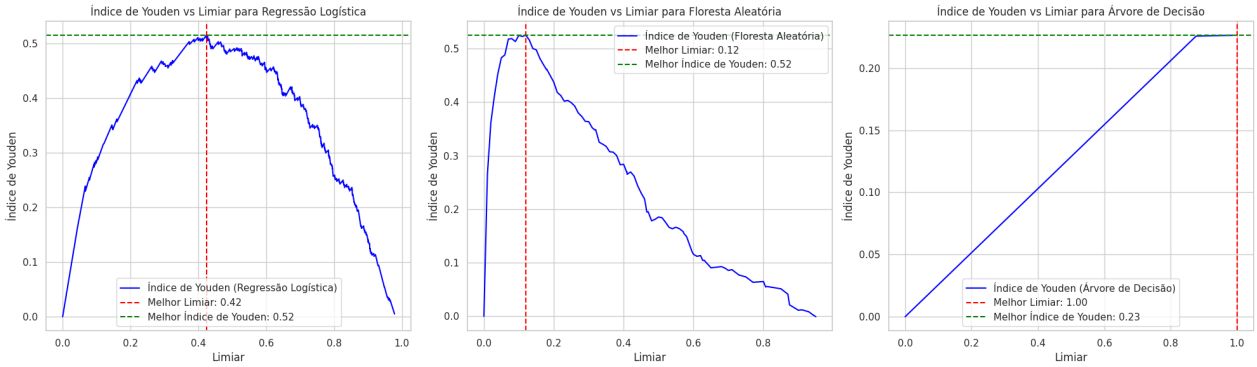
Após a preparação dos conjuntos de treino e teste e a aplicação dos pesos para lidar com o desbalanceamento das classes, três modelos de classificação foram treinados para prever o *status* de emprego dos empregados: regressão logística, floresta aleatória e árvore de decisão.

Após o treinamento dos modelos, foram calculadas suas métricas de desempenho. As métricas consideradas foram:

- Acurácia: Mede a proporção de previsões corretas.
- Precisão: Calcula a proporção de verdadeiros positivos entre todas as predições positivas.
- *Recall*: Avalia a proporção de verdadeiros positivos entre todos os casos reais positivos.
- *F1 Score*: representa a média harmônica entre precisão e *recall*.

Para garantir um melhor desempenho dos modelos de classificação, foi utilizada a técnica de otimização do limiar de decisão baseada no índice de *Youden*. Este índice é definido como a diferença entre a taxa de verdadeiros positivos (*true positive rate* - *True Positives Rate* (*TPR*)) e a especificidade menos 1, e é maximizado para encontrar o melhor limiar que separa as classes de forma mais eficaz. Conforme ilustrado na Figura 14, os gráficos dos índices de *Youden* para os modelos de regressão logística, floresta aleatória e árvore de decisão mostram diferentes limiares ótimos. A utilização do índice de *Youden* para determinar o melhor limiar é crucial, pois permite ajustar o ponto de corte de modo a otimizar a identificação correta das classes positivas e negativas, melhorando a utilidade prática e a eficácia preditiva dos modelos.

Figura 14 – Identificação do melhor limiar



Fonte: Elaborado pelo autor

Após a aplicação desta técnica, os melhores limiares para cada modelo foram determinados. Estes limiares foram então utilizados para calcular as métricas de desempenho, e os resultados obtidos foram apresentadas na Tabela 3:

Tabela 3 – Comparação de Métricas de Modelos Preditivos com Limiar Otimizado

Modelo	Acurácia	Precisão	*Recall*	*F1 Score*	Limiar Otimizado
Regressão logística	0.68	0.26	0.86	0.40	0.42
Floresta aleatória	0.74	0.30	0.80	0.43	0.12
Árvore de decisão	0.84	0.35	0.31	0.33	1

Fonte: Elaborado pelo autor

3.3.2 Ajuste de Parâmetros

Após ajustar os parâmetros dos modelos de classificação para melhorar o equilíbrio e evitar o *overfitting*, foram observadas melhorias nas métricas de desempenho. Utilizando uma regularização mais forte para a Regressão Logística (*C=0.6*), aumentando o número de estimadores e reduzindo a profundidade máxima para a Floresta Aleatória (*n_estimators=200, max_depth=10*), e limitando a profundidade máxima para a Árvore de Decisão (*max_depth=5*), os resultados ajustados são apresentados na Tabela 4.

Tabela 4 – Comparação de Métricas de Desempenho dos Modelos Ajustados

Modelo	Acurácia	Precisão	*Recall*	*F1 Score*
Regressão Logística	0.68	0.27	0.86	0.41
Floresta Aleatória	0.78	0.33	0.72	0.45
Árvore de Decisão	0.72	0.27	0.70	0.39

Fonte: Elaborado pelo autor

3.3.3 Aplicação da Validação Cruzada (*Cross Validation*)

Como última tentativa de melhorar as previsões dos modelos, aplicou-se a técnica de *Cross-Validation* (Validação Cruzada), especificamente usando a estratégia *Stratified KFold*. O *Cross-Validation* é uma técnica estatística que divide o conjunto de dados em várias partes, chamadas de *folds*, e então treina e testa o modelo múltiplas vezes, usando diferentes combinações de treinamento e teste. Essa abordagem ajuda a garantir que o modelo seja avaliado em todos os segmentos dos dados, reduzindo o risco de *overfitting* e proporcionando uma avaliação mais robusta, conforme demonstrado na Tabela 5.

Tabela 5 – Comparação de Métricas de Modelos de Classificação após Validação Cruzada

Modelo	Acurácia	Precisão	*Recall*	*F1 Score*
Regressão Logística	0.62	0.24	0.92	0.38
Floresta Aleatória	0.74	0.29	0.78	0.43
Árvore de Decisão	0.60	0.21	0.81	0.34

Fonte: Elaborado pelo autor

4 Resultados

Esta seção é voltada às análises realizadas através das tarefas de mineração de dados. Analisar e avaliar os resultados obtidos é a última etapa do processo KDD.

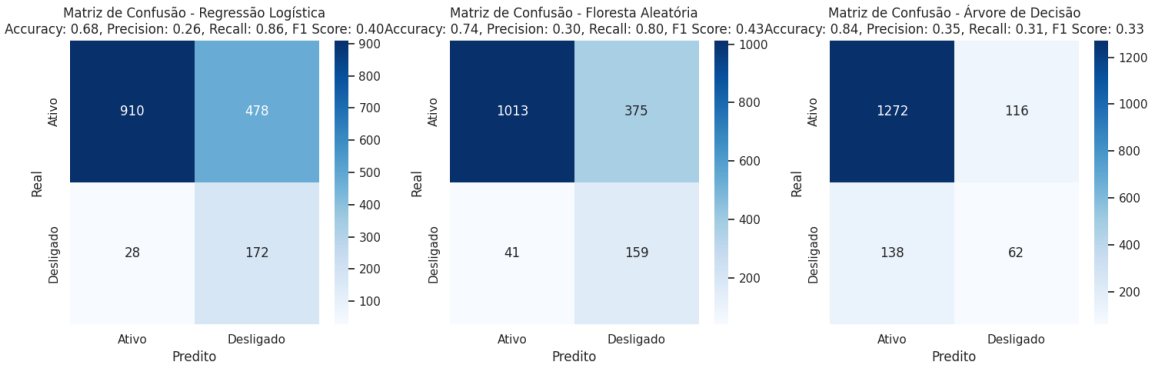
4.1 Avaliação dos Modelos Preditivos

A análise do desempenho dos modelos regressão logística, floresta aleatória e árvore de decisão revelou diferentes particularidades e adequações ao cenário empresarial. O modelo de regressão logística apresentou uma acurácia de 68%, com precisão de 26%, *recall* de 86% e *F1 score* de 0.40, utilizando um limiar de 0.42. Este modelo demonstrou uma alta capacidade de identificar corretamente os empregados que sairão (alto *recall*), mas com baixa precisão, indicando um número considerável de falsos ativos, como evidenciado pela matriz de confusão da Figura 15 que mostra 478 falsos ativos e 28 falsos desligados.

O modelo floresta aleatória apresentou resultados semelhantes, com uma acurácia de 74%, precisão de 30%, *recall* de 80% e *F1 score* de 0.43, com um limiar de decisão de 0.12. Este ajuste de limiar aumenta a sensibilidade, mostrando-se eficaz na identificação de empregados que sairão, com uma matriz de confusão que revela 375 falsos ativos e apenas 41 falsos desligados. Esta alta sensibilidade é um diferencial importante, embora ainda enfrente desafios de precisão.

Por outro lado, o modelo árvore de decisão alcançou a maior acurácia de 84%, com precisão de 35%, *recall* de 31% e *F1 score* de 0.33, utilizando um limiar de 1.00. Este modelo mostrou-se mais equilibrado entre os verdadeiros ativos e desligados, mas com uma sensibilidade relativamente baixa. A matriz de confusão para este modelo mostra 116 falsos ativos e 138 falsos desligados, indicando um desempenho mais equilibrado.

Figura 15 – Matriz de Confusão - Modelos individuais

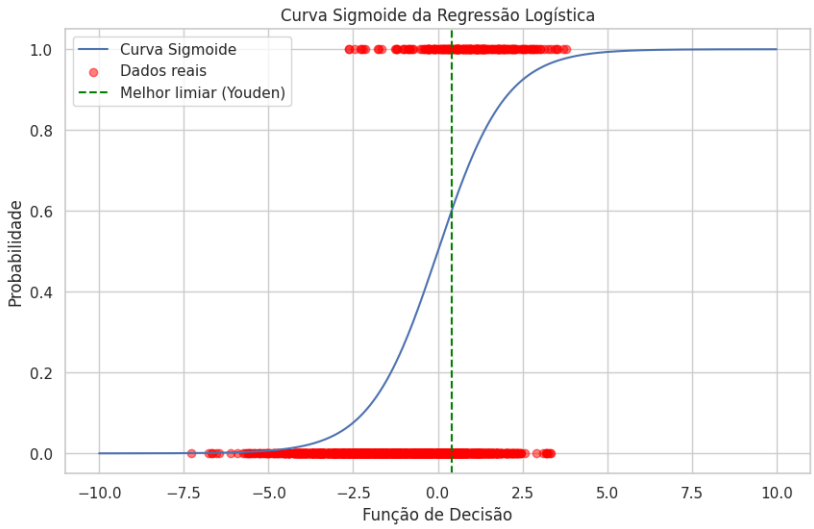


Fonte: Elaborado pelo autor

Para a empresa, o modelo mais interessante dependerá das prioridades estratégicas. Se a prioridade for minimizar a perda de talentos (maximizando o *recall*), o modelo floresta aleatória pode ser preferível devido à sua alta sensibilidade. Contudo, se a empresa busca um equilíbrio melhor entre precisão e sensibilidade, a árvore de decisão pode ser mais adequada, especialmente se a acurácia geral for um fator crítico. A regressão logística, com seu desempenho intermediário, pode ser uma opção inicial, mas ajustes adicionais podem ser necessários para melhorar a precisão. Cada modelo oferece informações valiosas, mas a decisão final deve considerar o impacto estratégico da predição de saída de empregados, balanceando entre a retenção de talentos e a eficiência operacional.

A Figura 16 mostra a curva *sigmoide* gerada pelo modelo de regressão logística com os dados reais sobrepostos. A curva *sigmoide* (linha azul) representa a probabilidade predita pelo modelo de um evento ocorrer à medida que a função de decisão varia. Os pontos vermelhos representam os dados reais de teste, com a probabilidade real de desligamento. A linha verde tracejada indica o melhor limiar de decisão encontrado utilizando o método de *Youden*, que no gráfico está em torno de 0.5. Observa-se que a maioria dos dados reais de *status* ativo (0) está agrupada na parte inferior da curva sigmoide, enquanto os dados de desligamento (1) estão na parte superior. A curva *sigmoide* fornece uma visualização clara de como o modelo de regressão logística separa as classes, atribuindo uma alta probabilidade de desligamento para valores altos da função de decisão e uma baixa probabilidade para valores baixos. A eficácia do modelo pode ser avaliada pela distância e distribuição dos pontos vermelhos em relação à curva e ao limiar de decisão.

Figura 16 – Curva S - Regressão Logística



Fonte: Elaborado pelo autor

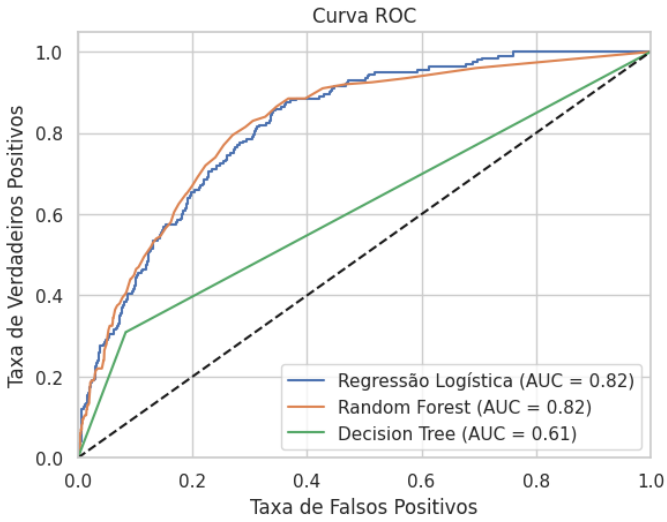
A Curva ROC (*Receiver Operating Characteristic*) é uma ferramenta gráfica utilizada para avaliar o desempenho de modelos de classificação. Ela é criada plotando a taxa de verdadeiros positivos - *TPR* contra a taxa de falsos positivos - *False Positives Rate* (*FPR*) em diferentes limiares de classificação. A *TPR*, também conhecida como sensibilidade ou *recall*, representa a proporção de positivos corretamente identificados pelo modelo. A *FPR* representa a proporção de negativos incorretamente identificados como positivos.

A área sob a curva (AUC - *Area Under the Curve*) é uma métrica de avaliação derivada da curva ROC. Valores de AUC mais próximos de 1 indicam um melhor desempenho do modelo, pois significam que o modelo tem uma maior capacidade de discriminar entre as classes.

O gráfico na Figura 17 compara o desempenho dos modelos de regressão logística, floresta aleatória e árvore de decisão na tarefa de classificação de desligamentos de colaboradores. A regressão logística e o floresta aleatória ambos apresentam um AUC de 0.82, indicando uma boa capacidade de discriminar entre as classes "desligado" e "ativo". Em contraste, o modelo árvore de decisão tem um AUC de 0.61, mostrando um desempenho significativamente inferior.

A linha tracejada preta representa a linha de referência de uma classificação aleatória (AUC = 0.5). Este gráfico demonstra claramente que os modelos de regressão logística e floresta aleatória são superiores ao árvore de decisão na classificação dos desligamentos.

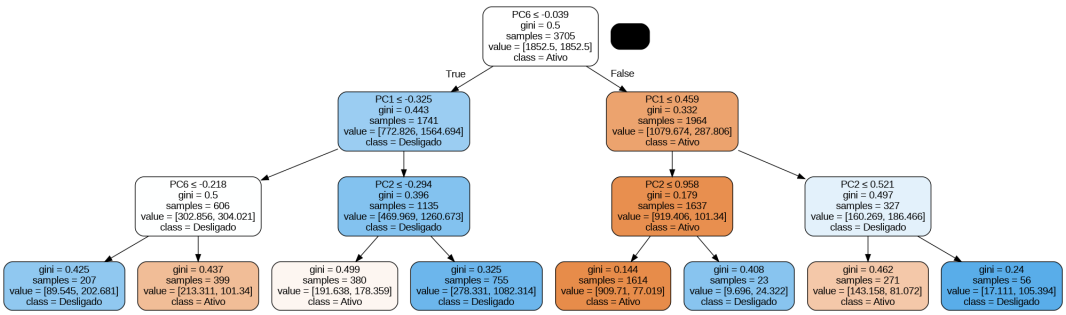
Figura 17 – Curva ROC - Modelos individuais



Fonte: Elaborado pelo autor

A Figura 18 ilustra uma árvore de decisão gerada para prever o risco de saída dos empregados utilizando componentes principais (PCs) derivados de uma Análise de Componentes Principais (PCA). Embora o uso de PCA possa melhorar a performance do modelo ao reduzir a dimensionalidade e eliminar redundâncias, a interpretação das divisões baseadas em PCs pode ser desafiadora. Cada nó da árvore de decisão mostra uma condição sobre um componente principal, e não diretamente sobre as variáveis originais, dificultando a compreensão de como essas variáveis influenciam a decisão final. Por exemplo, a primeira divisão em PC6 ≤ -0.039 e subsequentes divisões em outros componentes principais indicam combinações lineares específicas das variáveis originais que influenciam a classificação. No entanto, construir a árvore de decisão diretamente sobre as variáveis originais permitiria visualizar claramente como cada variável contribui para a classificação final de "Desligado" ou "Ativo", facilitando a identificação de fatores específicos que influenciam a rotatividade dos empregados.

Figura 18 – Árvore de decisão



Fonte: Elaborado pelo autor

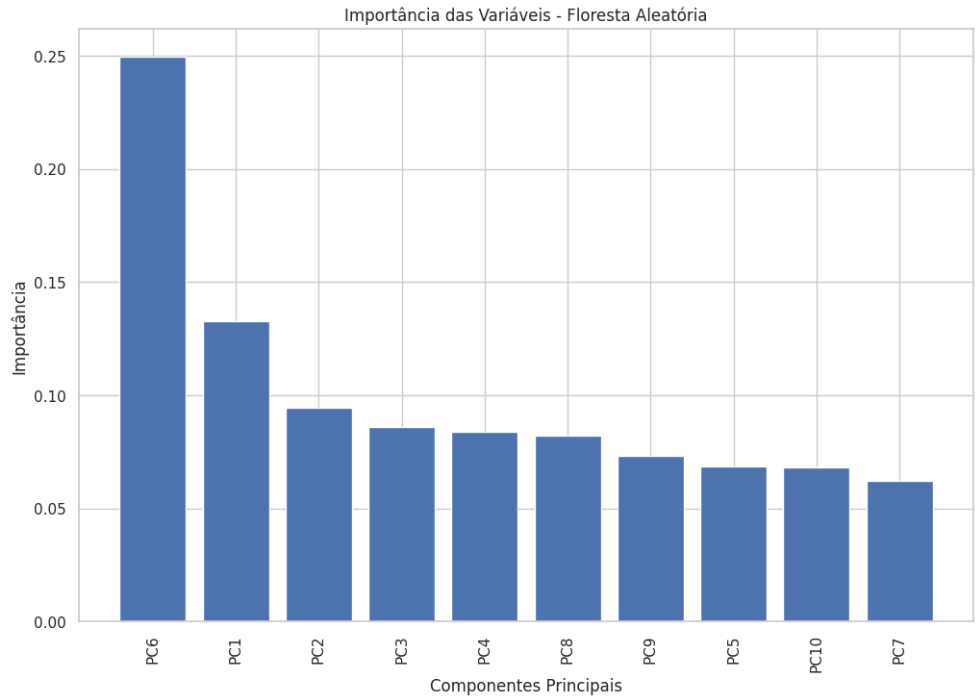
Apesar dessas limitações interpretativas, a árvore de decisão combinada com PCA mostrou-se eficaz, segmentando os empregados com boa precisão e pureza nas folhas finais, com algumas alcançando uma impureza de Gini muito baixa. Por exemplo, o nó raiz mostra um índice de Gini de 0.5, com 3705 amostras, dividindo-as em 1852.5 para cada classe, e subsequentes nós mostram índices de Gini ainda menores, como 0.144 e 0.179 em divisões mais profundas. Para uma aplicação prática, onde a interpretabilidade é crucial, utilizar as variáveis originais diretamente proporcionaria *insights* mais acionáveis e direcionados para intervenções estratégicas.

4.1.1 Importância de Variáveis

A análise da importância das variáveis nos modelos de floresta aleatória e árvore de decisão revela informações cruciais sobre os fatores que mais influenciam a predição do risco de saída de empregados. Conforme ilustrado nos gráficos, ambos os modelos destacam a relevância das componentes principais (PCs) na classificação.

No modelo de floresta aleatória, observa-se que a PC6 tem a maior importância, seguida pelas PCs 1 e 2, indicando que essas variáveis possuem um impacto significativo na decisão do modelo. As demais componentes, como PC3, PC4 e PC8, também apresentam importâncias consideráveis, mas em menor grau. Este modelo distribui a importância de forma relativamente equilibrada entre várias componentes principais, o que pode contribuir para sua alta sensibilidade e capacidade de generalização, conforme ilustrado na Figura 19.

Figura 19 – Importância das Variáveis - Floresta Aleatória

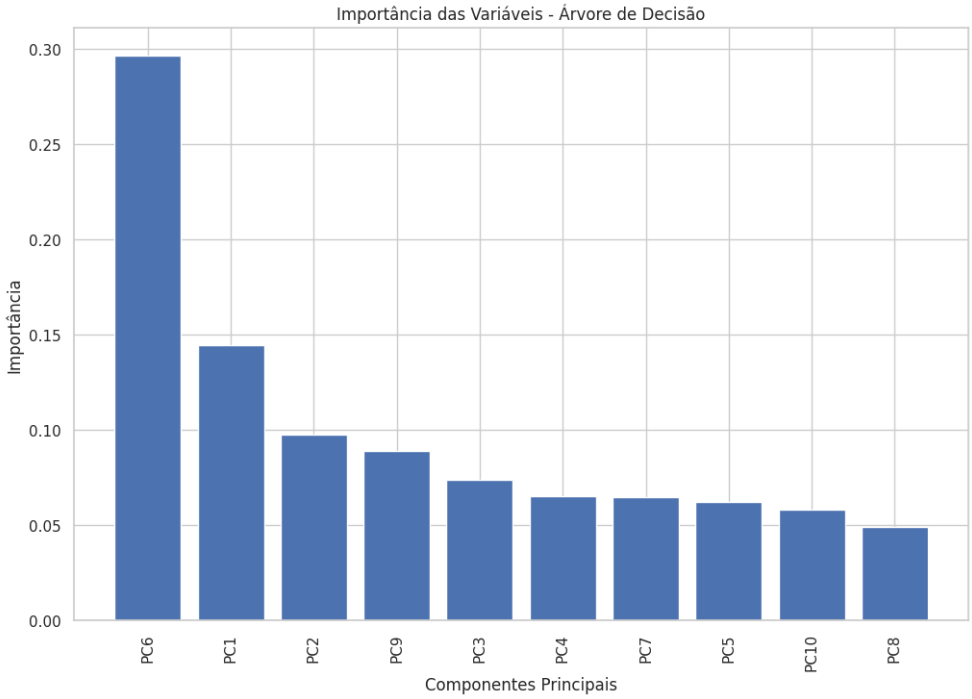


Fonte: Elaborado pelo autor

Por outro lado, o modelo de árvore de decisão apresenta uma concentração maior de importância nas primeiras componentes principais, especialmente PC6, PC1 e PC2, que juntas contribuem com uma parcela significativa da importância total. As componentes PC9 e PC4 também são relevantes, enquanto as demais componentes possuem importância menor. Este padrão sugere que o modelo de árvore de decisão depende fortemente de um número menor de variáveis para suas decisões, o que pode explicar sua maior acurácia, mas menor sensibilidade em comparação ao modelo de floresta aleatória, conforme ilustrado na Figura 20.

Em seguida, o modelo floresta aleatória previamente treinado foi utilizado para prever a probabilidade de desligamento dos empregados. As probabilidades da classe 1 (desligamento) foram calculadas e armazenadas como percentuais em um *DataFrame*, junto ao identificador ID2. O *DataFrame* foi exportado como um arquivo Excel, denominado "Probabilidades_Desligamento_Empregados_RF.xlsx,"e disponibilizado para download. Uma amostragem dos resultados pode ser conferida na Figura 21.

Figura 20 – Importância das Variáveis - Árvore de Decisão



Fonte: Elaborado pelo autor

Figura 21 – Resultado do modelo Floresta Aleatória

ID2	Probabilidade_Desligamento (%)	Previsao_Desligamento	Status Real
1038	0	Ativo	Desligado
3814	16	Desligado	Desligado
3331	18	Desligado	Desligado
11517	27	Desligado	Desligado
12980	32	Desligado	Desligado
1173	0	Ativo	Desligado
10502	26	Desligado	Desligado
12306	35	Desligado	Desligado
13533	32	Desligado	Desligado
1186	0	Ativo	Desligado
6998	15	Desligado	Desligado
1241	5	Ativo	Desligado

Fonte: Elaborado pelo autor

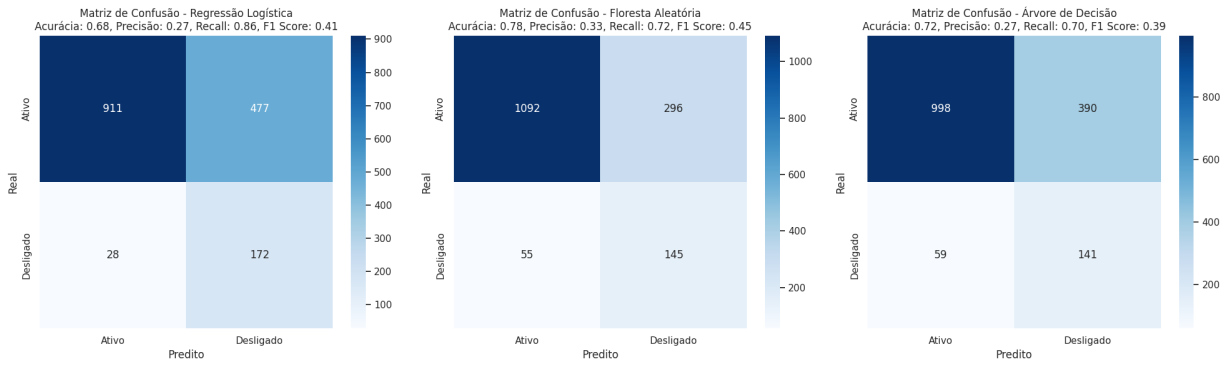
4.2 Avaliação de Modelos Após Ajuste de Hiperparâmetros

Comparando os resultados após o ajuste dos hiperparâmetros, a floresta aleatória, além de apresentar a maior acurácia (0.78), manteve um *recall* excelente (0.72) e um *F1 score* equilibrado (0.45), sugerindo que este modelo continua a ser o mais adequado para a tarefa de predição de saída de empregados. A alta sensibilidade (*recall*) é particularmente importante para a empresa, pois permite identificar a maioria dos empregados que têm alta probabilidade de sair, o que pode ser crucial para a implementação de medidas preventivas. Portanto, até agora, o modelo de Floresta Aleatória ajustado se destaca como a melhor opção, proporcionando um bom equilíbrio entre sensibilidade e precisão.

A Figura 22 apresenta as matrizes de confusão dos modelos de regressão logística, floresta aleatória e árvore de decisão após o ajuste dos hiperparâmetros. A matriz de confusão da floresta aleatória mostra que o modelo conseguiu identificar corretamente 1092 empregados ativos e 145 empregados desligados, resultando em um *recall* de 0.72. A precisão foi de 0.33, indicando que o modelo cometeu alguns erros ao classificar empregados desligados como ativos.

Por outro lado, a matriz de confusão do modelo de regressão logística apresenta uma acurácia de 0.68, com um *recall* de 0.86, mas uma precisão de 0.27, sugerindo que, embora o modelo seja bom em identificar empregados desligados, ele também comete muitos erros ao classificar empregados ativos como desligados. Já o modelo de árvore de decisão tem uma acurácia de 0.72, com um *recall* de 0.70 e uma precisão de 0.27, mostrando um desempenho inferior comparado aos outros modelos.

Figura 22 – Matriz de Confusão Ajustada



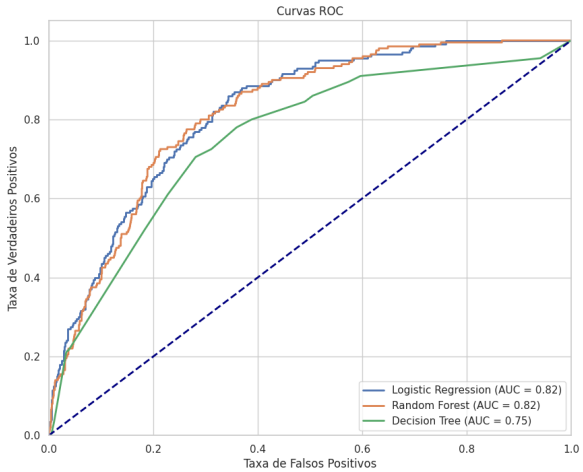
Fonte: Elaborado pelo autor

A Figura 23 apresenta as curvas ROC dos três modelos após o ajuste dos hiperparâmetros. As curvas ROC mostram que tanto a regressão logística quanto a floresta aleatória têm uma AUC de 0.82, indicando uma boa capacidade de discriminação entre as classes "desligado" e "ativo". Em contraste, o modelo de árvore de decisão apresenta uma AUC de 0.75, indicando um desempenho inferior.

Comparando com a Figura 17 apresentada anteriormente, observa-se que os ajustes de hiperparâmetros mantiveram a AUC da regressão logística e da floresta aleatória em 0.82, demonstrando consistência no desempenho desses modelos. No entanto, o modelo de árvore de decisão teve uma melhoria na AUC de 0.61 para 0.75 após os ajustes, embora ainda não seja tão eficaz quanto os outros dois modelos.

A linha tracejada preta representa a linha de referência de uma classificação aleatória (AUC = 0.5). Este gráfico demonstra claramente que os modelos de regressão logística e floresta aleatória são superiores ao modelo de árvore de decisão na classificação dos desligamentos. A proximidade das curvas ROC da regressão logística e da floresta aleatória à parte superior esquerda do gráfico indica que esses modelos têm uma alta taxa de verdadeiros positivos e uma baixa taxa de falsos positivos, o que é desejável em uma tarefa de classificação.

Figura 23 – Curva ROC Ajustada



Fonte: Elaborado pelo autor

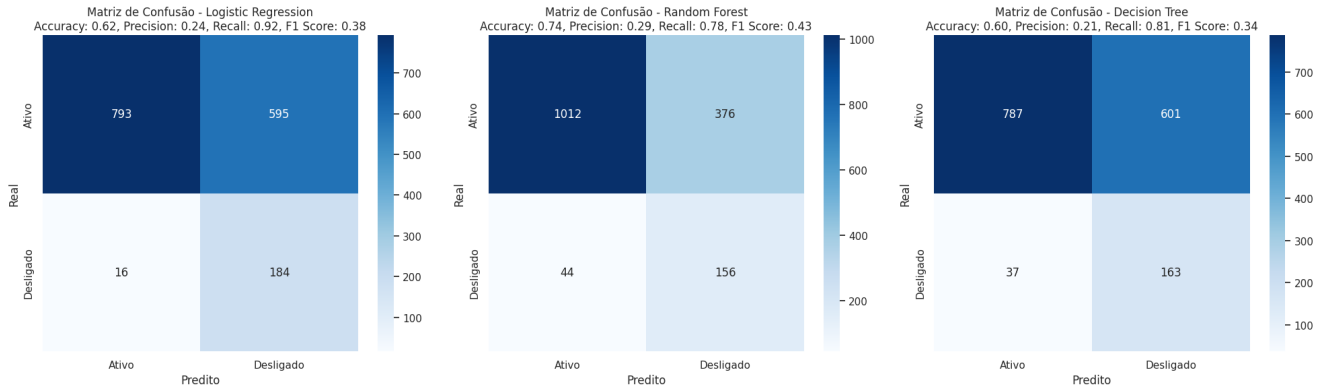
4.3 Avaliação de Modelos Após Aplicação de Validação Cruzada

A validação cruzada foi realizada com cinco *folds* estratificados, garantindo uma distribuição balanceada das classes entre os *folds*. Modelos de regressão logística, floresta aleatória e *árvore de decisão* foram treinados e avaliados durante o processo, recalculando os pesos das classes para cada conjunto de treinamento. Após a execução de todas as iterações, as métricas de precisão, *recall*, acurácia e *F1 Score* foram calculadas e registradas.

Os resultados mostraram que a Regressão Logística obteve uma acurácia de 62%, precisão de 24%, *recall* de 92% e *F1 Score* de 0.38. O modelo floresta aleatória apresentou uma acurácia de 74%, precisão de 29%, *recall* de 78% e *F1 Score* de 0.43. A árvore de decisão apresentou uma acurácia de 60%, precisão de 21%, *recall* de 81% e *F1 Score* de 0.34.

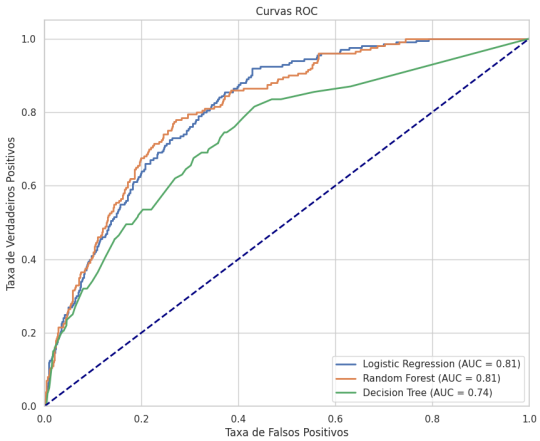
Além disso, a análise das matrizes de confusão na base de treinamento revelou que a regressão logística acertou 62% dos casos de empregados ativos e 92% dos casos de desligamento, indicando uma alta sensibilidade. O modelo de floresta aleatória teve um acerto de 74% para empregados ativos e 78% para desligados, destacando-se pelo equilíbrio entre *recall* e precisão. Por outro lado, a árvore de decisão acertou 60% dos empregados ativos e 81% dos desligados, demonstrando um desempenho inferior em comparação aos outros modelos.

Figura 24 – Matrizes de Confusão após Validação Cruzada



A Figura 24 apresenta as matrizes de confusão dos modelos de regressão logística, floresta aleatória e árvore de decisão após a aplicação da validação cruzada. Observa-se que a floresta aleatória manteve um *recall* alto (78%) para a classe de desligamento, confirmando sua eficácia em identificar empregados que têm maior probabilidade de sair. A precisão de 29% indica que o modelo ainda enfrenta desafios em termos de falsos positivos, mas a acurácia geral de 74% e o *F1 Score* de 0.43 mostram um equilíbrio razoável entre as métricas.

Figura 25 – Curvas ROC após Validação Cruzada



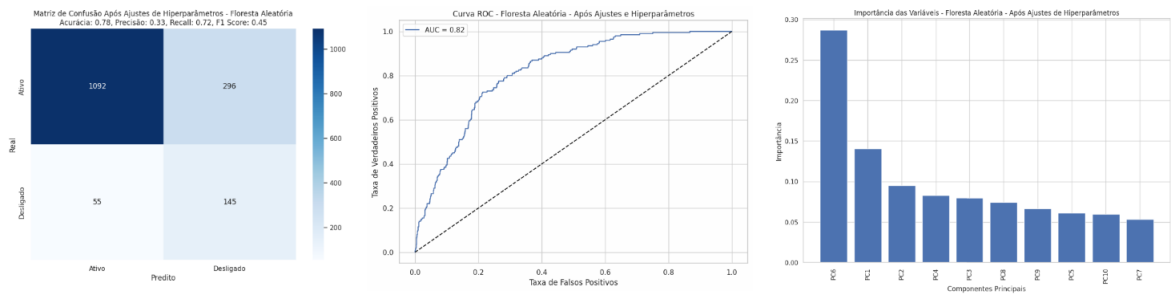
A Figura 25 mostra as curvas ROC dos três modelos após a validação cruzada. A AUC para a regressão logística é de 0.81, para a floresta aleatória é de 0.81 e para a árvore de decisão é de 0.74. Comparando com os resultados anteriores, a AUC da regressão logística e da floresta aleatória permanece alta, indicando consistência no desempenho desses modelos. A árvore de decisão, apesar de ter melhorado levemente sua AUC para 0.74, ainda apresenta um desempenho inferior comparado aos outros dois modelos.

Mesmo com a aplicação da validação cruzada, os resultados mostram que a floresta aleatória continua sendo o modelo com melhor desempenho, especialmente em termos de *recall* para a classe de desligamento, o que é crucial para identificar os empregados que possuem maior risco de sair da empresa. No entanto, ainda há espaço para melhorias, especialmente em termos de precisão para reduzir a quantidade de falsos positivos e melhorar a confiabilidade das previsões.

4.3.1 O melhor modelo

Após uma análise detalhada das três etapas de avaliação (otimização de limiar com o índice de *Youden*, ajustes de hiperparâmetros e validação cruzada), concluiu-se que o modelo de floresta aleatória apresentou o melhor desempenho geral. Especificamente, na etapa de ajustes de hiperparâmetros, o modelo de floresta aleatória atingiu uma acurácia de 0.78, precisão de 0.33, *recall* de 0.72 e *F1 Score* de 0.45, conforme ilustrado na Figura 26. Esses resultados demonstram um equilíbrio entre sensibilidade e precisão, tornando-o o modelo com melhor eficácia para identificar empregados com alto risco de desligamento, ao mesmo tempo em que mantém uma taxa de falsos positivos relativamente baixa. A capacidade de adaptação do modelo através dos ajustes de hiperparâmetros permitiu uma melhor adequação aos dados, resultando em um desempenho superior comparado às outras etapas e modelos. Portanto, a floresta aleatória se destaca como o modelo mais adequado para a predição de saída de colaboradores, oferecendo um balanço ideal entre precisão operacional e capacidade preditiva, essencial para a implementação de estratégias eficazes de retenção de talentos.

Figura 26 – Matriz de confusão, curva ROC e importância das variáveis para o modelo de Floresta Aleatória após ajustes de hiperparâmetros.



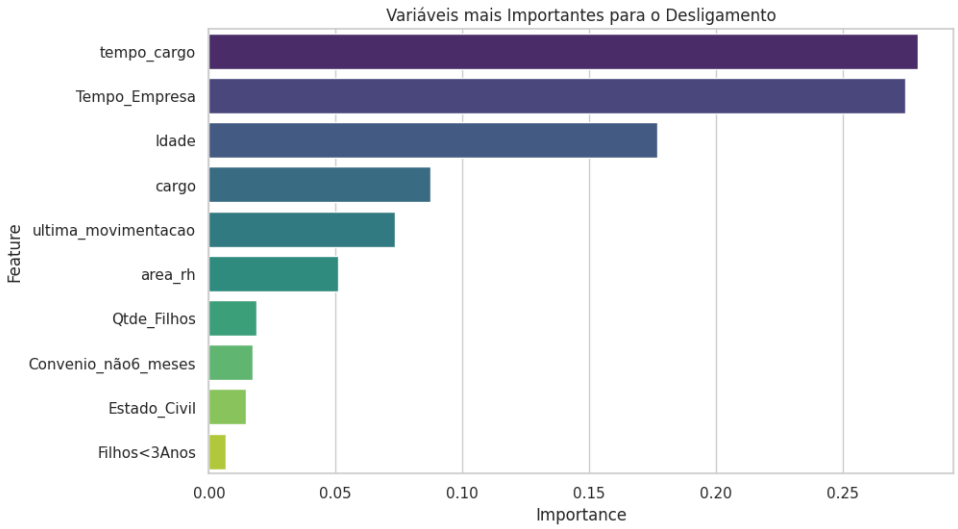
4.4 Análise das variáveis que mais influenciam no desligamento

Para uma análise das variáveis mais importantes para a previsão de desligamento dos empregados, foi utilizado o modelo ajustado de floresta aleatória. Inicialmente, as variáveis categóricas foram codificadas com um método ordinal e as numéricas foram normalizadas utilizando *MinMaxScaler*. Em seguida, selecionaram-se as dez variáveis que mais influenciam na predição de desligamento. Após o ajuste dos pesos das classes para tratar do desbalanceamento dos dados, treinou-se o modelo floresta aleatória com hiperparâmetros otimizados, limitando a profundidade máxima das árvores e aumentando o número de estimadores. A importância das variáveis foi então avaliada através do método de *feature_importance_* do modelo floresta aleatória ajustado.

Os resultados obtidos mostraram que as três variáveis mais importantes para a previsão de desligamento foram: *tempo_cargo* (27,97%), *Tempo_Empresa* (27,46%) e *Idade* (17,68%). A variável *cargo* apareceu como a quarta mais influente, seguida de *ultima_movimentacao*, que representa a última progressão salarial do empregado. Variáveis como *area_rh*, *Qtde_Filhos*, *Convenio_não6_meses*, *Estado_Civil* e *Filhos<3Anos* também apresentaram influência, mas de forma menos significativa. Este resultado destaca a importância do tempo de permanência na empresa e no cargo, bem como a idade, como fatores determinantes na decisão de desligamento dos empregados.

Conforme ilustrado na Figura 27, as variáveis *tempo_cargo*, *Tempo_Empresa* e *Idade* são as mais influentes, seguidas por *cargo* e *ultima_movimentacao*. Essa visualização ajuda a identificar quais fatores têm maior impacto na previsão de desligamento, auxiliando na formulação de estratégias de retenção de empregados.

Figura 27 – Variáveis Mais Importantes para o Desligamento



Fonte: Elaborado pelo autor

4.5 Oportunidades de Melhorias

A análise dos resultados obtidos através das diversas abordagens de modelagem sugere que é possível obter melhores resultados se obtivermos maior qualidade e quantidade de variáveis.

Para testar a eficácia do modelo, foi utilizada uma nova base de dados pública da IBM, composta por 35 variáveis e 1470 empregados, dos quais 1233 estavam ativos e 237 desligados. Nos modelos individuais, a floresta aleatória apresentou o melhor desempenho com uma acurácia de 0.98, precisão de 1.00, *recall* de 0.82 e *F1 Score* de 0.90. Esses resultados evidenciam que o modelo apresentou excelente desempenho com a base de dados da IBM.

No entanto, para alcançar resultados semelhantes na empresa estudada, é necessário revisar a base de dados utilizada. Problemas como a presença de variáveis irrelevantes ou ruidosas, a falta de variáveis críticas, ou mesmo a baixa qualidade dos dados (dados ausentes, inconsistentes ou desbalanceados) podem estar impactando negativamente a performance dos modelos. Além disso, a quantidade de dados pode ser um fator determinante, onde um volume insuficiente pode levar a modelos que não generalizam bem para novos dados. Portanto, uma revisão detalhada dos dados, principalmente no que tange a obtenção de dados adicionais, pode ser crucial para melhorar a performance dos modelos preditivos na empresa estudada.

5 Considerações Finais

Este capítulo final tem como objetivo refletir sobre os resultados alcançados em relação aos objetivos iniciais do estudo e discutir as perspectivas futuras para o aprimoramento da previsão de desligamento de colaboradores na empresa.

Inicialmente, o estudo visava implementar e avaliar diferentes modelos de Aprendizado de Máquina (ML) para prever a probabilidade de saída de colaboradores, buscando fornecer *insights* que auxiliassem nas estratégias de retenção de talentos. Para tal, foram utilizados diversos algoritmos de ML, incluindo regressão logística, árvores de decisão e floresta aleatória, aplicados a conjuntos de dados históricos de saída de colaboradores. Além disso, procurou-se identificar os principais fatores e variáveis influentes na predição da probabilidade de saída e ranquear os empregados de acordo com a probabilidade de desligamento.

Os modelos desenvolvidos apresentaram excelentes métricas de desempenho durante as fases de otimização de limiar, ajustes de hiperparâmetros e validação cruzada, com acurácia, precisão, *recall* e *F1-score* bastante elevados. No entanto, a análise detalhada revelou que o modelo de floresta aleatória se destacou como o melhor modelo. Especificamente, na etapa de ajustes de hiperparâmetros, a floresta aleatória atingiu uma acurácia de 0.78, precisão de 0.33, *recall* de 0.72 e *F1-score* de 0.45, demonstrando um equilíbrio robusto entre sensibilidade e precisão.

Porém, a fim de validar o modelo, independente da base utilizada neste projeto, uma base de dados pública da IBM foi utilizada. A aplicação dos modelos nesta nova base resultou em excelentes previsões, com a floresta aleatória alcançando uma acurácia de 0.98 e um *F1 Score* de 0.90. Além dos excelentes resultados com as métricas, a aplicação prática mostrou-se também com assertividade de 75% (a cada 10 pessoas que saem, o modelo identifica mais de 7). Este resultado revelou que a base de dados original pode conter variáveis irrelevantes, ruidosas ou problemas de desbalanceamento que impactam negativamente a performance dos modelos.

Comparando as variáveis das duas bases de dados, ficou evidente que as variáveis da base da IBM possuem uma maior capacidade preditiva. A base da IBM inclui variáveis como Satisfação com o Ambiente de Trabalho, Envolvimento com o Trabalho, Satisfação nos Relacionamentos no Trabalho, Equilíbrio entre Trabalho e Vida Pessoal, entre outras, que são diretamente relacionadas ao comportamento e satisfação do empregado. Em contraste, a base de dados original, apesar de possuir também variáveis importantes, como nível salarial, temporalidade de progressões salariais e quantidade de filhos, continha um volume maior de variáveis relacionadas a aspectos administrativos e demográficos, como Cargo, Centro de Custo, Região, Sindicato e Grupo de Avaliação. Embora relevantes, essas variáveis não capturam de maneira eficaz os fatores subjetivos e motivacionais que influenciam a decisão de um colaborador de deixar a empresa.

Diante desses *insights*, fica claro que a qualidade dos dados é crucial para a assertividade das previsões. Para melhorar a precisão dos modelos preditivos, é essencial revisar e enriquecer a base de dados atual, incorporando variáveis que capturem melhor os aspectos de satisfação e motivação dos empregados. Variáveis como Satisfação no Trabalho, Envolvimento no Trabalho, Quantidade de Empresas Trabalhadas, Formação Educacional e Frequência de Treinamentos são exemplos de dados que, se adicionados, podem aumentar significativamente a capacidade preditiva dos modelos.

Explorar técnicas de modelagem mais avançadas e integrar fontes adicionais de dados, como análise de sentimentos, podem fornecer uma compreensão mais profunda do contexto organizacional, elevando ainda mais a precisão das previsões. Assim, espera-se que com essas melhorias, seja possível alcançar predições mais assertivas, permitindo intervenções proativas e eficazes para a retenção de talentos, contribuindo para um ambiente de trabalho mais dinâmico e adaptável.

Em conclusão, este estudo demonstrou que o modelo preditivo desenvolvido tem potencial para alcançar excelentes resultados, uma vez que a base estiver mais consolidada. Com essas melhorias, será possível obter previsões mais precisas, auxiliando a equipe de recursos humanos a implementar ações estratégicas para reduzir a rotatividade e reter talentos essenciais para a empresa.

Referências

ANGRAVE, D.; CHARLWOOD, A.; KIRKPATRICK, I.; LAWRENCE, M. Hr and analytics: why hr is set to fail the big data challenge. **Human Resource Management Journal**, v. 26, p. 1–11, 01 2016. Disponível em: <https://www.researchgate.net/publication/292152333_HR_and_analytics_why_HR_is_set_to_fail_the_big_data_challenge>. Acesso em: 08 fev. 2023.

BERSIN, J. **People Analytics Market Growth: Ten Things You Need to Know**. 2016. Accessed: 2024-06-24. Disponível em: <https://joshbersin.com/2016/07/people-analytics-market-growth-ten-things-you-need-to-know/>.

BOHLANDER, G.; SNELL, S. **Administração de Recursos Humanos**. Tradução da 14ª edição norte-americana. [S.l.]: Cengage, 2016. 592 p. ISBN 978-8522106820.

BOUSHEY, H.; GLYNN, S. J. **There Are Significant Business Costs to Replacing Employees**. [S.l.], 2012. Disponível em: <http://www.americanprogress.org/issues/labor/report/2012/11/16/44464/there-are-significant-business-costs-to-replacing-employees/>. Acesso em: 09 fev. 2023.

CARLESS, S. A. Person–job fit versus person–organization fit as predictors of organizational attraction and job acceptance intentions: A longitudinal study. **Journal of Occupational and Organizational Psychology**, v. 78, n. 3, p. 411–429, 2005. Disponível em: <https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1348/096317905X25995>. Acesso em: 16 dez. 2023.

CHIAVENATO, I. **Introdução à teoria geral da administração: uma visão abrangente da moderna administração das organizações**. 7. ed. [S.l.]: Elsevier, 2003. ISBN 85-352-1348-1.

CHIAVENATO, I. **Comportamento Organizacional: A dinâmica do sucesso das organizações**. 3. ed. [S.l.]: Editora Manole Ltda., 2014. 492 p. ISBN 978-85-204-3798-8.

CHOUREY, A.; PHULRE, S.; MISHRA, S. A survey paper on employee attrition prediction using machine learning techniques. **Journal of Interdisciplinary Cycle Research**, v. 11, n. 12, p. 199–202, 2019.

CINTRA, J. C. **Psicologia organizacional e do trabalho**. Editora e Distribuidora Educacional S.A., 2018. v. 3. 280 p. ISBN 978-85-522-1171-6. Disponível em: <https://cm-kls-content.s3.amazonaws.com/201802/INTERATIVAS_2_0/PSICOLOGIA_ORGANIZACIONAL_E_DO_TRABALHO_I/U1/LIVRO_UNICO.pdf>. Acesso em: 03 jan. 2024.

COHEN, J. **Statistical Power Analysis for the Behavioral Sciences**. 2nd. ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine Volume 17 Number 3 (1996) (© AAAI)**, 1996.

FRANCO, T.; DRUCK, G.; SELIGMANN-SILVA, E. As novas relações de trabalho, o desgaste mental do trabalhador e os transtornos mentais no trabalho precarizado. **Revista Brasileira de Saúde Ocupacional**, Fundação Jorge Duprat Figueiredo de Segurança e Medicina do Trabalho - FUNDACENTRO, v. 35, n. 122, p. 229–248, 2010. ISSN 0303-7657. Disponível em: <https://doi.org/10.1590/S0303-76572010000200006>. Acesso em: 12 nov. 2023.

GéRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. [S.l.]: O'Reilly Media, Inc., 2019. 856 p. ISBN 9781492032649.

HONGYU, K.; SANDANIELO, V.; JUNIOR, G. AnÁlise de componentes principais: Resumo teÓrico, aplicaÇÃo e interpretaÇÃo. v. 5, p. 83–90, 07 2016.

KANE-SELLERS, M. L. **Predictive Models of Employee Voluntary Turnover in a North American Professional Sales Force Using Data-Mining Analysis**. Tese (Doctor of Philosophy Dissertation) — Texas A&M University, Texas, August 2007. Major Subject: Educational Human Resource Development.

KHERA, S. Predictive modelling of employee turnover in indian it industry using machine learning techniques. **Vision**, v. 23, p. 12–21, 2019.

LATTIN, J.; CARROLL, J. D.; GREEN, P. E. **Análise de Dados Multivariados**. [S.l.]: Cengage Learning, 2011. 475 p. ISBN 9788522109012.

LIAW, A.; WIENER, M. Random forest. **R News**, The R Foundation, v. 2, n. 3, p. 18–22, 2002.

LUCENA, M. da S. **Planejamento de recursos humanos**. Atlas, 2007. ISBN 9788522406197. Disponível em: <https://books.google.com.br/books?id=K1ByPgAACAAJ>. Acesso em: 09 fev. 2023.

MARLER, J.; BOUDREAU, J. An evidence-based review of hr analytics. **The International Journal of Human Resource Management**, v. 28, p. 1–24, 11 2016. Disponível em: <https://www.researchgate.net/publication/310048235_An_evidence-based_review_of_HR_Analytics>. Acesso em: 17 fev. 2023.

Ministério do Trabalho e Emprego. **Cadastro Geral de Empregados e Desempregados (CAGED) - Apresentação 2023**. 2023. Acessado em: 22 dezembro 2023. Disponível em: <http://pdet.mte.gov.br/images/Novo_CAGED/2023/202308/2-apresentacao.pdf>.

MITCHELL, T. **Machine Learning**. McGraw Hill, 2017. (McGraw Hill series in computer science). ISBN 9781259096952. Disponível em: <https://books.google.com.br/books?id=ifdcswEACAAJ>.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. [S.l.]: Massachusetts Institute of Technology, 2012. (Adaptive Computation and Machine Learning). ISBN 978-0-262-01802-9.

NORRMAN, F. **Predicting employee attrition with machine learning on an individual level and the effects it could have on an organization**. Dissertação (Master of Science Thesis) — KTH Royal Institute of Technology, School of Industrial Engineering and Management, Stockholm, Sweden, 6 2020. Approved June 1, 2020. Examiner: Anna Jerbrant. Supervisor: Gisela Bäcklander. Commissioner: Populum.

QUEIROZ, S. **Gestão das emoções no ambiente corporativo: descubra como o foco nas emoções individuais pode mudar a vida de lideres e liderados nas empresas**. Literare Books, 2020. v. 2. 352 p. ISBN 9788594552860. Disponível em: <https://books.google.com.br/books?id=hu_rDwAAQBAJ>. Acesso em: 03 jan. 2024.

RUBENSTEIN, A.; EBERLY, M.; LEE, T.; MITCHELL, T. Looking beyond the trees: A meta-analysis and integration of voluntary turnover research. **Academy of Management Proceedings**, v. 2015, p. 12779–12779, 01 2015.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: A Modern Approach (2nd Edition)**. [S.l.]: Prentice Hall, 2002. ISBN 0137903952.

SANTOS, E. A. P. dos; CRUZ, M. T. de S. (Ed.). **Gestão de pessoas no Século XXI: desafios e tendências para além de modismos**. Tiki Books PUC-SP/PIPEq, 2019. 234 p. ISBN 978-85-66241-18-1. Disponível em: <https://www.pucsp.br/sites/default/files/download/graduacao/cursos/administracao/Gestao-de-Pessoas-no-Seculo-XXI.pdf>. Acesso em: 10 nov. 2023.

SAYAH, F. **Decision Trees & Random Forest for Beginners**. 2021. <https://www.kaggle.com/code/faressayah/decision-trees-random-forest-for-beginners>. Accessed: 2023-05-14.

SUTTON, R.; BARTO, A. **Reinforcement Learning, Second Edition**. [S.l.]: The MIT Press, 2018. 552 p. ISBN 9780262039246.

YEDIDA, R.; REDDY, R.; VAHI, R.; JANA, R.; GV, A.; KULKARNI, D. **Employee Attrition Prediction**. 2018.