AVALIAÇÃO COMPARATIVA DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA PREVISÃO DE TURNOVER, COM USO DE INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

COMPARATIVE EVALUATION OF MACHINE LEARNING TECHNIQUES IN TURNOVER FORECASTING, USING EXPLAINABLE ARTIFICIAL INTELLIGENCE

Wagner Vidal Xavier da Silva[1] , Flávio Rosendo da Silva Oliveira[1]

[1]Análise e Desenvolvimento de Sistemas - Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco – Campus Paulista (IFPE) - Paulista - PE - Brazil

wvxs@discente.ifpe.edu.br, flavio.oliveira@paulista.ifpe.edu.br

---

Resumo. Este artigo realizou uma análise comparativa de dez algoritmos de Aprendizado de Máquina para prever cenários de *turnover* em organizações. Foram utilizadas duas bases de dados do site *Kaggle*: a primeira em relação a uma empresa americana, que não teve o nome divulgado, com 9.540 registros e a outra do trabalho de prevenção de rotatividade de funcionários da Portobello *Tech* com 11.995 registros, ambas com dez atributos cada. As métricas catalogadas foram precisão, *F1-Score*, *Recall* e Acurácia. Diante dos resultados dos algoritmos, o *Random Forest* se destacou, apresentando acurácia de 0,84 para a primeira base e 0,95 para a segunda em relação aos conjuntos testes das bases utilizadas. Após essa etapa, a metodologia *SHAP* foi utilizada para identificar os atributos mais impactantes na intenção de *turnover*. A média de horas trabalhadas mensais, a satisfação do funcionário e o tempo na empresa foram os atributos mais relevantes em ambas as bases para a decisão de pedir demissão.

Palavras-chave: *Turnover*; Aprendizagem de máquina; Metodologia *SHAP*.

Abstract. *This article carried out a comparative analysis of ten Machine Learning algorithms for predicting turnover scenarios in organizations. Two databases from the Kaggle website were used: the first for an American company, whose name was not disclosed, with 9,540 records and the other for Portobello Tech's employee turnover prevention work with 11,995 records, both with ten attributes each. The metrics cataloged were Precision, F1-Score, Recall and Accuracy. Given the results of the algorithms, Random Forest stood out, with an accuracy of 0.84 for the first base and 0.95 for the second in relation to the test sets of the bases used. After this stage, the SHAP methodology was used to identify the attributes with the greatest impact on turnover intention. The average number of hours worked per month, employee satisfaction and time with the company were the most relevant attributes in both bases for the decision to resign.*

Keywords*: Machine learning; Turnover; Random Forest; SHAP Methodology.*

---

1. Introdução

A demanda por funcionários altamente qualificados em um mercado competitivo leva os profissionais a buscarem aprimoramento contínuo e foco nos resultados organizacionais. Contudo, na mesma medida em que esses funcionários são cobrados, eles esperam que a organização ofereça um ambiente favorável ao desenvolvimento profissional (Moreira; Da Silva Nantes, 2024). Caso contrário, é muito provável que esse profissional peça demissão. Segundo Corrêa (2016), há uma estimativa de que as novas gerações de profissionais tendem a ser mais qualificadas e menos hesitantes em trocar o emprego atual por outro que ofereça melhores benefícios ou perspectivas de carreira. Quando essa expectativa realmente não é atendida, ocorre o fenômeno do *turnover*.

Segundo Boen (2022), *turnover* refere-se ao número de funcionários que saem da empresa em um período de tempo determinado, por vontade própria ou involuntariamente. Nesse sentido, a rotatividade se dá pelo fluxo de admissões e demissões de pessoas em uma organização. Silva (2012) descreve que o aumento do índice de *turnover* requer atenção dos gestores, pois se trata de um indicador do nível de satisfação dos colaboradores para com a organização e as demais políticas de gestão de pessoas da empresa.

Nesse sentido, prever esses índices de cenários de *turnover* organizacional é uma possibilidade que os gestores possuem de antecipar decisões a fim de que essas mudanças não provoquem problemas na dinâmica da organização. Além disso, entender a motivação dos funcionários e como isso pode atrapalhar na produtividade também se faz necessário a fim de criar um clima organizacional que favoreça o desenvolvimento do colaborador (Boen, 2022). A questão do custo financeiro de substituição de um funcionário deve ser levado em consideração. Haran e Nierderman (2022) apontam pesquisas que mostram que o custo para substituir um colaborador é aproximadamente 100% do salário orçado para a posição. Além disso, perde-se produtividade e abala a imagem da organização (Berger & Berger, 2017).

Considerando a importância da temática de *turnover*, o uso da tecnologia da informação pode desenvolver análises inovadoras de dados frente às diversas possibilidades desenvolvidas no campo da gestão de pessoas, sendo uma ferramenta aliada aos gestores.

Boen (2022) destaca que existe uma área de atuação do setor de RH, conhecida como *People Analytics*, a qual desenvolve estratégias baseadas em dados com o auxílio de aprendizagem de máquina, utilizando análises descritivas, visuais e estatísticas de dados relacionados a processos de RH, inclusive em relação à rotatividade de funcionários. Nesse sentido, a capacidade de mapear e entender os riscos associados ao *turnover* é fundamental para o desenvolvimento de estratégias eficazes de retenção de talentos e mitigação do *turnover* (Crisóstomo, 2010).

A aprendizagem de máquina, conforme definida por Gonçalves (2021), é uma área de estudo da computação que consiste em fornecer a computadores a capacidade de aprender a realizar tarefas sem serem explicitamente programados para tanto, através da exposição de dados. Considerando o cenário de *turnover*, com o apoio da aprendizagem de máquina, Gao (2019) demonstra que as empresas podem prever melhor quais funcionários irão deixar a organização no futuro, podendo planejar com antecedência e tomar medidas para reduzir esta probabilidade. Além disso, será utilizado no estudo a metodologia *SHAP* para interpretação dos resultados. Segundo Boen (2022), a metodologia *SHAP* usa conceitos do valor de *Shapley* para calcular a importância dos atributos do classificador. Nesse sentido, é uma técnica utilizada para compreender o resultado final dos classificadores e serve para verificar o quanto cada atributo impactou no resultado final do processamento do algoritmo (Lunderberg; Erion; Lee, 2017).

1.1 Objetivos

1.1.1 Objetivo Geral

Diante do contexto organizacional trazido pelo fenômeno do *turnover*, o objetivo geral do trabalho é realizar uma análise comparativa de algoritmos de aprendizagem de máquina para predição de *turnover*, viabilizando o uso de tecnologia nesse contexto organizacional. A ideia de comparar técnicas de aprendizagem de máquina é permitir, segundo Dietterich (1995), avaliar a capacidade de generalização dos modelos, observando seus comportamentos em novos dados não vistos durante o treinamento. Além disso, a comparação, de acordo com Domingos (2012), ajuda a compreender as possíveis vantagens e desvantagens de cada técnica, auxiliando na escolha do método mais adequado para

aplicações futuras. Para o presente estudo, serão utilizadas dez técnicas para comparação em duas bases de dados.

1.1.2 Objetivos Específicos

Para o presente estudo, são idealizados os seguintes objetivos específicos para a temática de aprendizagem de máquina e *turnover* organizacional:

● Interpretar os resultados com o uso da metodologia *SHAP*;

● Identificar os atributos de maior impacto na intenção de *turnover* com a metodologia *SHAP*.

2. Fundamentação Teórica

2.1 Aprendizagem de Máquina e Metodologia *SHAP*

O uso da aprendizagem de máquina (AM), também conhecida com *machine learning*, já é uma realidade em diversos cenários organizacionais. Nesse sentido, o objetivo é desenvolver técnicas computacionais sobre o aprendizado com o intuito de construir sistemas capazes de adquirir conhecimento automaticamente (Oliveira *et al.,* 2020). Entende-se a AM como um segmento subjacente da Inteligência Artificial (IA), de acordo com Deisenroth, Faisal e Ong (2020), a qual permite que as máquinas extraiam informações e aprendam a tomar decisões a partir de dados e exemplos. Diante desse cenário, os algoritmos de AM são construídos com o intuito de permitir que o computador tome decisões com base em conhecimento de dados prévios assim como os dados utilizados pelo usuário assim como prever resultados (Jakhar; Kaur, 2019). E essas decisões tentam minimizar a quantidade de erros e aumentar a probabilidade de suas previsões serem verdadeiras (Jakhar; Kaur, 2019). Além disso, sua habilidade de se modificar constantemente, quando exposto a mais dados, torna o desenvolvimento de alguns tipos de aplicações mais eficientes (Jordan; Mitchel, 2015).

O uso da AM tem diversas possibilidades e utilizá-las para prever resultados é uma delas. A predição de resultados objetiva predizer o valor de um determinado atributo (variável) baseado nos valores de outros atributos (Jordan; Mitchel, 2015). Nesse sentido, o atributo a ser predito é comumente conhecido como a variável preditiva,

dependente ou alvo, enquanto que os os atributos usados para fazer a predição são conhecidos com as variáveis preditoras, independentes ou explicativas (Jakhar; Kaur, 2019).

Considerando a ideia de predizer valores de um perfil de *turnover* em cenários organizacionais, utilizam-se técnicas de predição com intuito de promover a comparação na eficiência computacional. Define-se eficiência computacional, segundo Santos e Couto (2023), como a medida pelo tempo de execução das análises e da exatidão dos resultados obtidos através da comparação dos resultados. A primeira delas é a de *Random Forest*. Conforme definido por Lima (2021), o método *Random Forest* é um classificador que consiste em uma coleção grande de árvores de decisão estruturadas, nas quais os preditores são distribuídos de forma independente dentre as árvores. A segunda utilizada é a de Árvores de Decisão. Marim (2021) retrata que são representações simples que predizem classes baseadas nos valores de tributos de uma base de dados. Nesse sentido, utiliza-se da estratégia de dividir para conquistar: decompondo um problema complexo em subproblemas mais simples.

A terceira é a *K-Nearest Neighbors* (KNN). Destaca-se como característica central a qualidade de buscar classificar um dado indivíduo com base na classificação dentre os indivíduos mais próximos dele (LE, 2021). A Regressão Logística foi a quarta selecionada. Nela, aponta-se como um modelo estatístico comum para resolver problemas de classificação binária (Silva; Silva Neto, 2023). A quinta técnica é a *Support Vector Machines* (SVM). De acordo com Teramoto (2020), SVM é uma técnica de aprendizagem de máquina derivada de duas fundamentações sólidas: Teoria da Aprendizagem Estatística e Otimização Matemática.

Em relação à sexta técnica, aponta-se a *Naive Bayes*. Segundo Souza (2023), o *Naive Bayes* é considerado um dos algoritmos mais simples, porém completo, para classificação de dados. O seu principal conceito se baseia na teoria das probabilidades. O XGBOOST, a sétima técnica, baseia-se em aumento de gradiente utilizando para problemas de ranking em processamento de linguagem natural (Baldo, 2022). Já a oitava técnica, denominada ADABOOST *CLASSIFIER*, conforme Marques e Ara (2019), é um metaestimador que começa ajustando um classificador no conjunto de dados original e, em seguida, ajusta cópias adicionais do classificador no mesmo conjunto de dados.

Sobre a nona técnica, LIGHTGBM, observa-se um algoritmo mais eficiente, apresentando maior acurácia, principalmente no sentido de consumo de memória e

velocidade de treinamento (Jabeur; Mefteh-Wali; Viviani, 2024). Por fim, o *Perceptron* Multicamada, ou *MLP*, sendo uma arquitetura de rede neural artificial composta por múltiplas camadas de neurônios, incluindo uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada neurônio em uma camada está conectado a todos os neurônios da camada seguinte, proporcionado uma capacidade significativa de aprendizado de padrões complexos em dados (Santos, 2019)**.**

Além das técnicas de AM apresentadas, outro instrumento pertinente no processo de análise de *turnover* é a metodologia *SHAP*. Segundo Lundberg e Lee (2017), metodologia *SHAP* é uma técnica de interpretabilidade que usa conceitos do valor de *Shapley*, valor dado a uma das soluções para a teoria cooperativa dos jogos, para calcular a importância dos atributos do classificador. Boen (2022) acrescenta que essa metodologia tem como vantagem ser agnóstica, usando a mesma metodologia de interpretabilidade para qualquer tipo de classificador, e calculando a explicação a partir dos valores de entrada e saída do classificador. Além disso, essa metodologia está relacionada à explicação de modelos de aprendizagem de máquina, auxiliando a entender como a contribuição de cada variável afeta a classificação do modelo (Lundberg; Lee, 2017).

2.2 *Turnover*

O *turnover* é uma expressão em inglês que caracteriza o fluxo de entradas e saídas de profissionais associados a uma organização em um determinado período (Moreira; Da Silva Mantes, 2024). Essas movimentações podem ser motivadas por razões diversas: seja por descontentamento com alguma política da empresa, falta de motivação, problema nas relações interpessoais, ou busca de uma melhor colocação profissional. Além disso, as empresas têm o direito de buscar profissionais mais capacitados para integrar o seu quadro funcional ou ainda procurar pela inovação e reforma profissional de seus colaboradores, o que reflete também em *turnover* (Medeiros; Alves; Ribeiro, 2012). Nesse sentido, Pinheiro e Souza (2013) relata o turnover como um índice mensurável que relaciona o número de funcionários desligados da organização em um determinado período em relação ao número médio do quadro de pessoal efetivo, considerando apenas o número de desligamentos.

$$Turnover = \frac{Número\ de\ funcionários\ desligados}{Número\ médio\ do\ quatro\ de\ pessoal\ efetivo}$$

Devido ao aumento atual de investimentos das organizações na área de Recursos Humanos e Gestão de Pessoas, segundo Costa e da Silva (2020), é cada vez mais frequente a avaliação das principais causas que levam os colaboradores a saírem de um empresa e também os fatores que levam a organização a demiti-lo. Nesse sentido, entender as causas e os determinantes do *turnover* é imprescindível, uma vez que se trata de um fenômeno que gera cursos para as organizações, que, em geral, não são conhecidos e, por conseguinte, não são controlados (Pinheiro; Souza, 2013).

Além dos transtornos financeiros que os altos índices de *turnover* geram, seja com demissões ou admissões, a falta de mão de obra pode comprometer a produtividade da organização e, consequentemente, a eficiência operacional. Dessa forma, a rotatividade de pessoas não deve ser vista apenas como uma causa, porém, o efeito e consequências de alguns fenômenos internos e externos que condicionam atitudes e comportamentos dos colaboradores (Chiavenato, 2014). Portanto, o elevado índice de *turnover* reflete a desajustes na gestão organizacional que precisam ser melhorados.

Diante das consequências que o *turnover* pode propiciar dentro da organização, gerenciar suas taxas é fundamental para saúde e sucesso organizacional a longo prazo (Boen, 2022). Pesquisadores organizacionais mostram que o *turnover* voluntário de funcionários qualificados acarreta uma grande perda para a empresa, começando com todo o valor substancial que é perdido, reduzindo o desempenho financeiro assim como impactando diretamente na vantagem competitiva de ter funcionários diferenciados (Boen, 2022). Esse fenômeno, caso não seja identificado a tempo, pode ocasionar um *turnover* coletivo causando ainda mais danos (Hausknecht; Trevor; Howard, 2009).

Com o intuito de atenuar esse problema de *turnover*, algumas organizações têm conduzido pesquisas e técnicas de mineração de dados para avaliar o nível de satisfação e indicar potenciais razões por trás da insatisfação dos funcionários (Ajit, 2016). Dessa forma, a tecnologia se torna parceira nesse processo, modelando os dados com a função de auxiliar os gestores responsáveis na identificação da motivação de cada funcionário, possibilitando a elaboração de medidas para o planejamento de retenção do capital, em consonância às teorias da motivação e estratégia empresarial (Santos, 2020).

2.3 Trabalhos Relacionados

Diversos estudos têm investigado os fatores que impactam a questão da intenção de *turnover* nas empresas. O trabalho de Boen (2022) desenvolveu um projeto que engloba técnicas de aprendizado de máquina aplicadas no contexto de *People Analytics*, aplicando modelos preditivos supervisionados para classificação de *turnover*. Além disso, objetivou auxiliar o processo de tomada de decisão da empresa estudada em questão e mitigar a relação da probabilidade de *turnover* dos casos de falso positivo com o tempo até o *turnover* futuro desses casos (Boen, 2022). Nesse sentido, os dados adquiridos foram aplicados técnicas de processamento de classificação *Random Forest*, *Naive Bayes*, Regressão Logística e Árvores de Decisão.

Após a compilação e a análise, considerando as medidas de precisão, sensibilidade, *F1-score* e AUC, as técnicas de Regressão Logística e *Random Forest* apresentaram melhores desempenhos, com valores de acurácia de 0,72 e 0,82, respectivamente (Boen, 2022). Ao empregar a metodologia *SHAP* em seu trabalho, Boen (2022) avaliou a importância dos atributos de um classificador *Random Forest*, alcançando acurácia de 0,85. Com o uso dessa metodologia, observou-se, por exemplo, que o atributo *Relationship*, referente aos relacionamentos dentro da organização, impactou significativamente no aumento da probabilidade de intenção de turnover (Boen, 2022).

No trabalho da Jaisawal (2022) foi utilizado o modelo de *Random Forest* para investigar as características do capital humano que influenciam a remuneração em empresas de TI indianas. Observou-se que, nessa pesquisa, a autora buscou alternativas de retenção de talentos da força de trabalho indiana com o auxílio de aprendizagem de máquina. Para isso, foram testadas cinco técnicas, mas a de *Random Forest* apresentou melhores resultados com base nos parâmetros de Erro Quadrático Médio da Raiz (RMSE), coeficiente de correlação e acurácia de 0,93 (Jaisawal, 2022).

Alshehhi (2021) buscou entender as razões potenciais pelas quais os funcionários deixam seus empregos. Para isso, o autor empregou modelos de aprendizagem de máquina, utilizando algoritmos de *Random Forest, K-Nearest Neighbors* (KNN) e Regressão Logística para desenvolver um modelo preditivo que alcançasse os objetivos da pesquisa. Observou-se que o modelo de *Random Forest* foi o mais bem avaliado e aplicado no departamento de RH da empresa estuda na pesquisa, revelando que a maioria dos funcionários treinados permanece na empresa, indicando que as estratégias de retenção de talentos são boas (ALSHEHHI, 2021).

Em conformidade com as pesquisas mencionadas anteriormente, este estudo também utilizou ferramentas de aprendizagem de máquina e metodologia *SHAP*, porém, para avaliar , utilizou dez técnicas de algoritmos que foram *Random Forest*, Árvores de Decisão, KNN, Regressão Logística, SVM, *Naive Bayes*, XGBOOST, ADABOOST *Classifier*, *LightGBM* e *Perceptron* Multicamada (MLP), em duas bases de dados. Uma das principais preocupações deste estudo foi garantir a precisão dos resultados, por isso, cada técnica foi submetida a 30 rodadas de teste para minimizar discrepâncias nos resultados. Além da métrica de acurácia, amplamente utilizada em estudos de aprendizagem de máquina para avaliar a confiabilidade e validade dos algoritmos, este trabalho reforçou os resultados com as métricas de precisão, *f1-score* e *recall*.

3. Metodologia

A pesquisa se classifica, quanto a natureza dos dados trabalhados, como quantitativa. Segundo Will (2012), a pesquisa quantitativa permite classificar e realizar análise traduzindo os resultados em números, para serem classificados e consequentemente analisados. Além disso, conforme Prodanov (2013), a abordagem quantitativa requer o uso de recursos e técnicas de estatística, procurando traduzir em números os conhecimentos gerados na pesquisa. Nesse sentido, utilizam-se métricas estatísticas e ferramentas de aprendizagem de máquina a fim de responder a problemática em questão, seguindo os procedimentos metodológicos. Em relação ao objetivo do estudo, a intenção é que seja uma abordagem descritiva. Segundo Gil (2017), as pesquisas descritivas têm como objetivo a descrição das características de determinada população ou fenômeno. Além disso, podem ser elaboradas também com a finalidade de identificar possíveis relações entre variáveis.

Nas subseções seguintes, todo o preparo dos dados até a etapa final da aplicação dos algoritmos na predição de cenários organizacionais de *turnover* será descrito. Na subseção 3.1, é descrito o entendimento dos dados referente às duas bases de dados trabalhadas no trabalho. Na subseção 3.2, é detalhada os parâmetros utilizados para aprimorar as bases de dados para os processamentos dos algoritmos. A subseção 3.3 é descrita a modelagem das técnicas de aprendizagem máquina selecionadas, em consonância aos critérios de avaliação, descritos na subseção 3.4.

3.1 Entendimento dos dados

Para o presente trabalho, foram coletadas duas bases de dados do site Kaggle. A

primeira delas, intitulada de *Employee Churn Data* (2024), representa um banco de dados de uma grande empresa dos EUA, não sendo fornecido nenhuma informação que a identifique por motivos de privacidade. Nesse sentido, o departamento de RH reuniu dados sobre quase 10 mil funcionários que deixaram a empresa entre 2016-2020. Foram usadas informações de entrevistas de desligamento, avaliações de desempenho e registro de funcionários.

A base *Employee Churn Data* (2024) é descrita como binária, a qual a problemática inserida é de classificação, ou seja, pretende-se prever a qual classe determinados dados pertencem., que, de acordo com a pesquisa, seria identificar se o usuário, a partir das características apresentadas, pedirá ou não demissão. O conjunto de dados conta com 9540 registros, mensurados em nove colunas, conforme o dicionário de dados seguinte:

Tabela 1 - Descrição dos atributos da base de dados 01 - *Employee churn data*

Atributo	Descrição	Tipo de Dado
Bônus	Indicação se o funcionário recebia a mais em relação às metas. Valor 1 para sim, 0 para não.	Booleano
Houve promoção?	Indicação se o funcionário mudou de cargo no período trabalhado. Valor 1 para sim, 0 para não.	Booleano
Departamento	Subdivisão da área de atuação do funcionário em: suporte, vendas, RH, técnico, marketing, risco, contabilidade, gerenciamento de produto e gerencial.	String
Média de horas trabalhadas mensais	Quantidade média de horas trabalhadas mensalmente.	Inteiro
Quantidade de projetos trabalhados	Número de projetos atribuídos a determinado colaborador.	Inteiro
Faixa salarial	Média de valores recebidos na organização, categorizado em uma faixa salarial baixa, média e alta.	String
Satisfação do funcionário	Indicador da satisfação do colaborador com o trabalho.	Float
Tempo de empresa	Período do funcionário na organização.	Inteiro
Última avaliação de desempenho	Indicador atribuído pela organização acerca do desempenho funcional do colaborador.	Float

Fonte: dados da pesquisa, 2024.

Para a segunda base de dados, nomeada de *Employee Turnover* (2024), foi utilizada a base dados referente ao trabalho de predição de rotação de funcionários da empresa Portobello *Tech*, apresentando 11.995 registros. A base tem como problemática uma classificação binária, no intuito de identificar, a partir das características apresentadas, se o funcionário pedirá ou não demissão. Em relação às colunas da base de dados, pode-se observar uma semelhança com a primeira, apresentando as seguintes variáveis:

Tabela 2 - Descrição dos atributos da base de dados 02 - *Employee Turnover*

Atributo	Descrição	Tipo de Dado
Acidentes de Trabalho	Indicação se o funcionário recebia a mais em relação às metas. Valor 1 para sim, 0 para não.	Booleano
Departamento	Subdivisão da área de atuação do funcionário em: suporte, vendas, RH, técnico, marketing, risco, contabilidade, gerenciamento de produto e gerencial.	String
Média de horas trabalhadas mensais	Quantidade média de horas trabalhadas mensalmente.	Inteiro
Número de projetos	Número de projetos atribuídos a determinado colaborador.	Inteiro
Promoção nos últimos 5 anos	Indicação se o funcionário mudou de cargo nos últimos 5 anos. Valor 1 para sim, 0 para não.	Booleano
Salário	Média de valores recebidos na organização, categorizado em uma faixa salarial baixa, média e alta.	String
Satisfação do funcionário	Indicador da satisfação do colaborador com o trabalho.	Float
Tempo de empresa	Período do funcionário na organização.	Inteiro
Última avaliação	Indicador atribuído pela organização acerca do desempenho funcional do colaborador.	Float

3.2 Pré-Processamento

Para a realização dos experimentos, foram utilizados os dois conjuntos de dados mencionados na subseção 4.2. O primeiro passo foi a limpeza dos dados. Todas as análises e processamentos das bases de dados foram em linguagem *Python*. Nas bases propostas, não foram encontrados valores duplicados, ausentes, informações inválidas ou formatos fora do padrão. No entanto, para as variáveis categóricas 'departamento' e 'salário', foi necessário

criar variáveis *dummy*, a fim de converter esses dados em formato numérico e possibilitar sua incorporação nos modelos de aprendizado de máquina. De acordo com Abreu (2022), as variáveis dummy são variáveis binárias (0 ou 1) criadas para representar uma variável com duas ou mais categorias. Nesse sentido, para uma variável categórica de n categorias, serão criadas n-1 variáveis *dummy*. As *dummy* são representadas por 1 se a característica estiver presente e por 0 se não estiver (Abreu, 2022).

Além das variáveis *dummy,* foi necessário realizar um balanceamento na base de dados, realizando a técnica de *random undersampling*. Segundo Guimarães (2022), a técnica de *random undersampling* consiste em diminuir instâncias da classe majoritária até atingir a quantidade total de instâncias da classe minoritária. Uma das vantagens de utilizar essa técnica é evitar o *overfitting* assim como reduzir o número total de amostras no conjunto de dados, diminuindo o tempo de processamento e os recursos computacionais necessários para treinar modelos de aprendizado de máquina. Dessa forna, percebeu-se que, para a primeira base, havia 6.756 registros para a variável 0, que correspondia às pessoas que não haviam pedido demissão e 2.784, para a variável 1, a qual correspondia às pessoas que pediram demissão. Após o balanceamento, ambas variáveis ficaram com 2.784, totalizando 5.568 registros. Para a segunda base, também foi realizada o balanceamento. No primeiro momento havia 10.000 registros para a variável 0 e 1.995 para a variável 1. Após o balanceamento, ambas ficaram com 1.995, totalizando 3.990 registros. Dessa forma, a estratégia de *random undersampling* foi realizada a fim de mitigar possíveis tendências de resultados e ruídos que interferissem nas métricas.

3.2 Modelagem

Para ambas as bases de dados foram desenvolvidas a análise exploratória, com o intuito de verificar anomalias e alinhar uma compreensão preliminar dos dados. Não foram encontradas valores duplicados, formatados de forma diversa do esperado na base ou *outliers*. Após isso, as dez técnicas foram aplicadas juntamente com um *script* que salvasse as rodadas

de treinamento. Foram realizadas 30 rodadas para cada técnica, totalizando 600 rodadas de treinamento, respectivamente das duas bases de dados. A ideia de realizar essa quantidade de rodadas foi em consonância ao teorema do limite central. Para Bittencourt e Viali (2006), o teorema do limite central garante que o comportamento probabilístico de vários estimadores possa ser descrito com boa aproximação pela distribuição normal. Nesse sentido, os autores mostram que, para reduzir a variabilidade, deve-se realizar a estimativa de rodadas no número maior ou igual a 30, tendo a população distribuição,tornando as médias amostrais também com distribuição normal. Dessa forma, Casella e Berger (2011) reafirmam que se deve respeitar o teorema do limite central, realizando 30 rodadas de cada algoritmo para que se reduza a variabilidade e se obtenha uma estimativa mais estável da média.

Para a questão do treinamento e teste, foi utilizado o método *hold-out* para avaliar o desempenho de cada modelo. Conforme Monard e Baranauskas (2003), *hold-out* é um método mais simples de testar o modelo consiste em separar, de forma aleatória, uma parcela dos dados para testar o modelo, e utilizar o restante para treinamento. Para o presente trabalho, foi utilizado 30% dos dados para teste e 70% para treinamento.

3.3 Critérios de Avaliação

Após a modelagem, foram extraídas as médias das métricas de Precisão, F1-Score, Recall e Acurácia ao longo de 30 rodadas. Essas métricas, calculadas a partir da matriz de confusão, foram as principais medidas de avaliação utilizadas neste trabalho e são descritas a seguir:

A) Precisão: é a taxa de amostras de classe positiva classificadas corretamente (Monico *et al.*, 2009). O cálculo é realizado a partir da divisão do número de amostras classificadas corretamente (VP) dividido pela quantidade de verdadeiros positivos novamente mais os falsos positivos (FP).

$$Precisão = \frac{VP}{VP + FP}$$

B) F1- Score: é a média harmônica entre precisão (P) e sensibilidade (C). É utilizada quando não há preferência pela precisão e sensibilidade, desejando encontrar o ponto máximo para as duas medidas (Nogueira; Delamaro, 2020).

$$F1\text{-}Score = (2x \frac{P\ x\ C}{P + C})$$

C) Recall: é a proporção de instâncias que foram corretamente previstas. Nesse sentido, pode ser encontrada através da métrica de verdadeiros positivos (VP) divididos pela soma dos VP e falsos negativos (FN) (Zhang; Huang;Wang, 2017).

$$Recall = \frac{VP}{VP + FN}$$

D) Acurácia: é a razão da soma de todas as amostras corretamente classificadas, pelo número total do conjunto. Dessa forma, essa métrica, conforme Monico *et al*. (2009), apresenta o grau de proximidade de uma estimativa com seu parâmetro (ou valor verdadeiro).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

4. Experimentos e Resultados

4.1 Resultados dos algoritmos

Após uma análise dos resultados, destaca-se o desempenho do algoritmo de *Random Forest*. Na primeira base de dados, o F1-Score foi de 0,87 para a classe 0 (trabalhadores que não pediram demissão) e de 0,78 para a classe 1 (trabalhadores que pediram demissão). Na segunda base, os F1-*Scores* foram de 0,95 para ambas as classes. A acurácia foi de 0,84 na primeira base e de 0,95 na segunda, com as maiores médias em precisão e *recall*. As métricas de todos os algoritmos estão nas tabelas 3, para a base intitulada *Employee churn data,* e a tabela 4 para a segunda base, com o título *Employee Turnover,* referente à parcela da base de dados usada de testes. Além disso, as tabelas 5 e 6 apresentam as métricas relacionadas às porções dos conjuntos de dados que foram divididas para o treinamento. A Tabela 5 refere-se à base de dados intitulada *Employee Churn Data*, enquanto a Tabela 6 corresponde à base *Employee Turnover*. Em relação aos hiperparâmetros, para todas as técnicas aplicadas neste estudo, foram utilizados os hiperparâmetros padrão da versão 1.5.1 do *Scikit-Learn* (2024). Essa escolha visou garantir a padronização do processo de modelagem e facilitar a comparação entre os resultados obtidos com diferentes algoritmos. Além disso, não se fez necessária a otimização dos resultados dentro do escopo deste trabalho, uma vez que as métricas obtidas já indicavam resultados satisfatórios.

Tabela 3 - Resultados dos algoritmos da base de dados 01 - *Employee churn data* - Teste

Classificador	Hiperparâmetros	Classe	Precisão	F1-Score	Recall	Acurácia
Random Forest	n_estimators=50, max_depth=10, max_features=50, random_state=2	0	0,87	0,9	0,92	0,84
		1	0,78	0,72	0,66	
Árvores de Decisão	criterion='entropy', max_depth=3, random_state=0	0	0,86	0,87	0,88	0,81
		1	0,86	0,66	0,64	
KNN	n_neighbors=7, metric='minkowski', p=1	0	0,81	0,85	0,89	0,78
		1	0,66	0,57	0,5	
Regressão Logística	random_state=1, max_iter=600, penalty="l2", tol=0.0001, C=1,solver="lbfgs"	0	0,67	0,65	0,64	0,66
		1	0,65	0,67	0,69	
SVM	kernel='rbf', random_state=1, C = 2	0	0,56	0,46	0,39	0,54
		1	0,53	0,6	0,69	
Naive Bayes	GaussianNB()	0	0,63	0,6	0,58	0,62
		1	0,61	0,63	0,66	
XGBOOST	max_depth=2, learning_rate=0.05, n_estimators=250, objective='binary:logistic', random_state=3	0	0,86	0,89	0,93	0,84
		1	0,79	0,7	0,62	
ADABOOST *CLASSIFIER*	n_estimators=600, random_state=0	0	0,81	0,85	0,91	0,78
		1	0,68	0,56	0,47	
LIGHTGBM	num_leaves':250, 'objective':'binary', 'max_depth':2, 'learning_rate':.05, 'max_bin':100	0	0,85	0,89	0,93	0,84
		1	0,79	0,69	0,61	
Perceptron Multicamada (MLP)	hidden_layer_sizes': (200,),'max_iter': 300, 'random_state': 42	0	0,78	0,83	0,72	0,7
		1	0,45	0,04	0,52	

Fonte: Dados da pesquisa, 2024

Tabela 4 - Resultados dos algoritmos da base de dados 02 - *Employee Turnover* - Teste

Classificador	Hiperparâmetros	Classe	Precisão	F1-Score	Recall	Acurácia
Random Forest	n_estimators=50, max_depth=10, max_features=50, random_state=2	0	0,93	0,95	0,98	0,95
		1	0,98	0,95	0,92	
Árvores de Decisão	criterion='entropy', max_depth=3, random_state=0	0	0,93	0,94	0,95	0,94
		1	0,95	0,94	0,92	
KNN	n_neighbors=7, metric='minkowski', p=1	0	0,92	0,87	0,83	0,88
		1	0,85	0,89	0,93	
Regressão Logística	random_state=1, max_iter=600, penalty="l2", tol=0.0001, C=1,solver="lbfgs"	0	0,82	0,77	0,73	0,79
		1	0,76	0,8	0,84	
SVM	kernel='rbf', random_state=1, C = 10	0	0,77	0,71	0,66	0,73
		1	0,7	0,75	0,8	
Naive Bayes	GaussianNB()	0	0,73	0,47	0,35	0,61
		1	0,57	0,69	0,87	
XGBOOST	max_depth=2, learning_rate=0.05, n_estimators=250, objective='binary:logistic', random_state=3	0	0,94	0,94	0,94	0,94
		1	0,94	0,94	0,94	
ADABOOST *CLASSIFIER*	n_estimators=600, random_state=0	0	0,94	0,93	0,92	0,93
		1	0,92	0,93	0,94	
LIGHTGBM	num_leaves':250, 'objective':'binary', 'max_depth':2, 'learning_rate':.05, 'max_bin':100	0	0,94	0,94	0,93	0,94
		1	0,94	0,94	0,94	
Perceptron Multicamada (MLP)	hidden_layer_sizes': (200,),'max_iter': 300, 'random_state': 42	0	0,92	0,7	0,82	0,76
		1	0,82	0,79	0,92	

Fonte: Dados da pesquisa, 2024

Tabela 5 - Resultados dos algoritmos da base de dados 01 - *Employee churn data* - Treinamento

Classificador	Hiperparâmetros	Classe	Precisão	F1-Score	Recall	Acurácia
Random Forest	n_estimators=50, max_depth=10, max_features=50, random_state=2	0	0,89	0,92	0,94	0,87
		1	0,80	0,75	0,70	
Árvores de Decisão	criterion='entropy', max_depth=3, random_state=0	0	0,88	0,89	0,90	0,85
		1	0,88	0,69	0,67	
KNN	n_neighbors=7, metric='minkowski', p=1	0	0,81	0,85	0,89	0,82
		1	0,66	0,57	0,5	
Regressão Logística	random_state=1, max_iter=600, penalty="l2", tol=0.0001, C=1,solver="lbfgs"	0	0,69	0,67	0,68	0,69
		1	0,68	0,69	0,71	
SVM	kernel='rbf', random_state=1, C = 2	0	0,59	0,49	0,41	0,57
		1	0,57	0,63	0,71	
Naive Bayes	GaussianNB()	0	0,67	0,63	0,61	0,66
		1	0,64	0,66	0,69	
XGBOOST	max_depth=2, learning_rate=0.05, n_estimators=250, objective='binary:logistic', random_state=3	0	0,88	0,91	0,95	0,87
		1	0,83	0,74	0,66	
ADABOOST *CLASSIFIER*	n_estimators=600, random_state=0	0	0,83	0,87	0,93	0,80
		1	0,70	0,58	0,50	
LIGHTGBM	num_leaves':250, 'objective':'binary', 'max_depth':2, 'learning_rate':.05, 'max_bin':100	0	0,88	0,91	0,94	0,86
		1	0,82	0,73	0,64	
Perceptron Multicamada (MLP)	hidden_layer_sizes': (200,),'max_iter': 300, 'random_state': 42	0	0,80	0,87	0,74	0,74
		1	0,47	0,07	0,53	

Fonte: Dados da pesquisa, 2024

Tabela 6 - Resultados dos algoritmos da base de dados 02 - *Employee Turnover* - Treinamento

Classificador	Hiperparâmetros	Classe	Precisão	F1-Score	Recall	Acurácia
Random Forest	n_estimators=50, max_depth=10, max_features=50, random_state=2	0	0,95	0,96	0,98	0,98
		1	0,98	0,96	0,92	
Árvores de Decisão	criterion='entropy', max_depth=3, random_state=0	0	0,93	0,94	0,95	0,96
		1	0,95	0,94	0,92	
KNN	n_neighbors=7, metric='minkowski', p=1	0	0,92	0,87	0,83	0,90
		1	0,85	0,89	0,93	
Regressão Logística	random_state=1, max_iter=600, penalty="l2", tol=0.0001, C=1,solver="lbfgs"	0	0,83	0,79	0,74	0,83
		1	0,77	0,81	0,86	
SVM	kernel='rbf', random_state=1, C = 10	0	0,79	0,73	0,68	0,77
		1	0,72	0,76	0,81	
Naive Bayes	GaussianNB()	0	0,73	0,47	0,35	0,65
		1	0,57	0,69	0,87	
XGBOOST	max_depth=2, learning_rate=0.05, n_estimators=250, objective='binary:logistic', random_state=3	0	0,94	0,94	0,94	0,96
		1	0,94	0,94	0,94	
ADABOOST *CLASSIFIER*	n_estimators=600, random_state=0	0	0,94	0,93	0,92	0,95
		1	0,92	0,93	0,94	
LIGHTGBM	num_leaves':250, 'objective':'binary', 'max_depth':2, 'learning_rate':.05, 'max_bin':100	0	0,94	0,94	0,93	0,96
		1	0,94	0,94	0,94	
Perceptron Multicamada (MLP)	hidden_layer_sizes': (200,),'max_iter': 300, 'random_state': 42	0	0,92	0,7	0,82	0,80
		1	0,82	0,79	0,92	

Fonte: Dados da pesquisa, 2024

Essas métricas são cruciais, pois refletem a capacidade do modelo em equilibrar precisão e recall, dois aspectos fundamentais da classificação. Com base nessas avaliações, o algoritmo de *Random Forest* se mostrou como a escolha mais assertiva para ambas as bases de dados.

Agora, com essa constatação, avança-se para a última etapa da análise, onde aplicaremos a metodologia *SHAP* (Explicabilidade em Inteligência Artificial) para interpretar os atributos mais impactantes e traçar um perfil preciso do *turnover*.

Em relação aos hiperparâmetros da técnica de *Random Forest*, a técnica mais bem posicionada, para ambas as bases de dados estudadas e analisadas, observa-se o *n_estimators* (número de estimadores) igual a 50, sendo o número de árvores o qual acarreta um modelo mais robusto. Além disso, uma *max_depth* (profundidade máxima) igual a 10, limitando a complexidade das árvores a fim de evitar *overfitting*. Sobre o *max_features* (O número de características a considerar quando se procura o melhor *split*) igual a 50, destaca-se que apenas 50 *features* (características) serão consideradas para cada divisão em uma árvore. Dessa forma, fixar esse valor garante que o modelo produza resultados reproduzíveis sempre que for treinado com os mesmos dados (Probst; Wright; Boulesteix, 2019).

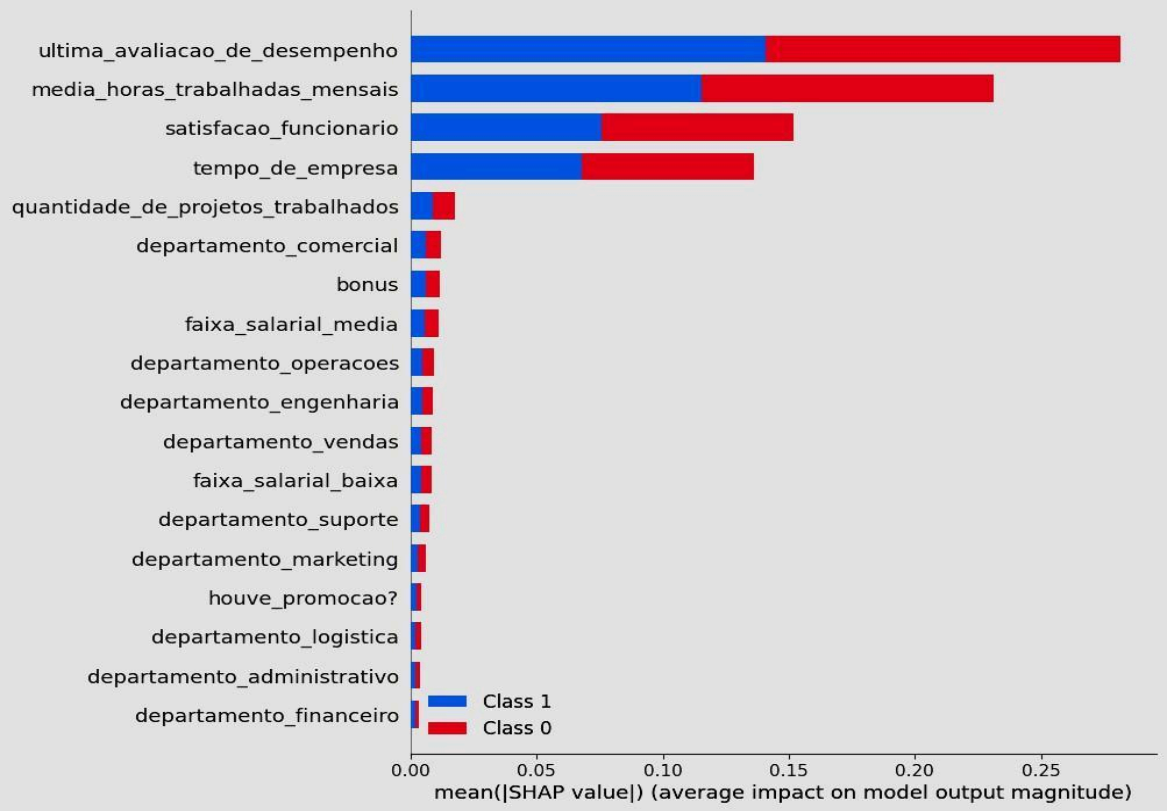
4.2 Resultados da Metodologia *SHAP*

Conforme observado, após os processamentos dos algoritmos de classificação, constatou-se que a técnica de *Randon Forest* apresentou métricas significativas para a última etapa do experimento: a interpretação utilizando a técnica *SHAP*, técnica que tem como objetivo criar um modelo mais simplificado e capaz de apresentar o comportamento geral do modelo complexo (Neiva, 2023). Para o presente estudo, foi utilizada a parametrização padrão do *SHAP*, na qual se calcula as importâncias das características em relação a um modelo específico que já foi treinado. O modelo pode ser qualquer algoritmo de aprendizado de máquina, como árvores de decisão, redes neurais, SVM, etc. Quanto ao tipo de explicador, foi utilizado o *TreeExplainer*, pois ele é otimizado para modelos baseados em árvores, como o *Random Forest.*

Na análise das figuras 1 e 2, observa-se que o atributo "última avaliação de desempenho" (Figura 1) e a "satisfação do funcionário" (Figura 2), respectivamente nas bases de dados 1 e 2, para *class* 0, para os colaboradores que não pediram demissão, e *class* 1, para os que pediram demissão, demonstraram ter uma alta importância no classificador. Esses atributos contribuíram principalmente para o aumento da probabilidade de pedido de demissão. Além desses destaques, outros parâmetros também se destacaram durante o processo de tomada de decisão na avaliação do perfil de *turnover*. Na base de dados 1, por

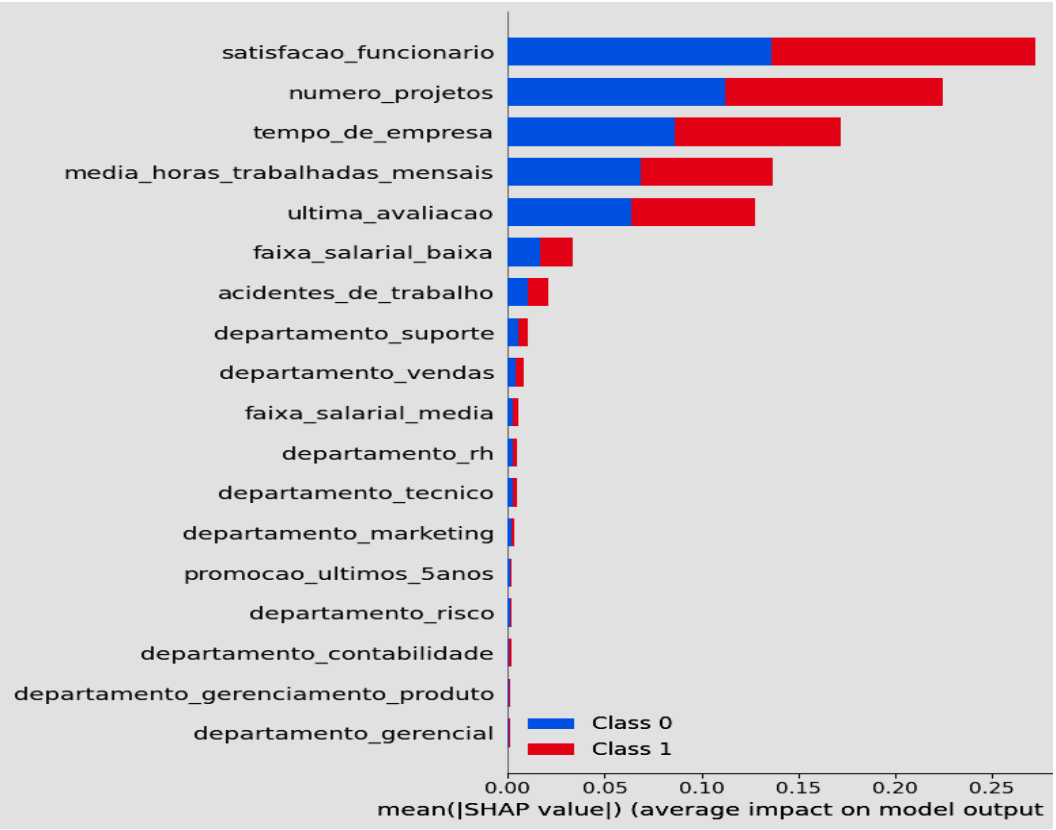
exemplo, destaca-se a média de horas trabalhadas, enquanto na base de dados 2, o número de projetos aos quais o funcionário está atrelado também mostrou relevância.

Figura 1 - Gráfico De Importância *Shap* Dos Atributos Do *Random Forest* Na Classificação De Demissões Da Base De Dados 01 - *Employee Churn Data*



Fonte: dados da pesquisa, 2024

Figura 2 - Gráfico De Importância *Shap* Dos Atributos Do *Random Forest* Na Classificação De Demissões Da Base De Dados 02 - *Employee Turnover*



Fonte: dados da pesquisa, 2024

É interessante notar que a variável departamento não demonstrou ter um impacto significativo na possibilidade de pedido de demissão. Pelo contrário, fatores mais relacionados à gestão de desempenho, satisfação do funcionário e média de horas trabalhadas mensais ganharam maior destaque na análise. Esses *insights* ressaltam a importância de considerar uma variedade de variáveis ao avaliar o *turnover*, fornecendo uma visão abrangente e precisa do cenário organizacional.

Um outro atributo que se destacou em ambas interpretações foi o tempo de empresa. Esse dado ressalta a importância de compreender não apenas o desempenho atual e a satisfação dos funcionários, mas também a trajetória e a estabilidade dentro da empresa, como indicativos cruciais para prever o *turnover*.

Na tabela 7, destaca-se a ordem dos atributos que mais impactaram as bases de dados estudadas.

Tabela 7 - Ordem dos Atributos com Maior Impacto nas Bases de Dados Estudadas

Ordem	Atributos - base 01	Atributos - base 02
1	ultima_avaliacao_de_desempenho	satisfacao_funcionario
2	media_horas_trabalhadas_mensais	numero_projetos
3	satisfacao_funcionario	tempo_de_empresa
4	tempo_de_empresa	media_horas_trabalhadas_mensais
5	quantidade_de_projetos_trabalhados	ultima_avaliacao

Fonte: dados da pesquisa, 2024

4.3 Discussão

Os resultados das métricas deste estudo apresentaram valores significativos frente aos estudos relacionados. Especificamente, a acurácia da segunda base ficou em torno de 0,95, estando em faixas de confiabilidade como os estudos de Boen (2022) e o de Jaisawall (2022), os quais obtiveram valores de 0,85 e 0,93, respectivamente, utilizando o algoritmo de *random forest*. A primeira base obteve 0,84 de acurácia, estando também em consonância a esses estudos. Além da acurácia, os resultados das métricas de precisão, *recall* e *F1-score* evidenciaram viabilidade no uso do *random forest*, conforme relatado no trabalho de Boen (2022). O trabalho também foi aderente ao de Alshehhi (2021), o qual destacou o algoritmo de *random forest* como o mais bem avaliado para ser utilizado na temática de intenção de *turnover*.

Sobre o uso da metodologia *SHAP*, o presente trabalho se alinha às expectativas trazidas pelo trabalho de Boen (2022), conseguindo observar os atributos que mais impactaram na intenção de turnover, referente às duas bases de dados. Um dos atributos que impactou em ambas as bases foi o de satisfação no trabalho, e Boen (2022) afirma que as empresas conseguiram reduzir a taxa de *turnover* após implementar um programa que aumentava a satisfação no trabalho. Dessa forma, evidencia-se a sinergia entre os trabalhos relacionados e o presente estudo sobre o uso da aprendizagem de máquina em relação à previsão de *turnover*.

5. Conclusão

Este estudo apresentou a comparação de dez técnicas de aprendizagem de máquina com intuito de investigar quais os atributos apresentariam mais impacto na intenção de *turnover* por parte dos funcionários. A partir dessa ideia, foram utilizadas duas bases de dados, com o intuito

de verificar a viabilidade do uso do algoritmo que fosse selecionado. Após a sistematização e balanceamento da base de dados, foram aplicadas as dez técnicas, sendo realizadas trinta rodadas para cada técnica. Depois dessa etapa, foram catalogados as métricas de precisão, *f1-score*, *recall* e acurácia, que foram os parâmetros estudados para se chegar no algoritmo de *Random Forest*, com melhor colocação para ambas as bases de dados.

Depois da análise das métricas, aplicou-se aos resultados do algoritmo de *Random Forest* a metodologia *SHAP* com a intenção de verificar o impacto das principais variáveis no processo de *turnover*. Os resultados da metodologia *SHAP* apontaram, para ambas as bases, que os atributos de tempo de empresa, última avaliação de desempenho, médias de horas trabalhadas mensais e satisfação do funcionário foram os destaques e apresentaram maior impacto na decisão do funcionário em pedir ou não demissão. Assim, com essas informações, os gestores podem programar estratégias organizacionais que mitiguem esses riscos de rotatividade de força de trabalho.

Este trabalhou enfrentou algumas limitações devido às bases de dados não apresentarem alguns atributos como idade, sexo, grau de instrução, distância da residência até o trabalho e senioridade dos funcionários, os quais poderiam fornecer mais detalhes ao perfil de intenção de *turnover.* Além disso, ter utilizado apenas uma metodologia de interpretação, que foi a *SHAP*, já que sua aplicação pode variar dependendo do contexto do problema.

Em relação aos direcionamentos futuros, pode-se atribuir mais atributos, conforme observado em uma das limitações do trabalho, tendo a possibilidade de nichar determinado segmento empresarial assim como uma delimitação geográfica. Além disso, explorar outros algoritmos de aprendizagem de máquina que não foram utilizados. Um outro desdobramento futuro é utilizar diferentes metodologias de interpretação. O presente trabalho utilizou a metodologia *SHAP*, contudo, em trabalhos futuros, outras metodologias como o *LIME* (*Local Interpretable Model-agnostic Explanations*) (ZAFAR; KHAN, 2021) ou ICE (*Individual Conditional Expectation*) (GOLDSTEIN, 2015) são possibilidades para novos *insights* sobre a temática.

REFERÊNCIAS

ABREU, Leonardo Evangelista de. People Analytics: uso de árvores de decisão na retenção de talentos. Orientador: Prof. Dr. Klaus Schlunzen Junior. 2022. 38 p. Trabalho de Conclusão de Curso (Graduação em Estatística) - FCT/Unesp, Presidente Prudente, 2022. Disponível em: https://repositorio.unesp.br/server/api/core/bitstreams/41a09be5-122b-4a10-976e-f9b6c1db97c2/content. Acesso em: 17 set. 2024.

AJIT, Pankaj. Prediction of employee turnover in organizations using machine learning algorithms. **algorithms**, v. 4, n. 5, p. C5, 2016.

ALSHEHHI, Khaled et al. Employee retention prediction in corporate organizations using machine learning methods. **Academy of Entrepreneurship Journal**, v. 27, p. 1-23, 2021.

BALDO, Fabiano et al. Adaptive fast xgboost for binary classification. In: SIMPÓSIO BRASILEIRO DE BANCOS DE DADOS, 37, 2022. Rio de Janeiro, Búzios, RJ. **Anais** […]. Rio de Janeiro: SBBD, 2022. p. 13-25. Disponível em: https://dblp.org/db/conf/sbbd/sbbd2022.html. Acesso em: 17 set. 2024.

BERGER, L. A., & Berger, D. **The Talent Management Handbook**: Making Culture a Competitive Advantage by Acquiring, Identifying, Developing, and Promoting the Best People.. 3.ed. Nova York: McGraw-Hill, 2017.

BITTENCOURT, Hélio Radke; VIALI, Lori. Contribuições para o ensino da distribuição normal ou curva de Gauss em cursos de graduação. *In*: **III Seminário Internacional de Pesquisa em Educação Matemática**, 2006, São Paulo, Águas de Lindoia, SP. Seminários [...], São Paulo: SBEM, 2006. Disponível em: https://www.researchgate.net/profile/Lori-Viali/publication/280444871_Contribuicoes_para_o_ensino_da_distribuicao_normal_ou_curva_de_Gauss_em_cursos_de_Graduacao/links/55b4fca508ae092e9655814d/Contribuicoes-para-o-ensino-da-distribuicao-normal-ou-curva-de-Gauss-em-cursos-de-Graduacao.pdf. Acesso em: 20 set 2024.

BOEN, Vinicius de Oliveira. **People Analytics: Aprendizado de máquina na gestão estratégica de pessoas, aplicando modelo preditivo de turnover.** Orientador: Prof. Dr. Luis Gustavo Nonato. 2022. 57p. Dissertação do Mestrado Profissional e Matemática, Estatística e Computação Aplicadas à Indústria. Universidade de São Paulo, São Carlos, 2022. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55137/tde-01122022-093014/publico/ViniciusdeOliveiraBoen_ME_revisada.pdf. Acesso em: 20 set. 2024.

CASELLA, G.; BERGER, R. L. Inferência estatística. 2.ed. São Paulo: Cengage Learning, 2011.

CHIAVENATO, Idalberto. **Gestão de Pessoas**: O Novo Papel dos Recursos Humanos Nas Organizações. 4.ed. Barueri-SP: Editora Manole Ltda, 2014. p.1-512

CORRÊA, Guido Machado. Motivação e desmotivação: elementos que podem influenciar a rotatividade em um hotel de luxo no Rio de Janeiro. 2016. 17.p. Niterói- RJ, Universidade Federal Fluminense, 2016. Disponível em: https://periodicoscientificos.ufmt.br/ojs/index.php/repad/article/view/17145/13595. Acesso em: 20 set 2024.

COSTA, Gabriel Resende; DA SILVA, Júlio Fernando. A influência das práticas de gestão de pessoas sobre a rotatividade de pessoal. **Cadernos de Gestão e Empreendedorismo**, v. 8, n. 1, p. 49-64, 2020.

CRISÓSTOMO, Israel. A motivação como ferramenta de crescimento. **Acedido a**, v. 17, 2010.

DEISENROTH, Marc Peter; FAISAL, A. Aldo; ONG, Cheng Soon. **Mathematics for machine learning**. Cambridge University Press, 2020.

DIETTERICH, Tom. Overfitting and undercomputing in machine learning. **ACM computing surveys (CSUR)**, v. 27, n. 3, p. 326-327, 1995.

DOMINGOS, Pedro. A few useful things to know about machine learning. **Communications of the ACM**, v. 55, n. 10, p. 78-87, 2012.

EMPLOYEE CHURN DATA. 2024. Disponível em: <https://www.kaggle.com/datasets/marikastewart/employee-turnover>. Acesso em 15 mar. 2024.

EMPLOYEE TURNOVER. 2024. Disponível em: <https://www.kaggle.com/datasets/akshayhedau/employee-turnover-analytics-dataset>. Acesso em 15 mar. 2024.

GAO, Xiang; WEN, Junhao; ZHANG, Cheng. An improved random forest algorithm for predicting employee turnover. **Mathematical Problems in Engineering**, v. 2019, n. 1, p. 4140707, 2019.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 6. ed. São Paulo: Atlas, 2017.

GOLDSTEIN, Alex et al. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. **journal of Computational and Graphical Statistics**, v. 24, n. 1, p. 44-65, 2015.

GONÇALVES, Paulo Henrik Ribeiro. **Um esquema rápido baseado em aprendizado de máquina para a predição interquadros do codificador de vídeo VVC**. Orientador: Prof. Dr. Marcelo Schiavon Porto. 2021. 90 p. Dissertação de Mestrado. Universidade Federal de Pelotas, 2021. Disponível em: https://guaiaca.ufpel.edu.br/bitstream/handle/prefix/7787/Dissertacao_Paulo_Henrrik_Ribeiro_Goncalves.pdf?sequence=1. Acesso em: 20 set. 2024.

GUIMARÃES, Eltoni Alves. **Um estudo para identificar e classificar ambiguidades em histórias de usuário usando aprendizagem de máquina.** Orientadora: Profª Drª. Márcia Jacyntha Nunes Rodrigues Lucena 2022. 118 p. Dissertação de Mestrado. Universidade Federal do Rio Grande do Norte, 2022.Disponível em: https://repositorio.ufrn.br/bitstream/123456789/50812/1/Estudoidentificarclassificar_Guimaraes_2022.pdf. Acesso em: 20 set. 2024.

HARAN, Vidya V.; NIEDERMAN, Fred. Social context of turnover-mixed methods study of Indian IT professionals. **Journal of Global Information Management (JGIM)**, v. 30, n. 1, p. 1-24, 2022

HAUSKNECHT, John P.; TREVOR, Charlie O.; HOWARD, Michael J. Unit-level voluntary turnover rates and customer service quality: implications of group cohesiveness, newcomer concentration, and size. **Journal of Applied Psychology**, v. 94, n. 4, p. 1068, 2009.

JABEUR, S. B.; MEFTEH-WALI, S.; VIVIANI, J.-L. Forecasting gold price with the xgboost algorithm and shap interaction values. **Annals of Operations Research**, Springer, v. 334, n. 1, p. 679–699, 2024.

JAISAWAL, Rachana. Machine Learning to Evaluate Important Human Capital (HC) Determinants Impacting IT Compensation. **Ramanujan International Journal of Business and Research**, v. 7, n. 2, p. 16-25, 2022.

JAKHAR, D., KAUR, I. 2019. Artificial intelligence, machine learning and deep learning: definitions and differences. **Clinical and Experimental Dermatology**, vol. 45, no. 1, pp. 131-132.

JORDAN, Michael I.; MITCHELL, Tom M. Machine learning: Trends, perspectives, and prospects. **Science**, v. 349, n. 6245, p. 255-260, 2015.

LE, Linh; XIE, Ying; RAGHAVAN, Vijay V. KNN loss and deep KNN. **Fundamenta Informaticae**, v. 182, n. 2, p. 95-110, 2021.

LIMA, Tiago Pessoa Ferreira et al. Previsão de óbito e importância de características clínicas em idosos com COVID-19 utilizando o Algoritmo Random Forest. **Revista Brasileira de Saúde Materno Infantil**, v. 21, p. 445-451, 2021.

LUNDERBG, S.M.; ERION, G. G.; LEE, S.-I. **A unified approach to interpreting model predictions. Advances in neural information processing systems**, v.30,2017.

MARIM, Mateus Coutinho et al. Caracterização e classificação do tráfego da Darknet com modelos baseados em árvores de decisão. *In*: **Anais do XXXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos,** 2021, Minas Gerais, Juiz de Fora, MG, Anais [...]. Minas Gerais: SBC, 2021. p. 127-140. Disponível em: https://www.researchgate.net/profile/Mateus-Coutinho-Marim/publication/354126142_Caracterizacao_e_Classificacao_do_Trafego_da_Darknet_com_Modelos_Baseados_em_Arvores_de_Decisao/links/61263feb2979ad5d6017a372/Caracterizacao-e-Classificacao-do-Trafego-da-Darknet-com-Modelos-Baseados-em-Arvores-de-Decisao.pdf. Acesso em: 20 set. 2024.

MARQUES, Mateus Maia; ARA, Anderson. SHINY ADABOOSTING: AN INTERACTIVE DASHBOARD TO ADAPTIVE BOOSTING ALGORITHM. **Revista do Seminário Internacional de Estatística com R**, v. 4, n. 1, p. 8-8, 2019.

MEDEIROS, Rochelle Reis; ALVES, Rafaela Cunha; RIBEIRO, Sidney Roberto. Turnover: Uma análise dos fatores que contribuem para a decisão de sair da empresa dos colaboradores da Alfa Comércio LTDA. **CONNEXIO-ISSN 2236-8760**, v. 2, n. 1, p. 115-126, 2012.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003.

MONICO, João Francisco Galera et al. Acurácia e precisão: revendo os conceitos de forma acurada. **Boletim de Ciências Geodésicas**, v. 15, n. 3, p. 469-483, 2009.

MOREIRA, Fabiano Greter; DA SILVA NANTES, Luana. Fatores que influenciam na rotatividade de pessoal nas organizações: um estudo bibliográfico. **Revista Estudos e Pesquisas em Administração**, v. 8, n. 1, 2024.

NEIVA, Davi Keglevich. **Interpretação de modelos complexos de aprendizado de máquina.** Orientador: Prof. Dr. Paulino Ribeiro Villas Boas. 2023. 76 p. Tese de Doutorado. Universidade de São Paulo, São Carlos, 2023. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55137/tde-15012024-160021/publico/DaviKeglevichNeiva_ME_revisada.pdf.Acesso em: 20 set. 2024.

NOGUEIRA, Lucas Lagôa; DELAMARO, Márcio Eduardo. Uma abordagem para redução do custo do Teste de Mutação utilizando Redes Neurais. *In*: **Anais Estendidos do XI Congresso Brasileiro de Software: Teoria e Prática.** Anais [...]. São Paulo: SBC, 2020. p. 22-28. Disponível em: https://repositorio.usp.br/directbitstream/53947e28-b676-4b86-acdb-551028cfed32/3033752.pdf . Acesso em: 20 set. 2024.

OLIVEIRA, Petronio Diego Silva de. Uso de aprendizagem de máquina e redes neurais convolucionais profundas para a classificação de áreas queimadas em imagens de alta resolução espacial. Orientador: Prof. Dr. Osmar Abílio de Carvalho Junior 2020. 34 p. Dissertação de Mestrado. Universidade De Brasília, Brasília, 2020. Disponível em: http://www.realp.unb.br/jspui/bitstream/10482/38234/1/2019_PetronioDiegoSilvadeOliveira.pdf . Acesso em: 20 set. 2024.

PINHEIRO, Ana Paula; SOUZA, Dercia Antunes. Causas e efeitos da rotatividade de pessoal/turnover: Estudo de caso de uma microempresa do setor de educação. *In*: **X Simpósio de Excelência em Gestão e Tecnologia– SEGeT**, v. 20, 2013. São Paulo. Anais [...]. São Paulo: SEGET, 2013. 13 p. Disponível em: https://www.aedb.br/seget/arquivos/artigos13/58618723.pdf. Acesso em: 20 set. 2024.

PROBST, Philipp; WRIGHT, Marvin N.; BOULESTEIX, Anne‐Laure. Hyperparameters and tuning strategies for random forest. **Wiley Interdisciplinary Reviews: data mining and knowledge discovery,** v. 9, n. 3, p. e1301, 2019.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do trabalho científico:** métodos e técnicas da pesquisa e do trabalho acadêmico. 2. ed. Novo Hamburgo: Feevale, 2013. Disponível em: https://www.feevale.br/Comum/midias/0163c988-1f5d-496f-b118-a6e009a7a2f9/E-book%20Metodologia%20do%20Trabalho%20Cientifico.pdf. Acesso em: 19 set 2024.

SANTOS, Luiz Felipe Alves dos; COUTO, Hudson Jean Bianquini. Estudo de comparação de softwares para processamento de imagens em estudos de distribuição de tamanho de bolha. In: JORNADA DO PROGRAMA DE CAPACITAÇÃO INSTITUCIONAL, 12, 2023. Cidade Universitária, Rio de Janeiro. Anais […]. Rio de Janeiro: PCI/CETEM, 2023. Disponível em: http://mineralis.cetem.gov.br/bitstream/cetem/2806/1/Luiz%20Felipe%20Alves%20dos%20Santos.pdf. Acesso em: 19 set 2024.

SANTOS, Michel Mozinho dos. Classificação de padrões de imagens: função objetivo para perceptron multicamada e máquina de aprendizado extremo convolucional. Orientador: Prof. Dr. Abel Guilhermino da Silva Filho. 2019. 133 p. Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, Recife. 2019. Disponível em: https://repositorio.ufpe.br/bitstream/123456789/36908/1/TESE%20Michel%20Mozinho%20dos%20Santos.pdf. Acesso em: 19 set 2024.

SANTOS, Vanessa Cristina Bissoli dos. Aprendizagem organizacional como instrumento de gestão de pessoas sob a ótica da competência em informação. Orientadora: Profa. Dra. Regina Célia Baptista Belluzzo. 2020. 344 p. Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Faculdade de Filosofia e Ciências da Universidade Estadual Paulista (UNESP) Campus de Marília, 2020. Disponível em; https://repositorio.unesp.br/server/api/core/bitstreams/26aaf465-2f9e-4791-9354-05ba712975f5/content. Acesso em: 19 set 2024.

SCIKIT LEARN.. 2024. Disponível em: <https://scikit-learn.org/stable/>. Acesso em 15 mar. 2024.

SILVA, Priscila Bez da. Turnover: um estudo de caso sobre as principais causas em uma empresa do ramo metal-mecânico. Orientador: Prof. Fabrício Machado Miguel. 2012. 98 p. Trabalho de Conclusão de Curso (Graduação em Ciências Contábeis) - Universidade Do Extremo Sul Catarinense - Unesc, Criciúma, 2012. Disponível em; http://repositorio.unesc.net/bitstream/1/1322/1/Priscila%20Bez%20da%20Silva%20.pdf. Acesso em: 19 set 2024.

SILVA, Risomario; SILVA NETO, Darcy Ramos da. Inteligência artificial e previsão de óbito por Covid-19 no Brasil: uma análise comparativa entre os algoritmos Logistic Regression, Decision Tree e Random Forest. **Saúde em Debate**, v. 46, p. 118-129, 2023

SOUZA, José Willian Santos et al. Aplicação Do Classificador Naive Bayes Para Detecção De Fraudes. **Ciência Da Computação: Avanços E Tendências Em Pesquisa**, v. 1, n. 1, p. 9-26, 2023.

TERAMOTO, Érico Tadao et al. Comparing different methods for estimating hourly solar ultraviolet radiation: Empirical Models, Artificial Neural Network and Support Vector Machine. 2020. **Revista Brasileira de Meteorologia**, v. 35, n. 1, p. 35-43, 2020.

WILL, D. E. M. **Metodologia da pesquisa científica**. 2.ed. Palhoça, SC: Unisulvirtual, 2011. Livro digital.

ZAFAR, Muhammad Rehman; KHAN, Naimul. Deterministic local interpretable model-agnostic explanations for stable explainability. **Machine Learning and Knowledge Extraction,** v. 3, n. 3, p. 525-541, 2021.

ZHANG, X.; HUANG, X.; WANG, F. The construction of undergraduate data mining course in the big data age. In: IEEE. **2017 12th International Conference on Computer Science and Education (ICCSE).** [S.l.], 2017. p. 651–654