UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

VICTOR THADEU BRUM SANSONE

CONSTRUÇÃO DE MODELO PREDITIVO DE DESLIGAMENTO DE COLABORADORES

PORTO ALEGRE

2020

AGRADECIMENTOS

Pelos livros infantis, livros acadêmicos, enciclopédias e coletâneas sempre acessíveis em nossa casa, agradeço aos meus pais, Ana e Pedro. Já na infância, vocês me permitiram conhecer o prazer da leitura, e a liberdade que apenas o conhecimento proporciona — mesmo que isso implicasse algumas páginas rabiscadas, recortadas, e até o sumiço de exemplares queridos. E, por não ter permitido que eu destruísse nosso acervo sozinho, por ter sido minha melhor amiga, além de ser um exemplo de determinação pessoal e profissional, agradeço à minha irmã, Ana Angélica. Muito obrigado. Este trabalho se iniciou com todos vocês.

Aos meus grandes amigos André Holthausen, Guilherme Peroni, Lauro Guilloux e Pedro Lima, agradeço por terem me proporcionado tantos momentos de alegria desde os nossos dias de colégio. Aos incríveis amigos Eduardo Martins, Gista Scarparo, Joaquim Barbosa, Leandro Bandeira, Maurício Borges, Maurício Folli e Plínio Cunha, agradeço pelos momentos felizes que passamos lado a lado no trabalho, com o Grêmio, e em nossas noites de *poker*. Ao amigo Guilherme Ziebell de Oliveira, agradeço por ter me acompanhado em tantos momentos marcantes, como quando meu pai faleceu, ou quando fizemos incontáveis festas, quando entramos na mesma faculdade e também saímos dela, ou ainda quando perdemos e ganhamos a Libertadores. Amigos, sem vocês ao meu lado, todos estes anos não teriam o mesmo brilho.

Por ter completado a minha vida e me fazer um homem feliz, agradeço à minha esposa Isis Medeiros dos Santos. Você acreditou em mim em todos os momentos. Quando o diploma parecia distante, você nos aproximou. Obrigado por me fazer sentir especial. Obrigado por todas as palavras de carinho e incentivo em todos estes anos. Obrigado também por dizer que sou inteligente, às vezes você até me convence. Amor, eu te amo, e este trabalho é nosso!

Aos professores Marcelo Cortimiglia e Rodrigo Dalla Vecchia, agradeço pelo tempo despendido, por me apoiarem e conduzirem durante a produção deste artigo. Ao colega Rodrigo Wegener, agradeço pela autonomia e confiança que sempre me depositou, possibilitando que este estudo fosse realizado. Muito obrigado, amigos.

Muito obrigado a todos vocês.

"Os adultos amam números. Quando você lhes fala de um novo amigo eles nunca lhe perguntam sobre o essencial. Eles nunca lhe dizem: "Como é o som da sua voz? Quais são os jogos que ele prefere? Ele coleciona borboletas?". Eles lhe perguntam: "Quantos anos ele tem? Quantos irmãos ele tem? Quanto ele pesa? Quanto o pai dele ganha? Somente então eles acham que o conhecem."

"... só vemos bem com o coração. O essencial é invisível aos olhos."

Antoine de Saint-Exupéry

CONSTRUÇÃO DE MODELO PREDITIVO DE DESLIGAMENTO DE COLABORADORES

Victor Thadeu Brum Sansone[*]
Marcelo Nogueira Cortimiglia[**]
Rodrigo Dalla Vecchia[***]

Resumo: Atualmente se observa a crescente necessidade das empresas em gerenciar a sua força de trabalho, visando à manutenção de profissionais qualificados e redução dos custos associados a processos demissionais. Somado a isso, constatam-se avanços no campo de investigação de Machine Learning, que possibilita a descrição de cenários futuros a partir de modelos preditivos orientados por dados. Essa combinação de fatores tem possibilitado às empresas o investimento em meios para prever quando seus funcionários estão mais propensos a deixar as organizações, antecipando-se à perda de talentos e reduzindo custos operacionais. Dessa forma, este estudo se propôs a construir um modelo preditivo de desligamento de colaboradores para uma instituição financeira no Brasil, além de compreender os principais fatores vinculados à rotatividade. O estudo foi conduzido testando-se o desempenho dos algoritmos *K-Nearest Neighbour*, Regressão Múltipla, *Naive Bayes* e *Random Forest* em uma base de dados contendo informações dos trabalhadores, coletada ao longo de um ano. Evidenciou-se que o melhor modelo preditivo foi construído a partir da técnica *Random Forest*, que apresentou acurácia de 78,3% e precisão de 81,5%. Observou-se também que as características pessoais, como idade e número de filhos, e profissionais, como remuneração e avaliação anual de desempenho, foram as variáveis mais relevantes para a classificação de um profissional como propenso ou não a deixar a empresa.

Palavras-chave: People Analytics. Recursos Humanos. Machine Learning. Modelo Preditivo. Demissão de Trabalhadores.

1 INTRODUÇÃO

A área de Gestão de Pessoas se desenvolveu como resposta às transformações relacionadas ao trabalho, a partir do desenvolvimento da industrialização e das novas configurações organizacionais ocorridas em meados do século XX (PEDRO, 2005). Frente às novas dinâmicas do mercado, que ao longo das últimas décadas têm exigido das empresas a capacidade de adaptar-se aos mais diversos cenários, a gestão do capital humano passa a ser um dos fatores de vantagem competitiva cruciais para o desempenho de uma organização (GARRIDO; SILVEIRA; SILVEIRA, 2018). A otimização da força de trabalho nas empresas se faz uma forte tendência, e a utilização de técnicas para análise de dados se faz relevante no contexto de uma sociedade digitalizada (TURSUNBAYEVA; DI LAURO; PAGLARI, 2018).

[*] victortbsv@gmail.com
[**] cortimiglia@producao.ufrgs.br
[***] rodrigovecchia@gmail.com

Como decorrência de uma necessidade das empresas em analisar dados de produtividade e obter predições associadas ao trabalho, surge o conceito de *People Analytics* (BODIE *et al.*, 2017). O conceito abrange diferentes aspectos, como estatística, programação computacional, *data mining*, criação de modelos e análise de dados, em uma abordagem voltada ao gerenciamento dos recursos humanos das companhias (BODIE *et al.*, 2017). Termos correlatos, como *HR Predictive Analytics* ou *Human Resource Data* ajudam a evidenciar a diversidade de domínios desse tema (TURSUNBAYEVA; DI LAURO; PAGLARI, 2018).

Uma das aplicações práticas de *People Analytics* está relacionada à identificação de fatores que influenciam na rotatividade no quadro de colaboradores das organizações (GRILLO; HACKETT, 2015). A rotatividade de pessoal está diretamente associada a custos como os de recrutamento e seleção, despesas com capacitações e indenizações, além de efeitos intangíveis vinculados à perda de mão de obra qualificada e alteração do clima organizacional (BORGES; RAMOS, 2011). A utilização de modelos, a fim de predizer o momento em que se dará o desligamento dos trabalhadores, permite que as empresas invistam em estratégias para melhoria das condições de trabalho e retenção de talentos (GRILLO; HACKETT, 2015).

No âmbito da previsão da rotatividade de funcionários, ou *turnover*, de seus empregados, muitas empresas têm desenvolvido modelos preditivos a partir de técnicas de *Machine Learning*. Os diferentes algoritmos que podem ser utilizados no aprendizado de máquina buscam encontrar padrões nas bases de dados contendo as entradas e saídas para o modelo a ser construído (SHEN, 2019). Os avanços em *Machine Learning* são potencializados pelo aumento na quantidade de dados gerados e armazenados diariamente, além dos aprimoramentos constantes nas áreas da computação. *Machine Learning*, portanto, pode ser compreendida como a intersecção entre a estatística, que fornece os fundamentos matemáticos necessários, e a ciência da computação, que disponibiliza as aplicações exigidas para a pesquisa (ALPAYDIN, 2011).

O estudo do tema supracitado justifica-se pela contínua necessidade das empresas em reduzir seus custos, manter seus talentos, obter ganhos em produtividade e melhorar o clima organizacional, de forma que o desligamento de colaboradores é um dos fatores determinantes para esses aspectos. Além disso, as empresas que não forem capazes de absorver e aplicar conhecimentos relacionados à análise de dados e *Machine Learning* a suas disciplinas internas, como Gestão de Pessoas, estarão abrindo mão de uma vantagem competitiva em relação ao mercado. Assim, o estudo relacionado ao problema de pesquisa de “Construção de Modelo

Preditivo de Desligamento de Colaboradores" é justificado pelas tendências tecnológicas atuais, bem como pela necessidade de sobrevivência das empresas.

Dessa forma, este artigo tem como objetivo a construção de um modelo para predição de desligamentos de colaboradores de uma empresa, a partir de técnicas de *Machine Learning*. Coletados os dados históricos do quadro de colaboradores de uma instituição financeira brasileira ao longo de um ano, pretende-se, além da construção do modelo que seja capaz de descrever a probabilidade de um trabalhador pedir demissão ou ser demitido, esclarecer quais são os principais fatores que influenciam na rotatividade de pessoal da corporação em questão. Decorre também deste estudo uma avaliação de desempenho de diferentes algoritmos de *Machine Learning*, elucidando o desempenho de cada um deles no contexto do *turnover* da empresa analisada. Estudos brasileiros nessa área não foram encontrados, o que demonstra uma lacuna que este artigo busca preencher.

Para atingir os objetivos, este artigo está dividido em quatro seções, além desta introdução. Na primeira seção, o referencial teórico é apresentado e discutido, demonstrando as definições relacionadas às realidades de desligamentos de colaboradores nas empresas, e as técnicas para construção de modelos preditivos a partir de *Machine Learning*. A segunda seção apresenta a metodologia utilizada na aplicação de *Machine Learning* para solução da situação-problema. A terceira seção apresenta os resultados obtidos com o método examinado, e a discussão sobre eles. Finalmente, na quarta, tem-se a conclusão.

2 REFERENCIAL TEÓRICO

Nesta seção serão discutidos os principais conceitos associados ao objetivo deste artigo. Desligamento de trabalhadores, métodos de *Machine Learning* e como a literatura vem trabalhando a utilização de aprendizado de máquinas para predição da rotatividade no quadro de colaboradores das empresas serão detalhados no transcorrer desta etapa.

2.1 DESLIGAMENTO DE COLABORADORES — *TURNOVER*

A rotatividade de funcionários, ou *turnover*, compreende a admissão e demissão de pessoas em uma organização, e é um dos aspectos pelos quais a empresa e o mercado de trabalho interagem, de forma que essa troca é, até certo ponto, um processo saudável (NASCIMENTO *et al.*, 2012). Entretanto, a recorrência nos desligamentos de colaboradores em um sistema organizacional pode ter as suas causas vinculadas à gestão ineficiente dos recursos humanos, reflexo das políticas e processos internos inapropriados (LIMA *et al.*, 2018).

Nesse sentido, o entendimento dos principais fatores que influenciam o *turnover*, bem como os efeitos associados a ele, são importante objetos de estudo.

O desligamento de indivíduos em uma organização pode ser classificado em dois tipos: aqueles voluntários — por iniciativa do empregado — e os involuntários — por iniciativa do empregador (SELDEN; SOWA, 2015). Os desligamentos voluntários estão relacionados à saída dos colaboradores como resultado de seu próprio desejo, em que as oportunidades vinculadas a deixar o atual emprego são maiores do que os benefícios de manter-se nele (PORTER; WOO; CAMPION, 2015). Os desligamentos por iniciativa da empresa, por sua vez, ocorrem para substituição de funcionários com desempenho incompatível com as expectativas do contratante, para redução e cortes de custos internos, ou para readequação do quadro de colaboradores (CHIAVENATO, 2014).

A compreensão dos fatores que motivam um empregado a deixar a empresa tem sido motivo de diversos estudos ao longo do último século, dada a sua relevância para a sobrevivência das empresas e para estudos relacionados a predições (RUBENSTEIN *et al.*, 2017). Com o objetivo de identificar os motivos que influenciam no aumento das taxa de desligamento das empresas, ferramentas como a pesquisa de clima — que objetiva a coleta de informações dos colaboradores empregados acerca de suas percepções sobre fatores internos da organização — e entrevistas de desligamentos (em que os motivos de desligamento são aprofundados junto aos empregados demitidos) podem ser muito relevantes (BORGES; RAMOS, 2011).

Para Holtom *et al.* (2008), as variáveis associadas ao *turnover* podem ser classificadas a parir de diferentes perspectivas. Os principais aspectos compreendidos são atributos individuais do trabalhador (idade, número de filhos, etc.), características do trabalho (salário, posição, rotinas, etc.), atitudes individuais para com o trabalho (envolvimento e satisfação com o emprego), novas condições pessoais (engajamento, estresse, etc.), contexto organizacional (clima interno, prestígio da empresa, etc.), interface contexto-indivíduo (influência individual, senso de justiça, etc.), fatores externos (oportunidades de trabalho externas), atitudes de partida (procura por novo emprego), comportamento do trabalhador (absenteísmo, atrasos, etc.), e *performance* individual (*feedbacks*, avaliações de desempenho, comportamento, etc.).

A rotatividade de pessoal implica elevados custos para a empresa, o que demonstra a necessidade de se acompanhar esse indicador. Alguns dos principais custos de reposição de funcionários são os decorrentes dos processos de recrutamento (propagandas, pesquisas de

mercado, etc.), seleção de profissionais (tempo dos recrutadores em entrevistas, aplicação de provas, etc.), treinamento do novo profissional contratado (programas de integração, tempo dos instrutores, baixa produtividade no início da jornada, etc.), além dos custos do próprio processo de desligamento em si (pagamento de direitos trabalhistas, entrevistas de desligamento, cargo vago improdutivo, etc.) (CHIAVENATO, 2014).

2.2 *MACHINE LEARNING*

A discussão sobre *Machine Learning* se dará por meio da sua conceituação, classificações típicas de sua aplicação e apresentação de alguns dos métodos mais utilizados na construção de modelos preditivos. Os métodos ilustrados também serão aqueles testados neste artigo.

2.2.1 Conceitos

O conceito de *Machine Learning* surgiu a partir de pesquisas de ponta nas disciplinas de Ciência da Computação e Inteligência Artificial, desde o advento dos computadores na década de 1940, construídos para realização de cálculos e inferências lógicas (LEWIS; DENNING, 2018). O aprendizado no contexto da computação pode ser compreendido pela seguinte definição: "um programa de computador aprende com experiência E, com respeito a uma classe T de tarefas com medida de desempenho P, se seu desempenho em tarefas T, medidas por P, melhora com experiência E" (MITCHELL, 1997, p. 2).

O propósito da utilização de *Machine Learning* é, portanto, a obtenção de um método suficientemente abrangente aos diversos tipos de dados existentes, que seja capaz de processá-los em bases para treino, e classificar com um nível aceitável de acurácia as bases de dados desconhecidas (JORDAN; MITCHELL, 2015). Os métodos de *Machine Learning* podem ser compreendidos nas seguintes classificações: aprendizado supervisionado, não supervisionado e de reforço (ALPAYDIN, 2011).

Mohri, Rostamizadeh e Talwalkar (2012) definiram os métodos supervisionados de aprendizado de máquina (ou *supervised learning*) como aqueles em que as variáveis de resposta do modelo são previamente classificadas a partir de suas variáveis de entrada. Esse conjunto de dados é utilizado para treinar o algoritmo para que sejam previstas as classificações de novos dados — cuja classificação é ainda desconhecida. Métodos supervisionados são utilizados, por exemplo, na predição de *e-mails* classificados como *spam* ou *não spam*.

O aprendizado do tipo não supervisionado (ou *unsupervised learning*) não apresenta classificação prévia da fonte dos dados, ao contrário do aprendizado supervisionado. Nesse caso, são descobertos padrões escondidos nos dados de entrada do modelo — o que pode ser uma vantagem ao descobrir padrões de classificação que não foram identificados anteriormente (SATHYA; ABRAHAM, 2013). A segmentação dos tipos de clientes de uma empresa de acordo com renda e informações demográficas é um exemplo de aprendizado não supervisionado (ALPAYDIN, 2010).

O aprendizado de reforço (ou *reinforcement learning*), por sua vez, diz respeito à sequência de ações que produzem um resultado positivo, de forma que apenas uma atividade que produza resultado satisfatório não é relevante no contexto. Um exemplo é o aprendizado de jogos, como o xadrez, em que, mais importante do que um movimento válido, como o avanço de um peão no tabuleiro, é a sequência de movimentos que produz um resultado satisfatório, como o xeque-mate no rei adversário. (ALPAYDIN, 2010).

2.2.2 Métodos de *Machine Learning*

O domínio do tema de *Machine Learning* compreende a existência de diversos métodos diferentes que se propõem a solucionar as mesmas famílias de problemas. Ao longo deste artigo, para construção do modelo preditivo, serão utilizados os seguintes métodos supervisionados: Regressão Múltipla, *Random Forest*, *Naive Bayes* e *K-Nearest Neighbour*. Esses algoritmos foram selecionados a partir da revisão da literatura, que será apresentada na seção 2.3 Predição de Desligamentos de Colaboradores a partir de *Machine Learning*, em que diferentes pesquisadores se utilizaram desses mesmos métodos para construir seus modelos.

2.2.2.1 Regressão Múltipla

A técnica de modelagem de Regressão Múltipla possui uma variável dependente Y, definida a partir de pelo menos uma variável independente X, de maneira a construir a melhor função Y que descreve essas relações entre variáveis (FLACH; MÜLLER, 2014). Essa equação apresenta ainda coeficientes **β** que representam a mudança média na variável de resposta para uma mudança na variável preditora, além de uma constante associada ao erro, dada por ∈ (FREEDMAN, 2009). De forma genérica, tem-se a seguinte equação:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

Figura 1 - Equação modelo de Regressão Múltipla
Fonte: Adaptada de Flach e Müller (2014).

2.2.2.2 *Random Forest*

Random Forest é um método que constitui um classificador h gerado a partir de um conjunto de outros classificadores estruturados por ramificações $\{h(\mathbf{x}, \Theta_k),\ k = 1, \ldots\}$ em que Θ_k são vetores randômicos independentes, e cada ramificação tem um único voto para definir a classe mais popular no vetor de dados de entrada x (BREIMAN, 2001). Em outras palavras, o *Random Forest* gera aleatoriamente um número suficientemente grande de árvores de decisão a partir das entradas do problema, comparando seus resultados para gerar uma classificação. A imagem a seguir exemplifica o conceito discutido, ilustrando a "classe azul" sendo definida em detrimento da "classe vermelho" dado o maior número de votos das árvores de decisão geradas:

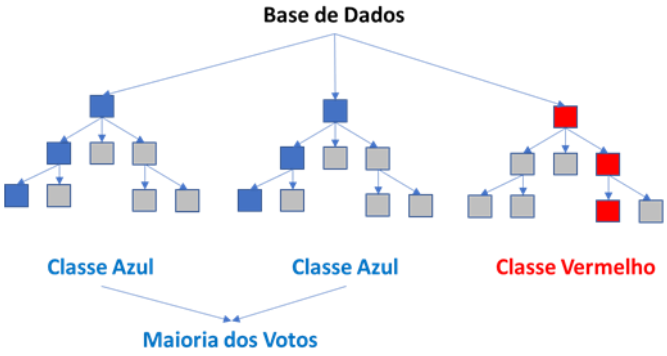


Figura 2 - Árvores de decisão geradas pelo *Random Forest*
Fonte: Elaborada pelo autor.

2.2.2.3 *Naive Bayes*

Aprendizagens por classificadores podem ser simplificadas assumindo, ingenuamente, que as variáveis são independentes das classes que venham a ser utilizadas (RISH, 2001). Dessa forma, ao se assumir que o evento B tenha diferentes classes (B1,...,Bn) e P(A|B) já conhecidas, pode-se descrever o teorema de Bayes, de acordo com Figueira e Deliberal (2013), como:

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \ldots + P(A \mid B_n)P(B_n)}$$

Figura 3 - Teorema de Bayes
Fonte: Adaptada de Figueira e Deliberal (2013).

Portanto, esse método é um classificador probabilístico que calcula as chances de um evento ocorrer a partir do conhecimento de outras probabilidades. Essas probabilidades são inferidas previamente pelos dados de entrada do problema em questão.

2.2.2.4 *K-Nearest Neighbour*

O *K-Nearest Neighbour* (KNN) é um método para classificação de padrões. Ele consiste em determinar a classe de exemplos não classificados a partir do padrão de seus vizinhos mais próximos. Tipicamente as distâncias entre os vetores de entrada são comparadas através da distância euclidiana entre eles, podendo-se variar de acordo com cada tipo de problema (WEINBERGER; SAUL, 2009). A imagem a seguir ajuda a compreender a relação entre as diferentes classificações possíveis para um mesmo espaço bidimensional de dados, classificado a partir de diferentes números de k de aproximações vizinhas (FORMANN-ROE, 2012).

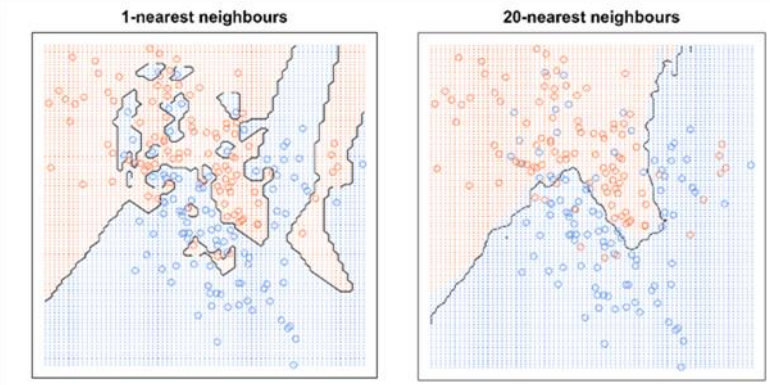


Figura 4 - Espaço bidimensional classificado a partir de diferentes valores para k
Fonte: Adaptada de Friedman, Hastie e Tibshirani (2008).

2.2.3 *Bias* e *Variance Tradeoff*

Uma das fontes de erros associada aos modelos de *Machine Learning*, e que deve ser minimizada, diz respeito a *bias* e *variance*, aspectos que reagem em direções opostas em função da flexibilidade e capacidade de generalizações em um modelo. Geurts (2002) compreende que *bias* é o fenômeno responsável por fazer com que uma função não seja satisfatoriamente flexível para abranger as diferentes classes de dados em uma base, e que, em contrapartida, *variance* é o fenômeno no qual a função é perfeitamente ajustada aos dados, sem capacidade de generalizações para outras bases.

Do *tradeoff* entre *bias* e *variance* depreendem-se os conceitos de *overfitting e underfitting*. *Overfitting* é o perfeito ajuste entre o modelo desenhado e a base de dados, considerando até ruídos e informações sem relevância na origem das informações. Já o *underfitting*, de forma análoga, ocorre quando os modelos não são capazes de detectar as tendências reais de um fenômeno (BRISCOE; FELDMAN, 2011).

A Figura 5 – Comparação entre Overfitting e Underfitting a seguir exemplifica os conceitos de *bias* e *variance* em relação a *overfitting* e *underfitting* (SINGH, 2018).

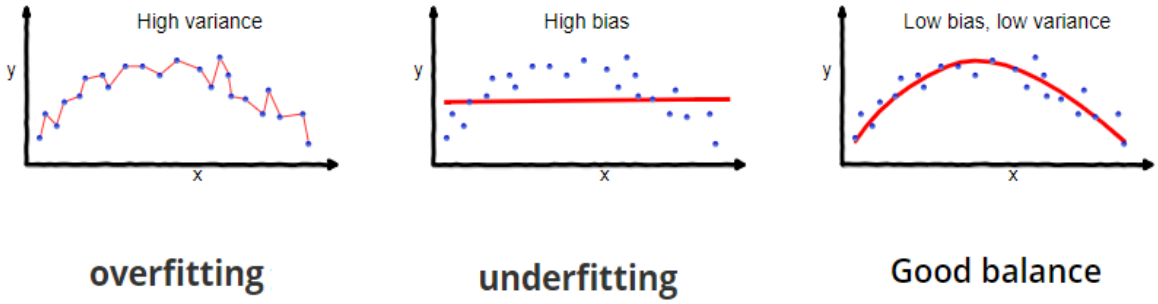


Figura 5 - Comparação entre *Overfitting* e *Underfitting*
Fonte: Adaptada de Singh (2018).

2.2.4 Matriz de Confusão

A fim de se comparar o desempenho de diferentes modelos preditivos, tradicionalmente é utilizada a técnica Matriz de Confusão. Através desse método, é possível determinar a acurácia de cada modelo, ou seja, sua taxa total de acertos *versus* o total de previsões realizadas (CASTRO; BRAGA, 2011). Outra métrica relevante que decorre da matriz de confusão é a precisão, que corresponde ao número de previsões corretas sobre o total de previsões feitas, em relação apenas à classe que se está tentando prever. Em matrizes de confusão, os resultados corretos são chamados de *verdadeiros positivos/negativos*, enquanto as previsões incorretas, de *falsos positivos/negativos*, em que o termo "*positivos*" é reservado aos elementos da classe que está se tentando prever (FAWCETT, 2005). A imagem abaixo ilustra uma matriz de confusão e as fórmulas para cálculo da acurácia e precisão:

	Observações Reais	
Previsões Realizadas	Classe que se tenta prever	Outra Classe
Classe que se tenta prever	Verdadeiro Positivo	Falso Positivo
Outra Classe	Falso Negativo	Verdadeiro Negativo

Acurácia = (Verdadeiro Positivo + Verdadeiro Negativo) / Total de Observações
Precisão = (Verdadeiro Positivo) / (Verdadeiro Positivo + Falso Positivo)

Figura 6 - Matriz de Confusão, Acurácia e Precisão
Fonte: Elaborada pelo autor.

2.3 PREDIÇÃO DE DESLIGAMENTOS DE COLABORADORES A PARTIR DE *MACHINE LEARNING*

A construção de modelos para predição de desligamento de colaboradores tem sido foco de diferentes estudos, de forma que diversas abordagens têm sido utilizadas para explorar o assunto. Novos métodos de aprendizagem de máquina são incorporados a esse desafio com o passar do tempo, assim como são mantidos aqueles que demonstram resultado satisfatório.

Há mais de uma década, Nagadevara, Srinivasan e Valk (2008) estudaram os fatores que influenciam a rotatividade voluntária de trabalhadores, na tentativa de construir um modelo preditivo. A partir de dados demográficos dos empregados, informações sobre posição ocupada e jornada de trabalho em uma empresa de TI, foram abordadas cinco técnicas de mineração de dados: *Artificial Neural Networks* (ANN), Regressão Logística, *Classification Trees*, *Regression Trees* e *Discriminant Analysis*. Os resultados encontrados demonstraram que a melhor técnica para predizer as pessoas que não deixaram a empresa foi a de ANN, embora todas tenham apresentado taxas de acurácia muito próximas. Por outro lado, a técnica que apresentou melhor desempenho para predizer os empregados que pediram demissão foi a *Discriminant Analysis*.

Aplicando métodos semelhantes, Saradhi e Palshikar (2011) trouxeram ainda outras técnicas de *Machine Learning* — tipicamente utilizadas na predição de evasão de clientes das empresas — para o universo de evasão voluntária de colaboradores, dada a similaridade entre os dois problemas. Os modelos preditivos testados foram: *Naive Bayes*, *Support Vector Machines* (SVM), *Random Forest* e Regressão Logística. A base de dados coletada era relativa a um período de 2 anos, em que 80% dos dados foram utilizados para treinar o modelo, enquanto os 20% restantes, para testá-lo. Os atributos selecionados para construção do modelo eram variáveis demográficas (idade, gênero, etc.) e situação na empresa (tempo de empresa, local de trabalho, etc.). Nesse estudo, a técnica de SVM se mostrou mais pertinente, pois teve a maior taxa de "verdadeiros positivos" em relação às outras técnicas, 81,16% no total.

Por sua vez, Fan *et al.* (2012) conduziram o problema de predição do *turnover* em uma empresa de tecnologia utilizando um modelo híbrido, mesclando as tecnologias de *Machine Learning* supervisionada e *clustering analysis*. A partir da aplicação de questionários que buscavam registrar a satisfação e o comprometimento dos trabalhadores na empresa, foram clusterizadas as respostas apuradas e definidos os grupos de empregados que estavam com maior predisposição a deixar a empresa.

Ribes, Touahri e Perthame (2017) testaram novos métodos de aprendizado de máquina em relação aos autores anteriores, para construir um modelo preditivo de *turnover*, além de propor políticas de retenção para solucionar esse problema. As variáveis utilizadas no estudo foram selecionadas a partir do método *Raking with Mutual Information* (MI), que selecionou dados como "tempo no time" e "nota de comportamento" em detrimento a dados de idade e gênero, por exemplo. Os métodos de *Machine Learning* utilizados foram: SVM, *Random Forest* e *Naive Bayes* e *Linear Discriminant Analysis* (LDA). Nessa comparação, o melhor resultado encontrado foi a partir da técnica *Random Forest*, e os fatores que mais contribuíram para o *turnover* foram "Unidade de Trabalho" e "*Performance* do empregado".

Com o mesmo propósito, Yedida *et al.* (2018) construíram um modelo de predição de demissões voluntárias utilizando o algoritmo *K-Nearest Neighbors* (KNN) — não testado nos trabalhos anteriores — para processar variáveis como *performance* de empregados, horas trabalhadas, anos na empresa, salários, etc. A partir da base de empregados selecionada, foram consumidos 70% dos dados para treinar o algoritmo, de forma que a acurácia alcançada foi em torno de 94%. Além disso, os resultados encontrados utilizando o algoritmo KNN foram comparados aos obtidos com outras três técnicas de *Machine Learning — Naive Bayes*, Regressão Logística e *Multi-Layer Perceptron Classifier* — de forma que o KNN se mostrou o melhor para a resolução do problema.

A pesquisa de Khera (2019) também buscou predizer a rotatividade em uma organização a partir de aprendizado de máquina. No estudo em questão, os métodos de *Support Vectors Machine* (SVM) foram utilizados, dada a sua baixa complexidade de implementação em uma empresa. Foram coletados dados de 1.650 empregados de 3 empresas diferentes pelo período de 3 anos. Informações demográficas e relacionadas ao trabalho foram utilizadas e relacionadas ao fato de um indivíduo ter deixado a empresa ou não. O modelo apresentou significativa acurácia de 85%, mas apenas em 65% dos dados predisse os verdadeiros positivos (indivíduos previstos para deixar a empresa e que de fato saíram).

Conclui-se, portanto, que a capacidade de predizer o desligamento de colaboradores é um tema relevante e muito discutido na literatura recente. Este artigo se assemelha com os demais supracitados em comparar diferentes técnicas preditivas. Em contrapartida, diferencia-se por aprofundar a discussão sobre a relevância dos fatores relacionados ao *turnover*, o que não ocorre naqueles trabalhos.

3 PROCEDIMENTOS METODOLÓGICOS

Nesta seção serão abordados os procedimentos metodológicos implicados no artigo. A empresa será brevemente contextualizada, além de ter descritas algumas particularidades do seu quadro de colaboradores. Logo após, o estudo será classificado de acordo com os critérios de natureza, abordagem, objetivos e procedimentos. Por fim, serão apresentadas as etapas realizadas para execução da pesquisa aplicada.

3.1 DESCRIÇÃO DO CENÁRIO

O modelo preditivo de *turnover* será construído para uma instituição financeira cooperativa sediada no Rio Grande do Sul, com atuação nacional. A empresa tem mais de 4 milhões de clientes correntistas (chamados de associados) distribuídos em 112 cooperativas de crédito, que podem ser atendidos em qualquer uma das 1,7 mil agências presentes em 22 estados brasileiros. Atualmente a organização possui R$ 15 bilhões em Patrimônio Líquido, mais de R$ 100 bilhões em ativos e R$ 2,7 bilhões de resultado líquido. Em 2018 a empresa foi reconhecida pelo segundo ano consecutivo como uma das melhores empresas para se trabalhar pela revista Você S/A (GRANATO, 2018).

A empresa, além das 112 cooperativas filiadas ao sistema, tem um Centro Administrativo que desenvolve os produtos financeiros ofertados pelas cooperativas aos seus associados e centraliza processos operacionais para que o sistema obtenha ganhos de eficiência em escala. No total, mais de 28 mil colaboradores trabalham em toda a organização. No Centro Administrativo, que será o foco deste artigo, trabalham aproximadamente 2 mil funcionários que executam atividades relacionadas a desenvolvimento de produtos, implementações de TI e áreas de suporte, como Gestão de Pessoas. Atualmente o Centro Administrativo apresenta um *turnover* mensal de aproximadamente 1,5%, 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.

3.2 CLASSIFICAÇÃO DA PESQUISA

Este estudo está classificado pelos aspectos de natureza, abordagem, objetivos e procedimentos, conforme sugerido por Silveira e Córdova (2009). Sob a perspectiva da sua natureza, esta é uma pesquisa aplicada, pois tem o propósito de fornecer uma ferramenta de gestão aos administradores da empresa, capaz de predizer os colaboradores que estão mais propensos a deixá-la. Dessa forma, em relação à abordagem, este é um estudo quantitativo, porque se utiliza de dados dos trabalhadores e diferentes conceitos de análises estatísticas para

construir o racional. A partir das informações coletadas, objetiva-se definir um modelo de predição de desligamentos, e compreender como as variáveis associadas ao trabalho — como dados contratuais e informações demográficas — se relacionam, o que por sua vez caracteriza este estudo como uma pesquisa explicativa. Por fim, pelo aspecto procedimental, este estudo de caso permite o aprofundamento sobre a realidade da rotatividade no quadro de pessoal da empresa em questão.

3.3 ETAPAS DO TRABALHO

Conforme estudos conduzidos por Alao e Adeyemo (2013) e Sajjadiani *et al.* (2019), um modelo preditivo para *turnover* pode ser desenvolvido a partir de quatro macroetapas: (i) coleta de dados e seleção de variáveis; (ii) preparação dos dados para posterior leitura do sistema; (iii) desenvolvimento do modelo preditivo; (iv) análise dos resultados obtidos.

Os dados referentes ao quadro de colaboradores da empresa são armazenados em diferentes sistemas internos. Essas informações foram acessadas via consultas em banco de dados e, uma vez extraídas, exigiram cruzamentos para construção de uma base única a ser utilizada para este estudo. O período histórico das informações extraídas foi relativo ao ano de 2018 (de Janeiro a Dezembro), ou seja, todos os colaboradores que estiveram ativos nesse intervalo de tempo, inclusive aqueles eventualmente desligados. A base de dados constitui-se de 2.059 colaboradores, dos quais 287 foram desligados de forma voluntária ou involuntária. Observado, portanto, o desbalanceamento entre a quantidade de colaboradores ativos contra os demitidos, selecionaram-se aleatoriamente apenas 287 colaboradores ativos, para houvesse equilíbrio nos dados a serem averiguados. A base de informações final foi composta por 574 colaboradores.

Definida, então, a origem única das informações, iniciou-se a etapa de preparação delas, para que pudessem ser lidas pelo modelo a ser desenvolvido. Dentre as centenas de aspectos registrados na base, foram selecionadas 14 variáveis de acordo com os trabalhos realizados por Khera (2019), Yedida *et al.* (2018) e Ribes, Touahri e Perthame (2017). As 14 variáveis escolhidas estão descritas e explicadas no quadro abaixo:

Nome da variável:	**Descrição da variável:**
Gênero	Informa se o profissional pertence ao gênero masculino ou feminino.
Faixa de Idade	Informa a idade do profissional.
Número de Filhos	Informa o número de filhos do profissional.
Tempo de Empresa	Informa o tempo que o profissional trabalha na empresa.

Grade	Informa o grau de senioridade do profissional na empresa. Quanto maior o Grade, mais experiente é o trabalhador (valores variam entre 4 e 15).
Gestor	Informa se o profissional ocupa um cargo de gestão de equipes.
Cidade Natal	Informa se o profissional é de Porto Alegre, Região Metropolitana ou outras cidades.
Remuneração	Informa a remuneração do profissional.
Remuneração de Mercado	Informa o quanto o cargo do colaborador recebe em média no mercado de trabalho.
Remuneração Média do Grade	Informa o valor da média salarial dos profissionais que ocupam o respectivo Grade na empresa.
Remuneração / Remuneração de Mercado	Informa a razão entre a remuneração do profissional e a remuneração que poderia receber se estivesse em outra empresa. Indica o quanto o trabalhador recebe em relação ao mercado de trabalho.
Remuneração / Remuneração Média do Grade	Informa a razão entre a remuneração do profissional e a remuneração que seus colegas de mesma experiência recebem em média. Indica o quanto o trabalhador recebe em relação aos seus colegas de mesma categoria.
Nota Avaliação de Desempenho	Informa o valor da avaliação de desempenho anual do colaborador, feita pelo seu superior imediato. Os valores da avaliação variam entre 0 e 1,2, em que avaliações inferiores a 1,0 são consideradas como insuficientes.
Desligamentos	Informa se o profissional deixou ou permaneceu na empresa.

Quadro 1 - Variáveis e suas definições
Fonte: Elaborado pelo autor.

Eventuais registros vazios foram consultados e preenchidos manualmente na base, bem como variáveis qualitativas, como “sim ou não”, transformadas em registros binários, (1 ou 0). Toda a base foi normalizada para que a diferença de magnitude entre os valores das variáveis não distorcesse os cálculos envolvidos na construção dos modelos (por exemplo, a variável número de filhos possui valores que variam entre 0 e 4, enquanto a variável remuneração, entre R$ 1.378,00 e R$ 30.825, na amostra selecionada).

A posse de uma base única e com dados preparados permitiu o avanço do estudo para a etapa de desenvolvimento do modelo preditivo de desligamento de colaboradores. Para essa implementação, utilizou-se o Rstudio, versão 3.6.1 (2019-07-05), Windows 64 bits. O R é uma linguagem para computação estatística de código aberto, enquanto o Rstudio é um *software*, também de código aberto, no qual realizam-se as codificações e instruções para se proceder à análise de dados. Os pacotes selecionados para execução dos métodos de *Machine Learning* propostos neste artigo foram: *Class Package*, para modelos de *K-Nearest Neighbour* (VENABLES; RIPLEY, 2002), *Caret Package*, para modelos a partir de Regressão Múltipla e *Naive Bayes* (KUHN, 2008), e *RandomForest Package* para modelos de *Random Forest* (BREIMAN, 2001).

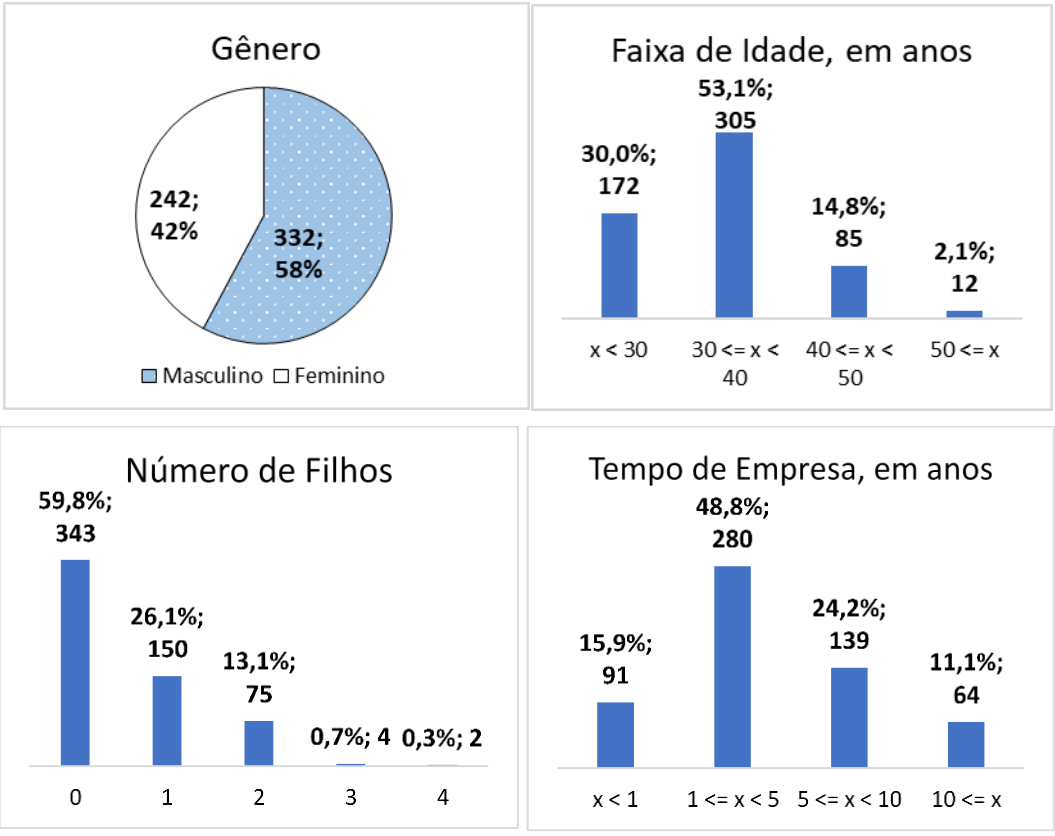
A etapa final compreendeu dividir a massa de dados construída em: bases de treino para os algoritmos (90% dos dados selecionados de maneira aleatória) e bases para testes de *performance* (10% dos dados restantes) (MATERA, 2019). Analisaram-se os resultados gerados pelos diferentes métodos definidos, avaliando sua acurácia e precisão através da aplicação de matrizes de confusão em suas previsões. Os resultados obtidos e conclusões constatadas estão descritos nos capítulos que seguem.

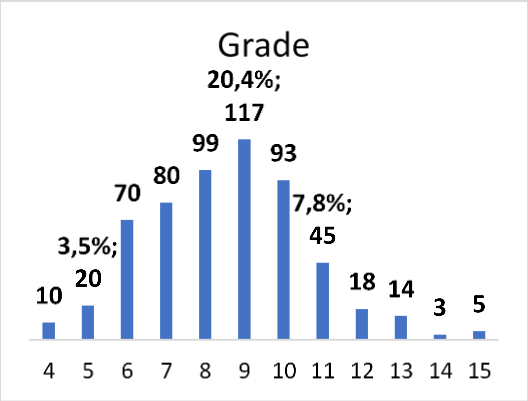
4 RESULTADOS

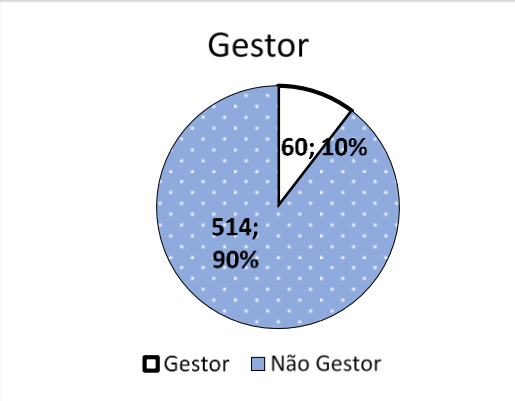
Nesta seção serão abordados os resultados obtidos no estudo, a partir da descrição da base de dados selecionada, evidenciando suas características, análise dos resultados obtidos por cada modelo gerado, e comparação entre os diferentes resultados obtidos. Por fim, será discutido sobre as características comuns às pessoas identificadas como potenciais trabalhadores que deixarão a empresa pelo melhor modelo identificado.

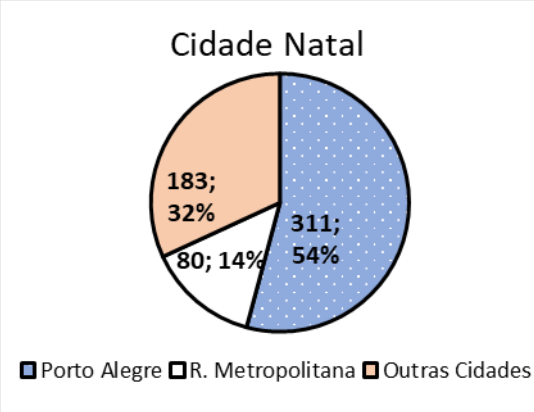
4.1 DESCRIÇÃO DA BASE DE DADOS

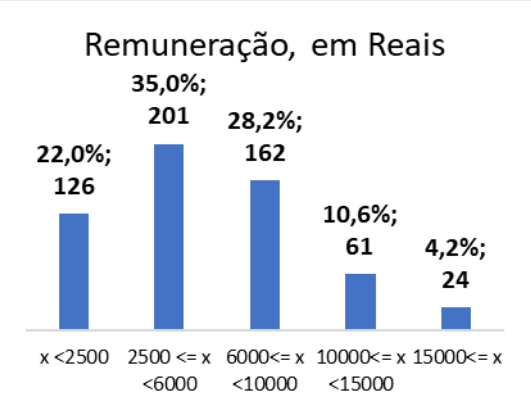
A base de 574 colaboradores selecionados aleatoriamente para construção do modelo possui características decorrentes das 14 variáveis utilizadas para a descrição dos trabalhadores. Abaixo está ilustrada a composição dessa base, e suas particularidades.

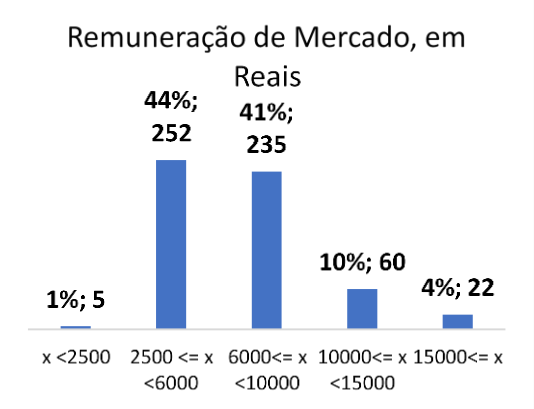


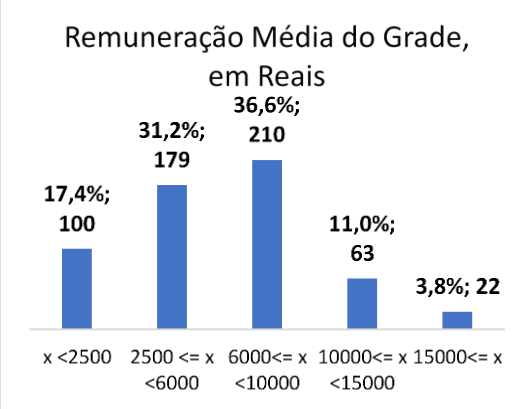

Grade
10
3,5%; 20
70
80
99
20,4%; 117
93
7,8%; 45
18
14
3
5
4 5 6 7 8 9 10 11 12 13 14 15

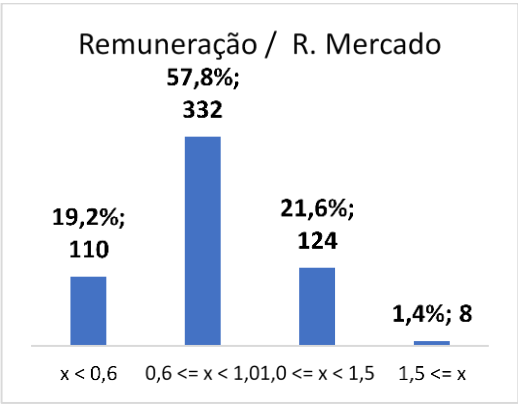

Gestor
60; 10%
514; 90%
Gestor
Não Gestor

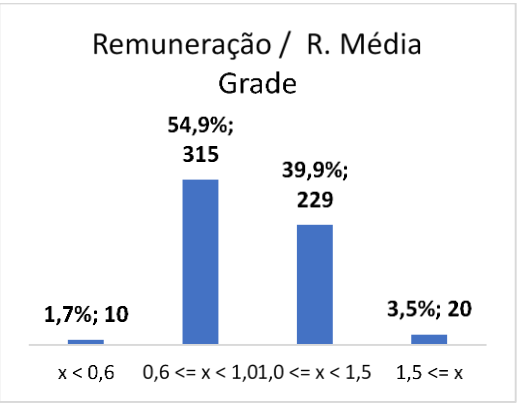

Cidade Natal
183; 32%
80; 14%
311; 54%
Porto Alegre
R. Metropolitana
Outras Cidades


Remuneração, em Reais
22,0%; 126
35,0%; 201
28,2%; 162
10,6%; 61
4,2%; 24
x <2500
2500 <= x <6000
6000<= x <10000
10000<= x <15000
15000<= x


Remuneração de Mercado, em Reais
1%; 5
44%; 252
41%; 235
10%; 60
4%; 22
x <2500
2500 <= x <6000
6000<= x <10000
10000<= x <15000
15000<= x


Remuneração Média do Grade, em Reais
17,4%; 100
31,2%; 179
36,6%; 210
11,0%; 63
3,8%; 22
x <2500
2500 <= x <6000
6000<= x <10000
10000<= x <15000
15000<= x


Remuneração / R. Mercado
19,2%; 110
57,8%; 332
21,6%; 124
1,4%; 8
x < 0,6
0,6 <= x < 1,0
1,0 <= x < 1,5
1,5 <= x


Remuneração / R. Média Grade
1,7%; 10
54,9%; 315
39,9%; 229
3,5%; 20
x < 0,6
0,6 <= x < 1,0
1,0 <= x < 1,5
1,5 <= x

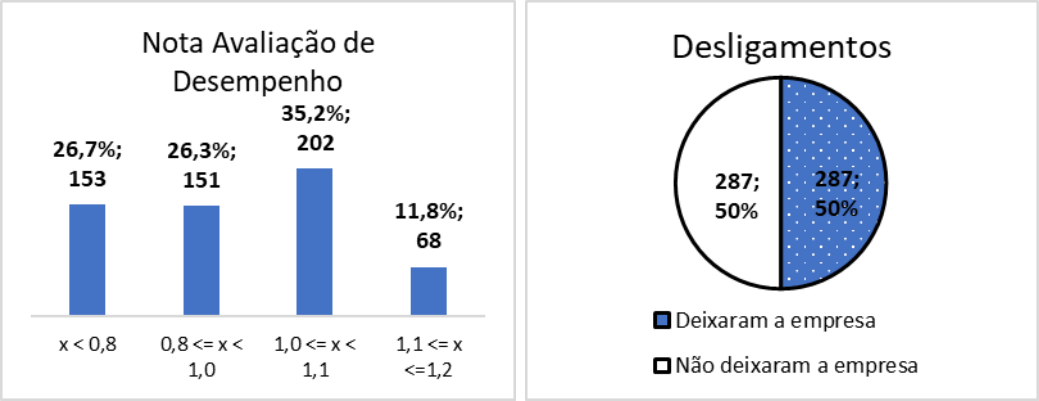


Figura 7 - Detalhamento da Base de Dados
Fonte: Elaborada pelo autor.

Dessa forma, observa-se que a base utilizada era composta majoritariamente por pessoas do gênero masculino, com idades entre 30 e 40 anos, sem filhos, não gestores, e nascidos em Porto Alegre. Adicionalmente, destaca-se que é de 126 o número de pessoas com remuneração inferior R$ 2.500,00, enquanto no Mercado apenas 5 desses profissionais se manteriam nessa faixa salarial inferior. Como já era esperado, 50% da base são formados por pessoas que permaneceram na empresa e 50% por pessoas que a deixaram, buscando um equilíbrio na formação da base de análise.

4.2 CONSTRUÇÃO DOS MODELOS E ANÁLISE DE RESULTADOS

O primeiro modelo preditivo foi construído a partir do algoritmo *K-Nearest Neighbour*, em que os colaboradores da base de teste foram categorizados em "deixará a empresa" ou "ficará na empresa" a partir dos votos fornecidos pelos seus "k" vizinhos mais próximos, na base de treino. A distância euclidiana foi utilizada como parâmetro para cálculo de distância. O modelo KNN exigiu, portanto, que fosse definido um valor de "k" vizinhos que deveriam ser utilizados no modelo. Desse modo, testaram-se valores entre k=1 (classificação pelo vizinho mais próximo) e k=574 (tamanho total da base de treino), de forma que o melhor modelo foi encontrado para k=25, conforme matriz de confusão abaixo:

Tabela 1 - Matriz de Confusão - KNN

	Dados Observados	
Predição	**Deixou a empresa**	**Ficou na empresa**
Deixará empresa	20	8
Ficará na empresa	10	22

Acurácia	70,0%
Precisão	71,4%

Fonte: Elaborada pelo autor.

O segundo modelo foi construído a partir da técnica de Regressão Múltipla, em que é descrita a melhor equação que se ajusta aos pontos definidos pelas variáveis da base de treino. A equação originada no modelo é representada por: y = -0,03 *gênero + 0,24 *Idade + 0,26 *Filhos – 0,14 *Tempo de Contrato + 1,71 *Grade + 0,03 *Gestor - 0,01 *Cidade Nascimento + 1,67 *Remuneração Mensal – 0,47 *Remuneração Mercado -2,41 *Remuneração Média Grade - 0,98 *(Remuneração Mensal sobre Remuneração Mercado) + 0,34 *(Remuneração Mensal sobre Remuneração Grade) + 0,48 *Nota Objetivo – 0,14. A base de testes foi então submetida ao modelo construído, de tal forma que os valores resultantes entre 0 e 0,5 foram classificados como "Deixará a empresa" e aqueles entre 0,5 e 1,0 como "Ficará na empresa". A matriz de confusão do modelo está descrita abaixo:

Tabela 2 - Matriz de Confusão - Regressão Múltipla

	Dados Observados	
Predição	**Deixou a empresa**	**Ficou na empresa**
Deixará empresa	18	10
Ficará na empresa	12	20

Acurácia	63,3%
Precisão	64,3%

Fonte: Elaborada pelo autor.

O terceiro algoritmo testado foi o *Naive Bayes*. As probabilidades *a priori* das classes "Deixou a empresa" e "Ficou na empresa" calculadas a partir da base de treino foram ambas de 50%, o que era esperado, visto que essa composição proposital visava evitar vieses nas predições. Os resultados preditivos do modelo quando submetido à base de testes está descrito abaixo:

Tabela 3 - Matriz de Confusão - *Naive Bayes*

	Dados Observados	
Predição	**Deixou a empresa**	**Ficou na empresa**
Deixará empresa	13	8
Ficará na empresa	17	22

Acurácia	58,3%
Precisão	61,9%

Fonte: Elaborada pelo autor.

Por fim, a técnica *Random Forest* foi utilizada para construir o modelo de previsões fundamentado no conceito de árvores de decisões aleatoriamente geradas. O número de variáveis testadas em cada árvore foi de 3 (*default* do algoritmo). Testou-se também o desempenho do modelo para um total de árvores variando entre 1 e 500 (acima de 500 árvores aleatórias o computador onde se realizou o estudo perdeu *performance*, e apresentou lentidão, inviabilizando os cálculos). O melhor resultado foi encontrado para 3 árvores, o que proporcionou o resultado ilustrado abaixo:

Tabela 4 - Matriz de Confusão - *Random Forest*

	Dados Observados	
Predição	**Deixou a empresa**	**Ficou na empresa**
Deixará empresa	22	5
Ficará na empresa	8	25

Acurácia	78,3%
Precisão	81,5%

Fonte: Elaborada pelo autor.

Para fins de melhor se comparar o resultado obtido pelos quatro modelos preditivos construídos, elaborou-se uma tabela comparativa em relação à acurácia e à precisão. A tabela está descrita abaixo:

Tabela 5 - Comparativo entre Métodos

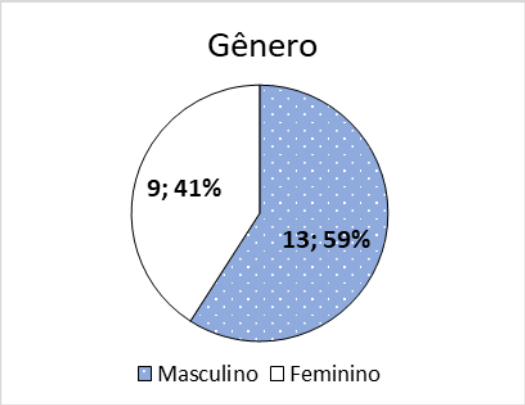
Modelo	**Acurácia**	**Precisão**
K-Nearest Neighbour	70,0%	71,4%
Regressão Múltipla	63,3%	64,3%
Naive Bayes	58,3%	61,9%
Random Forest	78,3%	81,5%

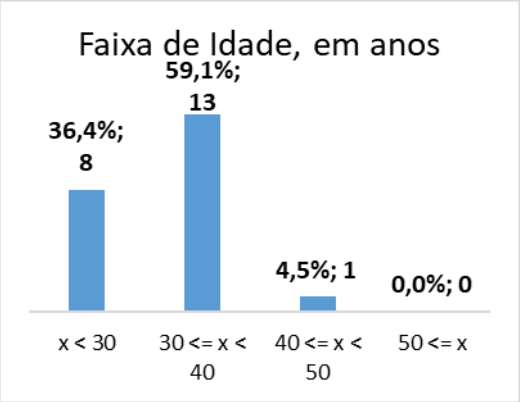
Fonte: Elaborada pelo autor.

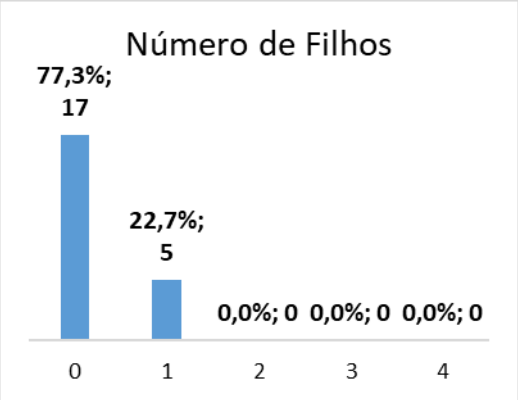
Constatou-se, portanto, que o melhor desempenho foi obtido pelo algoritmo *Random Forest*, que realizou 78,3% das previsões de forma correta, enquanto obteve 81,5% de acertos especificamente sobre as previsões daqueles que deixariam a empresa. O pior desempenho foi registrado pelo método *Naive Bayes*, que não conseguiu capturar efetivamente as particularidades dos dados do estudo, apresentando acurácia e precisão de 58,3% e 61,9%, respectivamente.

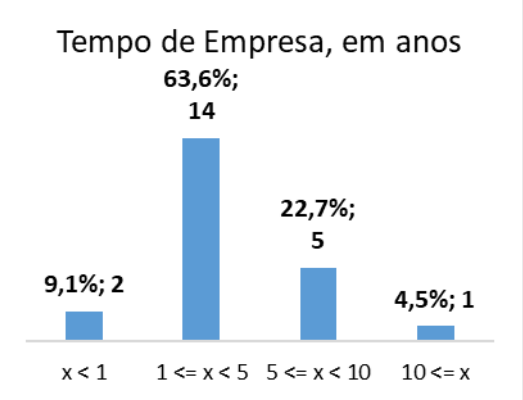
4.3 ANÁLISE DAS PREVISÕES NO MÉTODO *RANDOM FOREST*

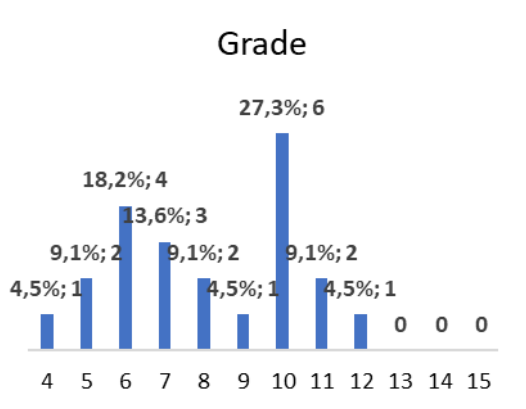
A partir das previsões realizadas pelo modelo *Random Forest*, método identificado como o de melhor desempenho preditivo, realizou-se uma análise das características comuns aos trabalhadores que foram corretamente previstos como aqueles que deixariam a empresa. Segue abaixo o detalhamento do perfil desses 22 empregados *verdadeiro positivos*, ou seja, que o algoritmo previu o desligamento satisfatoriamente:

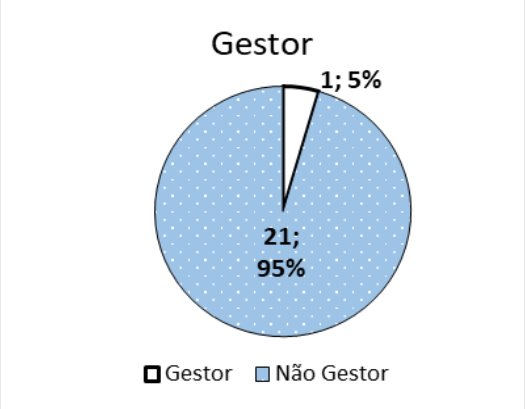

Gênero
9; 41%
13; 59%
Masculino
Feminino

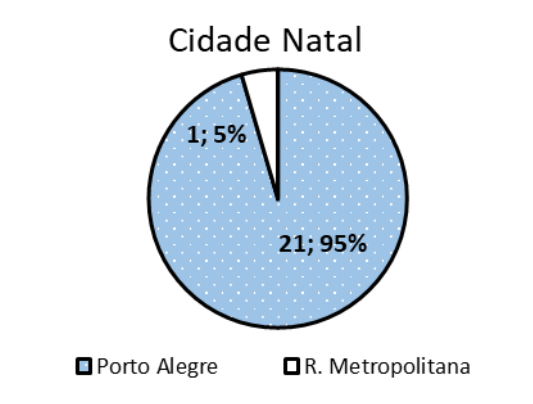

Faixa de Idade, em anos
36,4%; 8
59,1%; 13
4,5%; 1
0,0%; 0
x < 30
30 <= x < 40
40 <= x < 50
50 <= x

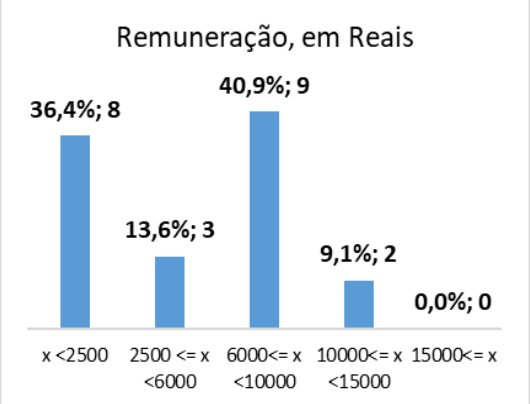

Número de Filhos
77,3%; 17
22,7%; 5
0,0%; 0
0,0%; 0
0,0%; 0
0
1
2
3
4


Tempo de Empresa, em anos
9,1%; 2
63,6%; 14
22,7%; 5
4,5%; 1
x < 1
1 <= x < 5
5 <= x < 10
10 <= x


Grade
4,5%; 1
9,1%; 2
18,2%; 4
13,6%; 3
9,1%; 2
4,5%; 1
27,3%; 6
9,1%; 2
4,5%; 1
0
0
0
4
5
6
7
8
9
10
11
12
13
14
15


Gestor
1; 5%
21; 95%
Gestor
Não Gestor


Cidade Natal
1; 5%
21; 95%
Porto Alegre
R. Metropolitana


Remuneração, em Reais
36,4%; 8
13,6%; 3
40,9%; 9
9,1%; 2
0,0%; 0
x <2500
2500 <= x <6000
6000<= x <10000
10000<= x <15000
15000<= x

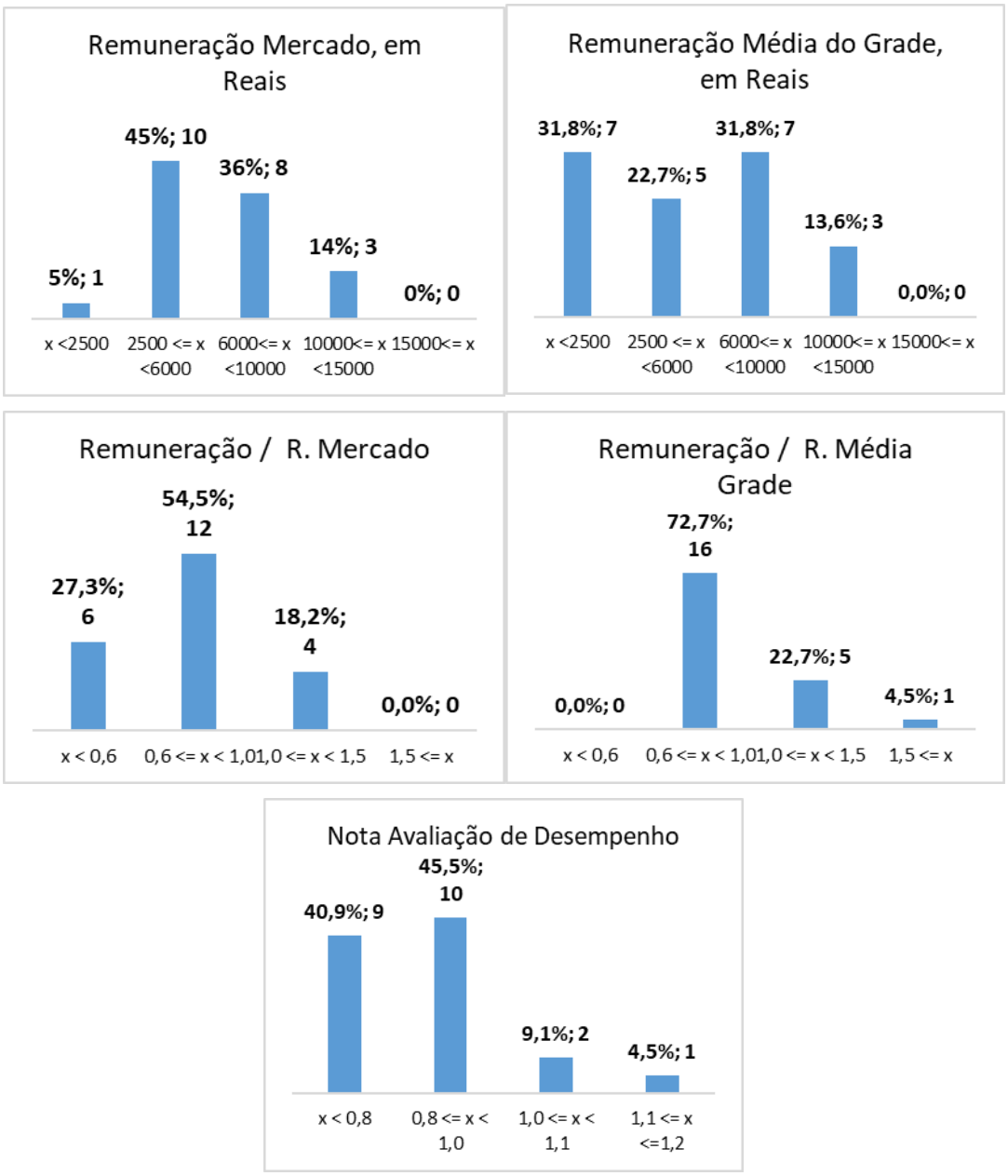


Figura 10 - Perfil das Predições Verdadeiro Positivos
Fonte: Elaborada pelo autor.

A partir dos dados acima, observou-se que, em relação à variável Gênero, as predições se mantiveram de acordo com a proporção do restante da base de dados, enquanto trabalhadores mais jovens e com menos filhos foram apontados com maior facilidade pelo modelo como sendo aqueles que deixariam a empresa. Em relação a colaboradores com nível de senioridade maior, evidenciados pelos Grades maiores ou pelo fato de serem Gestores, observou-se que estes não foram previstos em valores expressivos. Em relação ao quesito Remuneração, o modelo *Random Forest* identificou corretamente que deixariam a empresa o perfil de empregado que recebia menos do que a média salarial das pessoas do seu próprio Grade, ou

recebia menos do que a Remuneração de Mercado. Por fim, a variável Nota de Avaliação de Desempenho se mostrou muito relevante, pois mais de 86% das pessoas identificadas pelo algoritmo possuíram avaliação inferior a 1,0, ou seja, tiveram uma avaliação insuficiente.

5 CONCLUSÕES

A rotatividade no quadro de colaboradores de uma empresa está relacionada a fatores como custos, produtividade, e clima organizacional, de maneira que atuar proativamente nesse tema pode conferir vantagens competitivas para as organizações. Este artigo objetivou, portanto, a construção de um modelo preditivo para desligamentos de trabalhadores de uma instituição financeira, a partir de técnicas de *Machine Learning*. Pretendia-se também avaliar a *performance* desses algoritmos de aprendizado de máquina, comparando seus desempenhos quando implementados. Uma vez construído o modelo preditivo, buscava-se esclarecer quais eram os principais fatores organizacionais relacionados à rotatividade de pessoal na empresa analisada.

Para a construção do modelo, foram selecionadas quatro técnicas de *Machine Learning* a partir da revisão da literatura: *K-Nearest Neighbour*, Regressão Múltipla, *Naive Bayes*, e *Random Forest*. Para que fosse possível a comparação dos diferentes desempenhos obtidos, utilizou-se a ferramenta Matriz de Confusão e seus indicadores de Acurácia e Precisão. Selecionaram-se dados referentes às pessoas que trabalharam e deixaram a empresa pelo período de um ano, em que 90% das informações foram utilizadas para treinar os modelos, e os 10% restantes para testar suas *performances*.

Observou-se que o melhor modelo foi elaborado a partir do algoritmo *Random Forest* com 3 árvores, que obteve 78,3% de acurácia — capacidade de acertar o trabalhador que "deixou a empresa" ou "ficou na empresa" — e 81,5% de precisão — percentual de acertos das previsões feitas apenas sobre quem "deixou a empresa". Os algoritmos *K-Nearest Neighbour* e Regressão Múltipla tiveram desempenhos intermediários, enquanto o *Naive Bayes* apresentou o pior desempenho, com acurácia e precisão de 58,3% e 61,9%, respectivamente. Dentre as características preponderantes nas demissões, previstas corretamente pelo método *Random Forest*, observou-se a baixa média de idade, baixo número de filhos, não ocupar cargos de maior senioridade, remuneração inferior à média da própria empresa e mercado, além de avaliação de desempenho insuficiente.

O objetivo do artigo foi alcançado, desenvolvendo-se um modelo para predição de demissões, testando-se diferentes algoritmos de *Machine Learning*, e compreendendo os fatores

mais relevantes nas predições realizadas. Este trabalho seguirá sendo executado e aperfeiçoado na empresa analisada. Sugere-se como evolução deste estudo a utilização de outras variáveis associadas aos trabalhadores, além das 14 presentes neste artigo, bem como a testagem de outros algoritmos de aprendizagem de máquina. Outra discussão pertinente ao tema diz respeito às questões éticas associadas à utilização de dados pessoais em algoritmos cujos resultados, em algum grau, podem impactar a vida das pessoas.

REFERÊNCIAS

ALAO, D.; ADEYEMO A. B. Analyzing employee attrition using decision tree algorithms. **Computing, Information Systems & Development Informatics**, v. 4, n. 1, p. 17-28, 2013.

ALPAYDIN, E. **Introduction to machine learning**: adaptive computation and machine learning series. 2. ed. Cambridge, MA: MIT Press, 2010.

ALPAYDIN, E. **Machine learning**. Nova Jersey: John Wiley & Sons, Inc., 2011. v. 3.

BODIE, M. *et al*. **The law and policy of people analytics**. Colorado: University of Colorado Law Review, 2017.

BORGES, M.; RAMOS, N. Turnover: uma consequência de estratégias ineficientes de gestão empresarial? In: CONVIBRA ADMINISTRAÇÃO – CONGRESSO VIRTUAL BRASILEIRO DE ADMINISTRAÇÃO, 8., 2011, Rio de Janeiro. **Anais…** Rio de Janeiro: ENSP, 2011.

BREIMAN, L. R. F. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

BRISCOE, E.; FELDMAN, J. Conceptual complexity and the Bias/Variance tradeoff. **Cognition**, v. 118, p. 2-16, 2011.

CASTRO, C.; BRAGA, A. Aprendizado supervisionado com conjunto de dados desbalanceados. **Revista Controle & Automação**, v. 22, n. 5, p. 441-466, 2011.

CHIAVENATO, I. **Gestão de pessoas**: o novo papel dos recursos humanos nas organizações. Barueri, SP: Manole, 2014.

FAN, C. *et al*. Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. **Expert Systems with Applications**, v. 39, p. 8844-8851, 2012.

FAWCETT, T. An Introduction to ROC Analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861-874, 2005.

FIGUEIRA, M.; DELIBERAL, J. Aplicabilidade do teorema de bayes no monitoramento de redes sociais. In: MOSTRA DE INICIAÇÃO CIENTÍFICA, PÓS-GRADUAÇÃO, PESQUISA E EXTENSÃO, 13., 2013, Caxias do Sul. **Anais...** Caxias do Sul: Universidade de Caxias do Sul, 2013. (v. 2, p. 1-21).

FLACH, L.; MÜLLER, M. M. Apresentação de um modelo de regressão múltipla para o *disclosure* de ativos intangíveis. **Brazilian Journal of Quantitative Methods Applied to Accounting**, v. 1, n, 2, p. 36-51 2014.

FORTMANN-ROE, S. **Understanding the bias variance tradeoff**. 2012. Disponível em: <http://scott.fortmann-roe.com/docs/Biasvariance.html>. Acesso em: 20 set. 2019.

FREEDMAN, D. **Statistical models**. Cambridge: Cambrige University Press, 2009.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. [s.l] : Springer, 2008. v. 2.

GARRIDO, G.; SILVEIRA, R.; SILVEIRA, M. People analytics: uma abordagem estratégica para a gestão do capital humano. **Revista Eletrônica de Estratégia & Negócios**, v. 11, n. 1, p. 28-52, 2018.

GEURTS, P.; **Contributions to decision tree induction**: bias/variance tradeoff and time series classification. Systems and Modeling. Liège: University of Liège, Belgium, 2002.

GRANATO, L. Conheça as 150 melhores empresas para trabalhar de 2018. **Revista Você S/A**., n. 11, p. 116, 2018.

GRILLO, M.; HACKETT, A. **What types of predictive analytics are being used in talent management organizations?** 2015. Disponível em: <http://digitalcommons.ilr.cornell.edu/student/74>. Acesso em: 31 ago. 2019.

HOLTOM, B. *et al*. Turnover and retention research: a glance at the past, a closer review of the present, and a venture into the future. **The Academy of Management Annals**, v. 2, n. 1, p. 231-274, 2008.

JORDAN, M.; MITCHELL, T. Machine learning: trends, perspectives, and prospects. **Sciencemag.org**, v. 349, n. i.6245, p. 255, 2015.

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

KUHN, M. Building predictive models in r using the caret package. **Journal of statistic software**, v. 28, n. i.5, p. 1-26, 2008.

LEWIS, T.; DENNING, P. Learning machine learning. **Communications of ACM**, v. 61, n. 12, p. 24-27, 2018.

LIMA, K. *et al*. Rotatividade: percepção dos colaboradores sobre as causas de demissão voluntária. **Revista Expressão Católica**, v. 7, n. 2, p. 110-118, jul./dez. 2018.

MATERA. Machine Learning: bases de amostra e treino. Disponível em: http://www.matera.com/blog/post/machine-learning-dividir-amostra-em-treino-e-teste Acesso em: 05 jul. 2020.

MITCHELL, M. **Machine learning.** New York: McGrawl-Hill, 1997.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. Cambridge, MA: The MIT Press, 2012.

NAGADEVARA, V.; SRINIVASAN, V.; VALK, R. Establishing a link between employee turnover and withdrawal behaviours: application of data mining techniques. **Research and Practice in Human Resource Management**, v. 16, p. 81-99, 2008.

NASCIMENTO, K. *et al*. Rotatividade nas organizações: as causas dos desligamentos voluntários em uma empresa de serviços de juiz de fora. **Revista das Faculdades Integradas Vianna Júnior**, v. 3, n. 1, p. 9-29, 2012.

PEDRO, W. Gestão de pessoas nas organizações**. Revista Brasileira Multidisciplinar**, v. 9, n. 2, p. 81-86, 2005.

PORTER, C.; WOO, S.; CAMPION, M. Internal and external networking differentially predict turnover through job embeddedness and Job Offers. **Personnel Psychology**, v. 69, n. i.3, p. 635-672, 2015.

RIBES, E.; TOUAHRI, K.; PERTHAME, B. **Employee turnover prediction and retention policies design: a case study**. 2017. Disponível em: <https://arxiv.org/abs/1707.01377>. Acesso em: 15 nov. 2019.

RISH, I. An Empirical Study of the Naïve Bayes Classifier. International Joint Conferences on Artificial Intelligence - Work Empirical Methods in Artificial Intelligence. **ResearchGate**, v. 3, p. 41-46, 2001.

RUBENSTEIN, A. *et al*. Surveying the forest: a meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. **Personnel Psychology**, p. 1-43, 2017.

SAJJADIANI, S. *et al*. Using machine learning to translate applicant work history into predictors of performance and turnover. **Journal of Applied Psychology**, v. 104, n. 10, p. 1207-1225, 2019.

SARADHI, V.; PALSHIKAR, Girish. Employee churn prediction. **Expert Systems Applicantion**, v. 38, p. 1999-2006, 2011.

SATHYA, R.; ABRAHAM, A. Comparison of supervised and unsupervised learning algorithms for pattern classification. **International Journal of Advanced Research in Artificial Intelligence**, v. 2, n. 2, p. 34-38, 2013.

SELDEN, S.; SOWA, J. Voluntary turnover in nonprofit human service organizations: the impact of high performance work practices**, Human Service Organizations: Management, Leadership & Governance**, v. 39, p. 182-207, 2015.

SHEN, Cedric M. **People analytics & text mining with r**. São Paulo: Amazon, 2019.

SILVEIRA, D.; CÓRDOVA, F. A pesquisa científica. In: GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de pesquisa**. Porto Alegre: Editora da UFRGS, 2009. p. 31-42.

SINGH, S. **Understanding the bias variance tradeoff**. 2018. Disponível em: <https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229>. Acesso em: 08 set. 2019.

TURSUNBAYEVA, A.; DI LAURO, S.; PAGLARI, C. People analytics: a scoping review of conceptual boundaries and value propositions. **International Journal of Information Management**, v. 43, p. 224-247, 2018.

VENABLES, N.; RIPLEY, D. **Modern applied statistics with S, Fourth Edition**. [s.l.]: Springer, 2002.

WEINBERGER, K.; SAUL, K. Distance metric learning for large margin nearest neighbor classification. **Journal of Machine Learning Research**, v. 10, p. 207-244, 2009.

YEDIDA, R. *et al*. **Employee attrition prediction**. New York: Cornell University, 2018.