André Costa Nadalini

TECNOLOGIA DA INFORMAÇÃO E APRENDIZADO MÁQUINA: UMA ABORDAGEM PRÁTICA NA SELEÇÃO E CLASSIFICAÇÃO DE CARACTERÍSTICAS PARA CONTRATAÇÃO DE FUNCIONÁRIOS NA ERA DIGITAL.

Trabalho de Conclusão de Curso submetido à Coordenação do Curso de Engenharia de Produção e Sistemas para a obtenção do Grau de Engenheiro Eletricista com habilitação em Engenharia de Produção.
Prof. Dr. Ricardo Faria Giglio
Orientador:

Florianópolis

2019

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Nadalini, André Costa
Tecnologia da informação e aprendizado máquina: uma abordagem prática na seleção e classificação de características para contratação de funcionários na era digital. / André Costa Nadalini ; orientador, Ricardo Faria Giglio, 2019.
104 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Elétrica, Florianópolis, 2019.

Inclui referências.

1. Engenharia de Produção Elétrica. 2. Aprendizado máquina. 3. Modelos preditivos. 4. Tecnologia da informação. 5. Contratação. I. Giglio, Ricardo Faria. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Elétrica. III. Título.

André Costa Nadalini

TECNOLOGIA DA INFORMAÇÃO E APRENDIZADO MÁQUINA: UMA ABORDAGEM PRÁTICA NA SELEÇÃO E CLASSIFICAÇÃO DE CARACTERÍSTICAS PARA CONTRATAÇÃO DE FUNCIONÁRIOS NA ERA DIGITAL.

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de "Bacharel em Engenharia Elétrica com habilitação em Engenharia de Produção" e aprovado em sua forma final pelo Curso de Graduação em Engenharia de Produção e Sistemas na Universidade Federal de Santa Catarina

Florianópolis, 21 de novembro de 2019.

Prof. Guilherme Ernani Vieira, Dr.

Coordenadora dos Cursos de Graduação em Engenharia de Produção

Banca Examinadora:

Prof. Ricardo Faria Giglio, Dr.

Orientador

Universidade Federal de Santa Catarina

Prof., Enzo Morosini Frazzon Dr.

Universidade Federal de Santa Catarina

Letícia de Castro Rodrigues

Resultados Digitais

Um brinde aos meus pais, Fernando e Margarete, aos meus irmãos, Felipe e Elisa, e a todos os meus amigos e familiares, pelo incondicional apoio ao longo destes anos.

AGRADECIMENTOS

Agradeço aos meus pais, Fernando Nadalini e Margarete Costa, por todo suporte, amor e carinho. Sou privilegiado por tê-los.

Aos meus irmãos, Felipe e Elisa, que, mesmo distantes, nunca deixaram de estar perto.

À toda minha família, pelo amor e apoio incondicional.

Ao sempre querido Kiko Silva, que me mostrou que não podemos desistir sem nunca ter tentado.

Aos meus queridos amigos, que nos momentos mais felizes e difíceis, me deram força para continuar na trajetória.

À Resultados Digitais e XP Investimentos e todos os meus colegas de trabalho, por terem me proporcionado as primeiras experiências de estágios e vivência no mercado de trabalho.

A todo o departamento da Engenharia de Produção e Sistemas da UFSC, em especial ao programa PET, que me trouxe uma nova família, à professora Mirna de Borba e aos professores Antonio Cezar Bornia e Ricardo Faria Giglio.

"Tua caminhada ainda não terminou. A realidade te acolhe, dizendo que pela frente o horizonte da vida necessita de tuas palavras e de teu silêncio. Se amanhã sentires saudades, lembra-te da fantasia e sonha com tua próxima vitória. Vitória que todas as armas do mundo jamais conseguirão obter, porque é uma vitória que surge da paz e não do ressentimento.

É certo que irás encontrar situações tempestuosas novamente, mas haverás de ver sempre o lado bom da chuva que cai e não a faceta do raio que destrói.

Se não consegues entender que o céu deve estar dentro de ti, é inútil buscá-lo acima das nuvens e ao lado das estrelas. Por mais que tenhas errado e erres, para ti haverá sempre esperança enquanto te envergonhares de teus erros.

Tu és jovem. Atender a quem te chama é belo, lutar por quem te rejeita é quase chegar a perfeição. A juventude precisa de sonhos e se nutrir de lembranças, assim como o leito dos rios precisa da água que rola e o coração necessita de afeto.

Não faças do amanhã o sinônimo de nunca, nem o ontem te seja o mesmo que nunca mais. Teus passos ficaram. Olhes para trás... mas vá em frente pois há muitos que precisam que chegues para poderem seguir-te."

(Charles Chaplin)

RESUMO

Hodiernamente, tem-se explorado cada vez mais a tecnologia como um dos meios para se obter boa performance nos negócios. Em especial, no contexto das áreas de vendas de empresas de tecnologia da informação, ambiente que tem como principal função fazer o negócio crescer, mas possui a escassez de oferta de mão de obra qualificada, o que passa a ser um dos principais empecilhos para atingir esse objetivo, sem instrumentos tecnológicos. Corroborando com essa dificuldade, modelos de contratações tradicionais acabam por muitas vezes não considerando os quesitos de atração de bons vendedores tanto no quesito de performance, quanto também na sua retenção dentro da empresa. Com base nesse cenário, o presente projeto de pesquisa teve como objetivo propor um novo modelo de contratação para vendedores no contexto especificado, através do uso de aprendizado máquina. Para tanto, foram coletadas 76 variáveis de 103 vendedores do mercado de tecnologia da informação brasileiro ao longo de 24 meses para elaborar dois modelos preditivos: de regressão e de classificação. Seu objetivo consiste na priorização de candidatos durante processos seletivos, através da verificação de *features* que tenham alto grau de importância na preditividade de performance e retenção. Para tanto, as melhores métricas para avaliação de desempenho dos modelos preditivos foram determinadas através de metodologias de validação cruzada e tunagem de hiper-parâmetros. Ambos os modelos foram avaliados segundo as métricas mais pertinentes para cada qual. Para a classificação, utilizou-se MCC e AUC-ROC, enquanto que para regressão, o RMSE. Dentre as treze principais características resultantes do estudo, verificou-se que alunos de universidades públicas e que são contratados com menor tempo de formados, tendem a obter melhor performance nos dois cenários verificados.

Palavras-chave: Aprendizado máquina. Modelos preditivos. Contratação. Vendas. Tecnologia da informação.

ABSTRACT

Nowadays, technology has been more explored as a mean to achieve a good performance in business. In particular, looking deeper into the context of the sales areas from information technology companies, an environment whose main function is to growth and scale the business, but has a shortage of qualified labor supply, which turns out to be one of the main resources to achieve this goal, without technological tools. Corroborating to this difficulty, traditionally terminated hiring models often passes unnoticed when the subject is displayed both good salespeople performance and also focusing on their retention within the company. Based on this scenario, the present research project aimed to propose a new hiring model for salespeople in the specified context using the machine learnig. For this purpose, 76 variables were collected from 103 sales representatives from a brazillian information technological company, over 24 months to develop two predictive models: regression and classification. In order to prioritize selection process candidates by verifying features that have a high degree of importance in predicting performance and retention. To achieve this goal, it has been used the best. Therefore, the best metrics for performance evaluation of predictive models were determined through cross-validation methodologies and hyper-parameter tuning. Both models were evaluated according to the most relevant metrics for each one. For classification, MCC and AUC-ROC were used, while for regression, the RMSE. Among the thirteen main characteristics resulting from the study, it was found that students from public universities who are hired with the shortest time after graduation, obtained better performance in both scenarios verified.

Keywords: Machine learning. Predictable models. Hiring. Sales. Information technology.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

LISTA DE SÍMBOLOS

SUMÁRIO

1 INTRODUÇÃO

1.1 TEMA E CONTEXTUALIZAÇÃO

Cada vez mais as condições de crescimento econômico proporcionam às empresas oportunidades para melhorar seu desempenho. Contudo, durante crises financeiras estas oportunidades ficam mais raras e, muitas vezes, o risco de insucesso do negócio e a estabilidade do quadro de funcionários é comprometida. Por isto, gestores exploram cada vez mais novas estratégias para melhorar o desempenho das suas organizações, a fim de obter maior lucratividade e retenção de funcionários. A Tecnologia da Informação - TI - entra neste contexto para ajudar estas organizações a poderem se adaptar às condições de mercado e ganhar vantagens competitivas (TURBAN et al., 2010). Hodiernamente, muitos são os exemplos de TI vistos em empresas que possibilitaram algum tipo de alavancagem competitiva, seja, ou não, para fins lucrativos:

1. Os jogos Olímpicos de Inverno de 2010, no Canadá, tornou-se a primeira competição olímpica de mídia social. Isto porque, o Twitter e o *Facebook* eram plataformas digitais amplamente utilizadas por profissionais de marketing, atletas e fãs para obter informações sobre os jogos e enviar e receber promoções do evento;

2. Durante períodos de catástrofes ambientais, como os terremotos que ocorrem na década de 2010 no Chile o no Haiti, ferramentas como *Facebook* e *Skype* e diversos blogs foram essenciais para comunicar situações dos desastres, encontrar pessoas desaparecidas e compartilhar pedidos de doações para ajuda humanitária;

3. A *Amazon Web Service* - AWS - utiliza TI para coletar, armazenar, processar, analisar e destinar melhores produtos para seus clientes, através dos seus dados de navegação nas suas plataformas online. Com isto, é possível obter descobertas práticas e de alto valor com base nos seus ativos de dados. Idealmente, os dados são disponibilizados para as partes envolvidas através de inteligência de negócio de autoatendimento e ferramentas ágeis de visualização de dados que permitem a exploração rápida e fácil de conjuntos de dados, proporcionando aumento de vendas e consequentemente de receita para o negócio.

Alinhado com os fatores de inovação tecnológica, o fator humano também tem demonstrado cada vez mais o seu valor nas organizações

por ser o responsável pela criação, utilização, compartilhamento e uso deste conhecimento necessário para o alcance dos objetivos organizacionais (SHUKLA; SRIVASTAVA, 2016). Contudo, as dificuldades de contratação de mão de obra qualificada e estável acabam sendo dois dos principais fatores que comprometem o fator humano nestas organizações. Além das dificuldades, também apresentam diferentes tipos de custos para a empresa, como por exemplo novas contratações e treinamentos internos. Especialmente nas áreas de tecnologia, onde a rotatividade dos funcionários se mostra mais forte e é um problema que afeta a maioria das organizações (PINKOVITZ; MOSKAL; GREEN, 1997). Particularmente, o casamento das novas tecnologias com os funcionários se mostra como um enorme desafio, seja pelo contexto brasileiro de baixa qualificação da mão de obra (KATO; PONCHIROLLI, 2002), ou seja pela baixa abordagem de TI para contratação de funcionários.

Desta forma, alinhar as ferramentas e novos conceitos de TI, com o aumento da assertividade em contratação e consequente retenção de funcionários nas empresas se mostra um grande desafio e grande necessidade para os negócios. Por isto, este será este o tema central do presente trabalho.

1.2 DEFINIÇÃO DO PROBLEMA

Tendo em vista o cenário exposto, nota-se que o fator agilidade, ou seja, capacidade de se adaptar rapidamente, nunca foi tão importante para uma organização como nos dias atuais por conta do constante e rápido avanço das tecnologias (TURBAN et al., 2010). E dentro do contexto empresarial, o principal desafio é entender como as organizações podem se beneficiar desta tecnologia para obter vantagem competitiva perante seus concorrentes.

A falta de assertividade em contratações geram custos diretos e indiretos para a empresa. Os custos diretos resultantes da saída dos funcionários são mais fáceis de mensurar, porque envolvem gastos com o recrutamento, contratação, treinamentos, salários, benefícios e pagamentos de impostos, por exemplo (BROWN, 2000). Já os custos indiretos podem resultar do fato das competências do novo funcionário não serem adequadas à tarefa, o que pode provocar interrupções nos processos organizacionais (THATCHER; LIU; STEPINA, 2002), e na redução na qualidade do trabalho (GHAPANCHI; AURUM, 2011).

Contudo, a assertividade não é o único fator limitante em uma contratação. O fator retenção também se mostra tão importante quanto

o primeiro. Para Schuster (2008), o grande desafio enfrentado não é somente o da inserção no mercado de trabalho, mas também a sua manutenção dentro deste meio. Assim, o desafio proposto está em alinhar as novas tecnologias tanto com a assertividade nas contratações, quanto com a retenção de funcionários nas empresas. Desta forma, elas terão mais capacidade e possibilidade de prosperarem no mercado.

O problema do presente trabalho se insere no contexto da área de vendas de mercados de tecnologia brasileiro. O qual, por natureza, possui dificuldade na retenção dos seus quadros de funcionários (DARMON, 2004). Corroborando com o problema, Hemmans (2010) afirma que, a área de vendas das empresas possuem elevadas taxas de *turnover* e a retenção e contratação destes funcionários se torna um grande problema para o negócio. Tendo esta problemática em vista, a pesquisa aqui apresentada busca entender de que forma atributos pessoais de candidatos à processos seletivos, podem ser utilizados para prever o futuro desempenho dentro das organizações, aumentando a precisão em contratações. Além disso, também procurará mostrar de qual forma estas características podem ser classificadas e priorizadas para entender o comportamento de funcionários que ficam ou saem das empresas.

1.3 OBJETIVO

1.3.1 Objetivo geral

O principal objetivo do trabalho descrito consiste na elaboração de modelos, através de algoritmos de aprendizado máquina, que auxiliem o processo de contratação de funcionários de vendas no mercado de tecnologia brasileiro, permitindo trazer impactos de redução de custos para o negócio e aumento de retenção de funcionários.

1.3.2 Objetivos específicos

Para atingir o objetivo geral, são propostos os respectivos objetivos específicos:

1. Definir os atributos e parâmetros a serem analisados em funcionários de vendas.

2. Aplicar modelos de aprendizado de máquina supervisionados, estruturados em classificação e regressão, em uma empresa de tec-

nologia brasileira.

3. Identificar o desempenho dos modelos propostos, através de métricas de avaliação.

4. Classificar os principais atributos definidos, de modo a priorizá-los nas contratações.

5. Propor uma forma de implementação e acompanhamento do modelo para a empresa analisada, de forma que possa ser replicável em outros negócios.

1.4 JUSTIFICATIVA E IMPORTÂNCIA DA PESQUISA

Para Wagner (2017), a ênfase da produção das organizações no mundo empresarial hoje, passa de artigos baratos e descartáveis para bens e serviços de alta qualidade. Ainda para o autor, o processo exige maior flexibilidade nos seus processos para difenrenciá-los. Drucker (2012), complementa dizendo que as informações aumentariam a importância das empresas que buscam vantagem competitiva e, para ele, os funcionários são chave estratégicas para atingir esta posição. Assim, como o conhecimento especializado se torna mais crítico e, processos eficazes de busca e triagem para funcionários com tal conhecimento, tornam-se essenciais e potencialmente transformadores para muitas empresas (HAMILTON; DAVISON, 2018).

Embora os funcionários tenham grande importância para diferenciação competitiva, um possível indicador negativo é que, desde, pelo menos 2012, diversos executivos se queixaram sobre a incapacidade de encontrar trabalhadores qualificados no mercado (COOMBS, B, 2013; COX, J, 2012; DAVIDSON, A, 2012; MAURER, R, 2017). Hamilton e Davison (2018), ainda complementam dizendo que, como o mercado espera cada vez mais trabalhadores com habilidades gerais, tornar-se-á cada vez mais crucial para as empresas, a adoção de processos diferenciados para contratá-los. Assim, em um contexto de desenvolvimento dos conhecimentos, grandes mudanças nos processos de contratação das empresas, certamente serão vantagem competitiva.

Tendo em vista o que fora exposto, um questionamento pode ser levantado: "Como se pode propor uma metodologia que traga vantagem competitiva através do processo de contratação de uma empresa?"

Inúmeros autores utilizam conceitos e métodos de aprendizado de máquinas para prognosticar e classificar *features* em determinados problemas. Embora a abordagem seja pouco explorada na literatura

para contratações de funcionários, ela será testada através de algoritmos de aprendizado máquina para elencar quais são as principais características para contratação, aumentando o fator de competitividade das empresas. Além da determinação das principais variáveis, a acurácia e assertividade do modelo proposto também serão validadas ao longo da pesquisa.

1.5 LIMITAÇÕES E DELIMITAÇÕES

A primeira delimitação envolve a proposta e saída da pesquisa realizada. Ela se trata de levantamento de atributos com seus respectivos índices de importância para contratações e retenções dentro das organizações. Para este cálculo, o modelo utiliza somente técnicas de regressão e classificação em aprendizado máquina.

Além disto, modelagem proposta é focada em uma empresa de tecnologia brasileira, mais especificamente na área comercial. Embora existam áreas de produção, pós-venda, recursos humanos, financeiras e diversas outras encontradas dentro das organizações, elas não foram contempladas pelo modelo desenvolvido. Tais áreas e, também, outros segmentos do mercado, certamente são oportunidades de aplicações futuras da metodologia e modelo proposto.

Em relação à limitação do trabalho, o principal fator está nas obtenções das variáveis. Embora a definição e tratamento realizado em cada uma delas esteja descrita no capítulo 3, a limitação reside no número de variáveis levantadas e, a falta de dados para todos os componentes da amostra. Além disso, vale destacar que as saídas do problema de classificação, são limitadas apenas à funcionários ativos ou inativos, sem especificação do motivo pelo qual o funcionário saiu da empresa.

1.6 CARACTERIZAÇÃO METODOLÓGICA

Em termos do ponto de vista da natureza da pesquisa, ela pode ser classificada como aplicada. Segundo Moresi et al. (2003), esta classificação de natureza objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos. Com relação ao ponto de vista de como o problema é abordado ao longo da pesquisa, tem-se uma pesquisa quantitativa. Ou seja, para Morabito et al. (2018), um modelo é uma representação de uma situação ou realidade, sob o ponto

de vista de uma, ou mais, pessoas, construída para auxiliar o tratamento de uma situação de forma sistemática. Ainda para Morabito et al. (2018), o modelo visa identificar problemas, formular estratégias e oportunidades, bem como compreender melhor o ambiente em questão, fortalecendo a escolha pela pesquisa quantitativa.

Já do ponto de vista dos fins (ou objetivos) a pesquisa se qualifica como descritiva. Moresi et al. (2003), classifica a pesquisa descritiva como:

> "Expõe características de determinada população ou de determinado fenômeno. Pode também estabelecer correlações entre variáveis e definir sua natureza. Não tem compromisso de explicar os fenômenos que descreve, embora sirva de base para tal explicação. Pesquisa de opinião insere-se nessa classificação."

Finalmente, do ponto de vista dos processos técnicos, a pesquisa se enquadra na classificação de exploratória. Isto se dá pelo fato de que tem-se como propósito proporcionar maior familiaridade com fatores de sucesso e de fracasso de colaboradores individuais, tornando os condicionantes mais explícitos através da construção de hipóteses (GIL, 2002).

A construção da pesquisa se dará por meio do método que se encontra na figura 1, sob o seguinte formato:

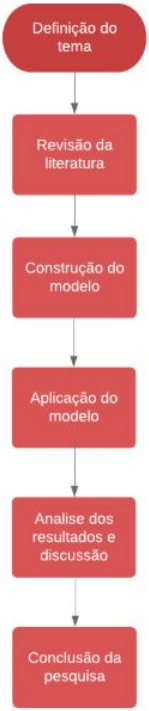
1.7 ESTRUTURA DO TRABALHO

O projeto de pesquisa é organizado em cinco capítulos, conforme indica a figura 1. Em que, no primeiro deles, são apresentados o tema e sua contextualização, a problemática a cerca do trabalho, os objetivos gerais e específicos, a justificativa e importância da pesquisa, as limitações e delimitações de escopo do trabalho, a caracterização metodológica e por fim sua estrutura.

No conseguinte capítulo, são apresentadas as fundamentações teóricas em cinco linhas: considerações iniciais, *turnover* de funcionários, TI, inovação e automação de operações de recursos nos negócios, classificação e regressão em modelos de aprendizado máquina (aqui esmiuçados em aprendizado máquina, árvores de decisão, florestas aleatórias e tunagem de hiper-parâmetros) e, finalmente, avaliação dos modelos de classificação e regressão.

O terceiro capítulo é formado por todos os procedimentos da

Figura 1: Estrutura metodológica



Fonte: autor

metodologia científica utilizada ao longo da pesquisa. Ela contém a descrição das variáveis do modelo e todos os procedimentos utilizados.

O quarto capítulo é orientado para a explicitação dos resultados da aplicação do modelo estimado e, por fim, o quinto capítulo conta com as conclusões finais da pesquisa, recomendações de trabalhos futuros e breve sugestão de implementação inicial nas empresas.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 CONSIDERAÇÕES INICIAIS

Este capítulo contempla toda a base teórica da pesquisa. Primeiramente serão abordados conceitos de *turnover* de funcionários, seguido pelas gestões de operações e recursos atualmente dentro do cenário de tecnologia da informação. Passando para o terceiro, que desenvolve conceitos sobre aprendizado máquina, classificação e regressão e tunagem de hiper-parâmetros as quais serão abordadas em quatro subseções distintas, pelas quais se dão o nome. Até que, finalmente, seja explorada a base teórica para métodos de avaliação de cada um dos modelos apresentados.

2.2 *TURNOVER* DE FUNCIONÁRIOS

O termo *turnover*, segundo Waldman e Arora (2004), representa a percentagem da alteração no quadro de funcionários de uma organização em um período de tempo, isto é, a movimentação (admissões e desligamentos) de contribuidores individuais. Além disso, para Ribeiro (2010), ele também pode ser analisado em uma escala macroeconômica, levando-se em consideração o volume de admissões e desligamentos no mercado de trabalho como um todo.

Para que as empresas tenham eficiência no seus objetivos de gerar receita, através de vendas, ela necessita crescer internamente de forma saudável e estruturada. Por isso, constantemente faz-se necessário, realizar pesquisas para acompanhar a taxa de turnover de seus profissionais e identificar as razões pelas quais eles deixaram (de maneira voluntária ou involuntária) a organização.

Já existem estimativas de que as perdas de executivos pode custar milhões de dólares por ano às organizações. Corroborando com este ponto, uma pesquisa com 5 mil executivos indicou que 46% deles têm a intenção de permanecer em seus cargos por apenas dois a cinco anos (PARISE; CROSS; DAVENPORT, 2006). Com o foco no setor de vendas em mercados de tecnologia, historicamente, esta profissão apresenta elevados índices de *turnover*. A retenção desses profissionais é um dos problemas mais latentes enfrentados pelas organizações de base tecnológica (HEMMANS, 2010). Diversos são os problemas que podem ser gerados por esta baixa adesão de funcionários, além dos elevados cus-

tos, também pode fazer com que os coordenadores não tenham o tempo suficiente para prospectar, atrair, selecionar e treinar um novo funcionário de modo que ele esteja apto para assumir a função de vendedor, pois, o alto nível de complexidade dos produtos demandam alta qualificação do mercado (SAMUELS, 2005).

De acordo com Chiavenato (1983), os custos das saídas são categorizados de três maneiras distintas: custos primários, que possuem uma ligação direta com a saída do funcionário, com recrutamento e seleção, integração e desligamento deste funcionário, além de, de acordo com Brown (2000), salários, benefícios e impostos; custos secundários, que se referem aos efeitos imediatos causados pela rotatividade de pessoal, são aqueles sentidos em curto prazo; e, os custos terciários que são estimáveis sentidos a médios e longo prazo. Com isto, evidencia-se que a retenção de funcionários é de extrema importância para as organizações (HAUSKNECHT; RODDA; HOWARD, 2009; PARISE; CROSS; DAVENPORT, 2006).

Analisando algumas causas raízes de demissões de vendas dentro do mercado de tecnologia, Cuffa, Floriani e Steil (2018) definem quatro fatores mais críticos que levam às saídas: fatores pessoais, ocupacionais, organizacionais e ambientais. A primeira categoria engloba fatores familiares e mudanças de carreira do funcionário. Já a segunda envolve o baixo desempenho, desalinhamento de cargos e o baixo relacionamento com os líderes. O fator organizacional, envolve baixa adesão de fatores culturais e climáticos da organização e, por fim, os ambientais envolvem oportunidades de carreira no exterior.

2.3 TI, INOVAÇÃO E AUTOMAÇÃO DE OPERAÇÕES E RECURSOS

Em mercados globalizados, o termo competitividade é amplamente explorado pelo senso comum nos dias de hoje. Uma maneira das empresas atingirem níveis competitivos é trabalhando com a redução dos seus custos e/ou aumento de sua qualidade. E para tal, técnicas de automação e modernização dos processos e recursos se tornam pontos chaves nas discussões (ACHARYA; SHARMA; GUPTA, 2018). Complementando a discussão, Rachwał (2011), diz que todas as empresas buscam adotar e implementar um conjunto de técnicas automatizadas para identificar e endereçar mudanças, que levem à melhoria contínua nos ciclos do negócio.

O campo da literatura sugere que a implementação deste tipo

de automação dos processos possui efeitos além das atividades e performances operacionais do negócio, e também grande efeito nos trabalhadores. A literatura aborda dois campos distintos das automações: o primeiro com relação a engenharia industrial e gerenciamento de operações, que olha para o negócio (VOSS, 1988), e o segundo, com relação ao estudo da sociologia industrial, voltada para o impacto nos trabalhadores (HUDSON, 1982). Entrando no campo da inovação, Alexe e Alexe (2018) diz que ela tem a capacidade de melhorar a performance das organizações. Por este motivo, as empresas incluíram este quesito na sua cultura interna. Fator que, segundo Rosenbusch, Brinckmann e Bausch (2011), é especialmente mais comum em empresas de tecnologia.

Entretanto, para que a inovação dos processos seja continuamente ligada ao sucesso do negócio, Rosenbusch, Brinckmann e Bausch (2011), questionam se existem empresas e ambientes mais propícios para o sucesso e alta performance da inovação. Segundo eles, a inovação possui efeito positivo nas pequenas e médias empresas (PME), contudo, com alguns fatores que possuem maior impacto neste desempenho. A inovação tem um impacto mais forte nas empresas mais jovens do que nas PME mais estabelecidas. Isto sugere que a responsabilidade frequentemente citada de novidade das empresas mais jovens também pode ser um trunfo para elas, ou seja, as novas empresas possuem capacidades únicas para criar valor apropriado através de inovações.

Através da revisão da literatura nota-se que a automação e inovação dos processos é um fator interdisciplinar por sua própria natureza. A falta de abordagem sobre o assunto no campo de contratações de funcionários, alinhado com a importância do mesmo, reforça a necessidade da compreensão de aplicações matemáticas deste tipo de tópico. Os efeitos desta modelagem, tanto nos negócios, quanto nos trabalhadores, levanta o questionamento de qual seria o efeito desta aplicação para contratação e seleção de funcionários para as empresas. Trazendo as considerações feitas por Pistono (2017), as pessoas cada vez mais estão prestes a se tornar obsoletas em um mundo automatizado. Sendo assim, por que não utilizar esta tecnologia, que tanto transforma o sitema sócioeconômico, para trazer cada vez mais pessoas adequadas aos *fits* das empresas e possibilitar continuar contratando pessoas e não somente máquinas?

2.4 CLASSIFICAÇÃO E REGRESSÃO EM MODELOS DE APRENDIZADO MÁQUINA

2.4.1 Aprendizado máquina

Durante as últimas décadas a humanidade obteve a possibilidade de levantar dados nas mais diversas áreas do campo científico. De forma geral, foi devido ao avanço da tecnologia que esta possibilidade, que outrora fora apenas um sonho do ser humano, se tornasse realidade. Hoje em dia é possível manipular simultaneamente centenas de milhares de dados, desde bancos de dados de compras de uma empresa, até mesmo os genes do corpo humano.

Como consequência, diversos bancos de dados foram cada vez mais acessíveis para a população, principalmente através da internet. Contudo, ter os dados não significa ter resultados, em outras palavras, eles precisam ser convertidos em informações para poderem ser interpretados e trazerem resultados concretos e avanços científicos. Neste contexto de desenvolvimento tecnológico que surge o Aprendizado Máquina - AM. Os modelos de AM possibilitaram estudar as interações entre dados e variáveis de forma rápida e organizada. Para isto, alguns algoritmos, ou teoremas, foram desenvolvidos. Como exemplos práticos do uso do AM, especificamente dentro do universo da Engenharia de Produção, podem ser citados, segundo a Data Science Academy (2018):

1. Previsão de Demanda: Indústrias obtêm a quantidade certa de produto na localização certa sendo fundamental para o sucesso do negócio. Os sistemas de aprendizagem de máquina podem usar dados históricos para prever as vendas de forma muito mais precisa e rápida do que os seres humanos podem por conta própria.

2. Logística: Para empresas de transporte, configurar horários e rotas é uma tarefa complexa e demorada. Os sistemas de aprendizagem de máquina podem ajudar a identificar a maneira mais eficiente e econômica de transportar bens ou pessoas do ponto A ao ponto B.

Segundo, Peres et al. (2012), a teoria de AM se baseia em modelos determinados a partir de um conjunto de dados ou representações de experiências. Isto quer dizer que ela se baseia nos fundamentos do aprendizado indutivo. Eles são implementados através de algoritmos que processam um conjunto de dados e extraem um modelo capaz de explicar ou representar os dados sob determinados aspectos. O modelo

construído pode ser utilizado para explicar ou representar um novo dado. O aprendizado indutivo é efetuado a partir de raciocínio sobre exemplos fornecidos por um processo externo ao sistema de aprendizado e podem ser melhorados através de novos conjuntos de informações.

Segundo, Monard e Baranauskas (2003), o aprendizado indutivo é a forma, que, através da lógica e inferência, permite-se obter conclusões genéricas sobre um conjunto de exemplos. Este raciocínio se caracteriza através de algo específico, que se generaliza. Em outras palavras, vai da parte para o todo, segundo o autor. Assim, as hipóteses geradas através do processo de inferência indutiva podem não representar fielmente a verdade, contudo, este é considerado um dos principais métodos utilizados para derivação do conhecimento novo e prognosticar futuros. Como forma de corroborar com a importância do aprendizado indutivo, Batista et al. (2003) diz que:

> "A inferência indutiva é um dos principais meios de criar novos conhecimentos e prever eventos futuros. O processo de indução é indispensável na obtenção de novos conhecimentos pelo ser humano. Foi por meio de induções que Kepler descobriu as leis do movimento planetário, que Mendel descobriu as leis da genética e que Arquimedes descobriu o princípio da alavanca. Pode-se ousar em afirmar que a indução é o recurso mais utilizado pelos seres humanos para obter novos conhecimentos. Apesar disso, este recurso deve ser utilizado com os devidos cuidados, pois se o número de observações for insuficiente ou se os dados relevantes forem mal escolhidos, as hipóteses induzidas podem ser de pouco valor."

Em suma, o AM é uma área da Inteligência Artificial - IA - cujo objetivo é desenvolver técnicas computacionais sobre determinado aprendizado e a construção de sistemas capazes de adquirir conhecimento de forma automática. Ele pode tomar decisões baseado nas experiências que se acumularam através da solução bem sucedida de problemas previamente resolvidos (MONARD; BARANAUSKAS, 2003, p. 69).

Segundo Monard e Baranauskas (2003), o aprendizado indutivo pode ser esmiuçado em dois: o supervisionado e o não supervisionado. No primeiro, alimenta-se o algoritmo com um conjunto de exemplos de treinamentos em que as classes são conhecidas. O objetivo deste modelo de indução é obter um classificador que determine de forma coerente a classe de amostras não rotuladas. Sua diferenciação se dá

em duas categorias: rótulos discretos (classificação) e rótulos contínuos (regressão) (MONARD; BARANAUSKAS, 2003). Na figura 2 é mostrada a hierarquia de aprendizado descrita, na qual os nós sombreados levam ao aprendizado supervisionado utilizando classificação.

Em contrapartida, o modelo não supervisionado analisa exemplos e busca determinar se eles podem ser agrupados de alguma maneira (CHEESEMAN et al., 1990). Estes agrupamentos necessitam de análise para entender o significado no contexto do determinado problema.

A importância dos dados de análise também são reforçadas por Monard e Baranauskas (2003), ao citar que caso os dados não possuam confiabilidade, as hipóteses obtidas podem ter pouco valor na sua generalização.

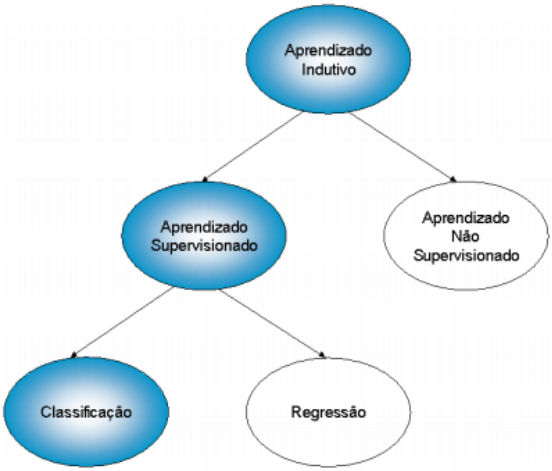
Neste projeto de pesquisa, ambos os métodos de aprendizado máquina indutivo supervisionados citados foram utilizados para o problema. O modelo de regressão procurou predizer quais características de funcionários teriam o maior impacto na sua performance durante o trabalho. Enquanto que o modelo de classificação trabalha com as mesmas características, porém, desta vez, para classificar funcionários que continuam ou não na empresa. Desta forma, o primeiro trabalha com fatores que influenciam no momento de uma contratação, enquanto o segundo, de retenção do funcionário. Ainda sobre os métodos, destacam-se alguns algoritmos que procuram resolver problemas dessa natureza, como por exemplo, *Neural Networks*, *Baysian Networks*, Árvore de decisão e *Random Forest*. Para a pesquisa, foram utilizados algoritmos de Árvore de decisão e *Random Forest* - Florestas aleatórias.

2.4.1.1 Aprendizado máquina supervisionado

Após a breve descrição de problemas de aprendizado supervisionados na seção anterior, o objetivo desta subseção é trazer uma análise aprofundada do assunto. Para tanto, algumas definições e conceitos imprescindíveis são apresentados para o entendimento da pesquisa e dos modelos de AM.

Como previamente destacado, o AM supervisionado tem a função de descobrir padrões, comportamentos e relacionamentos de informações, através do treinamento do modelo, a partir de um conjunto de dados. No geral, este treinamento se dá pelo fornecimento das respostas das variáveis conhecidas ao modelo de AM, para que ele possa aprender e trabalhar com os dados futuros (ainda sem respostas).

Figura 2: A hierarquia do aprendizado máquina



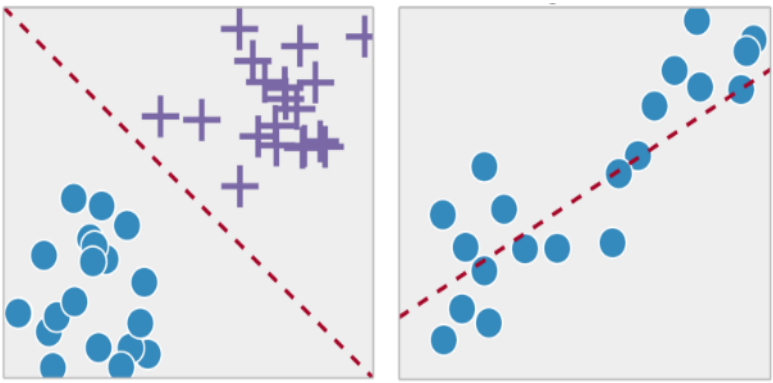
Fonte: (MONARD; BARANAUSKAS, 2003, p. 41)

Em suma, Barros, P (2016) aconselha o uso do aprendizado máquina supervisionado quando, a partir de um conjunto de variáveis de entrada (x), e desejando um conjunto de variáveis de saída (Y), é possível mapear funções que tornam possível a transformação da função $Y = f(x)$. A medida em que as funções de mapeamento são aperfeiçoadas, torna-se possível, a partir de uma nova entrada (x), prever seu correspondente (Y).

Os tipos de problemas de aprendizado de máquinas supervisionados são classificados como regressão e classificação, conforme a figura 2. A principal diferença entre eles é que, enquanto na regressão tenta-se prever resultados de saídas contínuas, ou seja, tenta-se mapear variáveis de entrada para alguma função contínua, no problema de classificação, o objetivo é prever resultados de saídas discretas. Isto é, a ideia é mapear variáveis de entradas em determinadas categorias distintas (BARROS, P, 2016).

Na figura 3 esta diferenciação fica mais clara. Ali pode-se prever idade de grupos (homem/mulher, por exemplo) a partir dos dados de exemplo da figura (b) e, prever se algum determinado tumor é benigno ou maligno pelo tamanho e idade do paciente, em um exemplo da figura (a).

Figura 3: Comparativo entre classificação e regressão



(a) Exemplo de Classificação (b) Exemplo de Regressão

Outro exemplo de classificação muito utilizado são os aceites, ou não, de empréstimos bancários para clientes. As instituições financeiras utilizam fatores de histórico de crédito pelos solicitantes do empréstimo, para embasar a decisão do aporte financeiro.

2.4.1.2 Aprendizado máquina não supervisionado

A Aprendizagem não supervisionada, por outro lado, não pode ser aplicada de forma direta em problemas de classificação e regressão. Isto se dá pelo fato do modelador abordar problemas com pouca ou nenhuma ideia do que os resultados devem aparentar. Podem derivar estruturas de dados onde não necessariamente se sabe o efeito das variáveis, impossibilitando a lógica de treinamento do algoritmo (MONARD; BARANAUSKAS, 2003). Por este motivo, o aprendizado máquina não supervisionado não será abordado nesta pesquisa.

2.4.1.3 Conceitos do aprendizado máquina supervisionado

Esta seção tem como objetivo descrever sucintamente algumas definições de termos amplamente usados na literatura de AM, de acordo com Monard e Baranauskas (2003).

1. **Indutor**: O objetivo de um indutor (ou programa de aprendizado ou algoritmo de indução) consiste em extrair um bom classificador a partir de um conjunto de exemplos rotulados. A saída do indutor, o classificador, pode então ser usada para classificar exemplos novos (ainda não rotulados) com a meta de predizer corretamente o rótulo de cada um. Após isso, o classificador pode ser avaliado considerando sua precisão, compreensibilidade, grau de interesse, velocidade de aprendizado, requisitos de armazenamento, grau de compactação ou qualquer outra propriedade desejável que determine quão bom e apropriado ele é para a tarefa em questão.

2. **Exemplo**: Um exemplo, também denominado de caso, registro ou de dado na literatura, é uma lista ordenada de valores de atributos (ou um vetor de valores de atributos). Um exemplo descreve o objeto de interesse, como por exemplo um paciente, dados médicos sobre uma determinada doença ou histórico de clientes de uma dada companhia.

3. **Atributo**: O atributo é utilizado para descrever certas características de um exemplo. Eles podem ser classificados como nominal, ou seja, quando não há uma ordem de valores (como as cores, por exemplo) e como contínuos, quando existe uma ordem (como nos números, por exemplo).

   Quando não é possível descrever algum atributo do exemplo, pelo fato dele ser desconhecido (ou não aplicável), utiliza-se os caracteres '!', ou '?'. Como exemplo para ilustrar seu uso, pode-se pensar no número de gestações para pessoas do sexo masculino. Nestes casos, os símbolos de não aplicável, ou desconhecido, seriam utilizados.

   Eles também são importantes no quesito de capacidade preditiva. No geral, isto quer dizer que no caso do objetivo de prever se uma rede de lojas vai ter sucesso, fatores como cor do cabelo, cor do olho, tamanho dos funcionários seriam fatores de baixo poder preditivo. Em contra partida, a localização da loja, o tamanho da cidade, o poder aquisitivo da região, seriam fatores de alto poder preditivo.

4. **Classe**: No aprendizado supervisionado, todo exemplo possui um atributo especial, chamados de classe. Ele descreve o fenômeno de interesse, ou seja, a meta que se deseja aprender e poder fazer as previsões a respeito. Como já mencionado, os rótulos são tipicamente pertencentes a um conjunto discreto (nominal) de classes

$C = \{C1, C2, ..., Ck\}$, no caso de classificação, ou de valores reais no caso de regressão.

5. **Ruído**: O processo de aquisição dos dados, de sua transformação, ou de classificação incorreta de classes pode gerar dados não perfeitos. Isto é muito comum nas análises cotidianas. Casos de dados com mesmos valores de atributos, mas classes diferentes são exemplos de dados imperfeitos, ou, com ruídos.

6. ***Bias***: Este é o nome dado para qualquer preferência de uma hipótese sobre outra, além da simples consistência com os exemplos. Devido ao fato que quase sempre existe um número grande de hipóteses consistentes, todos os indutores possuem alguma forma desta preferência. Na verdade, aprendizado máquina sem ele é impossível (MITCHELL, 1982).

7. **Modo de aprendizagem**: Sempre que todo o conjunto de treinamento esteja presente para o aprendizado, o modo de aprendizado de um algoritmo é não-incremental. Contudo, se o indutor não necessita construir a hipótese a partir do início, quando novos exemplos são adicionados ao conjunto de treinamento, ele é incremental. Portanto, no modo incremental o indutor apenas tenta atualizar a hipótese antiga sempre que novos exemplos são adicionados ao conjunto de treinamento.

8. **Erro e precisão**: as medidas de erro e precisão são diferentes para os casos de classificação e de regressão. Para a primeira, a taxa de erro de um classificador (h), ou, taxa de classificação incorreta (err(h)) são medidas de desempenho comumente utilizadas. Usualmente, a taxa de erro é obtida utilizando a equação 2.1, a qual compara a classe verdadeira de cada exemplo com o rótulo atribuído pelo classificador induzido. O operador retorna 1 se a expressão for verdadeira e zero caso contrário, e n é o número de exemplos. O complemento da taxa de erro, a precisão do classificador, denotada por acc(h) é dada pela equação 2.2.

$$err(h) = \frac{1}{n} \sum_{i=1}^{n} \|y_i \neq h(x_i)\| \tag{2.1}$$

$$acc(h) = 1 - err(h) \tag{2.2}$$

Para o segundo caso, nas medidas de regressão, o erro da hipótese (err) pode ser estimado calculando-se a distância entre o valor real com o atribuído pela hipótese induzida. Usualmente, duas medidas são comumente usadas: o erro médio quadrado (MSE) e a distância absoluta média (MAD), dadas pelas equações 2.3 e 2.4, respectivamente.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2 \tag{2.3}$$

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|y_i - h(x_i)| \tag{2.4}$$

9. **Distribuição de classes**: Para conjuntos de exemplos é possível calcular a distribuição de classes para casos de classificação. Para cada classe C_j sua distribuição $distr(C_j)$ será determinada como o percentual do total de exemplos n. Esta proporção é dada pela equação 2.5. As classes são classificadas em majoritárias, ou prevalentes, quando possuem maior percentual, ou, minoritária, quando possui a menor proporção.

$$distr(C_j) = \frac{1}{n}\sum_{i=1}^{n}\|y_i = C_j\| \tag{2.5}$$

A fim de ilustrar o cenário, adota-se um conjunto com 100 exemplos. Dentre eles 70 são de classe C_1, 20 de classe C_2 e 10 de classe C_3. Neste caso específico, tem-se $distr(C_1, C_2, C_3) = (0,60; 0,20; 0,10)$. No caso, a classe C_1 é a majoritária, ou prevalente, enquanto a C_3 é a minoritária.

10. **Erro majoritário**: O erro majoritário de um conjunto de exemplos independe do algoritmo de aprendizado. Ele irá fornecer um limite máximo abaixo do qual o erro de um classificador deve permanecer. Ele pode ser calculado pela equação 2.6.

$$maj - err(T) = 1 - \max_{i=1,\ldots,k} distr(C_j) \tag{2.6}$$

Para o exemplo anteriormente ilustrado na distribuição de classes, o erro majoritário é $maj - err(T) = 1 - 0,70 = 30\%$.

11. **Prevalência de classe**: O desbalanceamento de classes é um ponto crucial no AM. Por exemplo, supondo um conjunto de exemplos T, com a distribuição $distr(C_1, C_2, C_3) = (99\%; 0,25\%; 0,75\%)$, sendo majoritária a classe C_1. Novos exemplos pertencentes à classe C_1 teria precisão de 99% ($maj - err(T) = 1\%$). Um ponto de atenção, destacado por Monard e Baranauskas (2003), se dá quando as classes minoritárias possuem informações cruciais para o problema, como por exemplo C_1: paciente normal, C_2: paciente com doença x e C_3: paciente com doença y.

12. ***Under e Overfitting***: O *trade-off* entre a otimização e a eficiência de alogritmos é o ponto central deste tópico. Muitas técnicas de reconhecimento de padrões não foram originalmente projetadas para lidar com grandes quantidades de recursos irrelevantes, combiná-las com outras técnicas se tornou uma necessidade em diversas aplicações (GUYON; ELISSEEFF, 2003; LIU; YU, 2005). Monard e Baranauskas (2003), ressaltam que é possível induzir hipóteses que melhorem o desempenho de um modelo preditivo no conjunto de treinamento, contudo, pioram o desempenho em exemplos diferentes daqueles pertencentes ao conjunto de treinamento. Neste caso, as medidas de controle (erro, por exemplo) em um conjunto de teste independente evidencia um desempenho ruim da hipótese. Isto faz com que ela se ajuste em excesso ao conjunto de treinamento gerando o temido *overfitting*. No geral, o resultado que se observa no *overfitting* é uma boa performance do modelo nos dados de treinamento, contudo, não nos dados de avaliação. Isto porque o modelo memoriza os dados e se mostra incapaz de generalizar para quaisquer exemplos.

    Em contrapartida, é possível que poucos exemplos representativos sejam inseridos no sistema de treinamento, ou, o tamanho do classificador seja definido como muito pequeno, ou ainda, uma combinação dos dois casos. Neste cenário, diz-se que a hipótese ajusta-se muito pouco ao conjunto de treinamento, ocasionando o *underfitting*. Em outras palavras o modelo performa mal nos dados de treinamento, pois ele não é capaz de capturar o relacionamento entre os dados de entrada (normalmente chamados de x) e os dados alvo (geralmente chamados de y). A figura 4 ilustra graficamente algumas diferenças entre os modelos de *overfitting* e *underfitting*

13. ***Overtunning***: Ele se dá pelo excesso de ajuste de um algoritmo. Podendo ocorrer quando se ajusta o algoritmo de aprendizado, ou

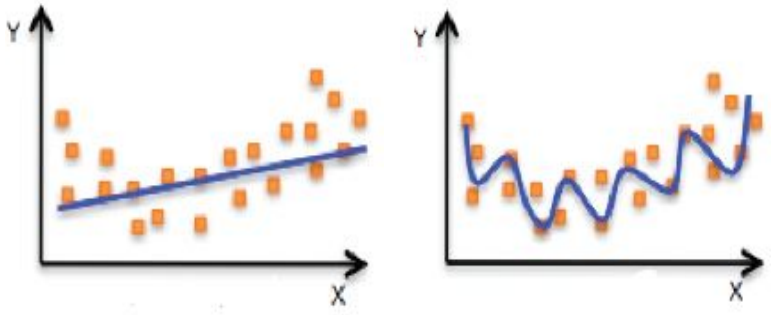
parâmetros, muito bem para otimizar seu desempenho em todos os exemplos. Caso os exemplos não sejam separados em conjuntos de treinamento e teste, permitindo uma avaliação final independente, o desempenho do sistema não pode ser utilizado de forma confiável como uma estimativa do desempenho preditivo do sistema.

14. **Poda**: Segundo Helmbold e Schapire (1997), o processo de poda simplifica uma árvore de decisão através da realização de sucessivos cortes de nós que representam baixa relevância preditiva para o modelo. Esta é uma técnica para lidar com o ruído e o *overfitting*. Sua essência consiste em lidar com o problema de *overfitting* através do aprendizado de uma hipótese genérica a partir do conjunto de treinamento de forma a melhorar o desempenho em exemplos não vistos (MONARD; BARANAUSKAS, 2003). Existem dois métodos para trabalhar com esta técnica: a pré-poda e a pós-poda. O primeiro caso é durante a geração da hipótese, em que alguns exemplos do treinamento são deliberadamente ignorados, de forma que, no final, não classifique todos os exemplos de treinamento de forma correta. Já o segundo, se trabalha inicialmente com uma hipótese de treinamento que explique os exemplos. Após isso, ela é generalizada através da eliminação de algumas partes, como o corte de alguns ramos da árvore de decisão, ou seja, é o processo de remover sub-nós de um nó de decisão.

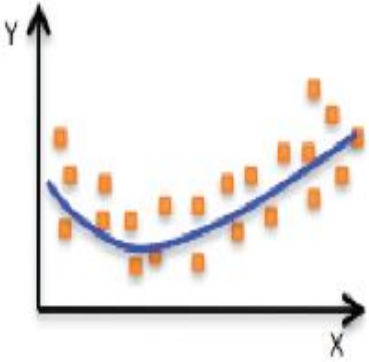
    Assim, ao mesmo tempo que diminui o tamanho da árvore, reduz sua complexidade e consequentemente o risco de *overfitting*. A importância deste processo ocorre por ser impossível determinar o momento certo de interromper o crescimento, de modo que não se consegue calcular o impacto da adição de um nó extra no aumento da capacidade preditiva da árvore sem antes tê-lo adicionado (GOMES, 2011).

15. **Matriz de confusão**: Esta matriz oferece uma medida efetiva do modelo de classificação, ao mostrar um comparativo entre as classificações preditas e as corretas para cada classe de um conjunto de exemplos. Ela será mais bem detalhada na seção 2.5 Avaliação do modelo de classificação e de regressão.

Figura 4: Comparativo entre modelos em *underfitting*, *overfitting* e balanceados



(a) *Underfitting* (b) *Overfitting*



(c) Balanceado

2.4.2 Árvores de decisão

Como previamente destacado, as árvores de decisão são modelos de aprendizado máquina supervisionados e, segundo Gomes (2011), podem ser utilizadas em dois grupos distintos: o de classificação e o de regressão. Witten et al. (2016), descreve as árvores como um modelo representado graficamente por nós e ramos. Monard e Baranauskas (2003) ainda definem algumas terminologias básicas utilizadas nas ár-

vores de decisão:

- **Nó raiz**: esta representa toda a população, ou a amostra em questão, a qual pode ser dividida em outros conjuntos homogêneos;
- **Divisão**: processo de divisão de um nó;
- **Nó de decisão**: quando o nó é dividido em sub-nós, os quais ainda podem ser sofrer divisão;
- **Nó Folha**: nós que não são mais divididos, ou seja, não possuem filhos;
- **Poda**: é o processo de remoção de sub-nós;
- **Ramificação**: é a sub-classificação de uma árvore a partir de um nó;
- **Nó pai**: nó que sofre uma divisão. Ele será "pai"em relação a sua divisão direta;
- **Nó filho**: é a nomenclatura dada à divisão direta de um determinado nó.

A árvore de decisão parte do macro (nó raiz) para o micro (folhas). Todas as terminologias podem ser verificadas na figura 5, a qual representa uma pequena árvore de decisão genérica conceitual.

Em uma perspectiva mais específica das árvores de decisão, seus galhos são uma pergunta de classificação e suas folhas são fatias dos conjuntos de dados, com as devidas classificações. No exemplo da figura 6, podem ser classificados clientes que não renovam os contratos telefônicos na indústria de telefones celulares.

Por uma perspectiva de negócios as árvores de decisões podem ser visualizadas como um novo segmento de conjuntos de dados originais, no qual cada segmento seria uma das folhas da árvore (GUIMARÃES et al., 2000). Por exemplo, segmentação de clientes, produtos e regiões de vendas são algo que os gerentes de marketing tem feito por muitos anos. Em 1999 essa segmentação foi feita para conseguir uma visualização de alto nível de um grande montante de dados - sem nenhuma razão particular para criar a segmentação exceto em que os registros dentro de cada segmentação fossem similares a qualquer outro (GUIMARÃES et al., 2000).

Ainda para Guimarães et al. (2000), a segmentação é dada para uma razão de predição de umas poucas informações importantes. Os

Figura 5: Árvore de decisão

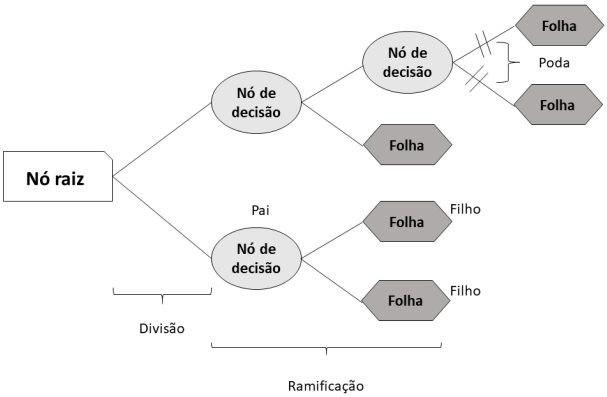
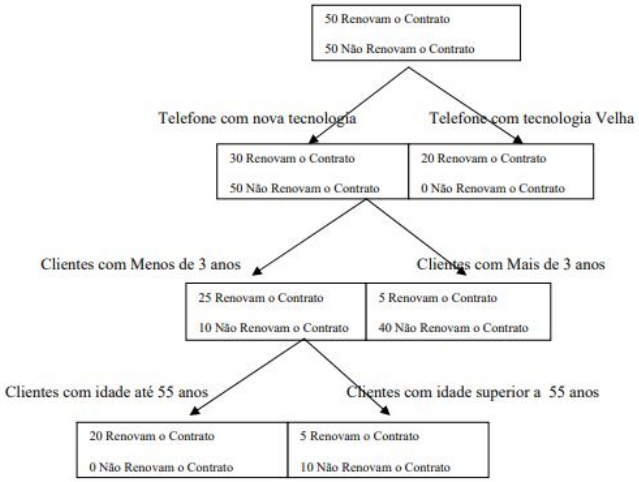


Figura 6: Árvore de decisão exemplo.



Fonte: (GUIMARÃES et al., 2000)

registros que recaem dentro de cada segmento ficam ali por causa de terem similaridade com respeito às informações que estão sendo preditas e não somente porque elas são similares - sem que a similaridade seja bem definida. Estes segmentos preditivos, que são derivados da

árvore de decisão, também veem com uma descrição das características que definem o segmento preditivo; sendo assim, as árvores de decisões e os algoritmos que as criam podem ser complexos, mas os resultados podem ser apresentados de uma forma fácil de serem entendidos e pode ser muito útil para o usuário comercial.

Existem diversos algoritmos para as árvores de decisão: CART, CHAID, QUEST, ID3, C4.5 e outros. A ideia por trás das árvores é de que, através de dados históricos, elas possam ser moldadas.

2.4.2.1 CART

O primeiro deles, CART, se trata de um modelo completo e utilizável em uma grande variedade de problemas diferentes. Ele foi desenvolvido pelos pesquisadores da Universidade de *Stanford*, na Califórnia, Leo Breiman, Jerome Friedman, Richard Olshen e Charles Stone em 1984.

Segundo a metodologia proposta por Breiman et al. (1984), o primeiro passo é a seleção dos preditores baseado em quão bem ele isola os registros com predições diferentes. Contudo, pelo fato do algoritmo se preocupar somente com as divisões atuais, as árvores acabam gerando os ruídos e ficam conhecidas como árvores "gulosas". Segundo Murthy e Salzberg (1995), quanto maior o tamanho ideal da árvore, menor será sua precisão.

O conceito de avaliação do processo de divisão se baseia em selecionar um atributo que irá produzir a melhor árvore. Assim, a avaliação precisa acontecer de uma forma heurística que realize boas decisões (não necessariamente ótimas) com as informações não completas. Ela terá a função de reduzir o grau de incerteza nos nós das árvores.

A medida de avaliação mais popular é a entropia (SHANNON; WEAVER, 1949). A equação 2.7 mostra o resultado derivado do trabalho de Claude Shannnon e Warren Weaver, em 1949.

$$-\sum_{j} p_j \log p_j \tag{2.7}$$

Na equação 2.7 e 2.8 , p_j é a probabilidade (p) do valor da predição ocorrer em um nó particular da árvore para a classe j. No exemplo da Figura 6, por exemplo, imagina-se um nó na árvore em que fosse tentado predizer quem não renovaria e quem renovaria em uma lista de clientes da companhia telefônica celular. Além disso, supõem-se que há 100 clientes no nó específico e, destes, 30 desistiram e o

complemento não. Sabe-se que as probabilidades de desistência p= 30% e não desistência p= 70%. Com isto, a equação de entropia seria: $-0.3\log(0.3) - 0.7\log(0.7) = (-0.3 * -1.74) + (-0.7 * -0.514) = 0.412$

Quanto menor for a medida de entropia, melhor a predição do nó. Ou seja, quanto mais próximo de 0 melhor e quanto mais próximo de 1 pior.

Outra medida de avaliação popular e muito utilizada é a função *gini* (WEISS; KULIKOWSKI, 1991). A equação 2.8 apresenta a função *gini*. No caso do *gini*, a pureza do nó ocorre quando o índice é igual a zero, e o nó torna-se impuro quando ele se aproxima do valor um (ou seja, quando há um aumento no número de classes uniformemente distribuídas no nó).

$$1 - \sum_{j} p_j^2 \tag{2.8}$$

O índice *gini* também pode ser calculado através das somas dos quadrados das probabilidades de sucesso e insucesso para cada sub-nó. A ideia nestes casos é verificar qual das divisões da árvore produzem nós mais homogêneos (sendo maior o valor, melhor para a divisão).

De volta para a metodologia proposta por Breiman et al. (1984), a segunda etapa consiste no processo de parada do crescimento da árvore. Este processo vai acontecer até que não se possa mais realizar as divisões, ou até definir algum critério de parada pré estabelecido. Na terceira e última etapa da metodologia, o autor propõem a realização do processo de poda da árvore. Como já explicado ao longo do projeto de pesquisa, este trabalha com cortes sucessivos nas árvores a fim de reduzir o problema do *overfitting*.

2.4.2.2 CHAID

Guimarães et al. (2000) destaca o CHAID como ferramenta complementar para construção das árvores de decisão. CHAID é similar ao CART na construção da árvore de decisão, mas difere na forma que escolhe suas divisões. Ao invés de utilizar métricas como a da entropia e a *Gini* para escolher divisões otimizadas, a técnica depende de um teste de chi-quadrado usado em tabelas de contingência para determinar que preditor de categoria é mais distante da independência com os valores de predição. Por CHAID depender das tabelas de contingências para formar seu teste de significância para cada preditor, todos os preditores devem ser ou categóricos ou conhecidos dentro de uma forma categórica

em agrupamentos (dividir a idade das pessoas em dez partes, de 0 a 9, de 10 a 19, de 20 a 29).

2.4.3 Florestas aleatórias

Provavelmente o modelo de classificação de renovação de contratos telefônicos na indústria de celulares, descrito na figura 6 contém erros. Além do *overfitting*, existem diversos fatores que devem ser levados em consideração em qualquer modelo de árvores de decisão e, portanto, as chances de acertos olhando para uma única árvore são baixos (GUIMARÃES et al., 2000).

Generalizando para qualquer problema, Donges (2018) nota que as pessoas possuem diferentes experiências e conhecimentos a respeito de diversos assuntos, fazendo com que a probabilidade delas darem a mesma resposta para o mesmo problema seja baixa. Agora, supondo um cenário em que milhares de indivíduos decidem dar suas opiniões, será que o comportamento do resultado médio seria o mesmo? Difícil responder esta pergunta, contudo, a probabilidade da resposta estar mais perto da correta é muito alta.

A fim de facilitar o entendimento desta lógica, Donges (2018) sugere uma analogia com o cotidiano de diversos estudantes universitários: a escolha de uma disciplina para o semestre. O aluno pede informações para vários conhecidos que já estudaram a determinada disciplina. Primeiro, é feita uma pergunta se um dos conhecidos já fez a matéria com determinado professor e gostou. Baseado na resposta obtida, ele recebe algumas sugestões. Depois disto, ele começa a pedir sugestões para mais amigos e cada um deles faz diferentes perguntas antes de dar alguma sugestão. No final, ele escolhe as disciplinas que receberam mais recomendações para embasar sua decisão final.

Este é o mesmo princípio utilizado pelas Florestas Aleatórios, ou *Random Forests*. De maneira simplista, Donges (2018) diz que o algoritmo de florestas aleatórias cria várias árvores de decisão e as combina para obter uma predição com maior acurácia e maior estabilidade. Breiman (2001) ressalta que, enquanto as árvores de decisão utilizam a melhor divisão dos nós entre todas as variáveis, as florestas aleatórias dividem os nós através dos melhores subconjuntos escolhidos aleatoriamente por nó. O algoritmo cria então um conjunto de árvores de decisão de maneira aleatória (daí o nome "florestas aleatórias") para aumentar a precisão e eficácia do modelo, reduzindo o problema do *overfitting*. Como o algoritmo não se preocupa individualmente com a

divisão dos nós entre todas as variáveis, ele evita o caráter guloso das árvores. A este processo dá-se o nome de *bagging* (CHEN, 2019).

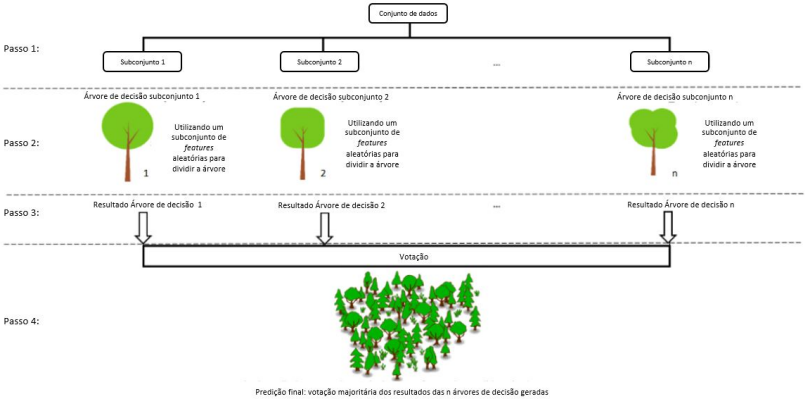
Corroborando com a conceitualização das florestas aleatórias, Moisen (2008), ressalta que no nó raiz, uma pequena amostra aleatória de variáveis são selecionadas e a melhor divisão é feita usando este conjunto limitado de variáveis. Para cada nó subsequente outra amostra aleatória é escolhida para continuar as sucessivas divisões. As árvores seguem esta lógica de crescimento até que atinjam o maior tamanho possível sem a poda. Todo este processo é refeito com um novo conjunto de dados diversas vezes e, a previsão final é a média das previsões das árvores em cada floresta para os casos de regressão, enquanto que para os casos de classificação, são as previsões com maiores números de votos.

A figura 7 ilustra o passo a passo de como o algoritmo de floresta aleatória funciona (CHEN, 2019). No primeiro passo são criadas n amostras de subconjuntos aleatórios para o treinamento do modelo, através do processo descrito como *bagging*. Este processo funciona como uma re-amostragem aleatória do conjunto de dados. No segundo passo, para cada uma destas amostras geradas, uma árvore de decisão é criada com as diferenciações citadas por Moisen (2008), em que para cada nó da árvore são escolhidas aleatoriamente um determinado número de variáveis limitadas para realizar a melhor divisão. Liaw, Wiener et al. (2002), ressalta que, por mais contraintuitivo que esta estratégia pareça, ela funciona muito bem contra o problema do *overfitting* e ainda ressalta que é uma metodologia de amigável para o usuário, tendo em vista que apenas dois parâmetros são sensíveis (o número de variáveis e o número de árvores da floresta).

No passo seguinte, cada árvore prediz, de forma independente, o resultado final, apenas dando sequência a lógica de divisão dos nós sem nenhuma realização de poda, neste momento. Por fim, no quarto, e último passo, é feita a previsão final. Para cada candidato teste do conjunto, a floresta aleatória usa a classe com maior número de votos para fechar esta previsão nos problemas de classificação e a média dos resultados da árvore, para os problemas de regressão.

Embora as Florestas Aleatórias tenham uma ótima reputação e funcionem muito bem na prática, existem parâmetros no algoritmo de funcionamento que devem ser tunados (*Tuning*) no modelo (ZHANG; MA, 2012). A seguir será ilustrada a devida importância deste tópico e a descrição de hiper parâmetros na biblioteca de *Scikit-Learn* utilizada durante o presente trabalho de pesquisa.

Figura 7: Processo de predição das florestas aleatórias.



Fonte: adaptado de (CHEN, 2019)

2.4.3.1 Tunagem de Hiper-parâmetros (*Hyperparameter Tuning*)

O processo de tunagem dos hiper-parâmetros se tornou um passo crucial dentro da prática de AM (BARDENET et al., 2013). Muitas vezes é mais vantajoso trabalhar com o processo de tunagem de parâmetros, do que buscar novos paradigmas e dados para um problema (PINTO et al., 2009; COATES; NG; LEE, 2011; BERGSTRA et al., 2011; SNOEK; LAROCHELLE; ADAMS, 2012; THORNTON et al., 2012).

De acordo com Koehrsen (2018), a melhor maneira de se pensar em um hiper-parâmetro é comparando os ajustes de um algoritmo para melhorar sua performance, assim como se ajusta um botão em um rádio para ajustar o som da saída. Estes hiper-parâmetros são ajustados antes do treinamento dos dados no algoritmo de Florestas Aleatórias. Neste processo de tunagem do modelo é onde o AM passa de uma ciência para um teste de tentativa e erro, se tornando um processo mais experimental do que teórico (KOEHRSEN, 2018).

O incremento dos hiper-parâmetros é verificado pelos resultados dos modelos de avaliação dos modelos, tratados na próxima seção, através do processo de validação cruzada e outras métricas. Dentro da biblioteca *Scikit-Learn*, Mohtadi, B (2017), Srivastava, T (2015) ressaltam quais os parâmetros mais importantes para Florestas Aleatórias e como eles impactam o modelo nos termos de *overfitting* e *underfitting*.

- *N_estimators*: representa o número de árvores que se deseja cons-

truir na floresta antes de realizar a previsão final dos resultados. Usualmente, quanto maior o número de árvores melhor. Contudo, adicionando muitas delas no modelo ele pode ficar muito lento e atrasar o processo de treinamento consideravelmente.

- *max_depth*: este parâmetro indica a profundidade de cada árvore na floresta. Quanto mais profunda uma árvore, mais divisões ela terá e, consequentemente, mais informações naquele conjunto de dados.

- *min_samples_split*: este representa o mínimo de amostras requeridas para que um nó se divida. Ele pode variar desde considerar pelo menos uma amostra em cada nó, até considerar todas as amostras por nó. Na prática, este parâmetro indica que quanto maior seu valor, mais restrita serão as árvores da floresta.

- *min_samples_leaf*: semelhante ao *min_samples_split*, este parâmetro descreve a quantidade mínima de amostras nas folhas. Quanto menor a folha, mais propenso a captar ruídos os dados de treinamento serão.

- *max_features*: este representa o número máximo de *features* que serão levadas em consideração no momento da divisão do nó.

2.5 AVALIAÇÃO DO MODELO DE CLASSIFICAÇÃO E DE REGRESSÃO

Embora os algoritmos de aprendizado máquina sejam ferramentas muito poderosas, não existe um único algoritmo que tenha o melhor desempenho para todos os tipos de problema. As métricas de avaliação de modelos desempenham papel fundamental no mundo de AM e, sem elas, os algoritmos podem não permitir descrever um problema de forma clara (NGUYEN; BOUZERDOUM; PHUNG, 2009). Por este motivo, surge a subseção atual: ela é responsável por descrever qual a limitação de cada algoritmo através da explicação de metodologias de avaliação que permitem entender qual a performance deles para os problemas de regressão e de classificação apresentados.

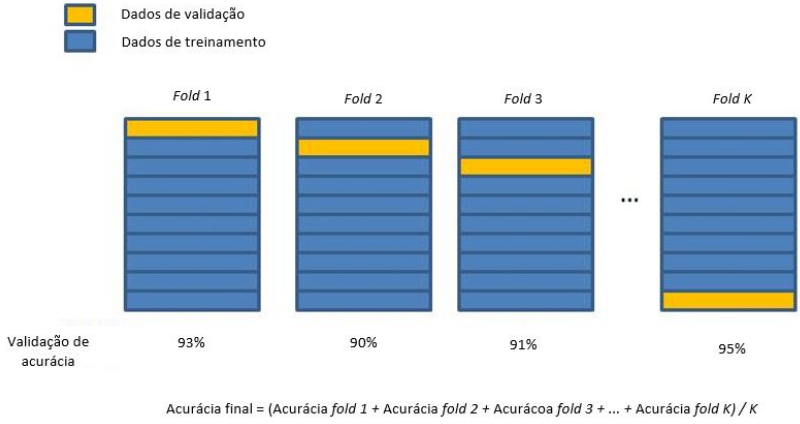
2.5.1 Validação Cruzada (*Cross Validation*)

A Validação Cruzada, ou *Cross Validation* - CV - é um estratégia amplamente utilizada para seleção e avaliação de algoritmos. A principal ideia por trás da CV é dividir os dados, uma ou diversas vezes, para estimar o risco de cada algoritmo: parte dos dados é utilizada para treinar cada algoritmo (a amostra de treinamento) e a parte restante é utilizada para estimar o risco do algoritmo (a amostra de validação) (ARLOT; CELISSE et al., 2010).

Contudo, Friedman, Hastie e Tibshirani (2001) destacam que a quantidade de dados disponíveis em certos casos são restritos, tornando-se um fator limitante para análises do modelo. Assim, surge a possibilidade de aplicar o método de Validação Cruzada *K-fold*. Este método divide todos os dados de treinamento em K número de divisões, chamadas de *folds*. De forma iterativa, o modelo é ajustado K vezes, cada vez treinando os dados com $K-1$ *folds* e validando no $K-$*ésimo fold*, também conhecido por dado de validação (KOEHRSEN, 2018). Além da figura 8, o exemplo a seguir ajuda a ilustrar o entendimento da metodologia: considerando um modelo com $K = 5$, a primeira iteração seriam treinados os primeiros quatro *folds* e avaliado no quinto. Na segunda, o treinamento seria realizado com o primeiro, segundo, terceiro e quinto *fold* e avaliado no quarto. Este procedimento aconteceria por mais três vezes, seguindo o mesmo princípio de que, a cada vez, a validação seria em um *fold* distinto. No final do treinamento, ou seja, de todas as K rodadas de treinamento, a performance final de acurácia do modelo seria medida pela média de cada uma individualmente.

O *trade-off* entre variância e viés deve ser sempre observado no uso da CV. Para o exemplo supracitado, em que $K = 5$, o viés pode ser um problema dependendo de como o desempenho do método varia com o tamanho do conjunto de treinamento, embora a variância seja baixa (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Desta forma, Garcia (2003) ressalta que o uso de $K = 10$ é um número considerado bom para se obter boa precisão do modelo gerado. A comparação entre este número de *folds* e a porcentagem de exemplos pode ser verificada na figura 9.

Figura 8: Exemplo de Validação Cruzada K-*fold*.



Fonte: adaptado de (DRAKOS, 2018)

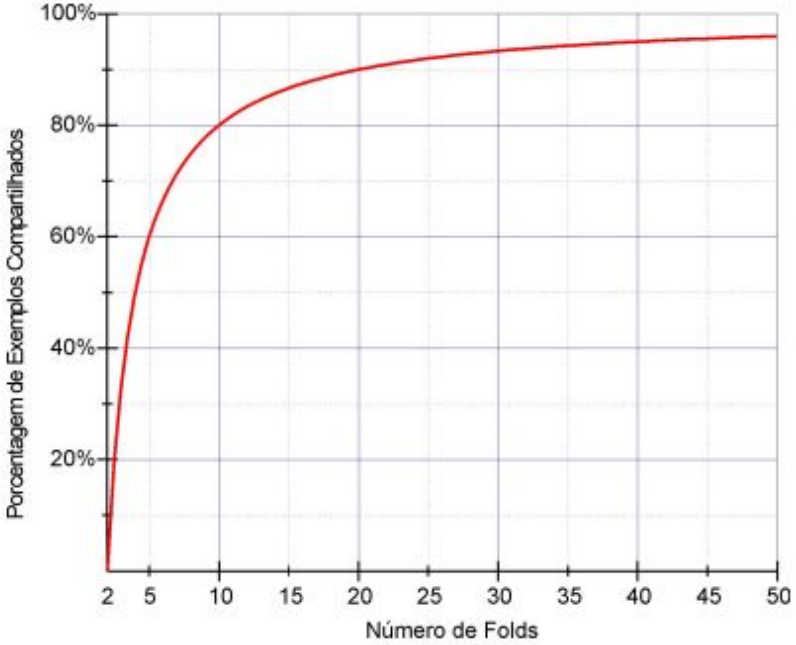
2.5.2 Avaliação dos modelos de classificação

2.5.2.1 Matriz de confusão

A matriz de confusão descreve os acertos e erros que acontecem durante o processo de classificação (GARCIA, 2003). Monard e Baranauskas (2003) descrevem a matriz de confusão de uma hipótese qualquer como uma maneira efetiva do modelo de classificação, ao mostrar o número de classificações corretas *versus* as classificações preditas para cada classe, sobre um conjunto de exemplos T. Como mostrado na figura 10, os resultados são totalizados em duas dimensões: classes verdadeiras e classes preditas, para k classes distintas $\{C1,C2,...,Ck\}$.

Nas matrizes de confusão, a diagonal principal da matriz mostra os casos em que a classificação foi corretamente realizada, enquanto que, as células fora desta diagonal, mostram as classificações erroneamente realizadas pelo modelo (GARCIA, 2003). No geral, segundo Botelho e Tostes (2010), por simplicidade, se considera problemas com duas classes, usualmente rotuladas como positivas e negativas. Nessa situação, a matriz apresenta a quantidade de casos distribuídos em quatro quadrantes: os previstos como sendo da classe 0 e que realmente são da classe 0, chamados de "verdadeiro negativo", ou seja, os dados originalmente eram falsos e foram preditos como tal. A quantidade de casos previstos como sendo da classe 0, mas que na verdade são da classe

Figura 9: Número de *folds versus* porcentagem de exemplos compartilhados em CV.



Fonte: (MONARD; BARANAUSKAS, 2003, p. 54)

Figura 10: Matriz de confusão de um classificador.

Classe	predita C_1	predita C_2	$\cdots$	predita C_k
verdadeira C_1	$M(C_1, C_1)$	$M(C_1, C_2)$	$\cdots$	$M(C_1, C_k)$
verdadeira C_2	$M(C_2, C_1)$	$M(C_2, C_2)$	$\cdots$	$M(C_2, C_k)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
verdadeira C_k	$M(C_k, C_1)$	$M(C_k, C_2)$	$\cdots$	$M(C_k, C_k)$

Fonte: (MONARD; BARANAUSKAS, 2003, p. 48)

1, denominados de ”falso negativo”. A quantidade de casos previstos como sendo da classe 1 e que realmente são da classe um, ”verdadeiro positivo”e, finalmente, os casos previstos como da classe 1, mas que na verdade são da classe 0, chamados de ”falso positivo”. A figura 11

mostra de forma instrutiva estas situações.

Figura 11: Matriz de confusão.

Fonte: autor, adaptado de (BOTELHO; TOSTES, 2010, p. 407)

Dentre as diversas métricas de avaliação de modelos de classificação, destacam-se a precisão, sensibilidade, especificidade, acurácia, média geométrica, F-*measure*, MCC e AUC-ROC (NGUYEN; BOUZERDOUM; PHUNG, 2009).

Tabela 1: Métricas de avaliação de modelos de classificação.

Tipos de métricas	Cálculo da métrica
Precisão	$\frac{VP}{VP+FP}$
Sensibilidade	$\frac{VP}{VP+FN}$
Especificidade	$\frac{VN}{VN+FP}$
Acurácia	$\frac{VP+VN}{VP+FP+VN+FN}$
Média geométrica	$\sqrt{\frac{VP}{VP+FN}\frac{VN}{VN+FP}}$
F-*measure*	$\frac{2\frac{VP}{VP+FP}\frac{VP}{VP+FN}}{\frac{VP}{VP+FP}+\frac{VP}{VP+FN}}$

Nota-se na tabela 1, que as métricas MCC e AUC-ROC não são abordadas. Elas são detalhadas na subseção seguinte. Agora, faz-se necessário entender as demais métricas levantadas, de acordo com Nguyen, Bouzerdoum e Phung (2009):

1. Precisão: é a porcentagem de predições positivas feitas por um

classificador que está correto. Ou seja, as medidas estão bem distribuídas nas classes.

2. Sensibilidade: é a porcentagem de verdadeiros positivos que foram corretamente detectados pelo classificador, ou seja, é uma espécie de acurácia dos exemplos positivos.

3. Especificidade: Ao contrário da sensibilidade, ela é, de certa forma, a acurácia nos exemplos negativos.

4. Acurácia: esta medida é utilizada quando as classes da variável estão razoavelmente distribuídas, trabalhando com a proporção de todas as predições verdadeiras, em relação ao total de previsões do modelo.

5. Média geométrica: medida sugerida em (KUBAT; MATWIN et al., 1997) e, desde então, utilizada por diversos pesquisadores na avaliação de classificadores em dados desbalanceados. Indica o equilíbrio entre a performance de classificação das classes minoritárias e majoritárias. Ela leva em consideração tanto a sensibilidade, quanto a especificidade.

6. F-*measure*: de acordo com Fawcett (2006), esta medida pode ser definida como a média harmônica da sensibilidade e precisão. Quanto maior for este valor, maior serão ambas as medidas (precisão e sensibilidade).

2.5.2.2 AUC-ROC

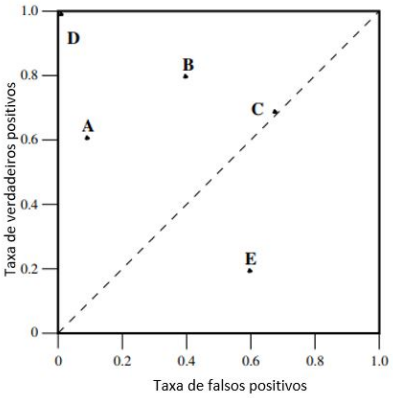
O gráfico ROC - *Receiver Operating Characteristic* - é uma técnica de visualização, organização e seleção de classificadores baseados em sua performance (EGAN, 1975). De acordo com Weiss (2004), o gráfico ROC e sua AUC - *Area Under the Curve* - são duas medidas das mais utilizadas para avaliar a performance geral de classificações.

Este gráfico permite mostrar a relação entre os "benefícios"(taxa correta de detecções ou taxa de verdadeiros positivos) e os "custos"(taxas de detecções falsas ou taxa de falsos positivos) conforme o limite de decisão varia. Além disso, ele também mostra que, para qualquer classificador, a taxa de verdadeiros positivos não pode aumentar sem que aumente também a taxa de falsos positivos (NGUYEN; BOUZERDOUM; PHUNG, 2009). De acordo com Fawcett (2006) os gráficos ROC tem sido cada vez mais utilizados pela comunidade de AM, devido ao fato

de que muitas vezes a precisão de classificação por si só, acaba sendo uma métrica ruim para medir o desempenho do algoritmo.

A curva ROC é um gráfico bi-dimensional e pode ser obtido tendo no seu eixo das ordenadas a "sensibilidade" (taxa de verdadeiros positivos que foram corretamente detectados pelo classificador) e no eixo das abscissas o complemento da especificidade, ou seja, 1–$especificidade$ (complemento da taxa de verdadeiros negativos, sobre o total de erros peditos pelo modelo) (GÖNEN, 2007).

Figura 12: Exemplo do gráfico ROC com cinco classificadores discretos.



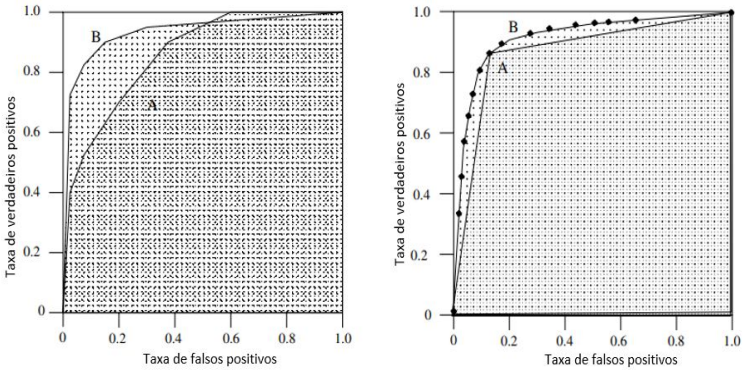
Fonte: autor, adaptado de (FAWCETT, 2006, p. 862)

Para comparar os classificadores, deseja-se reduzir a performance ROC para um único valor escalar que represente a performance esperada. Um método comum para se fazer isto é trabalhando com a área sob a curva ROC, denominada de AUC (BRADLEY, 1997; HANLEY; MCNEIL, 1982). Como o AUC-ROC é uma parte da área, seu valor será sempre entre 0 e 1. Contudo, pelo fato de que adivinhações aleatórias produzam uma linha diagonal entre os pontos $(0, 0)$ e $(1, 1)$, que representa uma área de 0,5, nenhum classificador realístico deve possuir um AUC-ROC menor do que este valor. Uma propriedade estatística importante do AUC-ROC, de acordo com Hanley e McNeil (1982), é sw que seu valor para algum classificador é equivalente à probabilidade do classificador ranquear uma instância positiva escolhida aleatoriamente superior a uma instância negativa escolhida aleatoriamente.

A figura 13a mostra o comparativo entre duas áreas sob a curva ROC denominadas de A e B. Nota-se que o classificador B possui maior área e, portanto, melhor performance média. A figura 13b mostra a

área sob a curva para um classificador discreto A e um classificador probabilístico B, quando B é usado com um único e fixo limite. Embora o desempenho de ambas as curvas sejam iguais no ponto limite, o desempenho de B passa a ser inferior a partir deste ponto (FAWCETT, 2006). Ainda segundo Fawcett (2006), é possível que um classificador com alto AUC desempenho, em uma região específica do espaço ROC, tenha performance abaixo do que um classificador de mais baixo AUC-ROC na média. Essa situação é ilustrada na figura 13a, na qual o classificador B é geralmente maior que A, exceto no ponto FP > 0,6, onde A passa a levar uma ligeira vantagem.

Figura 13: Área sob a curva ROC de exemplos A e B.



(a) Comparação entre AUC-ROC A e B

(b) Área sob a curva ROC de classificador discreto (A) e de classificador probabilístico (B)

Fonte: (FAWCETT, 2006)

2.5.2.3 MCC

O MCC - *Matthews Correlation Coefficient* - foi pela primeira vez introduzido por B.W. Matthews para avaliar a performance da estrutura secundária da proteína em predições. o MCC é frequentemente utilizado em problemas dicotômicos, ou seja, com apenas duas variáveis de saída envolvidas (MATTHEWS, 1975). Hoje em dia, embora seja amplamente utilizado, tem maior recorrência de uso na medicina (BOUGHORBEL; JARRAY; EL-ANBARI, 2017). Gorodkin (2004) refere-se

ao MCC como a versão discreta do coeficiente de correlação de Pearson e toma valores entre [-1;1]. O coeficiente assume valor de 1 se há correlação perfeita, 0 se não há correlação e -1 se há correlação perfeita negativa. Além disso, por natureza, o MCC é aplicado apenas para duas categorias. Nota-se que, no problema de classificação apresentado neste presente projeto de pesquisa, as variáveis de saída são dicotômicas, pois, obedecem apenas "Ativo", ou "Inativo"como resultado.

Embora aconselhado para problemas dicotômicos, o MCC é, algumas vezes, utilizado para realizar avaliação de modelos multi-categóricos, forçando-os a uma dicotomia, em que uma única categoria é isolada. Entretanto, esta indução leva à perda de informações e a avaliação final pode se tornar imprecisa (GORODKIN, 2004). Ressalva que, mesmo para problemas dicotômicos, a medida de MCC é insuficiente se uma das categorias contiver uma grande parte de todos os pontos de dados, de acordo com (GORODKIN, 2004). Se, por exemplo, 99% dos dados pertencem a uma categoria, um preditor poderia prever tudo como pertencente a essa categoria. Nesse caso, o coeficiente de correlação de Matthews seria zero, indicando que a previsão não se correlaciona com o que deve ser previsto.

De acordo com Matthews (1975), a expressão que define o MCC pode ser calculada segundo a equação 2.9, a seguir.

$$MCC = \frac{(VP * VN) - (FP * FN)}{\sqrt{(VP + FN)(VP + FP)(VN + FP)(VN + FN)}} \quad (2.9)$$

Com o intuito de corroborar com a importância e utilidade do MCC como métrica de avaliação de modelos preditivos de classificação, Boughorbel, Jarray e El-Anbari (2017) realizou uma simulação e utilizou quatro métricas distintas para comparar os resultados do modelo: MCC, AUC-ROC, Acurácia e F-*measure*. Observou-se que os classificadores geravam rótulos sem observar as informações transportadas, logo, não seria esperado que um dos classificadores superassem os outros, no caso de três classificadores, denominados de C1, C2 e C3, segundo o autor. Contudo, a acurácia e a F-*measure* apresentaram desempenhos variados para os classificadores denominados de C1 e C2, portanto, elas são sensíveis ao desequilíbrio dos dados. Em contrapartida, o autor cita que as métricas MCC e AUC-ROC demonstraram desempenho constante para os diferentes classificadores, sendo, portanto, mais robustos ao desequilíbrio dos dados. Por fim, a preferência pelo uso do MCC, sobre o AUC-ROC, para o autor, é que não existe uma fórmula explícita para o cálculo do AUC-ROC, enquanto que o MCC, pode ser calculado

pela fórmula 2.9.

2.5.3 Avaliação dos modelos de regressão

2.5.3.1 Testes de regressão - MSE, RMSE e MAE

Para problemas de regressão, uma das medidas mais utilizadas para estimar a distância entre o valor real ao valor atribuído pela hipótese induzida, geralmente é o erro quadrático médio (MSE) (MONARD; BARANAUSKAS, 2003). Enquanto isso, Laureano, Caetano e Cortez (2014) também sugerem o uso de métricas como o erro médio absoluto (MAE) e a raiz do erro quadrático médio (RMSE).

O cálculo do MSE pode ser verificado pela equação 2.3, previamente apresentada. A métrica MSE mede o erro quadrado médio das previsões. Para cada ponto, calcula-se a diferença quadrada entre as previsões e o alvo e, em seguida, calcula-se a média desses valores. Quanto maior esse valor, pior é o modelo. O MSE assume o valor zero para um modelo perfeito (LAUREANO; CAETANO; CORTEZ, 2014).

Para Wang e Bovik (2009), o MSE possui alguns benefícios que favorece sua utilização: fácil utilização; possui um significado físico muito claro: é a forma natural de medir a energia de um erro, ou seja, garantir a medição de que propriedades do algoritmo são mantidas ou não; é uma excelente métrica no contexto de otimização e; é uma métrica desejada para estatística e modelos de previsão. Contudo, a métrica pode apresentar também alguns aspectos negativos: caso seja feita uma única previsão muito ruim, o aspecto quadrático do modelo piorará o erro, distorcendo a métrica e finalmente superestimando a baixa qualidade do modelo. Esse é um comportamento particularmente problemático para casos de dados ruidosos (dados que por alguma razão não são totalmente confiáveis) - mesmo para ótimos modelos, o MSE pode ter um valor alto nessa situação, então torna-se difícil julgar quão bem o modelo está performando.

Contudo, se todos os erros são pequenos, em outras palavras, menores que 1, o efeito oposto é sentido: pode-se subestimar a qualidade do modelo. Já o RMSE é representado pela raiz quadrada do MSE e pode ser observado na equação 2.10. Ele possui uma unidade (dimensão), igual à dimensão dos valores observados e preditos. Interpreta-se seu valor como uma medida do desvio médio entre observado e predito (WANG; BOVIK, 2009).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2} = \sqrt{MSE} \tag{2.10}$$

Por fim, o MAE, equação 2.11, é muito semelhante a métrica MSE, contudo, toma o valor absoluto das diferenças entre previsão do modelo e valor real. De acordo com Laureano, Caetano e Cortez (2014), ele expressa na unidade de medida do atributo a prever, onde valores próximos de 0 traduzem melhor o modelo.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - Y_{a\,i}| \tag{2.11}$$

Willmott e Matsuura (2005) compara as medidas RMSE e MAE. Segundo o autor, medidas de erro médio baseado em somatório de erros quadráticos não descrevem o erro por si só. Embora o autor afirme que o RMSE tende a crescer muito mais rápido do que o MAE, ou seja, há uma maior penalização de erros grandes se comparado aos pequenos, estas medidas andam em concordância, ou seja, a análise do modelo não seria comprometida usando qualquer uma das duas métricas.

3 PROCEDIMENTO METODOLÓGICO

Os procedimentos metodológicos do presente trabalho de pesquisa são divididos em três seções. Na primeira delas, é conceituada as características e peculiaridades da empresa em que o modelo foi elaborado. A segunda é utilizada para discutir a respeito dos dados utilizados no modelo. Como, por exemplo, de onde eles vieram, como e em qual período foram feitas suas coletas e como eles foram tratados antes serem inseridos no algoritmo. Por fim, na terceira e última etapa, será discutido e apresentado como os conceitos do capítulo 2 foram utilizados na pesquisa, de forma precisa e esmiuçada, a fim de possibilitar possíveis replicações dos procedimentos, caso necessário.

3.1 A EMPRESA

A organização de TI pesquisada possuía, no momento da coleta das informações, aproximadamente 600 funcionários, e faz parte do emergente polo de tecnologia do sul do Brasil. Ela possui diversas áreas funcionais, das quais, uma delas é a área de vendas, cujo objetivo principal é aumentar a sua base de clientes de forma consciente, escalável e alinhada com os objetivos estratégicos organizacionais do negócio. Além disso, ela também é o foco da presente pesquisa.

Pelo alinhamento estratégico, a área de vendas está diretamente ligada à gestão executiva da organização e atende dois modelos de vendas distintos: as vendas diretas e as vendas via canais de parceria, em que ambas podem ser destrinchadas em operações Brasil e internacional. As vendas diretas atuam na aquisição de novos clientes sem o apoio ou a intervenção de terceiros, ou seja, depende apenas dos esforços dos colaboradores da própria empresa. Por outro lado, a área de vendas via canais de parceiros é a área responsável por adquirir novos parceiros na base, os quais farão o trabalho dos "vendedores diretos"para atrair novos clientes para a organização. A fim de delimitar o escopo de estudo, tanto por limitação de acesso aos dados das vendas via canais, quanto por representar menor parcela de receita do total de vendas, se comparado com a primeira, a pesquisa tem como foco a área de vendas diretas somente para o Brasil.

Tendo em vista o tópico de *turnover* de funcionários em mercados de TI, especialmente em áreas de contato com cliente, como vendas, já apresentados ao longo do capítulo 2, a organização estudada tam-

bém se justifica como objeto da pesquisa por apresentar características semelhantes às previamente descritas. Com o propósito de esclarecer estas semelhanças, dados da organização mostram que a taxa de *turnover* de 2017 foi superior (25,19%) aos dois precedentes anos (23,76% em 2016 e 18,06% em 2015). Corroborando com a hipótese levantada, a área de vendas se configurou como maioria neste cenário (37,82% ou 45 saídas).

3.2 DESCRIÇÃO DOS DADOS

Foram levantados para a pesquisa dados de 103 vendedores durante 24 meses, no período compreendido entre janeiro de 2017 e dezembro de 2018. A priori foram levantadas 76 dados de características (*features*) para cada um deles. Cada uma delas serão descritas a seguir, seja de forma individualizada, ou agrupada em famílias, quando necessário. É necessário destacar que a variável "salário"não foi levada em consideração do modelo, visto que não possuía grande amplitude nos valores verificados. Por este motivo, não fora considerada nos modelos.

- **Dados de performance:** as métricas de performance para um vendedor na empresa em questão, levam em consideração a quantidade de contas e receita que ele traz para o negócio em relação às suas metas. Pelo fato dos valores das metas serem diferentes para cada vendedor, dependendo dos seus níveis dentro da organização, os dados utilizados no presente modelo se baseiam nos percentuais médios de atingimento ($percentual = realizado/meta$) e, em seguida, é realizada a média aritmética entre o percentual de contas e o percentual de receita. Por exemplo, seja um vendedor com meta de 10 contas e $R\$10.000,00$ de receita, que tenha realizado 10 vendas, levantando $R\$8.000,00$, seus percentuais seriam $percentual\_contas = 100\%$ e $percentual\_receita = 80\%$. Sendo a média, e resultado final coletado para o algoritmo, neste exemplo, $percentual = 90\%$.

  Para cada um dos 103 vendedores, foram coletados os valores médios dos meses em que eles estavam trabalhando na empresa. Sendo que a contagem máxima para um vendedor são os 24 meses, o mínimo 3 meses, a média 9.5 meses e o desvio padrão de 6 meses.

- **Curso de formação:** este atributo, como o nome já sugere, indica o curso de formação superior do vendedor (concluído, cursando, ou inexistente). A saída esperada é o nome do curso, nos

Tabela 2: Top 10 cursos mais representativos na base de dados.

Cursos de formação	**Representatividade**
Administração	36,70%
Relações Internacionais	16,51%
Publicidade e Propaganda	6,42%
Direito	2,75%
Engenharia Mecânica	2,75%
Comunicação Social	1,83%
Engenharia Elétrica	1,83%
Gestão comercial	1,83%
Jornalismo	1,83%
Marketing	1,83%

Fonte: Autor.

casos de formados e estudantes, e "Não possui formação", para os que não realizaram ensino superior. Dos 39 diferentes cursos apresentados nos dados, 10 deles representam $74,31\%$, sendo que, também do total, Administração é o mais representativo, com $36,70\%$, conforme ilustrado na tabela 2.

- **Tempo de formado:** o tempo de formado, aqui representado em anos, começa a ser contado a partir do ano em que o vendedor se formou, levando apenas o mesmo em consideração. Por exemplo, caso um vendedor tenha se formado em agosto de 2015, em janeiro de 2017 o modelo conta como 2 anos de formado e não 1 ano e 5 meses, como seria na realidade. A necessidade de tal simplificação surge da limitação das bases consultadas, que constavam apenas o ano da formação.

  Na tabela 3 estão representados os percentuais da base para cada categoria levantada, que variam de 0 (mesmo ano de formação) até 14 anos de formado. Além disso, também conta com as categorias "Não havia se formado", para casos de pessoas que faziam curso superior, mas não estavam formados, e "Vazio"para casos de pessoas que, ou não possuíam graduação em ensino superior, ou não constavam os anos de formação na base de dados.

- **Tipo de universidade:** esta característica se refere à modalidade da universidade cursada pelo vendedor. Ela pode variar em quatro categorias: particular, pública, exterior e "sem formação superior".

Tabela 3: Percentual de vendedores por tempo de formação.

Tempo de formado	**Representatividade**
1	22,33%
Não havia formado	12,62%
2	7,77%
3	7,77%
Vazio	7,77%
5	6,80%
0	5,83%
6	5,83%
8	5,83%
4	4,85%
9	3,88%
7	1,94%
10	1,94%
15	1,94%
11	0,97%
12	0,97%
14	0,97%

Fonte: Autor.

Tabela 4: Percentual de vendedores por participação em entidades.

Variável de saída	**Representatividade**
Não	76,84%
Sim	23,16%

Fonte: Autor.

Tabela 5: Percentual de vendedores por primeiro emprego.

Variável de saída	**Representatividade**
Não	91,58%
Sim	8,42%

Fonte: Autor.

- **Participou de entidades estudantis:** ainda a respeito ao período de graduação no ensino superior, esta variável busca entender se o vendedor teve alguma participação em entidades durante a graduação. Embora binária a saída da variável, sendo "Sim", no caso de ter participado, ou "Não", no caso de não ter participado, alguns exemplos de entidades encontradas foram: Empresas Juniores, Programas de Educação Tutorial, Grupos de estudos aplicados e AIESEC. Na tabela 4 tem-se a comparação entre as variáveis de saída para a característica.

- **Primeiro emprego:** também binária, esta variável tem como saída "Sim", caso o emprego de vendas tenha sido a primeira experiência de trabalho da carreira, ou "Não", caso já tenha algum tipo de experiência prévia. Neste caso, experiências em entidades estudantis foram consideradas como uma experiência. A tabela 5 traz o comparativo entre as saídas.

- **Segmento do último trabalho:** dentre todos os vendedores, a maioria ressaltou que aquele não era o seu primeiro emprego. Desta forma, fez-se necessário também analisar qual era o segmento principal da empresa do último trabalho daquele vendedor. As categorias foram limitadas em indústria, comércio e serviço. A tabela 6 ilustra as respostas dos vendedores, sendo que a vasta maioria classifica o segmento como "comércio".

  Embora o destaque seja para o último trabalho, o fator não exclui possíveis participações dos vendedores em segmentos como comércio, por exemplo.

Tabela 6: Percentual de vendedores por segmento do último trabalho.

Variável de saída	Representatividade
Serviço	97,06%
Indústria	2,94%
Comércio	0,00%

Fonte: Autor.

Tabela 7: Percentual de vendedores com experiência em mercados de tecnologia.

Variável de saída	Representatividade
Não	62,11%
Sim	37,89%

Fonte: Autor.

- **Experiência com vendas:** se trata de uma métrica binária. Utilizada para identificar se determinado vendedor, já teve algum tipo de experiência na área. Caso afirmativo, a saída é "Sim", enquanto que em caso negativo, a saída é "Não". Ao todo, existiam 21 pessoas que não tinham trabalhado com vendas.

- **Experiência com tecnologia:** lógica idêntica ao conceito previamente apresentado. Contudo, desta vez, a *feature* tinha por objetivo identificar se o vendedor já havia trabalhado com mercado de tecnologia, mesmo da empresa em questão. Conforme a tabela 7, verifica-se que a maioria daqueles funcionários não havia prévia experiência no mercado de tecnologia.

- **Adaptação em pré vendas:** a área de vendas da empresa analisada é dividida em duas: pré-vendas e vendas. Nesta área de pré-vendas, o pré-vendedor, chamado de SDR - *Sales Development Representative* - possui a função de validar alguns requisitos mínimos para que a negociação possa ser repassada ao vendedor.

  Os vendedores analisados podiam, ou não, ter passado por esta área. Por isso, foram classificados em três maneiras distintas: "Não foi", para os vendedores que não passaram por pré-vendas; "1 mês", para os vendedores que passaram por um período de treinamento de adaptação antes de atuarem como vendedores e; "Mais de 1 mês", para os casos de pessoas que foram contratadas para serem SDRs e, em algum momento foram promovidas para vendas.

Tabela 8: Percentual de vendedores com participação, ou não, em pré-vendas.

Variável de saída	**Representatividade**
Não foi	39,81%
1 mês	39,81%
Mais de 1 mês	20,39%

Fonte: Autor.

Tabela 9: Percentual de atingimento de vendedores no período de pré-vendas.

Faixa de atingimento %	**Representatividade**
$\leq$71	17,31%
71<x$\leq$99,165	15,38%
99,165<x$\leq$117,57	15,38%
117,57<x$\leq$180	17,31%
Vazio	34,62%

Fonte: Autor.

O objetivo desta verificação é entender se existe algum tipo de relação em treinar em pré-vedas, ou subir SDRs para vendas, ou ainda, se não há necessidade de passar por esta etapa. A tabela 8 mostra a distribuição entre os três grupos analisados.

- **Atingimento em pré-vendas:** a meta de um pré-vendedor é baseada no percentual de oportunidades que ele gera para o vendedor. Basicamente, estas oportunidades se tratam de possíveis clientes com alguns critérios verificados pelos pré-vendedores e validados pelos vendedores. O item "atingimento em pré-vendas"se refere justamente ao percentual médio atingido pelo vendedor no seu período de pré vendas.

  Na tabela 9, pode-se verificar o percentual de representatividade dos vendedores em período de SDR por faixas de atingimento. Como primeiro exemplo, percebe-se que $17,31\%$ deles tiveram atingimento menor do que 71% em pré-vendas. Outro fator importante, se refere ao campo "Vazio". Nesse campo já foram desconsiderados aqueles vendedores que não foram SDRs, como verificado na tabela 8.

- **Percentual atingido no treinamento de vendas:** parte dos vendedores da base utilizada passaram por um período de treina-

Tabela 10: Percentual de atingimento de vendedores no período de treinamento.

Faixa de atingimento %	**Representatividade**
$\leq$56,33	14,05%
56,33<x$\leq$75	18,71%
75<x$\leq$95,5	23,83%
95,5<x$\leq$149	37,17%
Vazio	6,24%

Fonte: Autor.

Tabela 11: Percentual de vendedores com experiências em intercâmbios.

Variável de saída	**Representatividade**
Não	46,60%
Sim	39,81%
Vazio	13,59%

Fonte: Autor.

mentos e adaptação em um time específico. Contudo, esse time de treinamentos nem sempre existiu na empresa. Tendo em vista esse cenário, nota-se que $6,24\%$ da base não participou deste time.

Já para os vendedores que participaram, foram coletadas os percentuais de representatividade deles por faixa de atingimento nestes times. Vale destacar que não existia desempenho mínimo solicitado para que o vendedor "saísse"do treinamento, na época em questão. Na tabela 10 podem ser verificados os percentuais e representatividade.

- **Realizou algum intercâmbio:** característica binária, que indica a participação do vendedor em intercâmbio antes de entrar na empresa. Enquanto que a saída "Sim"indica que o ele já o realizou, a resposta "Não", indica o contrário, ou seja, não participou. A tabela 11, apresenta os percentuais de cada saída, bem como o chamado "Vazio", em que não constava no banco de dado se o vendedor já havia realizado algum intercâmbio ou não.

- **Mudança de cidade:** variável também binária, que representa se o vendedor precisou mudar de cidade especificamente para o trabalho. Em outras palavras, caso um vendedor tenha morado a vida inteira em uma cidade e tenha se mudado para a cidade

Tabela 12: Percentual de vendedores que mudaram de cidade para o trabalho.

Variável de saída	Representatividade
Sim	55,45%
Não	44,55%

Fonte: Autor.

do trabalho, sem saber que iria trabalhar lá, ele não será contabilizado como "Sim"nas variáveis saída de resposta. A tabela 12 indica os percentuais de representatividade em cada uma das duas categorias.

- **Idade na contratação:** as idades na contratação fazem parte de limites entre 19 e 43 anos. Na tabela 13 podem ser identificados os percentuais de representatividade por cada idade.

- **Testes de perfil, lógica, adesão cultural na empresa e fatores pessoais:** atualmente diversas empresas utilizam ferramentas de mapeamento de perfil de candidatos em processos seletivos. Isso é feito, através de ferramentas que utilizam combinações entre o perfil do candidato e a sua posição, perante à alguns testes situacionais, lógicos e psicométricos, que classificam as pessoas através de IA.

  O conjunto de saídas realizados por estas ferramentas, permite atribuir notas de 0 a 100 para quesitos como potencial, raciocínio, social, motivacional, adesão à cultura, capacidade analítica, pensamento conceitual, reflexão, pensamento criativo, ambição, facilitação, comunicação, assertividade, tomada de riscos, iniciativa, estabilidade emocional e diversos outros quesitos pessoais.

  Sabendo da importância destas ferramentas, a empresa analisada também fazia uso das mesmas para mapear o perfil dos candidatos. As notas atribuídas a cada quesito de cada funcionário é levada em consideração no modelo do presente projeto de pesquisa. Contudo, pelo fato da ferramenta ter sido implementada durante o período da coleta dos dados, 46% dos vendedores possuem notas atribuídas a cada um dos quesitos. Para os demais vendedores sem notas, as variáveis foram deixadas como vazias para não comprometer possíveis cálculos de média, por exemplo.

- **Permanece na empresa:** por fim, mas não menos importante, surge a variável que identifica se o vendedor ainda permanece na

Tabela 13: Percentual de vendedores por idade no momento da contratação.

Variável de saída	**Representatividade**
22	12,77%
24	12,77%
27	10,64%
26	9,57%
28	7,45%
23	6,38%
29	6,38%
31	6,38%
21	5,32%
25	4,26%
30	4,26%
34	3,19%
32	2,13%
35	2,13%
19	1,06%
20	1,06%
33	1,06%
36	1,06%
38	1,06%
43	1,06%

Fonte: Autor.

Tabela 14: Percentual de vendedores ativos ou inativos na empresa.

Variável de saída	**Representatividade**
Ativo	46,60%
Demissão em vendas involuntário	33,98%
Demissão em vendas voluntário	18,45%
Demissão após rotação	0,97%

Fonte: Autor.

empresa, ou se ele saiu. Sua saída é classificada em três maneiras distintas em um primeiro momento: demissão após rotação, demissão em vendas involuntário e demissão em vendas voluntário.

No primeiro caso, o vendedor mudou sua função dentro da empresa antes de ser desligado. No segundo, é o caso em que a empresa não visa mais contar com a colaboração do funcionário. Já no terceiro caso, é o funcionário que pede para sair da empresa. Na tabela 14 podem ser verificados os percentuais respectivos a cada categorização.

3.3 ETAPAS DA PESQUISA

O projeto de pesquisa apresentado foi dividido em cinco subetapas: levantamento dos dados, tratamento dos dados, *feature selection*, florestas aleatórias, tunagem de hiper-parâmetros e validação cruzada e, finalmente, as análises dos resultados. Para cada uma delas uma subseção é dedicada a fim de explorar com detalhes os procedimentos e práticas adotadas.

3.3.1 Levantamento dos dados

A primeira etapa fundamenta-se na coleta dos dados e características (*features*) que seriam utilizadas em cada um dos dois modelos estudados (classificação e regressão). Todas as características foram coletadas a partir do banco de dados da empresa contendo 103 vendedores durante 24 meses (entre janeiro de 2017 e dezembro de 2018). Deles, foram levantadas 76 características, as quais foram previamente explicadas na seção anterior, sejam individualmente, ou em família, como no caso dos testes de perfil.

A maior parte dos dados faz parte de um banco interno da em-

presa preenchidos pelos próprios funcionários de recursos humanos em três momentos distintos: durante o processo seletivo, no momento da admissão e no momento da demissão. A outra parte dos dados são referentes às entrevistas realizadas com os vendedores neste período.

Toda coleta dos dados visa a construção de modelos preditivos de classificação e de regressão. O primeiro modelo tem por objetivo analisar as diferenças entre os grupos que hoje estão ativos dos grupos dos que estão inativos na empresa, sendo esses as variáveis de saída Y do modelo de classificação.

Já o problema de regressão, tem como objetivo obter maior previsibilidade, através da preditividade do desempenho médio dos vendedores na empresa baseado em suas características levantadas no banco de dados. Como variável de saída Y, é utilizada a média de atingimento dos vendedores ao longo do período analisado.

Ambos os modelos buscam identificar as principais características dos vendedores que mais impactam as saídas atreladas a cada um dos problemas, através das suas variáveis de entrada. Para cada um deles, existem algumas diferenças entre as variáveis de entrada, as quais serão mais bem exploradas no decorrer deste capítulo. Além disso, todas as avaliações dos modelos são baseadas em intervalos de confiança para estimar suas precisões.

3.3.2 Tratamento dos dados

O processo de tratamento dos dados, consiste na lapidação de alguns dados para que possam aumentar a qualidade das modelagens utilizadas. Em certas situações, este tratamento viabiliza a inserção da variável no modelo, como por exemplo a substituição de ",", por ".", já que o padrão de codificação se baseia na linguagem americana na leitura dos dados. Além disso, nas variáveis binárias foram substituídos os "Sim"ou "Não, por "Verdadeiro"e "Falso", respectivamente.

Adotou-se, nesta etapa, o agrupamento dos dados de performance dos vendedores pelas suas médias de entregas. No caso de períodos em que o vendedor não possuía dados, eles foram considerados como vazios, para que não gerassem distúrbios nas médias, jogando-as para baixo, caso fosse optado pela substituição do zero.

3.3.3 Seleção dos dados - *Feature selection*

Das 76 *features* disponíveis no banco de dados, apenas 21 foram utilizadas para o problema de regressão e 22 para o modelo de classificação. A este processo de redução e seleção das *features*, dá-se o nome de *feature selection*. Existem diversos benefícios com a *feature selection*, como facilitação de visualização e compreensão dos dados, redução da capacidade de armazenamento, redução do tempo de treinamento e o aumento da capacidade preditiva do modelo (GUYON; ELISSEEFF, 2003).

Na prática, foram selecionados todos as *features* explicadas na seção 3.2, com limitação do uso de todas as citadas nos testes situacionais, lógicos e psicométricos, sendo selecionadas apenas as principais, segundo a descrição da empresa, denominadas de "Potencial Bruto", "Raciocínio", "Social", "Motivacional"e "Cultura". Já a diferença de uma *feature* entre as duas modelagens realizadas, regressão e classificação, se dá justamente na variável de saída do problema de classificação, a qual não faz parte dos dados de entrada na modelagem de regressão. Ou seja, a variável "Permanece na empresa"só é utilizada para classificar.

3.3.4 Florestas aleatórias, Tunagem de hiper-parâmetros e Validação Cruzada

Em ambos os modelos foram utilizados os algoritmos de Florestas aleatórias através de bibliotecas de Python scikit-learn e dos dados coletados, tratados e selecionados, como citado a priori. No caso de regressão, foi utilizado o *RandomForestRegressor*, enquanto que para a classificação, o *RandomForestClassifier*. A decisão pela escolha destes métodos se deu pela ótima capacidade preditiva do modelo, sem os caráteres gulosos, já citados durante o capítulo 2, presentes nas árvores de decisão.

Enquanto que na regressão a métrica utilizada para verificar o erro do modelo foi o RMSE, para o modelo de classificação foi utilizado o AUC-ROC e o MCC. Como já citado, a tunagem dos hiper-parâmetros possui um papel fundamental na redução desse erro, no caso do RMSE e aumento, no caso da AUC-ROC e MCC. Assim, a fim de otimizar as métricas de avaliação dos modelos, utilizou-se a validação cruzada para encontrar os melhores parâmetros do *tunning*.

Foram rodadas 36.000 vezes o modelo, escondendo de forma ale-

atória os parâmetros $k - fold$ em cada um deles e utilizando a amostra de treinamento contendo 33% do banco de dados. Desta forma, o programa retornou todas as possíveis configurações dos hiper-parâmetros com seus respectivos erros ordenados do melhor para o pior cenário de erro, em cada uma das modelagens. A melhor configuração, intitulada de *"Best config"*, foi inserida nas florestas aleatórias para rodar a validação cruzada dos dados de teste e obter, no caso de regressão, a dispersão entre os valores preditos e os valores reais. Já no caso da classificação, este processo permitiu gerar a matriz de confusão, o gráfico AUC-ROC e o valor do MCC. Finalmente, os parâmetros otimizados foram rodados pela última vez com todo o banco de dados, para encontrar dois resultados, que serão discutidos no capítulo a seguir: a *feature importance* e a distribuição dos erros.

3.3.5 Análise de resultados

Após a determinação da melhor configuração dos hiper-parâmetros (*"Best config"*), ambos os modelos trouxeram como resultados a *feature importance* e a distribuição dos erros. A opção por utilizar a distribuição dos erros e não somente um valor de RMSE ou MCC e AUC-ROC, permite maior confiabilidade do modelo com intervalos de confiança determinados.

Assim, as análises dos resultados são compreendidas nas importâncias das *features* e do AUC-ROC e MCC, para classificação. Já no caso da regressão, além dessa importância de *features* e o RMSE obtido, também foram realizadas análises de cenários, baseadas nas características para entender qual a correlação dela, com a variável de saída, a média da performance, nesse caso. Por fim, para ambos os modelos, são comparadas as principais *features* de cada modelo, a fim de entender variações de resultados entre cada um.

4 RESULTADOS

Compreendidos os conceitos apresentados pelo referencial teórico e todo o processo de procedimentos metodológicos, passam-se a analisar as principais saídas fornecidas pelos modelos de predição, envolvidos no contexto da empresa de tecnologia da informação. Para tanto, o presente capítulo é esmiuçado em três tópicos distintos. O primeiro deles, tem como objetivo explorar o modelo de classificação, enquanto que o segundo, o de regressão. Destacam-se do primeiro modelo as análises das *feature importances*, a matriz de confusão, a tunagem dos hiper-parâmetros o MCC e o AUC-ROC, bem como seus intervalos de confiança. Já para o modelo de regressão, destacam-se também as *feature importances*, a tunagem dos hiper-parâmetros e o RMSE dentro do seu intervalo de confiança.

Ao final das duas seções, no terceiro e último tópico, as *feature importances* dos dois modelos serão comparadas, a fim de encontrar semelhanças e diferenças em cada um deles. Este capítulo permite melhor compreensão dos mecanismos de funcionamento do setor de vendas do mercado estudado.

4.1 RESULTADOS DO MODELO DE CLASSIFICAÇÃO

4.1.1 Matriz de confusão

Para a construção da matriz de confusão, foram utilizados os dados referentes ao teste da validação cruzada. Na figura 14 podem ser verificados os resultados obtidos, a partir da amostra teste de validação, os verdadeiros positivos, os verdadeiros negativos, os falsos positivos e os falsos negativos. A partir deles, foram obtidos os resultados de precisão, sensibilidade, especificidade, acurácia, média geométrica e F-*measure* do modelo, verificados na tabela 15.

Destacam-se os resultados de precisão e F-*measure* obtidos do modelo. Ambos apresentaram os melhores resultados, dentre os especificados na tabela 15, com resultados de 80,00% e 79,78%, respectivamente. A precisão indica que a porcentagem de predições positivas feitas pelo classificador está sendo correta, ou seja, estão bem apresentadas. Já o F-*measure* mostram um equilíbrio harmônico entre a precisão e a sensibilidade.

Figura 14: Resultados da matriz de confusão a partir dos dados de teste da validação cruzada.

Fonte: autor.

Tabela 15: Métricas obtidas pela matriz de confusão.

Tipos de métricas	**Valor obtido**
Precisão	80,00%
Sensibilidade	63,16%
Especificidade	63,16%
Acurácia	70,59%
Média geométrica	71,08%
F-*measure*	79,78%

Fonte: Autor.

Contudo, faz-se necessário entender os resultados em que o modelo apresentou alguma falha na predição. Os três falsos positivos, que se caracterizam como vendedores ativos, mas que o modelo previu que estariam fora da empresa. Além disso, também foram previstos que sete vendedores que não são inativos, como ativos pelo modelo, sendo esses, os falsos negativos, conforme evidenciado na 14.Dos três falsos positivos a principal dificuldade na interpretação do resultado está na limitação da variável de saída do modelo, ou seja, não se sabe se os vendedores saíram da empresa de maneira voluntária ou involuntária. Já para os sete falsos negativos, pode-se levantar o questionamento de que esses vendedores podem estar para serem mandados embora da empresa e, a empresa, pode tomar ações preventivas nesses casos. Na realidade, dificilmente um modelo terá capacidade preditiva perfeita e todo modelo está suscetível à falhas. Nesse caso em questão, a limitação do tamanho da base pode ter sido um dos fatores mais críticos para a limitação do poder preditivo.

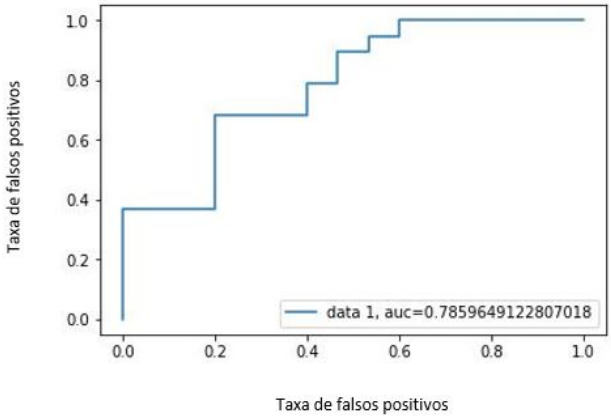
4.1.2 AUC-ROC e MCC

O AUC-ROC, foi dividido em duas análises. Na primeira delas, observou-se um valor de 0,786, conforme verificado na figura 15. O critério utilizado para a determinação desse valor, é o mesmo utilizado para determinar a matriz de confusão, ou seja, apenas com a amostra teste de 33%. Esse valor não demonstra um intervalo de confiança para o AUC-ROC, logo, surge a necessidade da segunda análise da métrica.

Como forma de obter maior confiabilidade na *area under the ROC curve*, calculou-se, para todo o conjunto de dados, o qual seria seu valor, dentro de um intervalo de 90% de confiança. A figura 16 demonstra essa distribuição, o qual, pelo intervalo determinado, está entre 0,591 e 0,828 o valor do AUC-ROC, com a média em 0,705, confirmando uma boa capacidade preditiva do modelo.

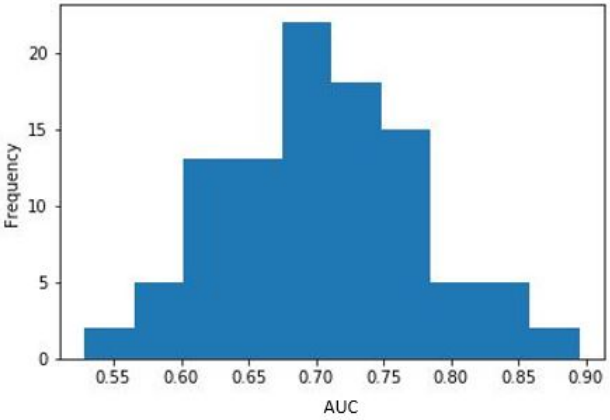
Também calculou-se o valor dentro de intervalo de confiança para o MCC. Sua distribuição pode ser verificada na figura 17. Com 90% de confiança, seus valores ficam entre 0,014 e 0,535, com média de 0,306.

Figura 15: Resultados do AUC-ROC a partir dos dados de teste da validação cruzada.



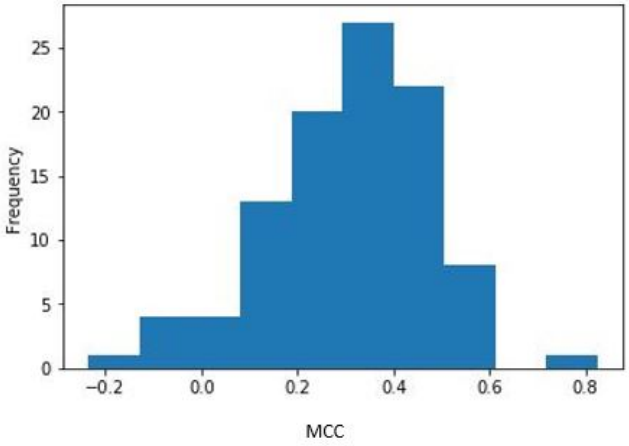
Fonte: autor.

Figura 16: Resultados da AUC-ROC para todo o banco de dados.



Fonte: autor.

Figura 17: Resultados da MCC para todo o banco de dados.



Fonte: autor.

4.1.3 Tunagem dos hiper-parâmetros - *Best config*

Para o processo de tunagem, foram levados em consideração cinco parâmetros do modelo, a fim de garantir a melhor configuração.

Tabela 16: Parâmetros de saída da tunagem, classificados pelo maior MCC.

Seq.	Max_ depth	Min_ split	Min_ leaf	Max_ features	ROC AUC	MCC
4292	44.0000	0.0644	0.1067	3	0.7471	0.4391
1360	10.6667	0.5000	0.1389	5	0.7600	0.4368
4867	44.0000	0.5000	0.1711	5	0.7476	0.4341
569	4.0000	0.4456	0.0422	5	0.7554	0.4284
778	10.6667	0.0644	0.0422	3	0.7495	0.4258

Fonte: Autor.

A melhor classificação, deu-se o nome de de "*Best config*", a qual possui os seguintes parâmetros ajustados: *max_depth*, *min_samples_split*, *min_samples_leaf*, *max_features* e *N_estimators*.

Com exceção do *N_estimators*, o qual é definido como 1000, todos os outros parâmetros de saída são ilustrados na tabela 16, por ordem decrescente do MCC referentes às cinco melhores sequências de classificação. Uma vez testado todas as possibilidades para os hiperparâmetros, foi escolhido o que trouxe a performance ótima para o modelo, referente à sequência 4292. Fez-se a escolha pelo melhor MCC, pois, de acordo com Chicco (2017), entre as métricas de avaliação de performance, o MCC é aquela que corretamente leva em consideração o tamanho da matriz de confusão, especialmente quando se trata de pequenos bancos de dados. Além disso, juntamente com o AUC-ROC, ambos são mais robustos e possuem estabilidade ao desequilíbrio de dados.

4.1.4 *Feature importance*

A lista de *features* que impactam de forma mais acentuada a predição da retenção de funcionários na empresa pode ser verificada na tabela 17. Nota-se que as seis primeiras *features* correspondem a aproximadamente 80% do total de importância. A média de entrega do vendedor, possui a maior importância na sua permanência na empresa, com 36,61% de importância.

A seguir, estão listadas as cinco principais *features* bem como suas características com relação à retenção do funcionário:

- **Média geral**: para a métrica mais importantes do modelo de classificação, notou-se que, funcionários ativos na empresa, pos-

suem médias de entregas maiores do que aqueles inativos. Enquanto a média dos que permanecem na empresa gira em torno de 81%, os que saíram, está em aproximadamente 61%. Fator esperado, uma vez que baixas performances pessoais culminam em demissões.

- **Tempo de formado**: funcionários que permanecem ativos na empresa possuem tempo de formado aproximadamente 1,5 anos menor do que os que já saíram.
- **Tipo de universidade**: existe maior concentração de vendedores ativos enquadrados na categoria de estudo em universidades públicas, com 52% da base total. Já para os inativos, existem cinco vezes mais funcionários nessa categoria que estudaram em universidades particulares do que os ativos.
- **Mudança de cidade**: observou-se que existe maior concentração de vendedores ativos que não mudaram de cidade para precisar trabalhar na empresa, em relação aos inativos. Complementando essa saída, para os vendedores que precisaram mudar de cidade, o índice de inativos chega a ser 50% maior do que os ativos, nesse caso.
- **Tecnologia**: o modelo demonstrou que não há necessidade do vendedor ter experiência no mercado de tecnologia para se manter ativo na função de vendedor.

4.2 RESULTADOS DO MODELO DE REGRESSÃO

4.2.1 RMSE

O modelo de regressão, além de ter como objetivo a encontrar as *feature importances* para predizer a performance dos vendedores, também tem como fim alcançar um bom erro esperado no modelo. Como forma de determinar o erro do modelo, utilizou-se o RMSE. Pelo erro estar em concordância com outros erros citados previamente, como o MSE e o MAE, os dois últimos não gerariam diferenças expressivas em relação ao primeiro.

Assim como no modelo de classificação, a determinação da métrica se deu em dois modelos distintos. O primeiro com os dados de teste de validação cruzada e o segundo com o intervalo de confiança

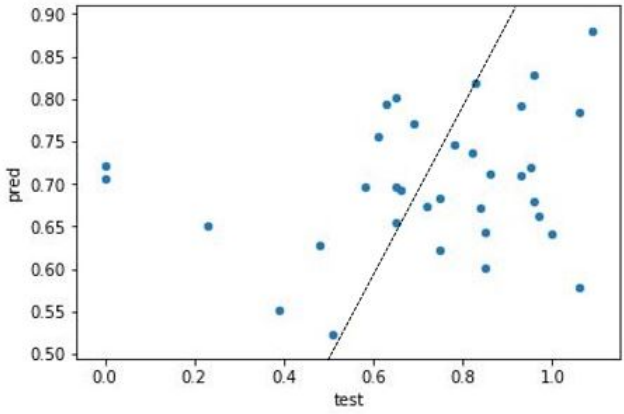
Tabela 17: *Feature importance* para o modelo preditivo de classificação.

Features	**Importância**
Média geral	0.366119
Tempo de formado (anos)	0.135121
Tipo de Universidade	0.109786
Houve mudança de cidade da última job?	0.097795
tecnologia	0.061471
Raciocínio	0.046108
Potencial Bruto	0.042175
Cultura pontuação	0.032617
intercambio	0.028288
Já possuia experiência com vendas?	0.027006
Motivacional	0.025910
Social	0.018603
aiesec e ej	0.009001

Fonte: Autor.

de 90% para todos os dados. A figura 18 representa a dispersão entre as predições realizadas pelo modelo em comparação com a amostra de testes utilizada na validação cruzada.

Figura 18: Gráfico de dispersão entre as predições realizadas e a amostra de teste.

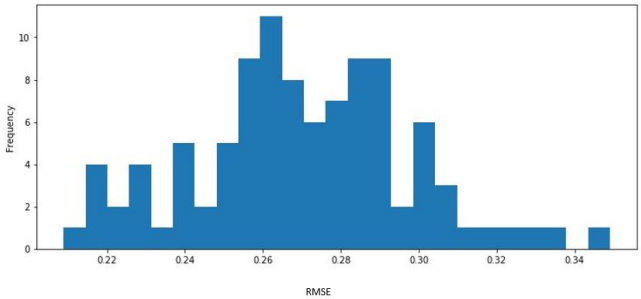


Fonte: autor.

Já a figura 19 representa a distribuição dos RMSE para a re-

gressão utilizando os parâmetros já tunados do modelo. Com 90% de confiança, o valor do erro se encontra entre 0,225 e 0,313, com média de 0,261.

Figura 19: Distribuição do RMSE para a melhor configuração.



Fonte: autor.

4.2.2 Tunagem dos hiper-parâmetros - *Best config*

Assim como no modelo de classificação, fez-se necessário obter as melhores configurações de parâmetros com o objetivo de minimizar o RMSE para o problema. De forma análoga ao que fora realizado na primeira modelagem, se encontrou os melhores parâmetros para a regressão, os quais também foram intitulados de *Best config*.

Como saída, a melhor configuração obteve um RMSE de 0,2609 e se caracterizou com os seguintes parâmetros durante a tunagem:

- $max\_depth = 44$;
- $min\_samples\_split = 0{,}173$;
- $min\_samples\_leaf = 0{,}01$;
- $max\_features = 3$;
- $N\_estimators = 1000$.

4.2.3 *Feature importance*

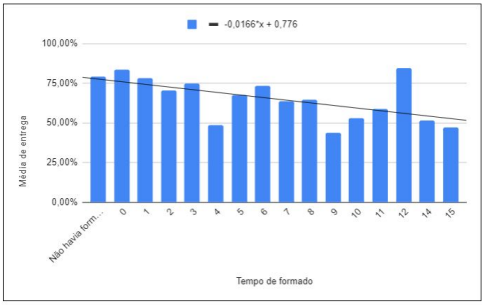
Na regressão, o resultado das *feature importances* foram mais homogêneos se comparados aos da classificação. Isto quer dizer que

não houve apenas uma característica com muito mais importância do que as demais, como a ”média geral”no primeiro modelo. Na tabela 18 verifica-se as treze principais características por ordem decrescente de importância.

Nota-se que, dentre as sete principais *features* descritas a seguir, cinco delas são referentes às notas de desempenhos pessoais.

- **Tempo de formado**: verificou-se que quanto maior o tempo de formado do vendedor no momento da sua contratação, menor tende a ser seu desempenho médio na empresa. Esta situação se ilustra na figura 20, com linha de tendência linear na distribuição.

Figura 20: Distribuição da média de entrega pelo tempo de formado.



Fonte: autor.

- **Tipo de universidade**: os desempenhos médios de vendedores que estudaram em universidades públicas giram em torno de 20% maior do que das universidades particulares. Com desempenhos médios de 79,7% e 66,4%, respectivamente.

- **Testes de cultura, de social, de raciocínio, de potencial e de motivacional**: Dos resultados de cada teste psicométrico realizado pela empresa no momento de processos seletivos, destacam-se os que demonstram resultados relacionados à cultura e ao social. Ambos estão diretamente relacionados com a média de entrega dos vendedores, ou seja, quanto maior seus resultados, maior também tende a ser a performance do vendedor. Os resultados podem ser verificados na figura 21. Com relação aos demais testes, raciocínio, potencial e motivacional, não possuem relação positiva com a performance. A principal interpretação desse resultado é que, provavelmente, essas competências não sejam essenciais para um funcionário de vendas no contexto espe-

Tabela 18: *Feature importance* para o modelo preditivo de regressão.

Features	**Importância**
Tempo de formado (anos)	0.174811
Tipo de Universidade	0.163294
Cultura pontuação	0.129299
Social	0.118457
Raciocínio	0.107870
Potencial Bruto	0.082358
Motivacional	0.061916
tecnologia	0.058213
Houve mudança de cidade da última job?	0.026032
intercambio	0.024414
aiesec e ej	0.019962
Primeiro emprego?	0.019492
Já possuia experiência com vendas?	0.013881

Fonte: Autor.

cificado. Contudo, podem ser características importantes para outras funções e cargos.

4.3 COMPARAÇÃO ENTRE OS MODELOS

Embora cada modelo tenha saídas diferentes, seus resultados estão atrelados. Isto quer dizer que, no melhor dos cenários, para um modelo de contratação de funcionários, faz-se necessário contratar funcionários que tenham boa performance e, concomitantemente, permaneçam ativos dentro da empresa. Assim, a presente subseção tem por objetivo comparar os resultados das *feature importances* de cada um dos modelos, a fim de obter a melhor combinação de cenários possível, equilibrando os dois modelos.

Para tal comparação, são expostas as principais *features* de cada modelo, com seus respectíveis níveis de importância lado a lado, conforme ilustrado na tabela 19.

Nota-se que, dentre as treze *features* levantadas para cada modelo, doze delas estão presentes nos dois. Além disso, a "Média geral"não poderia fazer parte do modelo de regressão, visto que, nessa modelagem, ela representa a variável de saída, Y. Outro ponto interessante observado, se refere à diferença entre as variáveis. Verifica-se que a maior diferença entre duas *features* presentes em ambos os modelos

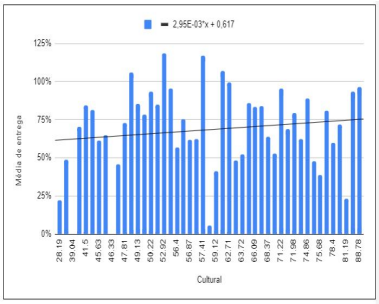
Tabela 19: Comparação entre as *feature importances* para os modelos de predição.

Features	**Importância regressão**	**Importância classificação**	**Diferença**
Tecnologia	0,0582	0,0615	0,0033
Intercâmbio	0,0244	0,0283	0,0039
AIESEC e EJ	0,0200	0,0090	0,0110
Experiência vendas	0,0139	0,0270	0,0131
Primeiro emprego	0,0195	-	-
Motivacional	0,0619	0,0259	0,0360
Tempo de formado	0,1748	0,1351	0,0397
Potencial Bruto	0,0824	0,0422	0,0402
Tipo de Universidade	0,1633	0,1098	0,0535
Raciocínio	0,1079	0,0461	0,0618
Mudança cidade	0,0260	0,0978	0,0718
Cultura pontuação	0,1293	0,0326	0,0967
Social	0,1185	0,0186	0,0999
Média geral	-	0,3661	-

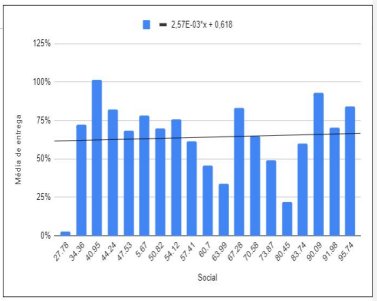
Fonte: Autor.

é de 0,0999, mostrando que, mesmo com objetivos distintos, há grande proximidade entre as importâncias das variáveis preditivas para cada modelo.

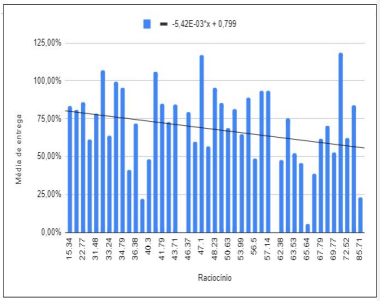
Figura 21: Distribuição das médias de entrega pelos resultados dos testes psicométricos.



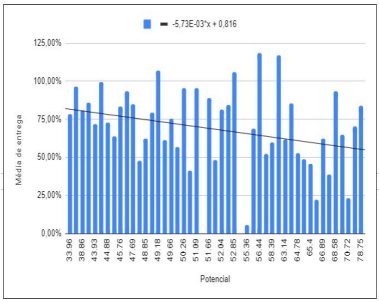
(a) Distribuição da média de entrega pelo teste de cultura



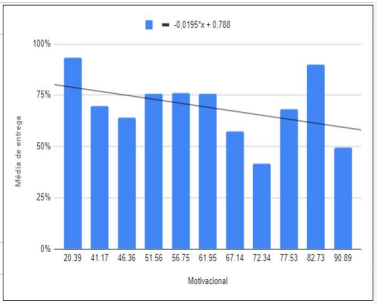
(b) Distribuição da média de entrega pelo teste de social



(c) Distribuição da média de entrega pelo teste de raciocínio



(d) Distribuição da média de entrega pelo teste de potencial



(e) Distribuição da média de entrega pelo teste de motivacional

Fonte: autor

5 CONCLUSÕES E RECOMENDAÇÕES

Com a crescente demanda das empresas por novas estratégias para melhorar seu desempenho, tem-se explorado cada vez mais a tecnologia como um dos meios para se obter esse tão esperado crescimento da performance. Contudo, ferramentas tecnológicas por si só não são suficientes para representar o total ganho de resultados. Especialmente no contexto das áreas de vendas de empresas de tecnologia da informação, em que se tem grandes dificuldades para contratar funcionários qualificados e também para retê-los no quadro das empresas. Assim, a harmonização entre o uso da tecnologia e informações disponíveis e a contratação de funcionários se torna um dos grandes desafios e necessidades para diversos negócios.

Neste contexto, faz-se necessário revisitar os objetivos propostos no presente projeto de pesquisa. Através do auxílio da aprendizagem máquina, a saída geral do trabalho, consistia na elaboração de modelos preditivos para auxiliar empresas de tecnologia a contratarem funcionários de vendas focando na performance e na retenção dos mesmos dentro da empresa, fora atingido. Para tal, levantou-se cinco saídas específicas que auxiliassem o atingimento do objetivo principal.

1. Definir os atributos e parâmetros a serem analisados em funcionários de vendas.
2. Aplicar modelos de aprendizado de máquina supervisionados, estruturados em classificação e regressão, em uma empresa de tecnologia brasileira.
3. Identificar o desempenho dos modelos propostos.
4. Classificar os principais atributos definidos, de modo a priorizá-los nas contratações.
5. Propor uma forma de implementação e acompanhamento do modelo para a empresa analisada, de forma que possa ser replicável em outros negócios.

Com a contextualização, problemática, objetivos definidos e esclarecidas e dados coletados, pode-se elaborar dois modelos de predição para o projeto. Embora cada modelo proposto contasse com saídas distintas, ambos se complementavam no geral. O primeiro modelo foi o de classificação, o qual objetivou-se levantar as principais *features* relacionadas com a retenção dos funcionários dentro da empresa. Já o segundo

modelo, o de regressão, também teve como objetivo levantar as *features* mais importantes, contudo, desta vez, relacionadas à performance dos vendedores dentro do contexto. Para tais modelos, utilizou-se dados de vendedores de uma empresa de tecnologia da informação em forte expansão no cenário brasileiro.

Uma vez que os modelos de aprendizado máquina estivessem rodando, fez-se necessário analisar os resultados encontrados. O primeiro passo, para se encontrar o melhor desempenho dos modelos foi tunar os hiper-parâmetros definidos, através de testes de validação cruzada e uso do maior AUC-ROC e MCC para classificação e menor RMSE para a regressão. Na classificação, tanto o AUC-ROC, quanto o MCC foram avaliados dentro de um intervalo de confiança de 90%, e se mantiveram dentro dos intervalos [0,59; 0,83] e [0,014;0,535], respectivamente. Embora o modelo tenha obtido boa capacidade preditiva, ele apresentou três casos de falsos positivos, caracterizando vendedores ativos, mas previstos como inativos e, além disso, sete falsos negativos, ou seja, vendedores que estão inativos, mas foram previstos como ativos.

Nos resultados encontrados para o modelo de regressão, adotou-se pela análise da métrica de RMSE para identificar o desempenho do modelo. De forma análoga à classificação, obteve-se um valor, com 90% de confiança entre [0,22.0,31], com destaque de que, neste caso, quanto menor o valor, melhor o desempenho do modelo. Com a performance do modelo validada, pode-se trabalhar com o quarto objetivo específico do projeto: classificar os principais atributos de desempenho dos modelos.

A priori, listou-se as treze principais *features* de cada um dos modelos com seus respectivos níveis de importância. Destaque para a "média geral de desempenho"na classificação e "Tempo de formado", "Tipo de formado"e as características psicométricas no modelo de regressão. Pode-se analisar como as principais *features* de cada modelo impacta os modelos, seja com maior retenção, ou com maior desempenho médio. Contudo, por se tratarem de modelos complementares, ou seja, a alta performance ter que andar junto com a alta retenção, comparou-se as treze *features* de cada modelo. Notou-se que doze, das treze *features* estavam presentes nos dois modelos e, além disso a maior diferença entre os níveis de importância da mesma *feature* para cada problema é de 0,0999, corroborando com o fator de proximidade entre os dois modelos.

Finalmente, o último objetivo específico busca ser contemplado no presente capítulo. Recomenda-se que empresas que busquem trabalhar com modelos preditivos na contratação de funcionários não migrem de forma abrupta dos seus modelos tradicionais de contratação

para o aqui apresentado. A recomendação é de que a empresa utilize as características mais importantes em grupos de teste e acompanhe a performance dos colaboradores em três ondas de implementação distintas: conhecimento das variáveis do modelo, aplicação em testes e, finalmente, aplicação em escala. O modelo não tem como objetivo excluir pessoas de processos seletivos, mas sim priorizar aquelas que possuem características que impactem positivamente sua performance. Também recomenda-se que a presente metodologia seja replicada em outros segmentos do mercado, seja em outras áreas do ambiente de tecnologia da informação, como marketing, pós-vendas e pré-vendas, por exemplo, ou, até mesmo, em setores mais tradicionais, como indústrias.

REFERĘNCIAS

ACHARYA, V.; SHARMA, S. K.; GUPTA, S. K. Analyzing the factors in industrial automation using analytic hierarchy process. *Computers & Electrical Engineering*, Elsevier, v. 71, p. 877–886, 2018.

ALEXE, C.-G.; ALEXE, C.-M. Similarities and differentiations at the level of the industries in acquiring an organizational culture in innovation. *Procedia Manufacturing*, Elsevier, v. 22, p. 317–324, 2018.

ARLOT, S.; CELISSE, A. et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, The author, under a Creative Commons Attribution License, v. 4, p. 40–79, 2010.

BARDENET, R. et al. Collaborative hyperparameter tuning. In: *International conference on machine learning*. [S.l.: s.n.], 2013. p. 199–207.

BARROS, P. *Aprendizagem de Maquina: Supervisionada ou Não Supervisionada?* 2016. Disponível em: <encurtador.com.br/lmtU5>. Acesso em: 10 jun. 2019.

BATISTA, G. E. d. A. P. et al. *Pré-processamento de dados em aprendizado de máquina supervisionado*. Tese (Doutorado) — Universidade de São Paulo, 2003.

BERGSTRA, J. S. et al. Algorithms for hyper-parameter optimization. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2011. p. 2546–2554.

BOTELHO, D.; TOSTES, F. D. Modelagem de probabilidade de churn. *RAE-Revista de Administração de Empresas*, v. 50, n. 4, p. 396–410, 2010.

BOUGHORBEL, S.; JARRAY, F.; EL-ANBARI, M. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, Public Library of Science, v. 12, n. 6, p. e0177678, 2017.

BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, Elsevier, v. 30, n. 7, p. 1145–1159, 1997.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. et al. Czussificu-tion and regression trees. wadsworth international. *Group*, 1984.

BROWN, J. Employee turnover costs billions annually. *Computing Canada Willowdale*, v. 26, p. 25, 2000.

CHEESEMAN, P. et al. Autoclass iii. *Program available from NASA Ames Research Center: Research Institute for Advanced Computer Science*, 1990.

CHEN. *Basic Ensemble Learning (Random Forest, AdaBoost, Gradient Boosting)- Step by Step Explained*. 2019. Disponível em: <encurtador.com.br/fLQX5>. Acesso em: 19 ago. 2019.

CHIAVENATO, I. *Recursos humanos: edição compacta*. [S.l.]: Ed. Atlas, 1983.

CHICCO, D. Ten quick tips for machine learning in computational biology. *BioData mining*, BioMed Central, v. 10, n. 1, p. 35, 2017.

COATES, A.; NG, A.; LEE, H. An analysis of single-layer networks in unsupervised feature learning. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. [S.l.: s.n.], 2011. p. 215–223.

COOMBS, B. *Filling the Skills Gap: Health Care Looks for More Data Workers*. 2013. Disponível em: <https://www.cnbc.com/id/100761846>. Acesso em: 31 mai. 2019.

COX, J. *Here's What's Really Holding Back the US Jobs Market*. 2012. Disponível em: <https://www.cnbc.com/id/49331167>. Acesso em: 31 mai. 2019.

CUFFA, D. de; FLORIANI, E. V.; STEIL, A. V. Turnover de profissionais de vendas: o caso de uma organização de tecnologia da informação. In: *International Congress of Knowledge and Innovation-Ciki*. [S.l.: s.n.], 2018. v. 1, n. 1.

DARMON, R. Y. Controlling sales force turnover costs through optimal recruiting and training policies. *European Journal of Operational Research*, Elsevier, v. 154, n. 1, p. 291–303, 2004.

DATA SCIENCE ACADEMY. *17 casos de uso de machine learning*. 2018. Disponível em: <http://datascienceacademy.com.br/blog/17-casos-de-uso-de-machine-learning/>. Acesso em: 08 ago. 2019.

DAVIDSON, A. *HR Staffs, Recruiters Overlook Qualified Job Seekers*. 2012. Disponível em: <https://www.cnbc.com/id/48081039>. Acesso em: 31 mai. 2019.

DONGES. *Aprendendo em uma floresta aleatória*. 2018. Disponível em: <encurtador.com.br/cpKRU>. Acesso em: 11 jun. 2019.

DRAKOS. *Cross-Validation*. 2018. Disponível em: <https://towardsdatascience.com/cross-validation-70289113a072>. Acesso em: 22 ago. 2019.

DRUCKER, P. *Management challenges for the 21st century*. [S.l.]: Routledge, 2012.

EGAN, J. *Signal detection theory and ROC analysis. Series in Cognition and Perception. 1975*. [S.l.]: Academic Press, New York, 1975.

FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics New York, 2001.

GARCIA, S. C. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. 2003.

GHAPANCHI, A. H.; AURUM, A. Antecedents to it personnel's intentions to leave: A systematic literature review. *Journal of Systems and Software*, Elsevier, v. 84, n. 2, p. 238–249, 2011.

GIL, A. C. Como elaborar projetos de pesquisa. *São Paulo*, v. 5, n. 61, p. 16–17, 2002.

GOMES, B. M. V. *Previsão de churn em companhias de seguros*. Tese (Doutorado), 2011.

GÖNEN, M. *Analyzing receiver operating characteristic curves with SAS*. [S.l.]: SAS Institute, 2007.

GORODKIN, J. Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, Elsevier, v. 28, n. 5-6, p. 367–374, 2004.

GUIMARÃES, W. S. A. et al. Data mining aplicado ao serviço público, extração de conhecimento das ações do ministério público brasileiro. Florianópolis, SC, 2000.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research*, v. 3, n. Mar, p. 1157–1182, 2003.

HAMILTON, R.; DAVISON, H. K. The search for skills: Knowledge stars and innovation in the hiring process. *Business Horizons*, Elsevier, v. 61, n. 3, p. 409–419, 2018.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, v. 143, n. 1, p. 29–36, 1982.

HAUSKNECHT, J. P.; RODDA, J.; HOWARD, M. J. Targeted employee retention: Performance-based and job-related differences in reported reasons for staying. *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management*, Wiley Online Library, v. 48, n. 2, p. 269–288, 2009.

HELMBOLD, D. P.; SCHAPIRE, R. E. Predicting nearly as well as the best pruning of a decision tree. *Machine Learning*, Springer, v. 27, n. 1, p. 51–68, 1997.

HEMMANS, D. *Comparing salary versus non-salary factors affecting the voluntary turnover of IT sales professionals*. [S.l.]: University of Phoenix, 2010.

HUDSON, C. A. Computers in manufacturing. *Science*, American Association for the Advancement of Science, v. 215, n. 4534, p. 818–825, 1982.

KATO, J. M.; PONCHIROLLI, O. O desemprego no brasil e os seus desafios éticos. *Revista da FAE*, v. 5, n. 3, 2002.

KOEHRSEN. *Hyperparameter Tuning the Random Forest in Python*. 2018. Disponível em: <encurtador.com.br/hnQXZ>. Acesso em: 21 ago. 2019.

KUBAT, M.; MATWIN, S. et al. Addressing the curse of imbalanced training sets: one-sided selection. In: NASHVILLE, USA. *Icml*. [S.l.], 1997. v. 97, p. 179–186.

LAUREANO, R.; CAETANO, N.; CORTEZ, P. Previsão de tempos de internamento num hospital português: aplicação da metodologia crisp-dm. *RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação*, Associação Ibérica de Sistemas e Tecnologias de Informação (AISTI), n. 13, p. 83–98, 2014.

LIAW, A.; WIENER, M. et al. Classification and regression by randomforest. *R news*, v. 2, n. 3, p. 18–22, 2002.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge & Data Engineering*, IEEE, n. 4, p. 491–502, 2005.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, Elsevier, v. 405, n. 2, p. 442–451, 1975.

MAURER, R. *HR Is Turning to Freelancers to Meet Talent Shortage*. 2017. Disponível em: <encurtador.com.br/knuV8>. Acesso em: 31 mai. 2019.

MITCHELL, T. M. Generalization as search. *Artificial intelligence*, Elsevier, v. 18, n. 2, p. 203–226, 1982.

MOHTADI, B. *In Depth: Parameter tuning for Random Forest*. 2017. Disponível em: <encurtador.com.br/DIR29>. Acesso em: 21 ago. 2019.

MOISEN, G. Classification and regression trees. *In: Jørgensen, Sven Erik; Fath, Brian D.(Editor-in-Chief). Encyclopedia of Ecology, volume 1. Oxford, UK: Elsevier. p. 582-588.*, p. 582–588, 2008.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003.

MORABITO, R. et al. *Metodologia de pesquisa em engenharia de produção e gestão de operações*. [S.l.]: Elsevier Brasil, 2018.

MORESI, E. et al. Metodologia da pesquisa. *Brasília: Universidade Católica de Brasília*, v. 108, p. 24, 2003.

MURTHY, S.; SALZBERG, S. Lookahead and pathology in decision tree induction. In: CITESEER. *IJCAI*. [S.l.], 1995. p. 1025–1033.

NGUYEN, G. H.; BOUZERDOUM, A.; PHUNG, S. L. Learning pattern classification tasks with imbalanced data sets. In: *Pattern recognition*. [S.l.]: IntechOpen, 2009.

PARISE, S.; CROSS, R.; DAVENPORT, T. H. Strategies for preventing a knowledge-loss crisis. *MIT Sloan Management Review*, Massachusetts Institute of Technology, Cambridge, MA, v. 47, n. 4, p. 31, 2006.

PERES, S. M. et al. Tutorial sobre fuzzy-c-means e fuzzy learning vector quantization: abordagens híbridas para tarefas de agrupamento e classificação. *Revista de Informática Teórica e Aplicada*, v. 19, n. 1, p. 120–163, 2012.

PINKOVITZ, W. H.; MOSKAL, J.; GREEN, G. How much does your employee turnover cost. In: *Small Business Forum*. [S.l.: s.n.], 1997. v. 14, n. 3, p. 70–71.

PINTO, N. et al. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computational biology*, Public Library of Science, v. 5, n. 11, p. e1000579, 2009.

PISTONO, F. *Os robôs vão roubar seu trabalho, mas tudo bem: Como sobreviver ao colapso econômico e ser feliz*. [S.l.]: Editora Companhia das Letras, 2017.

RACHWAŁ, T. Industrial restructuring in poland and other european union states in the era of economic globalization. *Procedia-Social and Behavioral Sciences*, Elsevier, v. 19, p. 1–10, 2011.

RIBEIRO, E. P. Fluxo de empregos, fluxo de trabalhadores e fluxo de postos de trabalho no brasil. *Brazilian Journal of Political Economy*, SciELO Brasil, v. 30, n. 3, p. 401–419, 2010.

ROSENBUSCH, N.; BRINCKMANN, J.; BAUSCH, A. Is innovation always beneficial? a meta-analysis of the relationship between innovation and performance in smes. *Journal of business Venturing*, Elsevier, v. 26, n. 4, p. 441–457, 2011.

SAMUELS, M. Management-flexibility is the key to keeping parents in it. how can it employers do more to retain the skills of working parents. *Computing*, p. 26–31, 2005.

SCHUSTER, M. E. Mercado de trabalho de tecnologia da informação: o perfil dos profissionais demandado. 2008.

SHANNON, C. E.; WEAVER, W. The mathematical theory of communication–univ. *Illinois press, Urbana, I*, v. 11, p. 117, 1949.

SHUKLA, A.; SRIVASTAVA, R. Meta analysis of the relationship between emotional intelligence and different behavioral intentions. *Research Journal of Business Management*, v. 10, n. 4, p. 58–73, 2016.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. Practical bayesian optimization of machine learning algorithms. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 2951–2959.

SRIVASTAVA, T. *Tuning the parameters of your Random Forest model*. 2015. Disponível em: <encurtador.com.br/foxLY>. Acesso em: 21 ago. 2019.

THATCHER, J. B.; LIU, Y.; STEPINA, L. P. The role of the work itself: an empirical examination of intrinsic motivation's influence on it workers attitudes and intentions. In: ACM. *Proceedings of the 2002 ACM SIGCPR conference on Computer personnel research*. [S.l.], 2002. p. 25–33.

THORNTON, C. et al. Auto-weka: Automated selection and hyper-parameter optimization of classification algorithms. *CoRR, abs/1208.3719*, 2012.

TURBAN, E. et al. *Tecnologia da Informação para Gestão-: Transformando os Negócios na Economia Digital*. [S.l.]: Bookman, 2010.

VOSS, C. A. Success and failure in advanced manufacturing technology. *International Journal of Technology Management*, Inderscience Publishers, v. 3, n. 3, p. 285–297, 1988.

WAGNER, J. *Comportamento organizacional-criando vantagem competitiva*. [S.l.]: Editora Saraiva, 2017.

WALDMAN, J. D.; ARORA, S. Measuring retention rather than turnover: a different and complementary hr calculus. *Human Resource Planning*, v. 27, n. 3, 2004.

WANG, Z.; BOVIK, A. C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, IEEE, v. 26, n. 1, p. 98–117, 2009.

WEISS, G. M. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, ACM, v. 6, n. 1, p. 7–19, 2004.

WEISS, S. M.; KULIKOWSKI, C. A. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. [S.l.]: Morgan Kaufmann Publishers Inc., 1991.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, v. 30, n. 1, p. 79–82, 2005.

WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016.

ZHANG, C.; MA, Y. *Ensemble machine learning: methods and applications*. [S.l.]: Springer, 2012.