RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

PREVENDO A TAXA DE CONVERSÃO DE LEADS NO SETOR DA EDUCAÇÃO COM TÉCNICAS DE APRENDIZAGEM DE MÁQUINA

PREDICTING LEAD CONVERSION RATE IN THE EDUCATION SECTOR WITH MACHINE LEARNING TECHNIQUES

Oberdan Santos da Costa[1], Luis Borges Gouveia[2]

Submetido em: 31/08/2021 e29712
Aprovado em: 10/10/2021 https://doi.org/10.47820/recima21.v2i9.712

RESUMO
A crise provocada pela COVID-19 acelerou processos de mudanças na economia global, levando a alterações nas empresas em estruturas, modelo de negócios e rotinas. Particularmente, Pequenas e Medias Empresas (PMEs) têm enfrentado desafios de encontrar caminhos para a jornada de transformação digital e adaptação na era da indústria 4.0, o que as leva a precisar de apoio para integrar suas transformações. O objetivo deste trabalho é prever a probabilidade de conversão de leads usando Aprendizagem de Máquina (ML) com o propósito de melhorar o processo das oportunidades de fechamento de matrículas nas PMEs do setor da educação. O trabalho tem fundamentação no Modelo de Transformação Digital para as PMEs (MTD_PMEs), abordagem específica na tecnologia ML e *Knowledge Discovery in Database* (KDD). A metodologia envolve uma sequência de três etapas do processo de KDD_AZ. Os dados foram coletados de um polo de uma universidade do sul do Brasil. Resultados indicam que os 8 atributos utilizados são significativos para prever a conversão de leads. A técnica de ML, Regressão Logística, chegou a uma precisão bruta de 100%, contribuindo assim para o aumento da taxa de conversão, ganho de tempo das equipes e filtragem de leads "improváveis", e ainda ajuda o marketing a melhorar sua mira para trazer leads qualificados/quentes.

PALAVRAS-CHAVE: Prevenção. Educação. Transformação digital. Aprendizagem de máquina. KDD

ABSTRACT
The crisis caused by COVID-19 accelerated processes of change in the global economy, leading to changes in companies' structures, business models and routines. Small and Medium Enterprises (SMEs) in particular have faced challenges of finding paths for the journey of digital transformation and adaptation in the industry 4.0 era, which makes them need support to integrate their transformations. The objective of the work is to predict the probability of conversion of leads using Machine Learning (ML) in order to improve the process of closing enrollment opportunities in SMEs in the education sector. The work is based on the Digital Transformation Model for SMEs (MTD_SMEs), specific approach in ML technology and Knowledge Discovery in Database (KDD). The methodology involves a three-step sequence of the KDD_AZ process. Data were collected from a university center in southern Brazil. Results indicate that the 8 attributes used are significant to predict lead conversion. The ML technique, Logistic Regression reached a raw accuracy of 100%, thus contributing to an increase in the conversion rate, saving time for teams and filtering "unlikely" leads, and also helps marketing improve its aim to bring qualified/hot leads.

KEYWORDS*: Prevention. Education. Digital transformation. Machine learning. KDD*

[1] Doutor em Ciência da Informação pela Universidade Fernando Pessoa em Porto-Portugal. Mestrado em GESTÃO DE EMPRESAS pela Universidade Lusófona de Humanidades e Tecnologias (2014-2015) Em Lisboa-Portugal. MBA Executivo em Gestão Empresarial pelas Faculdades de Ciências Gerenciais da Bahia (2011-2013). Especialização em Formação de Consultores Organizacionais - FCO pelo ISAN-FGV (2007), Especialização em gestão empresarial pelo ISAN-FGV (2003).

[2] Professor Catedrático na Universidade Fernando Pessoa. Publicou 66 artigos em revistas especializadas e 170 trabalhos em actas de eventos, possui 57 capítulos de livros e 17 livros publicados. Participou em 65 eventos no estrangeiro e 53 em Portugal. Orientou 8 teses de doutoramento e coorientou 2, orientou 21 dissertações de mestrado e coorientou 2. Actua nas áreas de Engenharia e Tecnologia com ênfase em Engenharia Electrotécnica, Electrónica e Informática e Ciências Exactas com ênfase em Ciências da Computação e da Informação

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

INTRODUÇÃO

Há pelo menos dois movimentos de mudança nas organizações provocados pela crise da COVID-19 que se alinham e adaptam à Era da Industria 4.0. O primeiro, diz respeito a um desencadeamento de pontos de inflexão nas organizações, levando seus líderes a refletir o atual cenário, tendo em vista que esse pode ser um bom momento para avaliarem seu atual modelo de negócios contra oportunidades emergentes de crescer, melhorar resultados e potencialmente adaptá-lo à Era da Industria 4.0. Para Ulas (2019), quando as mudanças no mundo dos negócios são revisadas para o resultado da Indústria 4.0, chama-se a atenção que a manufatura e processo de fabricação se torna prático. Nesse sentido, os estudos de Mirković, Lukić & Martin (2019) verificam que cada organização, independentemente do tamanho, mercado, história, tradição, localização, número de funcionários, número de produtos, número de clientes, percebe que precisa ser mais "digital" no sentido de agilidade, flexibilidade e capacidade de resposta a mudanças necessidades e expectativas dos clientes.

O segundo movimento, traz uma aceleração de múltiplos processos nas empresas em estruturas, modelo de negócios e rotinas, independentemente do tamanho, em diversos setores da economia global, entre eles os Processos de Aplicação de Tecnologias Digitais (PATDs). Os PATDs, envolvem o uso de muitos recursos tecnológicos (Business analytics, Aprendizagem de máquina, IoT, Inteligência Artificial, Big Data, dispositivos móveis, impressoras 3D, sensores inteligentes, realidade aumentada, virtualização, segurança cibernética, sistemas avançados de robótica e computação em nuvem e etc.) para várias finalidades, incluindo análise de dados, tomada de decisão e gestão empresarial. Em geral, esses processos têm como foco aumentar as capacidades das empresas, aprimorar as experiências e entregar valor para os clientes.

Com base no Modelo de Transformação Digital para as PMEs (MTD_PMEs), abordagem específica no recurso tecnológico método supervisionado de Aprendizagem de Máquina (ML), o estudo tem como objetivo prever a probabilidade de conversão de leads usando técnicas de aprendizagem de máquina, com o propósito melhorar o processo de conversão de oportunidades geradas na captação de leads nas PMEs do setor da educação. O estudo, justifica-se em pelo menos dois pontos. O primeiro, revela um conjunto de variáveis significativas que permitem apoiar a probabilidade de conversão de um lead durante uma conversa, pautada em dados da pré-inscrição. O segundo, eleva o ganho de tempo das equipes de conversão, tendo em vista que, os resultados dos leads filtrados como "improváveis de conversão", ajudam o marketing a melhorar sua mira e trazer leads qualificados, ou seja, quentes. O problema em foco neste estudo está em que medida as técnicas de aprendizagem de máquina contribuem para o aumento da taxa de conversão, ganho de tempo das equipes e filtragem de leads “improváveis.

O presente trabalho está estruturado da seguinte forma: na segunda seção é mostrado o referencial teórico. Na terceira seção, descreve a metodologia empregada. A seção 4 apresenta os resultados e discursões. Finalmente, seção 5 descreve as principais conclusões.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

1 REFERENCIAL TEÓRICO

Muitos modelos profissionais e acadêmicos foram desenvolvidos para a transformação digital. Este item do trabalho, além de fazer referências sobre o Modelo de Transformação Digital para as PMEs (MTD_PMEs), trata da aprendizagem de máquina e Modelo de descoberta de conhecimento ou Modelo KDD_AZ em transformação.

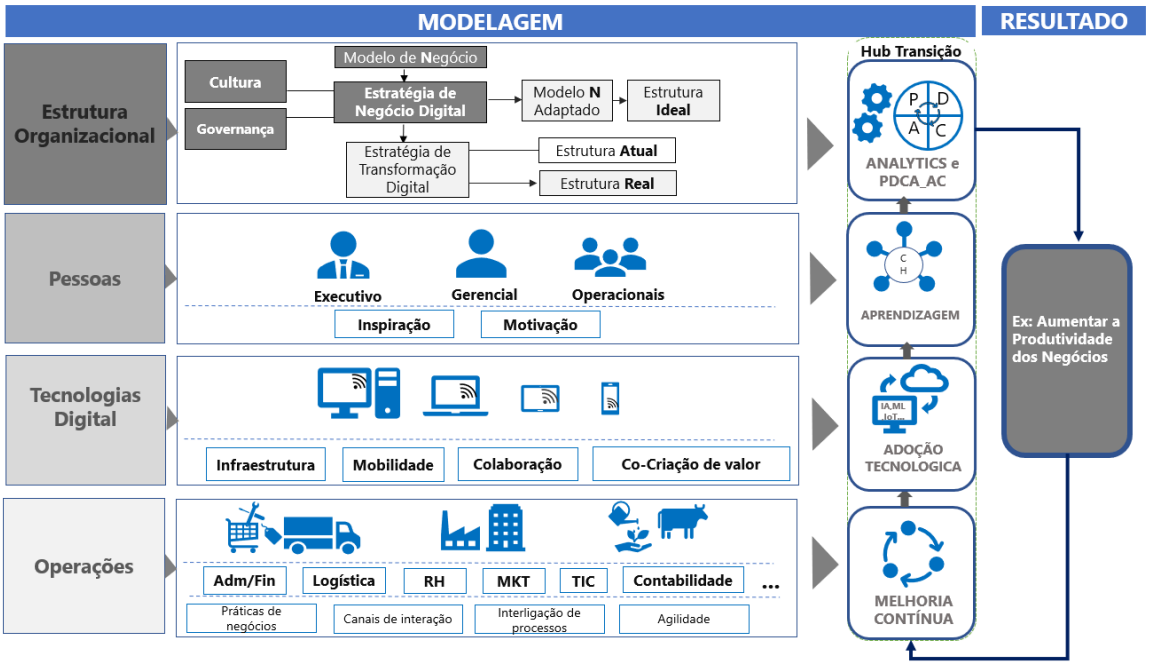
1.1 Modelo de Transformação Digital para as PMEs (MTD_PMEs)

O Modelo de Transformação Digital para as PMEs (MTD_PMEs), utilizado neste trabalho, usa dupla abordagem metodológica de caráter descritivo-qualitativo-comparativo e prescritivo-qualitativo. Parte desse modelo é representado pelo modulo modelagem e resultado – Figura 1. De acordo com Costa & Gouveia (2021) o MTD_PMEs, permite avaliar a maturidade digital, orientar e recomendar caminhos as empresas em sua jornada da transformação digital na era da indústria 4.0 por meio de dimensões consistentes, incluindo estrutura organizacional, pessoas, tecnologias (recursos tecnológicos) e operações em quatro níveis de avaliação, permitindo que as PMEs aprimorem seus negócios a um custo relativamente baixo em muitos setores.

Em seus estudos, Calipinar & Ulas (2013) resumem que as tecnologias digitais se tornam parte essencial da criação de valor para uma organização. Nessa mesma linha tecnológica Almeida, Cavalcante & Fettermann (2017) apontam que a dimensão tecnologia é a mais frequente nos modelos de maturidade analisados, sendo orientada para a mensuração do nível de adoção das tecnologias referentes à Indústria 4.0. Nesse sentido Costa & Gouveia (2021) corroboram e apontam que um reflexo positivo da atitude de uso tecnológico de uma empresa estar na adoção de tecnologias digitais acompanhada de sua exploração.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Figura 1 – Módulo Modelagem e Resultado



Fonte: Elaborado pelos autores

1.2 Aprendizagem de máquina (ML)

O aprendizado de máquina é a arte de investigar como os computadores podem aprender automaticamente, reconhecer padrões complexos e tomar decisões inteligentes com base nos dados. Por exemplo, um problema típico de aprendizado de máquina é programar um computador para que ele possa prever futuros comportamentos de compra de grupos de clientes após aprender com um conjunto de exemplos. Para Mitchell (1997) diz-se que um programa de computador aprende pela experiência E em relação a algum tipo de tarefa T e alguma medida de desempenho P se seu desempenho em T, conforme medido por P, melhora a experiência E.

O ML, é um subcampo da Inteligência Artificial (IA). No entanto, ambos os termos são chamados, simultaneamente e repetidamente, sempre que os tópicos de Big Data ou Data Analytics, ou alguns outros tópicos relacionados são discutidos.

Apesar do ML ser um recurso tecnológico poderoso e de grande influência para adquirir conhecimento automático, não existe um algoritmo exclusivo, capaz de resolver todos os problemas com maior desempenho sobre outros. Cada algoritmo possui limites e afincamento nos mais variados problemas a ele submetido (MONARD & BARANAUSKAS, 2003).

São várias as situações que envolvem aprendizagem de máquina. Em linhas gerais, quando tratamos de aprendizagem de máquina, faz-se necessário a compreensão de um ambiente lógico, ou seja, de um algoritmo, o qual é normalmente entendido como uma sequência de raciocínios, instruções ou operações para alcançar um objetivo**,** sendo necessário que os passos sejam finitos e operados

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

sistematicamente. O algoritmo aprende a partir de um grande volume de dados, ou seja, – suas experiências, para atingir um objetivo, que em geral, requer uma sequência de processos.

Existem quatro modos de aprendizagem usados na aprendizagem de máquina, cada um com características úteis para resolver tarefas diferentes: supervisionado, não supervisionado, semi-supervisionado, e aprendizagem por reforço como mostrado na Figura 2.

Figura 2 – Modos de aprendizagem de máquina



Fonte: Elaborado pelos autores

1.2.1 Métodos de ML supervisionado

Os métodos de ML supervisionados de ML são utilizados no desenvolvimento de sistemas inteligentes capazes de adquirir conhecimento de forma automática por meio da análise de um conjunto de dados (Ghotra, McIntosh & Hassan, 2015). De acordo com Machado (2011) algoritmos de ML podem ser vistos como mecanismos que extraem um padrão de comportamento a partir de experimentações. Essa experiência ocorre com a alimentação de dados e informações colhidas a partir de interações com o mundo real. Visto isto, faz-se necessário uma boa estratégia para que o modelo a ser desenvolvido possa aprender e melhorar por meio da observação dessas interações. Nesse sentido, Batista (2019) corrobora e aponta que algoritmos de ML são ferramentas computacionais poderosas para aquisição de conhecimento a partir da experiência. A autora destaca ainda, que o uso de técnicas com algoritmos de aprendizagem de máquina reduz o tempo computacional, removendo informações redundantes e irrelevantes, e que, a modo mais comum no ML é o Aprendizado Supervisionado (AS), cujo objetivo é aprender um modelo preditivo a parti de um conjunto de dados. As tarefas do modo AS mais comuns são regressão e classificação. A tarefa regressão envolve previsão de dados numéricos. Essa é entendida segundo Menezes, Farrar & Bower (2018) como um modelo de equação linear que descreve a relação entre duas ou mais variáveis. A tarefa de classificação envolve predizer a qual categoria pertence um exemplo. Para Han, Kamber & Pei (2011) aprendizagem supervisionada é basicamente sinônimo de

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

classificação. Uma aplicação importante do modo AS são os problemas de classificação, comumente encontrados na área da saúde, industrial, educação entre outras.

Um bom exemplo está na força sobrenatural do marketing para gerar leads para as equipes atuarem na conversão de desses leads, no entanto o que se ver são muitas perdas de oportunidades de conversão de leads em matrículas. Agora, imagine que você recebe do marketing uma lista de leads captados para fazer contato e queira classificar (prever) automaticamente a probabilidade de cada lead fechar o negócio (matrícula) ao final da conversa. Afinal, isso é possível e como? A resposta é sim e com o uso de técnicas de ML. Numa abordagem prática, com uso de ML supervisionado, para classificar (prever) uma determinada probabilidade de algo acontecer, deve-se efetuar pelo menos três etapas básicas (1) Obter um conjunto de dados e dividi-lo em treinamento, validação e teste; (2) usar os dados de treinamento (conjuntos de exemplos com entradas/exemplos, rotuladas com a saída esperada) e validação para informar um modelo de relação entre recursos/atributos e o alvo, obtendo a representação do conhecimento por meio desses exemplos; e (3) avaliar o desempenho do modelo com o conjunto de dados de teste (dados que o algoritmo nunca viu antes) para determinar como tão bem, ele classifica.

Existem vários classificadores (algoritmos) e suas respectivas famílias com variável de saída qualitativa ou categórica. O classificador refere-se a um conjunto de métodos de ML para prever uma classe ou categoria de indivíduos com base em um ou múltiplas variáveis preditoras (KASSAMBARA, 2017). Com base na literatura Wang, Wang, Huang, Wang & Liu (2009), Biagioni, Gerlich, Merrifield & Eriksson (2011), Altinkaya & Zontul (2013), Kormáksson, Barbosa, Viera & Zadrozny (2014), Tan & Shao (2015), Neves, Dias & Cordeiro (2018) observam que os classificadores dos métodos de ML mais utilizados são Árvore de Decisão, *Random forest*, K-Nearest *Neighbour*, Regressão logística, Regressão linear, Redes Neurais e Máquinas de Vetores Suporte.

1.2.2 Métodos de ML não supervisionado

Essa abordagem de aprendizado parece ser muito mais difícil, pois visa fazer com que o computador aprenda e faça algo que não lhe ensinamos como fazer. Uma característica desse método é que os exemplos de entrada não possuem classe rotulado. Nesse método os algoritmos buscam agrupar os exemplos históricos com base em características semelhantes que eles apresentam entre si, o que a torna poderosa por não assumir nenhuma classificação pré-descoberta de exemplos. Em síntese, exemplos com características mais semelhantes tendem a ficar no mesmo grupo, enquanto exemplos com características diferentes tendem a ser organizados em grupos distintos. O método tem como entrada um conjunto de exemplos, cada um descrito por um vetor de valores de atributo, mas sem rótulos de classe. Como resultado a saída é um conjunto de dois ou mais grupos de exemplos. Para Han et al. (2011) a aprendizagem não supervisionada é essencialmente um sinônimo de agrupamento. As tarefas mais comuns de aprendizagem não supervisionada são agrupamento, associação e detecção de anomalias. Essas tarefas e seus respectivos algoritmos de acordo com Heller, Williamson & Ghahramani

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

(2008) são projetadas para extrair estrutura de amostras de dados. A qualidade dessa estrutura é medida por uma função de custo que geralmente é minimizada para inferir parâmetros ótimos que caracterizam a estrutura oculta nos dados.

1.2.3 Métodos de ML semi-supervisionado

Esse modo de aprendizado de máquina usa um conjunto de observações/exemplos rotulados e não rotulados ao aprender um modelo. Isso permite a ele utilizar dois conjuntos para encontrar uma hipótese capaz de classificar novos exemplos entre as classes já existentes. Em uma abordagem simples, exemplos rotulados são usados para aprender modelos de classe e exemplos não rotulados são usados para refinar os limites entre as classes.

O setor da medicina, por exemplo tem feito muito uso desse aprendizado, particularmente na área de imagem, onde um médico pode rotular um subconjunto de dados de imagens e usá-las para treinar um modelo. O modelo treinado classifica o resto das imagens não rotuladas no conjunto de dados. Nesse sentido, de acordo com a NVIDIA (2018) o conjunto de dados rotulado resultante é então usado para treinar um modelo de trabalho que deve, em teoria, superar modelos não supervisionados.

1.2.4 Métodos de ML por esforço

O modo de aprendizagem por esforço traz uma abordagem diferente dos demais modos de aprendizagem, isso porque o método treina um algoritmo com base no conhecimento sobre o estado do indivíduo no ambiente, no qual está inserido. Em seus estudos, Oh et al. (2021) afirmam que a aprendizagem por reforço tem um objetivo maximizar as recompensas cumulativas esperadas (ou recompensas médias), que é simples, mas geral o suficiente para capturar muitos aspectos da inteligência. Eles destacam ainda que embora o objetivo da aprendizagem por esforço simples, desenvolva algoritmos eficientes para otimizar tal objetivo, em geral envolve muito esforço de pesquisa, desde a construção de teorias até investigações empíricas.

1.3 Modelo de descoberta de conhecimento ou Modelo KDD_AZ em transformação

O modelo KDD_AZ em transformação ou somente KDD_AZ, exposto na Figura 3 é uma adaptação da estrutura básica do *Knowledge Discovery in Database* (KDD) de Fayyad, Piatetsky-Shaporo e Smyth (1996) pelos autores. A sequência de três etapas do processo de KDD_AZ compreende: pré-processamento, mineração de dados e pós-processamento, cada uma com suas respectivas fases de operações de tarefas (impactos x complexidade).

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Figura 3 – Modelo KDD_AZ adaptado de Fayyad et al. (1996)



Fonte: Elaborado pelos autores

As três etapas do processo KDD_AZ, incluem: (1) pré-processamento que compreende as funções que se relacionam a obtenção, seleção, preparação, exploração e transformação de dados, cujo objetivo é preparar os dados para os algoritmos da etapa seguinte; (2) Mineração de Dados (MD) que realiza a busca efetiva de conhecimentos úteis no contexto proposto para a aplicação do KDD_AZ; (3) pós-processamento abrange o a interpretação e avaliação, e tratamento do conhecimento. Na literatura, diversos pesquisadores trabalharam com a ideia de que conhecimento pode ser automaticamente detectado, validado e usado de forma inteligente para inúmeras finalidades (WILLING & JOHNSON, 2004).

2 METODOLOGIA

Para aplicar aprendizado de máquina no contexto das PMEs no setor da educação, serão utilizados algoritmos que executam diferentes tarefas sobre um conjunto de dados de um polo de educação de uma universidade do sul do Brasil no período de captação 2020 e 2021. O conjunto de dados utilizado reúne dados da pré-inscrição (nome do curso), dados do entendimento da composição e motivações no momento do contato. O procedimento adotado nesse trabalho envolve uma sequência de três etapas do processo de KDD_AZ, usando de técnicas de ML com métrica simples e objetiva para prever a probabilidade de fechamento da matrícula do lead com a maior precisão possível. Todas as implementações deste trabalho foram desenvolvidas usando linguagem Python e ferramentas disponíveis nos pacotes pandas e *scikit-learn*.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

2.1 Etapa 1: Pré-processamento

Essa etapa do KDD_AZ envolve cinco fases, e tratam das propriedades dos recursos, que dizem respeito a pré-inscrição (nome do curso), dados do entendimento da composição e motivações no momento do contato com o lead. Essas fases são descritas na Tabela 1.

Tabela 1 – Fases do KDD_AZ

Fases dos dados	Tarefas
Obtenção	Importados dos dados de um polo de educação de uma universidade privada no sul do Brasil, no período 2020 e 2021, no formato csv.
Seleção	Seleção de dados estruturados composto por 8 atributos/recursos e uma amostra de tamanho com 1596 leads
Preparação	Limpeza (correção/remoção dados inconsistentes, exclusão de valores ausentes ou substituía-os pôr NA, verificação de dados ausentes ou incompletos e identificação de anomalias (*outliers*)) e engenharia de recursos (integração e construção de dados)
Exploração	Análise Exploratória de Dados (EDA), que compreendeu a sumarização dos dados de forma descritiva e a variação/dispersão dos dados, distribuição de frequência e análise de correlações de forma gráfica.
Transformação	Normalização e Conversão/codificação dos dados em formas apropriadas para entregar a etapa de mineração de dados

Fonte: Elaborada pelos autores

2.2 Etapa 2: Mineração de dados (Modelagem)

Como o objetivo do trabalho é prever a probabilidade de fechamento da matrícula, utilizamos o método de aprendizado supervisionado para classificar a classe de conversão dos leads. Nesse sentido, aplicamos três técnicas de ML: KNN, Random Forest e Regressão Logística. A escolha dessas técnicas dar-se em função delas serem amplamente utilizadas nas tarefas de classificação. Diferentes configurações de parâmetros foram testadas para cada uma das técnicas a fim de melhorar os resultados e reduzir a ocorrência de *overfitting* nos testes do modelo.

A técnica k-Nearest Neighbors (KNN) é normalmente usada para categorizar uma nova observação com base em observações passadas. A ideia básica do modelo KNN é que ele classifica novas observações identificando seus vizinhos K mais próximos, atribuindo o valor da maioria. Geralmente, a técnica KNN alcança bons resultados, mas pode apresentar um custo computacional considerável (BRUCE & BRUCE, 2017).

A técnica Random Forest (RF), é um método baseado em árvore de decisão que usa um conjunto de árvores para reduzir a propensão ao ajuste excessivo. RF é poderoso e comumente usados nas tarefas de classificação e regressão. Sua abordagem como classificador gera um modelo que é basicamente um agrupamento de Árvores de Decisão, onde cada árvore possui características próprias (BREIMAN, 2001).

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

A técnica Regressão Logística (RL) é usada para prever a classe (ou categoria) de indivíduos com base em uma ou mais variáveis preditoras (x). RL, em geral, é usada para modelar um resultado binário, ou seja, com apenas dois valores possíveis: 0 ou 1, sim ou não, doente ou não doente, evadiu ou não evadiu. De acordo com Kassambara (2017), a técnica RL não retorna diretamente a classe de observações. Isso nos permite estimar a probabilidade (p) da participação na classe. A probabilidade varia entre 0 e 1.

2.3 Etapa 3: Pós-processamento

Após construirmos o modelo de classificação preditivo na modelagem, avaliaremos o desempenho do modelo**,** ou seja, o quão bom o modelo é em prever o resultado de novos dados de teste de observações que não foram usados para treinar o modelo. Em outras palavras, precisamos estimar a *precisão da* previsão do modelo e os erros de previsão usando um novo conjunto de dados de teste. Como já sabemos o resultado real de observações nos dados de teste, o desempenho do modelo preditivo pode ser avaliado comparando valores de resultado previstos com valores de resultado conhecidos.

2.3.1 Métricas de avaliação de desempenho

As métricas e método comumente usadas para avaliar o desempenho dos modelos de classificação preditiva são: Acurácia, Matriz de Confusão, Precisão, Recall, F1-Score e curva ROC. Um resumo das fórmulas de cada umas das métricas de desempenho é mostrado na Tabela 3 - Resumo das métricas de desempenho.

Precisão bruta da classificação ou acurácia

Essa métrica representa a taxa de acertos em relação ao total de amostras. De acordo com Kassambara (2017), a acurácia não é uma boa métrica de um classificador binário, isso porque ela pode cometer dois erros: Primeiro, ela pode atribuir incorretamente um indivíduo positivo à uma categoria negativa e por segundo ela pode atribuir incorretamente um indivíduo negativo à uma categoria positiva. Para corrigir esse problema aplica-se a matriz de confusão e dela extrai-se as métricas de precisão, sensibilidade e F1-Score.

Matriz de confusão

De forma rápida a matriz de confusão permite visualizar classificações erradas e corretas. Assim, para avaliar os resultados obtidos pelos classificadores treinados e identificar aquele que apresentou os melhores resultados na previsão de conversão, a Matriz de Confusão (MC) foi calculada para cada caso.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

A MC separa as previsões em quatro valores: verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN).

A terminologia "positivo", no caso do trabalho, representa um caso de conversão. Os verdadeiros positivos são os valores que realmente correspondem a uma conversão e foram previstos como tal. Os falsos positivos são os valores que correspondem a uma não-conversão e foram previstos como conversão. Os verdadeiros negativos são os valores que correspondem a uma não-conversão e foram previstos como tal. Os falsos negativos são os valores que correspondem a uma conversão, mas foram previstos como não-conversão.

Precisão

Essa métrica é particularmente útil, pois avalia a precisão de um modelo na previsão de rótulos positivos, limitando o número de falsos positivos. Ela responde à pergunta, dentre o número de vezes que um modelo previu positivo, com que frequência ele estava correto?

Recall ou Sensibilidade

Também conhecida como Sensibilidade, essa métrica avalia a "Taxa de Verdadeiros Positivos" reais que o modelo identificou corretamente. Como resultado temos uma mostra de quantos dos dados positivos foram capturados pelas previsões positivas. É uma métrica importante, tendo em vista que é problemático investir tempo em leads que provavelmente não fecharão.

F1-Score

Essa métrica mede a média harmônica entre as medidas de Precisão e Recall. Isso porque o F1-Score dar mais peso aos valores mais baixos. Essa métrica proporciona uma boa capacidade de avaliação em casos de classificação binária. Ela é adequada em casos que precisam da análise de todo o quadro com a mesma importância e não apenas de um de seus aspectos.

Resumo das métrica de desempenho do modelo de classificação preditiva

Tabela 2 – Resumo das métricas de desempenho

Métrica de desempenho	**Fórmula**
Acurácia	(VP + VN) / (VP + VN + FP + FN)
Precisão	V P/ (VP + FP)
Recall (Sensibilidade)	VP / (VP + FN)
F1-Score	(2* Recall *Precisão) / (Recall + Precisão)

Fonte: Elaborado pelos autores

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Curva ROC

A curva Característica de Operação do Receptor (ROC) é uma medida gráfica do desempenho geral do modelo, mostrando a proporção de verdadeiros positivos e falsos positivos (proporção total de observações classificadas corretamente) em todos os valores possíveis de corte de probabilidade popular para avaliar o desempenho geral do modelo ou a precisão de um classificador, que corresponde à proporção total de observações classificadas corretamente. A área sob a curva (AUC) resume o desempenho geral do classificador, em todos os cortes de probabilidade possíveis, representando assim a capacidade de um algoritmo de classificação de distinguir 1s de 0s (ou seja, eventos de não eventos ou positivos de negativos). Por exemplo, um modelo cujas previsões estão 100% erradas tem uma AUC de 0, enquanto um modelo cujas previsões são 100% corretas tem uma AUC de 1.

3 DISCUSSÃO DE RESULTADOS

Dois pontos merecem destaques nos resultados e discursões. O primeiro, trata da significância de um atributo/variável preditora para o estimador, que também deve ser avaliada, a fim de se obter um conjunto de recursos significativos para obtenção dos melhores resultados no modelo. Por segundo, a utilização das diferentes técnicas de classificação teve como objetivo alcançar o melhor resultado possível para prever a probabilidade de conversão de oportunidades geradas na captação de leads. Para analisar o melhor estimador/classificador, comparamos os resultados da capacidade de previsão do modelo versus sua capacidade de aprender e observar se há *overfitting,* pois é comum mostrar as estatísticas de treinamento do modelo.

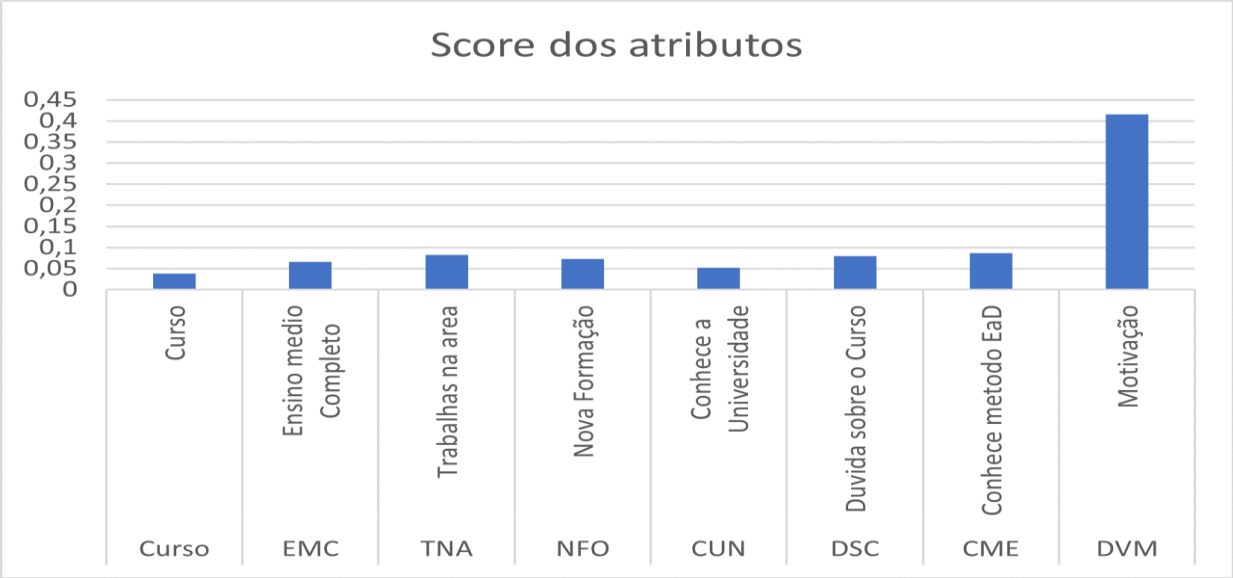
3.1 Atributos mais importantes

A significância de um atributo/variável preditora para o estimador também deve ser avaliada, a fim de se possa obter um conjunto de recursos significativos para obtenção dos melhores resultados no modelo. Uma boa medida usada para comparar variáveis em modelos preditivos é chamado de importância variável (VI).

A Figura 4 mostra os 8 atributos mais importantes para a previsão de conversão de leads conforme recursos/atributos levantados.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Figura 4 – Scores dos atributos



Fonte: Elaborado pelos autores

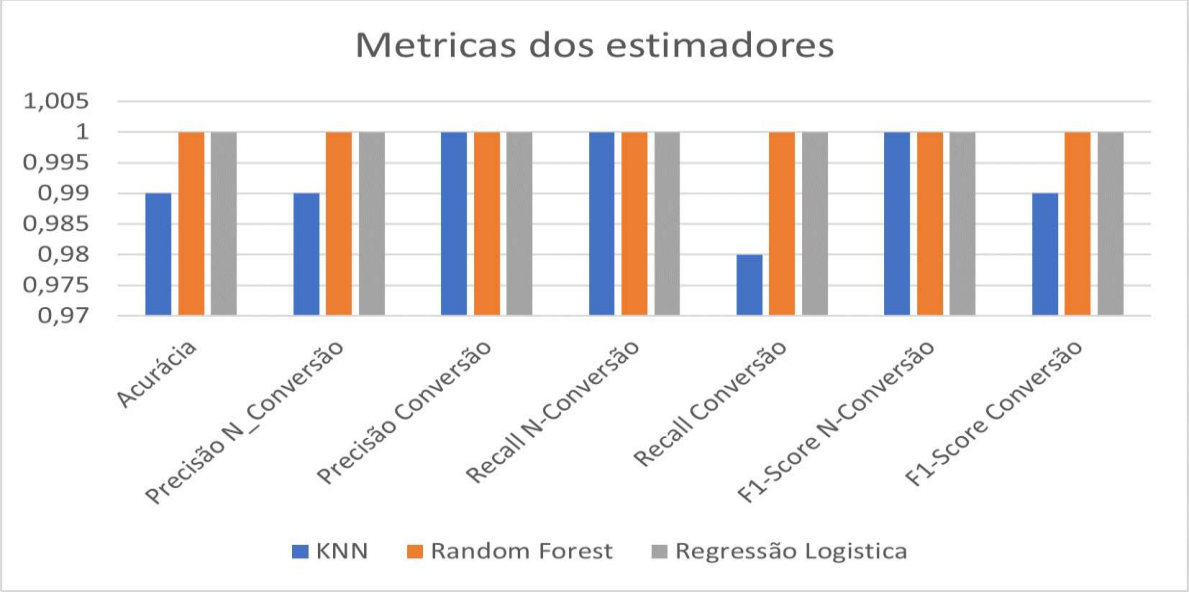
Juntos, esses atributos chegam a 87,43% de importância, de acordo com o modelo construído usando técnica RL da ML. Podemos afirmar que os atributos listados na Figura 4 são os principais responsáveis pela previsão de conversão de leads.

3.2 Desempenho dos Classificadores

Uma boa maneira de avaliar o desempenho de classificador/estimador é olhar para a sua Matriz de Confusão (MC). Os resultados dos classificadores foram muito próximo olhando pela MC e fez-se um resumo das métricas (Precisão, Recall e F1-score) derivadas da MC. Essas estatísticas das métricas de avaliação de desempenho dos estimadores são mostradas na Figura 5.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Figura 5 – Métricas de avaliação de desempenho dos estimadores



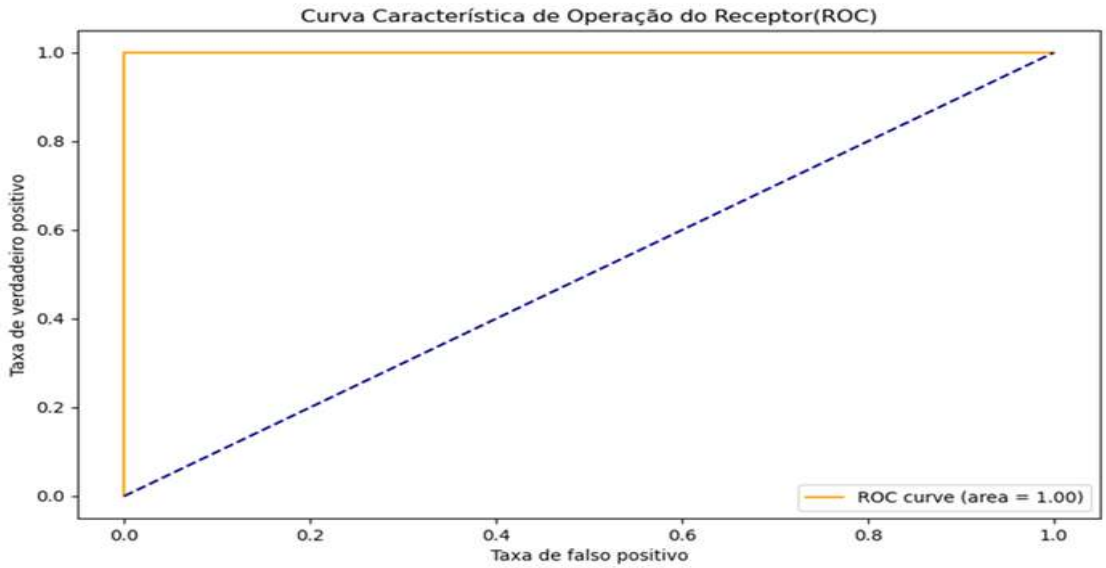
Fonte: Elaborado pelos autores

Ao analisarmos as métricas de avaliação mostradas na Figura 5, infere-se que os métodos Random Forest e Regressão Logística foram capazes de fazer as melhores previsões de verdadeiros positivos, ou seja, dos casos nos quais o lead provavelmente irá fechar a matrícula. Por outro lado, dentre todas as técnicas aplicadas, o método KNN apresentou também um bom resultado com relação a previsão de conversão do lead. Embora os três estimadores apresentem bons resultados optamos por prosseguir com o estimador Regressão Logística (RL) por ter padrões de parâmetros que se encaixam perfeitamente na modelagem de resultados binário para este trabalho. O classificador RL nesse resultado é perfeito, tendo em visto, que alcançou 100% de classificação correta de verdadeiros positivos e verdadeiros negativos.

Considerando que o objetivo de analisar o poder preditivo de um modelo é garantir que ele detecte o máximo possível de verdadeiros positivos enquanto minimiza os falsos positivos, utilizamos a ferramenta curva de ROC e AUC para demonstrar o desempenho do modelo de classificação por meio da relação da Taxa de Verdadeiro Positivo (Sensibilidade) e da Taxa de Falso Positivo (1-Especificidade), variando o *threshold* (ponto de corte na probabilidade estimada). Dessa maneira, a Curva ROC permite encontrar o *threshold* em que existe otimização da sensibilidade em função da especificidade. O ponto onde ocorre esta otimização encontra-se mais próximo do canto superior esquerdo do gráfico da curva ROC como mostrado na Figura 6.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Figura 6 – Curva ROC e AUC



Fonte: Elaborado pelos autores

Resultado apresentado na Figura 6 indica um classificador realmente de alto desempenho, pois a curva sobe abruptamente para o canto superior esquerdo, ou seja, o método identifica corretamente muitos positivos sem classificar incorretamente muitos negativos como positivos. A AUC é 1,00, portanto, o classificador pode ser considerado como excelente.

Agora tomando o estimador RL como base para as operações de implementação, executamos o conjunto de dados de testes, parametrizado para 16 exemplos e obtivemos os seguintes resultados, como mostrado na Tabela 3.

Tabela 3 – Estimador RL parametrizado

ESTIMADOR RL PARAMETRIZADO	
LEADS	
LEAD1=[1.38705442 -0.52917332],	Previsto=[0.01096455 0.98903545]
LEAD2=[1.25761298 -0.39234234],	Previsto =[0.02316462 0.97683538]
LEAD3=[-0.90816361 0.51411561],	Previsto =[0.94106479 0.05893521]
LEAD4=[1.6327847 -0.15942508],	Previsto =[0.04233817 0.95766183]
LEAD5=[1.03877541 0.03504071],	Previsto =[0.17138649 0.82861351]
LEAD6=[1.03760658 -0.59207386],	Previsto =[0.01224267 0.98775733]
LEAD7=[2.16199969 0.38161715],	Previsto =[0.21653096 0.78346904]
LEAD8=[0.11076434 0.34126079],	Previsto =[0.69963233 0.30036767]
LEAD9=[-0.77944662 0.60757554],	Previsto =[0.95457127 0.04542873]
LEAD10=[0.91802345 0.57984128],	Previsto =[0.73237132 0.26762868]
LEAD11=[-0.52820267 1.11515053],	Previsto =[0.99357463 0.00642537]
LEAD12=[1.58148912 -0.08150395],	Previsto =[0.06232312 0.93767688]
LEAD13=[0.31562106 1.01722554],	Previsto =[0.97466543 0.02533457]
LEAD14=[0.61245334 0.98701273],	Previsto =[0.96005853 0.03994147]
LEAD15=[0.03394077 0.19344227],	Previsto =[0.5667005 0.4332995]
LEAD16=[0.84044667 0.5330707],	Previsto =[0.7076786 0.2923214]

Fonte: Elaborado pelos autores

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Ao analisarmos a Tabela 4, destacamos 3 pontos importantes, são eles:

Ponto 1: Os leads de número 1,2,4,5,6,7 e 12 têm baixa probabilidade de conversão**,** portanto podemos afirmar que esses leads são potenciais candidatos a não matricular.

Ponto 2: Os leads de números 3,8,9,10,11,13,14 e 16 têm boas probabilidade de conversão, portanto podemos afirmar que esses leads são potenciais candidatos a matricular.

Ponto 3: Ainda podemos identificar situações em que não temos certeza do movimento de matrícula do lead ou não, como é o caso do lead 15, onde as probabilidades de matricular e não matricular são muito semelhantes (56,67% não matrícula e 43,32 % matrícula).

CONSIDERAÇÕES FINAIS

Para extrair conhecimento da base de leads e aumentar as taxas de conversão das oportunidades geradas na captação, faz-se necessário identificar recursos da pré-inscrição e processar recursos importantes durante a comunicação com o lead. A identificação de padrões de um lead que provavelmente irá fechar a matrícula com uma instituição de ensino é um problema de grande importância, que precisa ser resolvido. Isso porque sabemos que há um alto custo das operações de marketing e esforços das equipes de conversão para fechar matrícula. Os resultados desse trabalho revelam os principais pontos característicos encontrados antes e durante uma conversa com um lead e sua probabilidade de conversão de matrícula. Além de permitir um ganho de tempo maior para as equipes de conversão, esses resultados filtram leads como "improváveis de conversão", ajudando o marketing a melhorar sua mira e trazer leads qualificados. O trabalho tomou como base o Modelo de Transformação Digital para as PMEs (MTD_PMEs) com foco especificamente na dimensão tecnologia, em função desse permitir que as PMEs aprimorem seus negócios a um custo relativamente baixo em muitos setores.

O procedimento metodológico adotado nesse trabalho envolveu uma sequência de três etapas do processo de KDD_AZ, usando de técnicas de ML com métrica simples e objetiva para prever a probabilidade de conversão da matrícula do lead com a maior precisão possível. Ressalta-se ainda, que todas as implementações deste trabalho foram desenvolvidas usando linguagem Python e ferramentas disponíveis nos pacotes pandas e *scikit-learn*.

A técnica de Regressão Logística do método supervisionado da aprendizagem de máquina foi fundamental em pelo menos dois pontos para o resultado de descobertas deste estudo. O primeiro está na identificação dos atributos mais significativos para a classificação correta da conversão de leads. Esses, juntos, chegam a 87,43% de significância para o modelo. O segundo ponto destaca a precisão de previsão da conversão leads. Essa é constatada por meio da avaliação de desempenho do modelo - matriz de confusão, a qual, alcançou 100% de classificação correta de verdadeiros positivos e verdadeiros negativos. Em confirmação a esses resultados, a curva ROC e AUC demonstram o desempenho do modelo de classificação por meio da relação da Taxa de Verdadeiro Positivo (Sensibilidade) e da Taxa de Falso Positivo (1-Especificidade), variando o *threshold* (ponto de corte na probabilidade estimada) conforme figura 6.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Acreditamos que grande parte desse resultado se dá em função do tratamento feito com os dados (ciência de dados) antes de utilizá-lo na modelagem.

A principal contribuição do trabalho vem na formação de um conjunto de variáveis significativas para prever a probabilidade de conversão de lead, ou seja, do fechamento da matrícula de um lead nas instituições de ensino. As relações encontradas entre as variáveis preditoras identificadas e a conversão foram estudadas, de maneira que o trabalho possa servir de apoio às equipes de conversão, bem como a área de marketing das instituições. Diante desses resultados, espera-se que este trabalho abra fronteiras para o ML alcançar Inteligência Artificial não somente no setor da educação, mas também ajude a reduzir o tempo e esforços das equipes de conversão, e no alinhamento de prospecção de leads qualificados/quentes pela área de marketing das instituições de ensino.

REFERÊNCIAS

ALMEIDA, T. D.; CAVALCANTE, C. G. S.; FETTERMANN, D. C. Indústria 4.0: Tecnologias e nível de maturidade de suas aplicações. *In.:* **11º Congresso Brasileiro de Inovação e Gestão de Desenvolvimento do Produto**, São Paulo, 2017.

ALTINKAYA, M.; ZONTUL, M. Urban bus arrival time prediction: a review of computational models. **International Journal of Recent Technology and Engineering**, v. 2, n. 4, p. 164-169, 2013.

BATISTA, M. R. S. **A utilização de algoritmos de máquinas em problemas de classificação**. 2019. 112 p. Dissertação (Mestrado Profissional em Matemática, Estatística e Computação Aplicada à Indústria) - Universidade de São Paulo, Instituto de Ciências Matemáticas de Computação, São Paulo, 2019. Disponível em: http://www.teses.usp.br/teses/disponiveis/55/55137/tde-25032019-141126/en.php. Acesso em: 30 abr. 2021.

BIAGIONI, J.; GERLICH, T.; MERRIFIELD, T.; ERIKSSON, J. Easytracker: automatic transit tracking, mapping, and arrival time prediction using smartphones. *In.:* **Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems**, 2011. p. 68-81.

BREIMAN, L. Random Forests. (R. E. Schapire, Ed.) **Machine Learning**, v. 45, p. 5-32, Oct. 2001. doi:10.1023/A:1010933404324.

BRUCE, P.; BRUCE, A. **Practical Statistics for Data Scientists**. Massachusetts: O'Reilly Media, 2017.

CALIPINAR, H.; ULAS, D. Model suggestion for SMEs economic and environmental sustainable development. *In.:* NWANKWO, S.; NDUBISI, N. OLY (Ed.). **Enterprise Development in SMEs and Entrepreneurial Firms.** USA: Business Science Reference, IGI Global, 2013. p. 270-290.

COSTA, O. S.; GOUVEIA, L. B. Industria 4.0: Uma Proposta de Modelo de Transformação Digital para as Pequenas e Médias Empresas. *In.:* MARTINS, E. R. **Engenharia de Produção Planejamento e Controle da Produção em Foco**.Guarujá, SP: Cientifica Digital, 2021. Vol. 1, cap. 9, p. 117-133. doi: 10.37885/201102037.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 1-6, 1996.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

GHOTRA, B.; MCINTOSH, S.; HASSAN, A. E. Revisiting the impact of classification techniques on the performance of defect prediction models. *In.:* **Proceedings of the 37th International Conference On Software Engineering**, Florence, Italy: IEEE Press, 2015, May. p. 789-800.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**: Concepts and Techniques. 3nd. Burlington: Morgan Kaufmann Publishers, 2011.

KASSAMBARA, A. **Machine Learning Essential**: Practical guide in R. Inglaterra: STHDA, 2017.

HELLER, K. A.; WILLIAMSON, S.; GHAHRAMANI, Z. Statistical models for partial membership. *In.:* **Proceedings of the 25th International Conference on Machine Learning**, 2008. p. 392-399.

KORMÁKSSON, M.; BARBOSA, L.; VIERA, M. R.; ZADROZNY, B. Bus travel time predictions using additive models. *In.:* **Proceedings of the 14th International Conference On Data Mining**. Shenzhen, China, IEEE. 2014 p. 875-880.

MACHADO, V. P. **Inteligência Artificial**. Teresina: EDUFPI, 2011.

MENEZES, V. V.; FARRAR, J. T.; BOWER, A. S. Westward mountain-gap wind jets of the northern Red Sea as seen by QuikSCAT. **Remote Sensing of Environment**, v. 209, p. 677-699, 2018. https://doi.org/10.1016/j.rse.2018.02.075

MIRKOVIĆ, V.; LUKIĆ, J.; MARTIN, V. Key characteristic of organizational structure that supports digital transformation. Strategic Management and Decision Support Systems in Strategic Management. *In.:* **Proceedings of the 24th International Scientific Symposium**. Subotica, Republic of Serbia, 2019. p. 255-261.

MITCHELL, T. M. **Machine Learning**. Porto Alegre: McGraw-Hill, 1997.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos de aprendizado de máquina. In: REZENDE, S. O. **Sistemas inteligentes:** fundamentos e aplicações. Barueri, SP: Manole, 2003. p. 89-94

NEVES, D. V.; DIAS, F. C. A.; CORDEIRO, D. **Uso de aprendizado supervisionado para análise de confiabilidade de dados de crowdsourcing sobre posicionamento de ônibus**. 2018. Dissertação (Mestrado) - Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

NVIDIA. **Blog**: Supervised Vs. Unsupervised Learning. The Official NVIDIA Blog, 2018. Disponível em: https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/. Acesso em: 30 abr. 2021.

OH, J.; HESSEL, M.; CZARNECKI, W. M.; XU, Z.; VAN HASSELT, H.; SINGH, S.; SILVER, D. Discovering reinforcement learning algorithms. In.: **34th Conference on Neural Information Processing Systems (NeurIPS 2020)**, Vancouver, Canada, 2021.

TAN, M.; SHAO, P. Prediction of student dropout in e-learning program through the use of machine learning method. **International Journal of Emerging Technologies in Learning (IJET**), v. 10, n. 1, p. 11-17, 2015.

ULAS, D. Digital Transformation Process and SMEs. **Procedia Computer Science**, v. 158, p. 662-671, 2019. doi: 10.1016/j.procs.2019.09.101.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

WANG, Y.; WANG, F.; HUANG, J.; WANG, X.; LIU, Z. Validation of artificial neural network techniques in the estimation of nitrogen concentration in rape using canopy hyperspectral reflectance data. **Int. J. Remote Sens**., v. 30, p. 4493-4505, 2009. doi: 10.1080/01431160802577998

WILLING, P.; JOHNSON, S. Factors that influence students' decision to dropout of online courses. **Journal of Asynchronous Learning Network**, Nashville, v. 8, n. 4, p. 105-118, 2004.