2021 | Volume 5 | Nº 4 | Pág. 766 a 787

DOI: https://doi.org/10.15536/reducarmais.5.2021.2417

Revista Educar Mais

Mineração de dados educacionais com aprendizagem de máquina

Educational data mining with machine learning

Minería de datos educativos con aprendizaje máquina

Vanessa Faria de Souza[1]

RESUMO

Com a o aumento da disponibilidade de dados, sobretudo no contexto educacional, surgiram áreas específicas para extração de informações relevantes, como a Mineração de Dados Educacionais (MDE), que integra inúmeras técnicas que dão suporte a captação, processamento e análises desses conjuntos de registros. A principal técnica associada a MDE é a Aprendizagem de Máquina, que vem sendo empregada a décadas no processamento de dados em diversos contextos, mas com o advento do Big Data houve uma intensificação na aplicação dessa técnica no intuito de extrair informações relevantes de uma enorme quantidade de dados. Nesse sentido, esse estudo tem como objetivo realizar a previsão do desempenho de alunos, em um conjunto de dados públicos, ademais indicar quais os principais atributos preditores para o desempenho dos alunos. Para isso foi implementado um processo de MDE baseado em 4 etapas: 1) Coleta de dados; 2) Extração de recursos e limpeza de dados (Pré-Processamento e Transformação); e 3) Processamento analítico e algoritmos; 4) análise e/ou interpretação dos resultados. Como resultado foi identificado que para o conjunto de dados utilizado neste estudo o algoritmo Árvores de Decisão foi o mais preciso – com uma acurácia de 87% – bem como foi constatado que atributos relacionados as atividades escolares são mais preditores para o desempenho dos alunos do que dados de características demográficas e socioeconômicas.

Palavras-chave: Mineração de Dados Educacionais; Previsão de Desempenho; Aprendizagem de Máquina.

ABSTRACT

With the increase in the availability of data, especially in the educational context, specific areas have emerged for the extraction of relevant information, such as Educational Data Mining (MDE), which integrates numerous techniques that support the capture, processing and analysis of these sets of records. The main technique associated with MDE is Machine Learning, which has been used for decades in data processing in different contexts, but with the advent of Big Data there was an intensification in the application of this technique in order to extract relevant information from a huge amount of data. In this sense, this study aims to predict the performance of students, using a set of public data, and to compare which of the Machine Learning algorithms used was the most effective, in addition to indicating which are the main predictive attributes for student performance. For this, an EAW process based on 4 steps was implemented: 1) Data collection; 2) Extraction of resources and cleaning of data (Pre-Processing and Transformation); and 3) Analytical processing and algorithms; 4) analysis and / or interpretation of results. As a result, it was identified that for the data set used in this study, the Decision Trees algorithm was the most accurate - with an accuracy of 87% - as well as it was found that attributes related to school activities are more predictive of student performance than demographic and socioeconomic characteristics data.

Keywords: *Educational Data Mining; Performance Prediction; Machine Learning.*

[1] Docente na área de Informática do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS) – Campus Ibirubá, Ibiruba/RS - Brasil. Doutoranda no Programa de Pós-Graduação em Informática na Educação (PPGIE) da UFRGS. Mestre pelo Programa de Pós-Graduação em Informática (PPGI) da UTFPR. E-mail: vanessa.souza@ibiruba.ifrs.edu.br.

RESUMEN

Con el aumento de la disponibilidad de datos, especialmente en el contexto educativo, han surgido áreas específicas para la extracción de información relevante, como Educational Data Mining (MDE), que integra numerosas técnicas que apoyan la captura, procesamiento y análisis de estos conjuntos. de registros. A principal técnica associada a MDE é a Aprendizagem de Máquina, que vem sendo empregada a décadas no processamento de dados em diversos contextos, mas com o advento do Big Data houve uma intensificação na aplicação dessa técnica no intuito de extrair informações relevantes de uma enorme quantidade de dados. En este sentido, este estudio tiene como objetivo predecir el rendimiento de los estudiantes, utilizando un conjunto de datos públicos, y comparar cuál de los algoritmos de aprendizaje automático utilizados fue el más efectivo, además de indicar cuáles son los principales atributos predictivos del rendimiento de los estudiantes. . Para ello, se implementó un proceso de ODE basado en 4 pasos: 1) Recolección de datos; 2) Extracción de recursos y limpieza de datos (Pre-Procesamiento y Transformación); y 3) Procesamiento analítico y algoritmos; 4) análisis y / o interpretación de resultados. Como resultado, se identificó que para el conjunto de datos utilizado en este estudio, el algoritmo de árboles de decisión fue el más preciso, con una precisión del 87%, y se encontró que los atributos relacionados con las actividades escolares son más predictivos de los estudiantes. desempeño que los datos de características demográficas y socioeconómicas.

Palabras clave: *Minería de datos educativos; Predicción de rendimiento; Aprendizaje automático.*

1. INTRODUÇÃO

Nos últimos anos a educação tem se modificado, em decorrência do avanço tecnológico disponível que direcionou a uma instrumentação do setor educacional, tanto em softwares voltados para o ensino, como na administração digital dos registros acadêmicos pelos gestores das instituições, bem como no uso da internet para a aprendizagem, em especial pela popularização do *e-learning*. Todos esses fatores impulsionaram um crescimento exponencial no volume de dados educacionais, e para se analisar uma grande quantidade de dados, é imprescindível contar com recursos computacionais, caso contrário a tarefa torna-se impraticável.

Dessa forma, as técnicas de mineração de dados estão ganhando cada vez mais importância no setor educacional, pois são uma forma de acompanhar, analisar e avaliar o processo de aprendizagem. Provavelmente, as técnicas de mineração de dados podem fornecer aos formuladores de políticas educacionais modelos para apoiar seus objetivos de aprimorar a eficiência e a qualidade do ensino e da aprendizagem. Além disso, o uso de diferentes técnicas de mineração de dados pode ser visto como base para uma mudança sistêmica, capaz de impactar de maneira positiva nas soluções de problemas específicos das Instituições de Ensino, por exemplo, viabilizando soluções que envolvam a personalização dos ambientes educacionais ou fornecendo suporte para o processo de tomada de decisão no ambiente educacional.

Nesse cenário, destaca-se a Mineração de Dados Educacionais (MDE) que utiliza as técnicas da Mineração de Dados (MD) para extrair informações relevantes de conjuntos diversificados de dados educacionais. Segundo a Sociedade Internacional de *Educational Data Mining*[2], a MDE pode ser definida da seguinte forma:

> É uma disciplina emergente, preocupada com o desenvolvimento de métodos para explorar dados únicos e cada vez mais em larga escala, provenientes de contextos

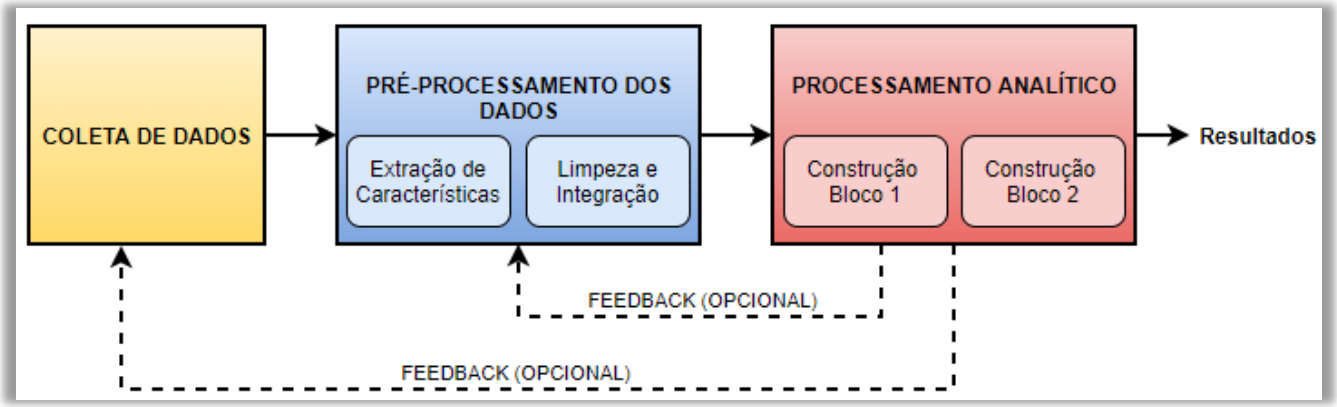
[2] http://educationaldatamining.org/

educacionais e usa esses métodos para entender melhor os alunos e as configurações em que aprendem (EDM, 2020).

Em outras palavras, a MD refere-se a um conjunto de técnicas computacionais para extrair informações de grandes massas de dados, e quando os dados analisados são provenientes de contextos educacionais, chama-se MDE (ROMERO; VENTURA, 2013). Igualmente, De Los Reyes et al. (2019) define MDE como uma área voltada ao desenvolvimento de métodos para explorar dados oriundos de ambientes educacionais e utilizá-los para compreender melhor os processos de ensino e aprendizagem. Nessa acepção, Baker, Isotani e Carvalho (2011) alegam que a MDE é definida como a área de pesquisa que tem como finalidade aperfeiçoamento e amadurecimento de técnicas para investigar conjuntos de dados obtidos em cenários educacionais. Conforme os autores, a natureza destes dados é mais diversa do que a observada nos dados tradicionalmente utilizados em tarefas de mineração, demandando adaptações e novas abordagens. Ao mesmo tempo, essa diversidade nos dados representa um potencial de implementação de recursos fundamentais para auxílio na melhoria da educação (BAKER; ISOTANI; CARVALHO, 2011; DE LOS REYES et al., 2019; RIGO et al., 2014). Sendo assim, necessita-se de técnicas e ferramentas que auxiliem na tarefa de verificar, interpretar e relacionar esses dados, com o intuito de gerar conhecimento útil e relevante, o que, segundo De Los Reyes et al. (2019) já era um objetivo das técnicas de DM, empregadas para identificar padrões de comportamento e encontrar insights que provoquem melhorias em produtos e serviços.

No que se refere ao processo de aplicação da MDE, este é similar ao da MD. Para Aggarwal (2015), o fluxo de trabalho de um processo típico de Mineração de Dados contém as seguintes etapas: 1) Coleta de dados; 2) Extração de recursos e limpeza de dados (Pré-Processamento e Transformação) – para tornar os dados adequados para processamento; 3) Processamento analítico e algoritmos – projetar métodos analíticos eficazes para extrair informações e conhecimentos relevantes a partir dos dados processados e o autor ainda sugere que os resultados precisam ser analisados e/ou interpretados, por isso cabe ao pesquisador verificar a melhor forma de realizar essa análise. A sequência das etapas do processo proposto por Aggarwal (2015) é apresentada na Figura 1, na qual pode-se observar que o processo de MD pode ser iterativo.

Figura 1 – Processo de Data Mining proposto por Aggarwal (2015)



Fonte: Adaptado de Aggarwal (2015).

Com relação a etapa do processamento analítico, há algumas técnicas mais proeminentes que podem ser utilizadas para este fim, algumas delas são: a Estatística Descritiva e Inferencial, algoritmos de Aprendizagem de Máquina (*Machine Learning*) e as redes neurais profundas associadas à Aprendizagem Profunda (*Deep Learning*).

Diante desse contexto, esse estudo tem como objetivo realizar a previsão do desempenho de alunos, por meio da aplicação de algoritmos de Aprendizagem de Máquina em um conjunto de dados público do repositório *UCI Machine Learning*[3]*;* e analisar a precisão desses diferentes algoritmos. Ademais, este estudo pretende verificar se os atributos que compõe a base de dados são suficientes para realizar a geração de modelos eficazes na previsão do desempenho dos alunos, bem como avaliar quais desses atributos são mais preditivos do rendimento desses estudantes. Cabe ainda destacar que a manutenção de modelos para prever o desempenho de alunos pode auxiliar professores na obtenção de informações que apoiem no desenvolvimento de ações preventivas, que possam evitar o baixo rendimento dos estudantes, bem como apoiar em intervenções para a diminuição da evasão Baker (2000).

2. APRENDIZAGEM DE MÁQUINA NO CONTEXTO EDUCACIONAL

Desde a popularização do *e-learning* e do surgimento da análise automatizada de dados educacionais, muitos esforços têm sido realizados para aprimorar a experiência da aprendizagem, por esse motivo a MDE ganhou notoriedade, pois um de seus interesses é explorar a maneira como as pessoas aprendem (ROMERO; VENTURA, 2013). Avanços nessa área permitiram coletar e analisar dados sobre os alunos e seus ambientes e explorar o comportamento das pessoas enquanto aprendem e umas das principais técnicas empregadas para essa finalidade são os modelos de AM, que podem dar suporte a uma transformação da educação dita por vezes tradicional para uma educação otimizada e personalizada.

A AM consiste em extrair informações dos dados, é uma área de pesquisa formada pela interseção da Estatística, Inteligência Artificial e Ciência da Computação, muitas vezes é referenciada como análise preditiva ou aprendizado estatístico e muitos pesquisadores defendem que ela é um campo da Inteligência Artificial (BISHOP; PATTERN, 2006; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; MITCHELL, 1997). Pode-se dizer que a AM basicamente investiga como os computadores podem melhorar seu desempenho com base em dados.

A primeira definição de AM foi elaborada por Samuel (1959) que a definiu como "Campo de estudo que permite que os computadores efetuem operações sem serem explicitamente programados" (SAMUEL, 1959). Uma das definições mais amplamente utilizadas para definir AM é a feita por Mitchell (1997, p. 2): "Um programa de computador aprende a partir da experiência E, em relação a uma classe de tarefas T, com medida de desempenho P, se seu desempenho em T, medido por P, melhora com a experiência E" (MITCHELL, 1997). Com essa afirmação o autor quer dizer que o termo "aprendizado" empregado a essa técnica é uma forma de explicar que o sistema faz a mesma tarefa, ou tarefas, sobre um mesmo conjunto de dados de uma maneira mais eficiente a cada execução. Mitchell (1997) complementa dizendo que AM é uma área que se ocupa por investigar métodos computacionais adequados para a aquisição de novos conhecimentos, novas habilidades e novas formas de organização do conhecimento já existente. Dessa forma, o campo do aprendizado de máquina é norteado pela questão de como produzir programas que automaticamente melhoram com a sua experiência.

Nesse sentido, a AM é um campo de pesquisa que se ocupa do desenvolvimento de algoritmos que podem realizar previsões sobre grandes volumes de dados, esses algoritmos utilizados no contexto

[3] https://archive.ics.uci.edu/ml/index.php

educacional podem auxiliar em inúmeras possibilidades. Nessa perspectiva, logo no início da disseminação e popularização da Inteligência Artificial e da Aprendizagem de Máquina, Baker (2000) as visualizou como um conjunto promissor de mecanismos de software e tecnologias a serem empregadas na melhoria do processo educacional, antes mesmo da consolidação da MDE como um campo de pesquisa. Nesse sentido, o autor intensificou suas pesquisas sobre isso e agrupou as possíveis contribuições de modelos de AM ao processo educacional em três papeis principais:

1. *Modelo como ferramenta científica* – utilizado como um meio para compreender e prever algum aspecto de uma situação educacional. Por exemplo, um modelo é desenvolvido para entender porque muitos alunos têm desistido da disciplina de Cálculo, para prever o desempenho acadêmico, ou para realizar análise do comportamento de alunos.
2. *Modelo como componente* – empregado como algum aspecto do processo de ensino ou aprendizagem, sendo usado como um componente de um artefato educacional. Por exemplo, um modelo poderia ser integrado ao um ambiente de aprendizado baseado em computador, esse dessa forma se adaptar e interagir com os alunos, dando suporte em dúvidas frequentes e poderia ir refinando a forma como ajuda esses estudantes, conforme vai interagindo com eles.
3. *Modelo como base para o design* – utilizado como componente básico para projetar sistemas específicos para os processos educacionais, formando a base para o design de uma ferramenta de computador para a educação. Por exemplo, um modelo de diálogo orientado a tarefas formaria a base do design e implementação de ferramentas para comunicação mediada por computador entre alunos e professores, em um ambiente de aprendizado colaborativo suportado por computador.

Quando a Aprendizagem de Máquina é empregada como técnica de MDE, considera-se que ela está incluída na primeira possibilidade prevista por Baker (2000), com a finalidade de entender situações educacionais e com isso dar apoio no processo de tomada de decisão, nesse sentido há muitos estudos que foram e estão sendo desenvolvidos, entretanto alguns podem ser destacados, como: Cortez e Silva (2008), Chui *et al.* (2020), Zhang e Wu (2019) e Rodrigues *et al.* (2016), que são detalhados na seção de trabalhos relacionados.

Para finalizar esta seção, que tratou das definições de AM, cabe destacar que essa área possui uma extensa terminologia e para um melhor entendimento deste estudo, é preciso conhecer alguns desses termos. No caso desta pesquisa, é importante compreender os conceitos: Treinamento; Teste e Modelo. Treinamento é a fase em que o algoritmo de AM é aplicado a uma base de dados e este deve buscar padrões sobre os dados; Teste é a fase em que o modelo é avaliado; e Modelo é o produto da submissão dos dados a um algoritmo, por vezes modelo é confundido com o algoritmo, apesar de serem estruturas codificadas não são iguais. Depois da fase de treinamento de um algoritmo é gerado um modelo que pode ser salvo para ser posteriormente aplicado a uma nova base de dados, desde que essa possua os mesmos atributos.

3. TRABALHOS RELACIONADOS

Nesse sentido, é importante destacar um estudo que utilizou a mesma base de dados desta pesquisa: Cortez e Silva (2008). Neste estudo os dois conjuntos de dados, foram modelados com uma classificação de cinco níveis e foi utilizada a regressão, pois o atributo a ser previsto era a nota (numérico). O objetivo de Cortez e Silva (2008) era analisar o desempenho dos alunos sob uma

perspectiva de quais atributos mais influenciam na previsão do desempenho. Para isso, os autores utilizaram quatro algoritmos: Árvores de decisão, *Random Forest*, Redes Neurais Simples e *Suporte Vector Machines*. Os resultados dos autores mostraram que uma boa precisão preditiva pode ser alcançada, desde que estejam disponíveis as primeiras e/ou segundas séries do período escolar. Cortez e Silva (2008) ressaltam ainda que o desempenho do aluno é altamente influenciado por avaliações anteriores e pelo número de faltas. Como resultado direto desta pesquisa os autores relatam que ferramentas mais eficientes de previsão do aluno podem ser desenvolvidas, melhorando a qualidade da educação e aprimorando a gestão dos recursos escolares.

No estudo desenvolvido Chui *et al.* (2020) o objetivo foi desenvolver um modelo para prever o desempenho acadêmico de alunos da graduação. Nesse contexto, os autores afirmaram que a partir de uma perspectiva global, se um número considerável de estudantes deixarem a universidade devido ao fracasso acadêmico, não apenas a reputação das próprias universidades sofre, mas as aspirações sociais dos alunos também são prejudicadas. Por isso há uma necessidade de desenvolver meios precisos de prever graduandos que estejam em risco acadêmico. Tendo esse objetivo em vista, o problema a ser tratado dizia respeito a identificação de padrões de alunos com tendência a ter um baixo desempenho. A base de dados utilizada nessa pesquisa foi a *Open University Learning Analytics* coletada entre 2013 e 2014 pela *Open University* do Reino Unido, essa base compreende 7 cursos de graduação e contém atributos de mais de 32 mil alunos sobre: atividades avaliativas, notas, perfil demográfico e uma apresentação pessoal desses estudantes. Durante a formatação da base foram definidas duas classes para categorizar os estudantes pelos autores, Aprovação e Falha, o que configura um problema de classificação binária. Na sequência foi aplicado o algoritmo *Suport Vector Machines* (SVM) sobre a base de dados, o algoritmo foi modificado com várias otimizações para melhorar os resultados de previsões já relatados no estado da arte e diminuir o tempo de treinamento, que segundo os autores foi 60% menor que para o algoritmo tradicional. Quanto aos resultados das classificações, o modelo gerado por Chui *et al.* (2020) alcançou entre 92% e 94% de acurácia sobre os dados de teste. Concluindo, os autores relataram que com previsões precisas de alunos que tendem a ter um desempenho ruim, ações podem ser projetadas para amenizar as desistências desses alunos e motivá-los a conclusão.

Zhang e Wu (2019) consideraram o contexto de cursos *e-leaning* do tipo *Massive Open Online Courses* (MOOCs) para o desenvolvimento de sua investigação, os autores contextualizaram o problema de sua pesquisa destacando que com o rápido desenvolvimento de cursos desse tipo, tornou-se uma questão importante na pesquisa educacional explorar as características de aprendizagem *on-line* e fornecer apoio à melhoria dos métodos de ensino e das atividades acadêmicas. Nesse cenário, os autores introduziram o objetivo da pesquisa que foi prever o desempenho em MOOCs, mais especificamente a previsão de notas dos alunos. Para isso, os autores utilizaram dados dos alunos que cursaram o MOOC de Programação em Linguagem C, os atributos eram basicamente sobre: informações básicas do perfil dos alunos, pontuação nas atividades avaliativas, número de questões solucionadas, pontuação final, postagens nos fóruns de discussão. Quanto a abordagem para a solução do problema, embora fosse esperado que os autores utilizassem Regressão, pois as notas são valores contínuos, os autores dividiram os resultados dos alunos em classes de 85 a 100 – A, de 70 a 85 – B, de 60 a 70 – C e >60 – D, portanto tornando-se um problema de classificação. Após a formatação e classificação da base de dados os autores realizaram a aplicação de 3 algoritmos de AM para geração de modelos de previsão de notas: ID3, C4.5 e CART – todos baseados em Árvore de Decisão. As precisões alcançadas pelos modelos sobre a base de dados de teste foram: IDE3 – 81%, C4.4 – 75%, CART – 76%. Zhang e Wu (2019) afirmaram que os modelos

baseados em árvores de decisão são consideravelmente simples de serem implementados, e têm precisão relativamente satisfatória, por isso devem ser empregados para apoiar ações que induzam a permanência de alunos em MOOCs, ou em outros cenários educacionais.

Por fim, Rodrigues *et al.* (2016) desenvolveram sua pesquisa também no contexto de MOOCs, o estudo desenvolvido tinha como objetivo reconhecer perfis de engajamento de alunos nesse tipo de curso. Para esse fim, os autores utilizaram uma base de dados formada por atributos de 5 mil alunos de um curso MOOC com o tema Nova Gramática da Língua Portuguesa da plataforma *Openredu*. Os dados utilizados pelos autores diziam respeito a 15 atributos que retratavam a frequência de diferentes categorias de postagens, assiduidade e notas. Na busca pelas categorias de engajamento, os autores utilizaram o Agrupamento, por meio dos algoritmos *K-mens* e *Ward Clustering*, que possibilitaram a identificação de três perfis de engajamento no MOOC investigado: Engajados – 16% do total de alunos; Esporádicos – 26% do total; e Desengajados – 58% do total. Em conclusão Rodrigues et al. (2016) destacaram que a abordagem de Agrupamento pode ser usada para ajudar pesquisadores a identificar perfis comportamentais dos alunos em relação ao envolvimento em interações via fórum e durante atividades no decorrer de um MOOC, ainda relataram que entender os perfis comportamentais de alunos em cursos desse tipo pode servir como indicação para os designers atenderem a essa diversidade de padrões e orientar o desenho de estratégias adaptativas, que permitam aumentar o comprometimento dos alunos em conjunto com uma melhor experiência de aprendizado.

As pesquisas evidenciadas correspondem a exemplos de aplicação do processo de MDE, por meio da técnica de AM, todas com propósitos de entender acontecimentos no âmbito educacional e com isso ter mais subsídios para desenvolver soluções ou tomar decisões.

4. METODOLOGIA: PROCESSO DE MDE APLICADO NA PREVISÃO DE DESEMPENHO DE ALUNOS

Devido ao aspecto desta pesquisa, em que se trata basicamente da análise de exemplos que estimulam a compreensão (GIL, 2008), esta se caracteriza como uma pesquisa exploratória. De acordo com Gil (2008) uma pesquisa exploratória possui de forma geral o intuito de proporcionar maior familiaridade com um problema de pesquisa, isto posto a problemática que se pretende aprofundar nesse estudo é como realizar um processo de MDE. Neste sentido, este estudo tem como principal objetivo realizar a previsão do desempenho de alunos, em duas disciplinas, utilizando técnicas de MDE. Com a realização desse processo é também possível identificar qual dos algoritmos aplicados foi o mais eficaz na predição do desempenho desses estudantes, considerando os atributos que compõe a base de dados. À vista disso, formulou-se as questões de pesquisa que nortearam este estudo:

- *Questão de pesquisa 1 (QP1)* – Qual a eficácia de modelos gerados a partir de algoritmos baseados em Aprendizagem de Máquina na previsão do desempenho de alunos?
- *Questão de pesquisa 2 (QP2)* – Quais algoritmos de Aprendizagem de Máquina geram modelos mais eficazes?
- *Questão de pesquisa 3 (QP3)* – Qual conjunto de atributos mais influência na previsão do desempenho de alunos?

Para responder essas questões foram empregados alguns procedimentos que configuram a metodologia adotada nesta pesquisa, que em termos gerais resume-se na realização do processo de Mineração de Dados proposto Aggarwal (2015). Como descrito na introdução esse processo é composto basicamente por 3 etapas: 1) Coleta de dados; 2) Extração de recursos e limpeza de dados (Pré-Processamento e Transformação); e 3) Processamento analítico e algoritmos; e por fim, de acordo com Aggarwal (2015) uma quarta etapa de análise e/ou interpretação dos resultados deve ser inserida, conforme os objetivos do processo de mineração de dados realizado. Essas etapas compõem basicamente os procedimentos metodológicos adotados nessa pesquisa que são detalhados na sequência.

4.1 Primeira Etapa: Coleta dos Dados

Na primeira etapa os dados foram coletados do repositório de dados público o *UCI Machine Learning*. Estes dados abordam o desempenho dos alunos no ensino secundário de duas escolas portuguesas. Os atributos dos alunos incluem notas, características demográficas, (sociais e escolares) e foram reunidos por meio de relatórios escolares e questionários. Foram fornecidos dois conjuntos de dados de 1044 alunos relativos ao desempenho em duas disciplinas distintas: Matemática e Língua Portuguesa. Os atributos constantes na base de dados extraída estão descritos no Quadro 1. Os dados sistematizados formaram um *Data Frame*[4] com 1044 linhas e 33 colunas.

Quadro 1 – *Atributos da Base de dados*

ID	Atributos	Descrição
1	Escola	Escola do aluno (binário: 'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira)
2	Gênero	Gênero do aluno (binário: 'F' - feminino ou 'M' - masculino)
3	Idade	Idade do aluno (numérico: de 15 a 22)
4	Endereço	Tipo de endereço residencial do aluno (binário: 'U' - urbano ou 'R '- rural)
5	Famsize	Tamanho da família (binário:' LE3 '- menor ou igual a 3 ou' GT3 '- maior que 3)
6	Pstatus	Status de coabitação dos pais (binário:' T '- morando junto ou' A '- à parte)
7	Medu	Escolaridade da mãe (numérico: 0 - nenhum, 1 - ensino fundamental (4ª série), 2 - 5ª a 9ª série, 3 - ensino médio ou 4 - ensino superior)
8	Fedu	Escolaridade do pai (numérico: 0 - nenhuma, 1 - ensino primário (4º ano), 2 - 5º ao 9º ano, 3 - ensino secundário ou 4 - ensino superior)
9	Mjob	Trabalho da mãe (nominal: 'professor', 'saúde' relacionado, 'serviços' civis (por exemplo, administrativo ou policial), 'em_casa' ou 'outro')
10	Fjob	Trabalho do pai (nominal: 'professor', 'saúde' relacionado, civil 'serviços' (por exemplo, administrativo ou policial), 'em_ casa' ou 'outro')
11	Razão	Razão para escolher esta escola (nominal: perto de 'casa', escola 'reputação', 'curso' preferência ou 'outro')
12	Tutor	Tutor do aluno (nominal: 'mãe', 'pai' ou 'outro')
13	Tempo de Viagem	Tempo de viagem de casa para a escola (numérico: 1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. A 1 hora, ou 4 -> 1 hora)
14	Horas de Estudo	Tempo de estudo semanal (numérico: 1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas ou 4 -> 10 horas)

[4] Um data frame é semelhante a uma matriz mas as suas colunas têm nomes e podem conter dados de tipo diferente

15	Reprovações	Número de reprovações anteriores nas aulas (numérico: n se 1 <= n <3, senão 4)
16	Schoolup	Suporte educacional extra (binário: sim ou não)
17	Famsup	Suporte educacional familiar (binário: sim ou não)
18	Pago	Aulas extras pagas dentro da disciplina (matemática ou português) (binário: sim ou não)
19	Atividades	Atividades extracurriculares (binário: sim ou não)
20	Creche	Cursou creche (binário: sim ou não)
21	Superio	Deseja cursar o ensino superior (binário: sim ou não)
22	Internet	Acesso à internet em casa (binário: sim ou não)
23	Romântico	Com um relacionamento romântico (binário: sim ou não)
24	Famrel	Qualidade das relações familiares (numérico: de 1 - muito ruim a 5 - excelente)
25	Tempo Livre	Tempo livre depois da escola (numérico: de 1 - muito baixo a 5 - muito alto)
26	Gooout	Saindo com os amigos (numérico : de 1 - muito baixo a 5 - muito alto)
224	Dalc	Consumo de álcool durante o trabalho (numérico: de 1 - muito baixo a 5 - muito alto)
28	Walc	Consumo de álcool no fim de semana (numérico: de 1 - muito baixo a 5 - muito alto)
29	Saúde	Estado de saúde atual (numérico: de 1 - muito ruim a 5 - muito bom)
30	Faltas	Número de faltas na escola (numérico: de 0 a 93)
32	G1	Nota do primeiro período (numérico: de 0 a 20)
32	G2	Nota do segundo período (numérico: de 0 a 20)
33	G3	Nota final (numérico: de 0 a 20, meta de saída)

Fonte: UCI – *Machine Learning*.

4.2 Segunda Etapa: Extração de recursos e limpeza de dados (Pré-Processamento e Transformação)

Com relação a segunda etapa os dados foram pré-processados para se adequar a aplicação das técnicas de MDE – algoritmos de AM – utilizadas na previsão do desempenho dos alunos, para isso foi utilizada a linguagem de programação e ciência de dados R, e várias tarefas foram realizadas:

1. Junção dos dados dos alunos das duas disciplinas em uma única base, para todos os alunos – como os dados disponíveis no UCI Machine Learning estavam divididos em duas bases de dados (uma para os dados da disciplina de matemática e outra para os dados de português), foi elaborado um script para unir os dados em uma única base.

2. Transformação do atributo G3 de numérico para níveis de classificação – nesse procedimento cada faixa de notas recebeu um valor no formato de caractere, atribuindo uma categoria/classe para os registros: notas entre 20 e 16 = "A", notas entre 15 e 11 = "B", notas entre 10 e 4 = "C", notas entre 4 e 0 = "D". Essa abordagem foi inspirada no estudo desenvolvido por Zhang e Wu (2019), os autores que também tinham como objetivo prever o desempenho de alunos, mais especificamente a previsão de notas dos alunos. Como detalhado na seção de trabalhos relacionados os autores dividiram os resultados dos alunos em classes: de 85 a 100 – A; de 70 a 85 – B; de 60 a 70 – C; e >60 – D, portanto tornando-se um

problema de classificação. Nesse sentido, considerou-se essa abordagem bastante eficaz e aderente ao estudo aqui desenvolvido, por isso optou-se por fazer de forma semelhante, realizando a categorização das notas, em quatro classes A, B, C e D.

3. Formatação dos dados como *Data Frame*: como a base de dados possui atributos numéricos e categóricos a única opção de formatação na linguagem R é o Data Frame.

4. Divisão da base de dados em treinamento e teste, em que 85% dos dados da base foram definidos para treino (888 registros para treino e 156 para teste): como a base de dados utilizada neste estudo é relativamente pequena (apenas 1044 linhas de informação), optou-se por usar mais dados para o treinamento dos algoritmos, e menos dados para sua avaliação. Existem autores na literatura que corroboram este tipo de adaptação quando se possui base de dados pequenas como, Japkowicz e Shah (2014).

5. Transformação do atributo a ser previsto (atributo meta – o desempenho final do aluno) para Factor (G3): para a aplicação de algoritmos de AM e AP com a linguagem R, o atributo meta deve estar configurado com o formato Factor, isso é um pré-requisito das bibliotecas do R. Este atributo é importante, pois é a nota final do aluno que é prevista no processo de predição de desempenho.

4.3 Terceira Etapa: Processamento Analítico e Algoritmos

No que tange a processamento analítico e algoritmos, em Cortez e Silva (2008), os dois conjuntos de dados, utilizados nessa pesquisa, foram modelados com uma classificação de cinco níveis e foi utilizada a regressão, pois o atributo que os autores queriam prever era a nota (numérico). Em contraste, neste estudo as notas de desempenho dos alunos foram classificadas como A, B, C ou D, e, portanto, foram empregados algoritmos de classificação para realização das previsões.

Tanto a classificação como a regressão são tarefas de Aprendizagem de Máquina Supervisionada, nesse tipo de técnica a base de dados possui colunas com categorias que servem para treinar o modelo, que deve, na próxima etapa, identificar as categorias de cada linha. A Aprendizagem Supervisionada é utilizada para resolver dois tipos diferentes de problemas: Classificação e Regressão. A Classificação refere-se ao processo de previsão de valores de categorias, ou no caso desse estudo uma faixa de notas que foi classificada com uma categoria. Problemas de Regressão buscam prever um valor numérico, por exemplo prever o preço de uma casa com base no endereço, ou valor de ações no mercado (KUBAT, 2017).

Alguns dos algoritmos de AM para classificação mais conhecidos são: *Naïve Bayes*; Árvores de Decisão, *Random Forest* (RF) e *Suport Vector Machines* (SVM). Estes algoritmos foram empregados neste estudo para previsão do desempenho dos alunos. Assim foi possível comparar quais desses algoritmos geram modelos mais eficazes. Para realização da aplicação desses algoritmos foram utilizadas bibliotecas específicas do R: *Naïve Bayes* – biblioteca "e1071"; Arvores de Decisão – biblioteca "rpart"; *Random Forest* – biblioteca "randomForest"; *Suport* Vector Machine – biblioteca "e1071".

Além das bibliotecas outra importante questão a ser considerada na aplicação de algoritmos de AM é a configuração de seus principais parâmetros, elementos que influem diretamente na eficácia dos modelos gerados. As configurações feitas nos algoritmos utilizados neste estudo foram: *Naïve Bayes*: Configuração *Default*; Árvores de Decisão: Configuração *Default*; *Random Forest:* Configuração

Default, e foi definido uma floresta com 30 árvores; SVM: Foi definido o kernel "radial" e um valor de custo de 5.0. Os valores dos parâmetros dos algoritmos *Randon Forest* e SVM diferentes do padrão, foram determinados, por meio de testes e consulta a documentação do R. Tais testes consistiram em definir os valores, rodar o algoritmo e verificar os resultados alcançados, estes foram repetidos até chegar a um conjunto de solução satisfatória, com relação às métricas de avaliação utilizadas; tais métricas são descritas na próxima seção. Todos os detalhes sobre as configurações padrão (*default*) podem ser analisadas no Apêndice A.

De acordo com Kubat (2017); Igual & Seguí (2017); e Aggarwal (2015), os algoritmos utilizados nesta pesquisa podem ser descritos da seguinte forma:

1. *Naïve Bayes* – é um algoritmo supervisionado de AM baseado no teorema de Bayes e fundamentado no princípio de independência de recurso, que afirma que os recursos de um conjunto de dados não têm relação entre si. Devido a essa suposição de independência, o algoritmo tem essa denominação de ingênuo e é o mais simples de todos os algoritmos de aprendizado de máquina e, no entanto, é muito aplicado por ser eficaz.

2. A Árvore de Decisão – é um algoritmo de AM baseado em entropia, o princípio por trás de seu trabalho é que cada atributo no conjunto de dados é tratado como um nó na árvore de decisão. Em cada nó é tomada uma decisão sobre qual caminho escolher na árvore, dependendo do valor do atributo nesse nó específico, o processo continua até que o nó da folha seja alcançado, porque esse contém a decisão final sobre a classificação da instância.

3. *Random Forest* – uma única Árvore de Decisão pode ser enviesada, dependendo dos dados, uma abordagem que pode melhorar essa falha é utilizar várias Árvores de Decisão que fazem sua própria previsão e a previsão final é encontrada calculando a média de todas as previsões feitas por todas as árvores. Essa abordagem é conhecida como ensemble learning (aprendizado em conjunto). No aprendizado em conjunto, vários algoritmos de tipos iguais ou diferentes são unidos para criar uma maior capacidade para o modelo de AM. O Random Forest é um tipo de modelo de aprendizado em conjunto, esse algoritmo une vários algoritmos de Árvore de Decisão, criando uma floresta.

4. *Suport Vector Machine* (SVM): O algoritmo SVM se originou nos anos 60 e é um dos mais famosos algoritmos de AM, e tem sido muito utilizado desde então, antes das Redes Neurais Artificiais se popularizarem ele era considerado o algoritmo de AM mais preciso. Para classificar uma nova instância, limites de decisão diferentes podem ser utilizados, dessa forma, objetivo do algoritmo SVM é encontrar o limite de decisão que classifica os registros de tal maneira que as chances de a classificação ser incorreta seja minimizada. O algoritmo faz isso maximizando a distância entre os atributos de instâncias mais próximos de todas as classes na base dados, e ele consegue encontrar esse limite com a ajuda de vetores de suporte, por isso o seu nome. Os vetores de suporte passam pelos atributos de dados mais próximos das classes para classificação, o trabalho do algoritmo é maximizar a distância entre esses vetores, traçando uma linha paralela no meio deles, esse limite de decisão é considerado o limite de decisão ideal. Uma das razões, pela qual o SVM é tão amplamente aplicável é que ele pode ser facilmente estendido para bases de dados complexas que não são linearmente separáveis. Isso é feito mapeando os registros de treinamento para um espaço de maior dimensão, onde eles se tornam um conjunto linearmente separável, essa técnica é denominada truque do kernel (kernel trick).

4.4 Quarta Etapa: Análise dos Resultados

Por fim, no intuito de complementar o processo proposto por Aggarwal (2015), acrescenta-se mais uma etapa no processo de Mineração de Dados que de acordo com o autor depende do propósito do estudo a ser desenvolvido, que é a interpretação dos resultados alcançados. Essa etapa nesse estudo pretende avaliar a eficácia de previsão do desempenho dos modelos gerados pelos algoritmos submetidos às bases de dados. Para realizar a verificação dos resultados de um modelo de classificação são necessários dois itens os métodos de avaliação e as métricas de interpretação, os dois devem ser aplicados em conjunto para que seja possível observar se um modelo é eficaz ou não. Os métodos indicam como esse modelo será avaliado, e as métricas traduzem os resultados da aplicação desses métodos em números que possam ser interpretados.

Para este estudo o método de avaliação empregado foi o de Treinamento e Teste, em que a base de dados é dividida de forma aleatória em duas porções, uma para treinamento e outra para teste, de acordo com Japkowicz e Shah (2014) geralmente fica 85% das instâncias para treinamento e 15% para teste. O algoritmo ao ser aplicado sobre a base de treinamento recolhe informações sobre os atributos das instâncias e gera um modelo de classificação ou regressão com base nesses atributos e informações, após isso esse modelo é aplicado sobre a base de teste (que contém registros diferentes da base de treinamento) e então as métricas de avaliação são calculadas sobre essa aplicação.

Apenas a aplicação do método de avaliação não indica se o modelo é eficaz ou não, para isso devem ser utilizadas métricas que possibilitem interpretação do quanto o modelo foi preciso em suas classificações, em outras palavras quantificar o seu desempenho. Algumas das métricas mais utilizadas no contexto de avaliação de modelos de Aprendizagem Supervisionada estão resumidas no Quadro 2.

Quadro 2 – Métricas de Avaliação de Algoritmos de ML e DL

MÉTRICA	DEFINIÇÃO
Precisão da classificação (Acurácia)	A precisão é uma métrica de avaliação comum para problemas de classificação. É o número de previsões corretas feitas como uma proporção de todas as previsões realizadas sobre a base de testes. Em outras palavras, é a porcentagem de instâncias classificadas corretamente de todas as instâncias, pode ser considerada mais útil em uma classificação binária do que em problemas de classificação de várias classes, porque pode ser menos claro exatamente como a precisão se divide nessas classes.
Intervalo de Confiança (IC)	Corresponde a uma métrica que indica que há uma probabilidade de 95% que a verdadeira precisão do modelo algorítmico testado esteja dentro desse intervalo.
Taxa de não informação	Essa é a precisão alcançável, sempre prevendo a categoria da classe majoritária. Portanto, corresponde a melhor escolha, sem outras informações.
Valor de P	Consiste em um teste unilateral para verificar se a *precisão* é melhor que a *taxa de não informação*, considerando a maior porcentagem da classe dos dados.
Kappa	Corresponde a uma medida de concordância usada em escalas nominais que fornece uma ideia do quanto as observações se afastam daquelas esperadas, fruto do acaso, indicando assim o quão legítimas as interpretações são. É parecida com a à precisão, excetuando por ser normalizada na linha de base do acaso no conjunto de dados. É passível de considerada uma medida mais utilizada para problemas com desequilíbrio nas classes.

Fonte: Adaptado Japkowicz e Shah (2014) e Documentação do R (2021).

5. RESULTADOS

Os resultados desse estudo foram sistematizados de acordo com as três questões de pesquisa, e são apresentados na sequência. Os resultados das demais métricas de avaliação de algoritmos de classificação, além das descritas no Quadro 3, podem ser visualizadas no Apêndice B.

QP1 – Qual a eficácia de modelos gerados a partir de algoritmos baseados em Aprendizagem de Máquina na previsão do desempenho de alunos no ensino tradicional?

Foi gerado um modelo para cada algoritmo selecionado para esse estudo, a partir de sua aplicação na base de dados, os resultados referentes ao método de avaliação Treinamento/Teste e as métricas mais proeminentes para a análise da eficácia do modelo estão disponíveis na Tabela 1.

Tabela 1 – Resultados das Análises dos algoritmos

Métrica	Naïve Bayes	Árvores de Decisão	*Random Forest*	SVM
Acurácia	0,66	0,87	0,83	0,82
Intervalo de confiança de 95%	58-74%	81-92%	77-89%	75-88%
Taxa de não informação	0,38	0,5	0,51	0,51
Valor de p	9.791e-13	2.2e-16	2.2e-16	8.754e-16
Kappa	0,51	0,79	0,73	0,71

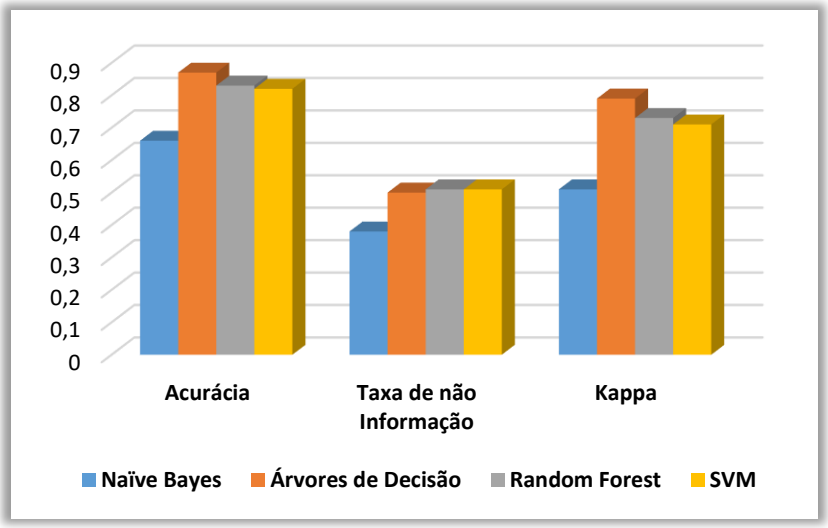
Fonte – Autora (2021)

Com esses resultados percebe-se que excluindo o algoritmo *Naïve Bayes,* que realmente tem uma estrutura bastante simplificada, os demais algoritmos, geraram bons desempenhos com precisão de classificação acima de 80% em todos os casos, sendo boas opções para a previsão do desempenho de alunos, em bases de dados compostas por registros demográficos, sociais e de rendimento escolar. Para uma melhor visualização dos resultados das análises dos algoritmos pode-se observar de forma ilustrativa os resultados da acurácia, taxa de não informação e *Kappa*[5] na Figura 2.

No que diz respeito as métricas apresentadas indicada na Tabela 1, a primeira é a acurácia – número de previsões corretas divididas pelo número total de previsões – correspondente a 87% no algoritmo de Árvore de Decisão, com um intervalo de confiança de 81-92%, o que significa que há uma probabilidade de 95% que a verdadeira precisão desse modelo esteja dentro desse intervalo. Logo após, encontra-se a taxa de não informação que corresponde a 50% para o modelo de Árvores de Decisão, essa métrica indica a precisão alcançável sempre prevendo a categoria da classe majoritária. Conforme o valor de *p,* pode-se afirmar que o modelo gerado por esses algoritmos oferece um desempenho significativamente melhor sobre a taxa de não informação. Na sequência, a estatística *Kappa*, que apresentou valor de 79%, mostra quão bem as previsões do modelo corresponderam as categorias reais da classe. De acordo com as diretrizes propostas por Landis e Koch, (1977) a *Kappa* nada mais é que uma concordância justa entre o modelo e as verdadeiras categorias de uma classe, uma vez que a precisão aleatória é controlada.

[5] Foram inseridas apenas essas métricas no gráfico, porque essas têm a mesma unidade de medida – a porcentagem.

Figura 2 – Resultados das Análises dos algoritmos



Fonte: Autora (2021)

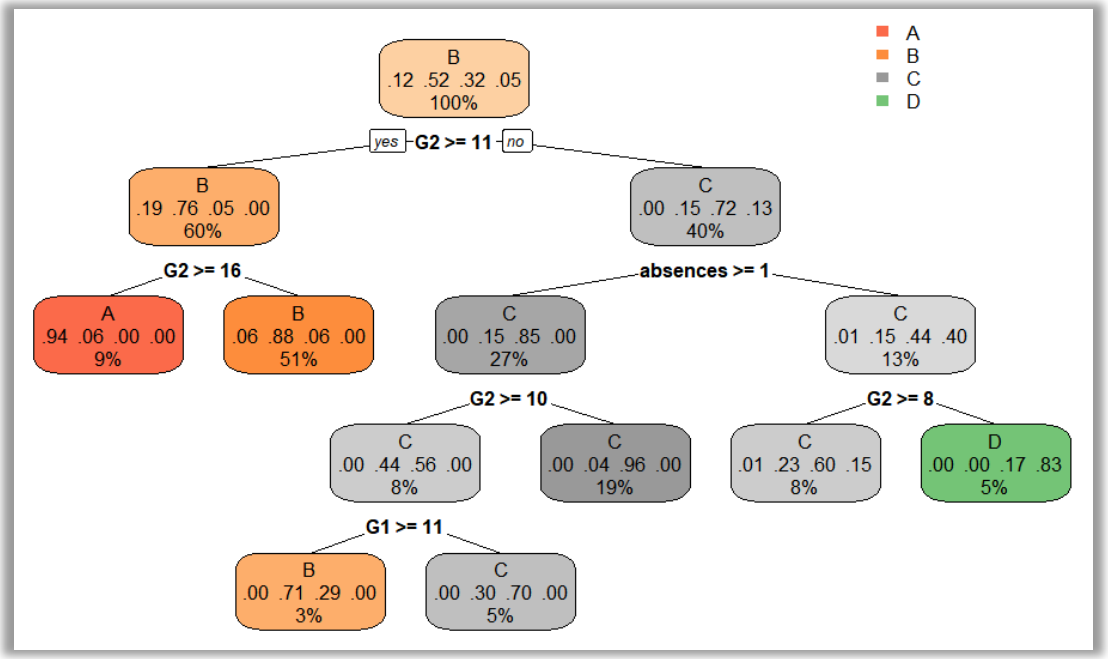
QP2 – Quais algoritmos de Aprendizagem de Máquina geram modelos mais eficazes?

De acordo com os resultados da avaliação dos modelos apresentados na seção anterior o algoritmo Árvore de Decisão gerou o modelo com maior acurácia na previsão do desempenho de alunos para a base de dados analisada. Embora o algoritmo *Random Forest* também tenha apresentado um bom desempenho, levando em conta a precisão, que foi maior para o modelo de Árvore de Decisão; a simplicidade de configuração, a qual o pesquisador pode até mesmo utilizar o padrão da linguagem de ciência de dados; recursos computacionais requeridos, que são mínimos para bases de dados como a utilizada neste estudo; e tempo de processamento, que é apenas maior que do *Naïve Bayes*, no geral duas vezes menor que do *Random Forest* e SVM; o algoritmo Árvore de Decisão seria a melhor escolha para realizar a predição nesse conjunto de registros. O que confirma o estudo de Zhang e Wu (2019) que salientaram que os modelos baseados em Árvores de Decisão são relativamente simples de serem implementados, e têm precisão relativamente satisfatória.

QP3 – Qual conjunto de atributos mais influência na previsão do desempenho de alunos?

Para definir quais os elementos mais influentes na previsão do desempenho, foi gerado o gráfico – Figura 3 – de Árvore de Decisão (com a biblioteca "rpart.plot"), este propicia visualizar quais são os atributos que estão mais no topo da árvore. Tais atributos, devido aos cálculos de entropia realizados, para geração da árvore de decisão, são os mais importantes para prever o atributo meta. Nesse sentido, não houve grandes descobertas, pois de acordo com a Figura 3 os principais atributos para prever o desempenho dos alunos são as notas (G2 na raiz da árvore e nos níveis 2 e 3; e G1 no nível 4), e a quantidade de faltas (*absences* no nível 2), dessa forma demonstrando que os atributos vinculados as atividades escolares são mais preditores para o desempenho que dados de características demográficas e socioeconômicas.

Figura 3 – Árvore de Decisão



Fonte: Autora

Todavia, não se pode descartar a influência desses elementos no desempenho dos alunos, pois é de conhecimento que estudantes podem ter seu rendimento escolar prejudicado, ou abaixo do esperado, por estar enfrentando alguma adversidade em casa, o que impacta em suas notas e pode ocasionar muitas faltas. Como não há registros de questionários aplicados para entender melhor esses elementos, não há uma confirmação dessas suposições, que são consideradas adequadas, mas não foram verificadas. Devido a essa falta de evidências, com base nos indicadores gerados pela aplicação das técnicas de MDE, os principais atributos que influenciam são os relacionados ao desempenho escolar.

6. CONSIDERAÇÕES FINAIS

Este estudo teve como principal objetivo realizar a previsão do desempenho de alunos utilizando um conjunto de dados público e comparar a eficácia da previsão dos modelos gerados pelos algoritmos de AM, processo que constatou que o Algoritmo Árvore de Decisão foi o mais preciso. Com isso, foi possível verificar que uma base de dados composta por registros de notas, frequência e características demográficas, (sociais e escolares) é suficiente para realizar a geração de modelos eficazes na previsão do desempenho dos alunos. Ademais, foi possível identificar que os atributos referentes as notas e as faltas dos alunos são os mais preditivos para o desempenho que os registros de características demográficas. Com relação à previsão de desempenho a técnica de MDE aplicada foi adequada, em que os resultados alcançados são os seguintes: *Naïve Bayes* com uma acurácia de 66%; Árvores de Decisão com 87%; *Random Forest* com 83%; *Suport Vector Machine* com 82%. Esses resultados apontam que o algoritmo Árvore de Decisão teve o melhor desempenho sobre esse conjunto de dados, o que confirma estudos mais amplos como o desenvolvido por Zhang e Wu (2019).

O estudo de Zhang e Wu (2019) apresentou uma abordagem metodológica similar a desenvolvida neste estudo, embora fosse esperado que os autores utilizassem Regressão, pois as notas eram valores contínuos, os autores dividiram os resultados dos alunos em classes, portanto tornando-se

um problema de classificação. Após a formatação e classificação da base de dados os autores realizaram a aplicação de 3 algoritmos de AM para geração de modelos de previsão de notas: ID3, C4.5 e CART – todos baseados em Árvore de Decisão. As precisões alcançadas pelos modelos sobre a base de dados de teste foram: IDE3 – 81%, C4.4 – 75%, CART – 76%. Nesse sentido, corroborando os resultados encontrados nesta pesquisa, Zhang e Wu (2019) afirmaram que os modelos baseados em Árvores de Decisão são consideravelmente simples de serem implementados, e têm precisão satisfatória, por isso devem ser empregados para apoiar ações que induzam a permanência de alunos nos diversos cenários educacionais.

Em relação conjunto de atributos com maior influência na previsão do desempenho de alunos não foi identificada uma descoberta relevante, pois de acordo com a Árvore de Decisão gerada, os atributos referentes as notas e as faltas dos alunos são os mais preditivos para o desempenho, fato que não provocou uma surpresa. Embora possa levantar questionamentos sobre o que ocasionou um desempenho abaixo do esperado em alguns alunos.

Por fim, esta pesquisa apresenta como principal contribuição mostrar a aplicação do processo de MDE, em um conjunto de dados público, que pode ser replicado por outros pesquisadores. Além disso, os resultados das avaliações dos algoritmos exibidos podem dar suporte na escolha de métodos mais eficazes para a aplicação em conjuntos de dados educacionais. Outra contribuição foi detalhar a possibilidade de manter modelos para previsão do desempenho armazenados, que podem ser aplicados sobre os dados dos alunos, auxiliando na resolução de problemas como a evasão e o baixo desempenho de forma antecipada, apoiando ações preventivas principalmente quando os alunos apresentam comportamentos aquém do esperado.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul.

7. REFERÊNCIAS

AGGARWAL, Charu C. **Data Mining: The Textbook**. 1. ed. New York, USA: Springer, 2015. v. 1*E-book*. Disponível em: https://doi.org/10.1007/978-3-319-14142-8

BAKER, Michael J. The roles of models in Artificial Intelligence and Education research : a prospective view. **Journal of Artificial Intelligence and Education**, *[S. l.]*, v. 11, p. 122–143, 2000.

BAKER, Ryan; ISOTANI, Seiji; CARVALHO, Adriana. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, *[S. l.]*, v. 19, n. 02, p. 3–13, 2011. Disponível em: https://doi.org/10.5753/rbie.2011.19.02.03

BISHOP, Christopher M.; PATTERN. **Pattern Recognition and Machine Learning**. 1. ed. Nova York, USA: Springer, 2006. *E-book*.

CHUI, Kwok Tai *et al.* Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. **Computers in Human Behavior**, *[S. l.]*, v. 107, n. December 2017, p. 105584, 2020. Disponível em: https://doi.org/10.1016/j.chb.2018.06.032

CORTEZ, P.; SILVA, A. Usando Data Mining para prever o desempenho dos alunos do ensino médio. **In A. Brito e J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference**

(FUBUTEC 2008). Porto, Portugal: EUROSIS, ISBN 978-9077381-39-7, 2008. p. 5-12.

DE LOS REYES, Daniel A. Guimarães *et al.* Predição de sucesso acadêmico de estudantes: uma análise sobre a demanda por uma abordagem baseada em transfer learning. **Revista Brasileira de Informática na Educação**, *[S. l.]*, v. 27, n. 1, p. 1–25, 2019. Disponível em: https://doi.org/10.5753/rbie.2019.27.01.01

DOCUMENTAÇÃO R. **Language R: Documentation**. *[s. l.]*, 2021. Disponível em: <https://www.r-project.org/other-docs.html>. Acesso em 28 de jun. 2021.

EDM. **Educational Data Mining**. *[s. l.]*, 2020. Disponível em: http://educationaldatamining.org/. Acesso em: 31 maio. 2020.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2008.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. California, USA: Springer, 2009. *E-book*.

IGUAL, Laura; SEGUÍ, Santi. **Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications**. 1. ed. *[S. l.]*: Springer, 2017. *E-book*. Disponível em: https://doi.org/10.1007/978-3-319-50017-1

JAPKOWICZ, Nathalie; SHAH, Mohak. **Evaluating Learning Algorithms: A Classification Perspective**. 1a Ed. ed. Cambridge: *[s. n.]*, 2014. *E-book*.

KUBAT, Miroslav. **An Introduction to Machine Learning**. 2. ed. Coral Gables, FL, USA: Springer, 2017. *E-book*. Disponível em: https://doi.org/10.1007/978-3-319-63913-0

LANDIS, J. Richard; KOCH, Gary G. This content downloaded from 185.2.32.58 on Tue. *[S. l.]*, v. 33, n. 2, p. 363–374, 1977.

MITCHELL, Tom M. **Machine Learning**. 1. ed. Nova York, USA: McGraw-Hill Science/Engineering/Math, 1997. *E-book*.

RIGO, Sandro José *et al.* Minerando Dados Educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. **Revista Brasileira de Informática na Educação**, *[S. l.]*, v. 22, n. 01, p. 168–177, 2014. Disponível em: https://doi.org/10.5753/RBIE.2014.22.01.132

RODRIGUES, Rodrigo Lins *et al.*Discovery engagement patterns MOOCs through cluster analysis. **IEEE Latin America Transactions**, *[S. l.]*, v. 14, n. 9, p. 4129–4135, 2016. Disponível em: https://doi.org/10.1109/TLA.2016.7785943

ROMERO, Cristobal; VENTURA, Sebastian. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, *[S. l.]*, v. 3, n. 1, p. 12–27, 2013. Disponível em: https://doi.org/10.1002/widm.1075

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, *[S. l.]*, v. 3, n. 3, p. 210–229, 1959. Disponível em: https://doi.org/10.1147/rd.33.0210

ZHANG, Yaling; WU, Bei. Research and application of grade prediction model based on decision tree algorithm. *In*: 2019, Chengdu, China. **Turing Celebration Conference (ACM TURC 2019)**. Chengdu, China: ACM, 2019. p. 1–6. Disponível em: https://doi.org/10.1145/3321408.3322857

Submissão: 17/05/2021

Aceito: 28/06/2021

APÊNDICE A – DESCRIÇÃO DAS CONFIGURAÇÕES DE PARÂMETROS DEFAULT DOS ALGORITMOS UTILIZADOS

CLASSIFICADOR NAÏVE BAYES

Biblioteca: e1071 – Método: naiveBayes
Descrição: Calcula as probabilidades a-posteriores condicionais de uma variável de classe categórica dadas variáveis preditoras independentes usando a regra de Bayes. **Detalhes**: O classificador Naïve Bayes padrão (pelo menos esta implementação) assume independência das variáveis preditoras e distribuição Gaussiana (dada a classe de destino) dos preditores métricos. Para atributos com valores ausentes, as entradas de tabela correspondentes são omitidas para previsão.

Aplicação

```
## S3 method for class 'formula'
naiveBayes(formula, data, laplace = 0, ..., subset, na.action = na.pass)
## Default S3 method:
naiveBayes(x, y, laplace = 0, ...)
## S3 method for class 'naiveBayes'
predict(object, newdata,type = c("class", "raw"), threshold = 0.001, eps = 0)
```

Para conhecer todas as descrições dos parâmetros default do classificador Naïve Bayes, acesse: https://cran.r-project.org/web/packages/e1071/e1071.pdf

---

ÁRVORE DE DECISÃO

Biblioteca: rpart– Método: rpart

Aplicação

```
rpart(formula, data, weights, subset, na.action = na.rpart, method, model = FALSE, x = FALSE, y = TRUE,
        parms, control, cost)
```

Para conhecer todas as descrições dos parâmetros default do algoritmo Árvore de Decisão, acesse: https://cran.r-project.org/web/packages/rpart/rpart.pdf

---

RANDON FOREST

Biblioteca: randomForest – Método: randomForest
Descrição: randomForest padrão do R implementa o algoritmo de floresta aleatória de Breiman (baseado no código Fortran original de Breiman e Cutler) para classificação e regressão. Também pode ser usado no modo não supervisionado para avaliar proximidades entre pontos de dados.
Aplicação

```
## S3 method for class 'formula'
randomForest(formula, data=NULL, ..., subset, na.action=na.fail)
## Default S3 method:
randomForest(x, y=NULL, xtest=NULL, ytest=NULL, ntree=500, mtry=if (!is.null(y) && !is.factor(y))
             max(floor(ncol(x)/3), 1) else floor(sqrt(ncol(x))), replace=TRUE, classwt=NULL, cutoff,
             strata, sampsize = if (replace) nrow(x) else ceiling(.632*nrow(x)), nodesize = if (!is.null(y)
             && !is.factor(y)) 5 else 1, maxnodes = NULL, importance=FALSE, localImp=FALSE,
             nPerm=1, proximity, oob.prox=proximity, norm.votes=TRUE, do.trace=FALSE,
             keep.forest=!is.null(y) && is.null(xtest), corr.bias=FALSE, keep.inbag=FALSE)
## S3 method for class 'randomForest'
print(x, ...)
```

Para conhecer todas as descrições dos parâmetros default do algoritmo Random Forest, acesse: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

---

SUPORT VECTOR MACHINES

Biblioteca: e1071 – Método: SVM
Descrição: O SVM é usado para treinar uma máquina de vetores de suporte. Ele pode ser usado para realizar regressões e classificações gerais (do tipo nu e épsilon), bem como estimativas de densidade.

Aplicação

```
## S3 method for class 'formula'
svm(formula, data = NULL, ..., subset, na.action = na.omit, scale = TRUE)
## Default S3 method:
svm(x, y = NULL, scale = TRUE, type = NULL, kernel = "radial", degree = 3, gamma = if (is.vector(x)) 1 else
    1 / ncol(x), coef0 = 0, cost = 1, nu = 0.5, class.weights = NULL, cachesize = 40, tolerance = 0.001,
    epsilon = 0.1, shrinking = TRUE, cross = 0, probability = FALSE, fitted = TRUE, ..., subset, na.action =
    na.omit)
```

Para conhecer todas as descrições dos parâmetros default do algoritmo SVM, acesse: https://cran.r-project.org/web/packages/e1071/e1071.pdf

---

APÊNDICE B – DEMAIS MÉTRICAS E RESULTADOS DOS ALGORITMOS AVALIADOS.

Quadro 4 – Métricas de Avaliação de Algoritmos de ML e DL

MÉTRICA	DEFINIÇÃO
Matriz de confusão	Uma matriz de confusão fornece uma análise mais detalhada das classificações corretas e incorretas para cada classe. Uma breve explicação de como interpretar uma matriz de confusão é a seguinte: os elementos da diagonal principal representam o número de pontos para os quais o rótulo previsto é igual ao rótulo verdadeiro, enquanto qualquer coisa fora da diagonal principal foi rotulada incorretamente pelo classificador. Portanto, quanto mais altos os valores presentes na diagonal principal da matriz de confusão, melhor, indicando muitas previsões corretas.
Precisão da classificação (Acurácia)	A precisão é uma métrica de avaliação comum para problemas de classificação. É o número de previsões corretas feitas como uma proporção de todas as previsões realizadas sobre a base de testes. Em outras palavras, é a porcentagem de instâncias classificadas corretamente de todas as instâncias, pode ser considerada mais útil em uma classificação binária do que em problemas de classificação de várias classes, porque pode ser menos claro exatamente como a precisão se divide nessas classes.

Intervalo de Confiança (IC)	Corresponde a uma métrica que indica que há uma probabilidade de 95% que a verdadeira precisão do modelo algorítmico testado esteja dentro desse intervalo.
Taxa de não informação	Essa é a precisão alcançável, sempre prevendo o rótulo da classe majoritária. Portanto, corresponde a melhor escolha, sem outras informações.
Valor de P	Consiste em um teste unilateral para verificar se a *precisão* é melhor que a *taxa de não informação*, considerando a maior porcentagem da classe dos dados.
Kappa	Corresponde a uma medida de concordância usada em escalas nominais que fornece uma ideia do quanto as observações se afastam daquelas esperadas, fruto do acaso, indicando assim o quão legítimas as interpretações são. É parecida com a à precisão, excetuando por ser normalizada na linha de base do acaso no conjunto de dados. É passível de considerada uma medida mais utilizada para problemas com desequilíbrio nas classes.
Área sob curva (AUC) – Taxas de Sensibilidade e Especificidade	A área sob a curva é uma métrica de desempenho para medir a capacidade de um classificador binário de discriminar entre classes positivas e negativas. Exemplos: 1) Uma área de 1,0 representa um modelo que fez todas as previsões perfeitas; 2) Uma área de 0,5 representa um modelo tão bom quanto aleatório. A AUC pode ser dividida em *Sensibilidade* e *Especificidade*. *Sensibilidade* é a verdadeira taxa positiva, são as instâncias numéricas da classe positiva que realmente foram previstas como positivas. A *Especificidade* é a verdadeira taxa negativa, ou seja, é o número de instâncias da classe negativa que foram realmente previstas como negativa.
Valores Preditivos Positivo e Negativo	*Valor preditivo positivo* – mostra o número da classe positiva prevista corretamente como uma proporção do total de previsões da classe positiva realizadas. *Valor preditivo negativo* – mostra o número da classe negativa prevista corretamente como uma proporção do total de previsões da classe negativa realizadas. Esses parâmetros descrevem o desempenho de um teste de diagnóstico.
Prevalência	Mostra com que frequência a classe positiva realmente ocorre na amostra.
Taxa de detecção	Denota o número de previsões positivas corretas da classe feitas como uma proporção de todas as previsões realizadas.
Prevalência de detecção	Apresenta o número de previsões positivas de classe feitas como uma proporção de todas as previsões realizadas.
Precisão Balanceada	Atribui essencialmente a média das taxas reais positivas e negativas, isto é – (sensibilidade + especificidade) / 2.

Fonte: Adaptado Japkowicz e Shah (2014) e Documentação do R.

1. Naïve Bayes

Confusion Matrix and Statistics

previsoes

	A	B	C	D
A	17	1	0	0
B	9	49	13	10
C	0	10	30	9
D	0	0	0	8

Overall Statistics

Accuracy : 0.6667

95% CI : (0.5868, 0.74)

No Information Rate : 0.3846

P-Value [Acc > NIR] : 9.791e-13

Kappa : 0.5138

Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.6538	0.8167	0.6977	0.2963
Specificity	0.9923	0.6667	0.8319	1.0000
Pos Pred Value	0.9444	0.6049	0.6122	1.0000
Neg Pred Value	0.9348	0.8533	0.8785	0.8716

```
Prevalence             0.1667   0.3846   0.2756  0.173
08
Detection Rate         0.1090   0.3141   0.1923  0.05
128
Detection Prevalence   0.1154   0.5192   0.3141  0.
05128
Balanced Accuracy      0.8231   0.7417   0.7648  0.
64815
```

2. Árvores De Decisão

```
Confusion Matrix and Statistics
   previsoes
    A  B  C  D
 A 14  4  0  0
 B  0 72  9  0
 C  0  2 47  0
 D  0  0  4  4

Overall Statistics
               Accuracy : 0.8782
                 95% CI : (0.8164, 0.9251)
    No Information Rate : 0.5
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.7996

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: A Class: B Class: C Class: D
Sensitivity          1.00000   0.9231   0.7833  1.0000
0
Specificity          0.97183   0.8846   0.9792  0.9736
8
Pos Pred Value       0.77778   0.8889   0.9592  0.5
0000
Neg Pred Value       1.00000   0.9200   0.8785  1.0
0000
Prevalence           0.08974   0.5000   0.3846  0.02
564
Detection Rate       0.08974   0.4615   0.3013  0.0
2564
Detection Prevalence 0.11538   0.5192   0.3141  0.
05128
Balanced Accuracy    0.98592   0.9038   0.8812  0.
98684
```

3. Random Forest

```
Confusion Matrix and Statistics
   previsoes
    A  B  C  D
 A 12  6  0  0
 B  0 71 10  0
 C  0  3 46  0
 D  0  0  6  2

Overall Statistics
               Accuracy : 0.8397
                 95% CI : (0.7726, 0.8935)
    No Information Rate : 0.5128
P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.7326

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: A Class: B Class: C Class: D
Sensitivity          1.00000   0.8875   0.7419  1.0000
0
Specificity          0.95833   0.8684   0.9681  0.9610
4
Pos Pred Value       0.66667   0.8765   0.9388  0.2
5000
Neg Pred Value       1.00000   0.8800   0.8505  1.0
0000
Prevalence           0.07692   0.5128   0.3974  0.01
282
Detection Rate       0.07692   0.4551   0.2949  0.0
1282
Detection Prevalence 0.11538   0.5192   0.3141  0.
05128
Balanced Accuracy    0.97917   0.8780   0.8550  0.
98052
```

4. Suport Vector Machines

```
Confusion Matrix and Statistics
   previsoes
    A  B  C  D
 A 13  5  0  0
 B  3 70  8  0
 C  0  6 42  1
 D  0  0  4  4

Overall Statistics

               Accuracy : 0.8269
```

95% CI : (0.7583, 0.8827)
No Information Rate : 0.5192
P-Value [Acc > NIR] : 8.754e-16

Kappa : 0.7154

Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: A	Class: B	Class: C	Class: D
Sensitivity	0.81250	0.8642	0.7778	0.8000
0				
Specificity	0.96429	0.8533	0.9314	0.9735
1				
Pos Pred Value	0.72222	0.8642	0.8571	0.5
0000				
Neg Pred Value	0.97826	0.8533	0.8879	0.9
9324				
Prevalence	0.10256	0.5192	0.3462	0.03
205				
Detection Rate	0.08333	0.4487	0.2692	0.0
2564				
Detection Prevalence	0.11538	0.5192	0.3141	0.
05128				
Balanced Accuracy	0.88839	0.8588	0.8546	0.
88675				