2022 | Volume 6 | Pág. 183 a 198

DOI: https://doi.org/10.15536/reducarmais.6.2022.2691

Revista Educar Mais

Mineração de dados educacionais com algoritmos de regressão: um estudo sobre a predição do desempenho

Educational data mining with regression algorithms: a study on performance prediction

Minería de datos educativos con algoritmos de regresión: un estudio sobre predicción del rendimiento

Vanessa Faria de Souza[1]; Sílvio Cézar Cazella[2]

RESUMO

Com o aumento da disponibilidade de dados, sobretudo no contexto educacional, a Mineração de Dados Educacionais (MDE) está tornando-se cada vez mais importante para a tomada de decisão neste contexto. Um dos principais objetivos da MDE é a predição do desempenho, pois quando se sabe de forma antecipada sobre a performance dos alunos, é possível intervir evitando reprovações, e até mesmo a evasão. Nesse sentido, esse estudo tem como objetivo realizar a previsão do desempenho de alunos, em um conjunto de dados públicos dividido em duas bases de dados menores, uma referente a disciplina de Português e outra com dados relativos à disciplina de Matemática, utilizando algoritmos de Regressão. Ademais, pretende-se 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); 3) Processamento analítico e algoritmos; e 4) Análise dos Resultados. Como resultados foram identificados que no conjunto de dados da disciplina de Matemática o algoritmo de Árvores de Decisão foi o mais preciso, com uma acurácia de 90%, e o Random Forest teve o melhor desempenho para os dados referentes a disciplina de Português, 80% de precisão. Além disso, constatou-se que atributos relacionados as atividades escolares são mais preditores para o desempenho dos alunos, todavia alguns atributos decorrentes de características demográficas e socioeconômicas também influenciam no desempenho.

Palavras-chave: Mineração de Dados Educacionais; Aprendizagem de Máquina; Algoritmos de Regressão; Predição do Desempenho.

ABSTRACT

With the increase in data availability, especially in the educational context, Educational Data Mining (EDM) is becoming increasingly important for decision making in this context. One of the main objectives of the MDE is the prediction of performance, because when it is known in advance about the performance of students, it is possible to intervene preventing failures, and even evasion. In this sense, this study aims to predict the performance of students, in a set of public data, divided into two smaller databases, one referring to the Portuguese subject and the other with data related to the Mathematics subject, using algorithms of Regression. Furthermore, it is intended to indicate which are the main predictors of student performance. For this, an MDE process was implemented based on 4 steps: 1) Data collection; 2) Resource extraction and data cleaning (Pre-Processing and Transformation); 3) Analytical processing and algorithms; and 4) Analysis of Results. As a result, it was identified that in the dataset of the Mathematics discipline, the Decision Trees algorithm was the most

[1] Licenciatura em Matemática, graduada em Sistemas de Informação, Mestre em Informática e Doutoranda no Programa de Pós-Graduação em Informática na Educação (PPGIE) da Universidade Federal do Rio Grande do Sul (UFRGS). Professora do Instituto Federal do Rio Grande do Sul (IFRS), Ibirubá/RS - Brasil. E-mail: vanessa.souza@ibiruba.ifrs.edu.br

[2] Graduado em Informática, Mestre e Doutor em Ciência da Computação e Professor da Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre/RS - Brasil. E-mail: silvioc@ufcspa.edu.br

e-ISSN 2237-9185

accurate, with an accuracy of 90%, and Random Forest had the best performance for the data referring to the Portuguese discipline, 80% of precision. In addition, it was found that attributes related to school activities are more predictors of student performance, however some attributes resulting from demographic and socioeconomic characteristics also influence performance.

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

RESUMEN

Con el aumento de la disponibilidad de datos, especialmente en el contexto educativo, la Minería de Datos Educativos (EDM) se está volviendo cada vez más importante para la toma de decisiones en este contexto. Uno de los principales objetivos del MDE es la predicción del desempeño, ya que cuando se conoce de antemano el desempeño de los estudiantes, es posible intervenir previniendo los fracasos, e incluso la evasión. En este sentido, este estudio tiene como objetivo predecir el desempeño de los estudiantes, en un conjunto de datos públicos, divididos en dos bases de datos más pequeñas, una referida a la asignatura de portugués y otra con datos relacionados con la asignatura de Matemáticas, utilizando algoritmos de regresión. Además, se pretende señalar cuáles son los principales predictores del rendimiento de los estudiantes. Para ello se implementó un proceso MDE basado en 4 pasos: 1) Recolección de datos; 2) Extracción de recursos y limpieza de datos (Preprocesamiento y Transformación); 3) Procesamiento analítico y algoritmos; y 4) Análisis de Resultados. Como resultado, se identificó que en el conjunto de datos de la disciplina Matemáticas, el algoritmo Decision Trees fue el más preciso, con una precisión del 90%, y Random Forest tuvo el mejor desempeño para los datos referentes a la disciplina portuguesa, 80%. de precisión Además, se encontró que los atributos relacionados con las actividades escolares son más predictores del rendimiento de los estudiantes, sin embargo, algunos atributos resultantes de las características demográficas y socioeconómicas también influyen en el rendimiento.

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

1. INTRODUÇÃO

O desempenho acadêmico dos alunos é fonte de grande preocupação e interesse para a maioria das pesquisas na área da educação, instituições de ensino e setores governamentais dedicados a educação. É essencial que as instituições de ensino monitorem o desempenho acadêmico de seus alunos e tomem as respectivas medidas de melhoria (SINGH; PAUL, 2020). Os educadores devem avaliar o desempenho dos alunos para atender aos objetivos definidos e promover um ambiente de melhoria contínua (TALAL; SAEED, 2019), pois o desempenho acadêmico dos alunos é um dos principais fatores que são considerados para avaliar a performance de uma instituição de ensino, sendo um dos mais proeminentes, um fator-chave nas classificações de qualidade dessas instituições (YAACOB; NASIR; YAACOB, 2019). Devido à importância deste aspecto os educadores podem obter insights sobre os obstáculos encontrados pelos alunos para alcançar um desempenho acadêmico superior, traçando estratégias para aprimorar as habilidades dos alunos e implementar ações corretivas para melhorar o mau desempenho, e uma das estratégias mais utilizadas atualmente para se obter esses insights é por meio da aplicação da Mineração de Dados Educacionais (MDE) (ALMASRI; CELEBI; ALKHAWALDEH, 2019).

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,

e-ISSN 2237-9185

caso contrário a tarefa torna-se impraticável (BAKER, 2015). 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 (ROMERO; VENTURA, 2020).

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 (ROMERO; VENTURA, 2020). 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 (BAKER; INVENTADO, 2014; BAKER, 2015; ROMERO; VENTURA, 2013; ROMERO; VENTURA, 2020). A Mineração de Dados Educacionais (MDE) utiliza as técnicas da Mineração de Dados (MD) para extrair informações relevantes de conjuntos diversificados de dados educacionais, e pode ser definida da seguinte forma, segundo a Sociedade Internacional de Mineração de Dados Educacionais:

> É uma disciplina emergente, preocupada com o desenvolvimento de métodos para explorar dados únicos e cada vez mais em larga escala, provenientes de contextos educacionais e usa esses métodos para entender melhor os alunos e as configurações em que aprendem (MDE, 2020).

Um dos principais objetivos da MDE é a previsão do desempenho de alunos, isso pode ser visto no mapeamento sistemático de literatura de Souza e Perry (2020), neste foram encontrados 18 estudos que abordam esta temática, sendo a segunda mais investigada entre os pesquisadores (perdendo apenas para Análise de Comportamento). Pois, como já salientado, este é um fator que demanda muita atenção no setor educacional, visto que um dos critérios para que escolas e universidades sejam consideradas de alta qualidade é baseado no histórico de realizações acadêmicas dos alunos (SHAHIRI; HUSAIN; RASHID, 2015). Nesta tarefa a MDE é utilizada para identificar com antecedência como será a performance do aluno no decorrer do curso, para poder intervir caso necessário e assim melhorar seu processo de aprendizagem (SOUZA; PERRY, 2020). Dessa forma, Shahiri, Husain e Rashid (2015) afirmam que prever o desempenho dos alunos é muito útil para ajudar educadores e alunos a melhorar o processo de ensino e aprendizagem. Os autores ainda salientam, 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, e apoiar em intervenções para a diminuição da evasão.

Diante desse contexto, esse estudo tem como objetivo realizar a previsão do desempenho de alunos, por meio da aplicação do processo de MDE, utilizando como técnica algoritmos de Aprendizagem de Máquina (AM), em um conjunto de dados público do repositório UCI – *Machine Learning*[3]; e analisar a precisão desses algoritmos. A abordagem utilizada foi a aprendizagem de máquina supervisionada para regressão, em que são previstas as notas dos alunos, sendo considerados seus valores numéricos. 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, e sobretudo avaliar quais desses atributos têm maior influência sobre as notas finais.

Para esse fim, este documento está organizado da seguinte forma: na seção 2 são abordados alguns trabalhos relacionados, destacando três pesquisas que empregaram a MDE com algoritmos de AM na

[3] https://archive.ics.uci.edu/ml/datasets/Student+Performance

e-ISSN 2237-9185

previsão do desempenho; a seção 3 apresenta uma breve descrição da amostra, a seção 4 trata da metodologia, em que é desenvolvido o processo de Mineração de Dados proposto por Aggarwal (2015); na seção 5 são expostos os resultados alcançados; por fim, na seção 6 são descritas as conclusões da autora.

2. TRABALHOS RELACIONADOS

Muitas pesquisas têm sido conduzidas com o objetivo de predizer o desempenho de alunos, tanto nos anos iniciais de formação, na educação básica, como na graduação. Nesta seção são abordados alguns destes estudos, que similarmente a este, utilizaram a MDE para realização da previsão do desempenho e como técnica principal empregaram algoritmos de AM.

Em seu estudo Dabhade *et al.* (2021) tinham como objetivo compreender os vários fatores dos quais o desempenho dos alunos depende, e prever o desempenho dos estudantes para o próximo semestre com base nos dados de perfis e no desempenho anterior. Ademais, os autores buscaram identificar os atributos que mais afetam o desempenho dos alunos, isso foi feito de forma análoga ao presente estudo, em que os autores utilizaram a matriz de correlação, para identificação desses atributos. O conjunto dos registros dos alunos utilizados pelos autores foi composto por dados de perfil, coletados por meio de um questionário; e dados relacionados ao curso realizado, coletados na seção acadêmica da instituição. Assim, a base de dados foi composta por 112 atributos, relativos à dados: pessoais, educacionais, comportamentais e extracurriculares. Para gerar os modelos de predição do desempenho dos alunos os autores utilizaram dois algoritmos para regressão: Regressão Linear Múltipla e *Suport Vector Machines* (SVM), todavia o SVM foi aplicado com 3 diferentes tipos de núcleos (*linear, rbf e poly*). Para o desenvolvimento dos modelos foi utilizado o Python, e a avaliação foi realizada com o método de Treinamento/Teste, com 80% da base para treino e 20% para teste. Como resultados Dabhade *et al.* (2021) obtiveram que o modelo mais eficiente foi o gerado pelo algoritmo SVM com núcleo linear, chegando a 83,44% de precisão na predição do desempenho dos alunos. Além disso, os modelos propostos forneceram evidências de que o desempenho passado recente é mais importante para a previsão do desempenho futuro, bem como existe uma relação entre as características de comportamento dos alunos e o desempenho acadêmico.

Na pesquisa realizada por Malini e Kalpana (2021) foi utilizada a mesma base de dados pública analisada no presente estudo, e o objetivo dos autores foi empregar a MDE para caracterizar os distintos fatores (atributos) que afetam as notas finais dos alunos, e com base nestes realizar uma predição do desempenho. Para isso, os autores realizaram uma redução da dimensionalidade da base para chegarem a um conjunto de atributos ideais para o processo de previsão, em que foram utilizados alguns tipos de aprimoramentos de algoritmos para a geração dos modelos: *Boosting* e *Bagging*, e o algoritmo de Redes Neurais Artificiais do tipo *Multilayer Perceptron* (MLP). Os autores utilizaram o método Treinamento/Teste para avaliação dos modelos e diferentes métricas como: Exatidão, Taxa de Verdadeiro Positivo, Taxa de Falso Positivo, Acurácia/Precisão, *Recall*, *F-Measure* e Matriz de confusão. Para o desenvolvimento das aplicações os autores empregaram o software WEKA. Como resultados Malini e Kalpana (2021) encontraram de forma geral que as Redes Neurais tipo MLP tiveram 72% de precisão, os algoritmos *Bagging* alcançaram 88% e os *Boosting* atingiram 86% de precisão, além disso os autores ressaltaram que os dados do contexto econômico desempenham um papel importante na vida do estudante que os afeta de várias maneiras, sobretudo na aprendizagem.

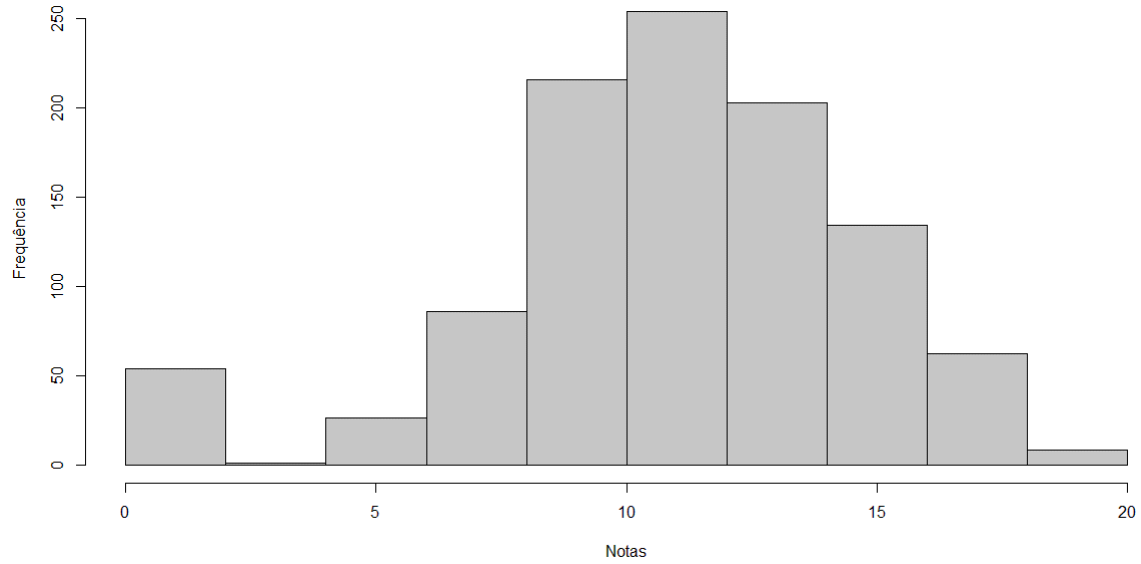
e-ISSN 2237-9185

Por fim, Riestra-González, Paule-Ruíz e Ortin (2021) utilizaram algoritmos de Aprendizagem de Máquina para criar modelos para a previsão inicial do desempenho dos alunos na resolução de tarefas em Ambientes Virtuais de Aprendizado (AVA), analisando os arquivos de logs gerados até o momento da previsão. Os autores relataram que os modelos eram independentes de curso, porque os conjuntos de dados foram criados com todos os cursos da universidade analisada para um ano letivo. Dessa forma, os autores previram o desempenho dos alunos em 10%, 25%, 33% e 50% da duração de um curso, ofertado. Para realização das previsões os autores empregaram os algoritmos: Árvore de decisão, *Naïve Bayes*, Regressão Logística, SVM e Redes Neurais Artificiais do tipo MLP. Para avaliação os autores dividiram os dados em um conjunto com 80% das instâncias para treinamento e 20% para teste, usando um método de amostragem aleatória estratificada, então repetiram o processo de treinamento e teste 30 vezes, medindo a média da precisão, desvio padrão, intervalos de confiança de 95% de precisão, medida de F1 e valores da Área sob a Curva (AUC). Para a construção e aplicação dos modelos os autores utilizaram o Python. Riestra-González, Paule-Ruíz e Ortin (2021) relataram que a exatidão de todos os modelos aumenta à medida que avança o momento da previsão. Como resultados os autores identificaram que as Redes Neurais do tipo MLP obtiveram o melhor desempenho: sendo de 80,1% de precisão quando 10% do curso foi entregue; e 90,1% quando a metade do curso foi realizado.

3. ANÁLISE DESCRITIVA DOS DADOS

As bases de dados extraídas do UCI – *Machine Learning* são referentes aos dados de 1044 alunos de duas escolas do Ensino Médio de Portugal (a nomenclatura utilizada em Portugal é Ensino Secundário), com relação as disciplinas de Matemática e Português. Devido as altas taxas de reprovação identificadas, 383 alunos, cerca de 37% dos alunos analisados, a realização da predição do desempenho se torna uma ferramenta importante para idealização e desenvolvimento de intervenções sempre que necessário, amenizando reprovações, e a evasão. Sobretudo, no que tange a disciplina de Matemática, se tratando especificamente sobre esta porção dos dados, há uma taxa de 47% de reprovações, quase a metade dos alunos analisados. Na Figura 2 é possível visualizar o histograma do desempenho dos alunos (nota final). Cabe destacar que a escala de notas é a utilizada na Europa, para serem aprovados os alunos precisam tirar notas finais iguais ou superiores a 10, como mostra o Quadro 1.

Figura 1 – Histograma das Notas Finais



Fonte: Autores (2021)

e-ISSN 2237-9185

Quadro 1 – Escala das Notas Finais

PAÍS	I Excelente	II Bom	III Satisfatório	IV Suficiente	V Reprovação
Portugal/França	16-20	14-15	12-13	10-11	0-9
Escala da Maioria dos Países Europeus	A	B	C	D	F

Fonte: Cortez e Silva (2008)

Os alunos investigados tinham idade entre 15 e 22, e maior parte estava concentrada entre 16 e 18 sendo 74% da amostra. A grande maioria dos alunos nunca havia tido uma reprovação, cerca de 82%. Quando questionados sobre o tempo que dedicam aos estudos fora de sala de aula, a maioria dos alunos, 48%, disseram estudar duas horas por dia (em uma escolha de 1 a 4 horas). Cerca de 91% dos alunos têm a intenção de realizar um curso superior, e 49% se dedicam a atividades extracurriculares.

4. METODOLOGIA

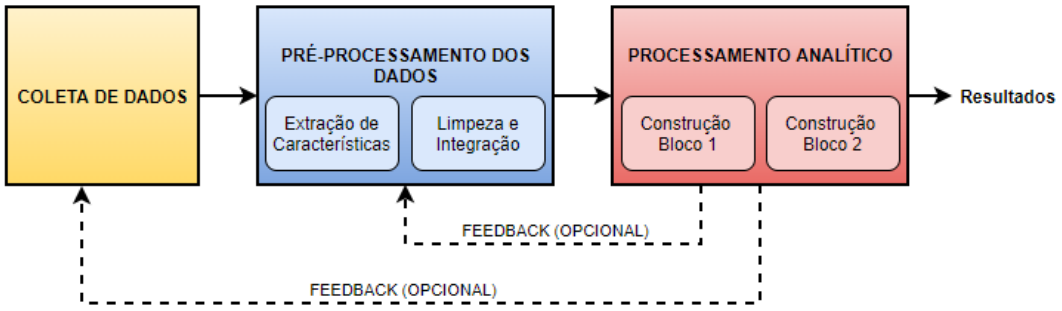
Este estudo tem como principal objetivo realizar a previsão do desempenho de alunos, em duas disciplinas, utilizando técnicas de MDE. Com a realização desse processo é também possível identificar qual dos algoritmos de regressão aplicados foi o mais eficaz na predição do desempenho, considerando os atributos 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 o desempenho de modelos gerados a partir de algoritmos baseados em Regressão na previsão do desempenho dos alunos?
- ✓ Questão de pesquisa 2 (QP2) – Qual dos algoritmos de Regressão gera o modelo com melhor desempenho?
- ✓ Questão de pesquisa 3 (QP3) – Qual conjunto de atributos tem maior influência sobre as notas finais dos 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). No que se refere ao processo de aplicação da MDE, este é similar ao da Mineração de Dados. 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, inserindo uma quarta etapa no processo. A sequência das etapas do processo proposto por Aggarwal (2015) é apresentada na Figura 2, na qual pode-se observar que o processo de MD pode ser iterativo.

e-ISSN 2237-9185

Figura 2 – Processo de Data Mining proposto por Aggarwal (2015)



Fonte: Adaptado Aggarwal (2015)

4.1 Primeira Etapa: Coleta de Dados

Nesta primeira etapa os dados foram coletados do repositório de dados público: UCI – *Machine Learning*. Estes dados abordam o desempenho dos alunos no Ensino Médio de duas escolas portuguesas, como já salientado. Os atributos dos alunos incluem notas, características demográficas, (sociais e escolares) e foram reunidos por meio de relatórios escolares e questionários e podem ser divididos em 3 seguimentos: (1) *Atributos pessoais* – sex, age, address, health; (2) *Atributos comportamentais* – romantic, Dalc, Walc, higher, goout, famrel; (3) *Antecedentes acadêmicos* – studytime, failures, paid, activities, nursery, absences; e (4) *Antecedentes econômicos* – pstatus, fedu, famsize, fjob, medu, reason, guardian, mjob, famsup, traveltime, freetime, internet, schoolsup. Foram fornecidos dois conjuntos de dados com 33 atributos sobre 1044 alunos em duas disciplinas: Matemática (395 registros) e Língua Portuguesa (649 registros). Os dados sistematizados formaram dois Data Frames um com os dados da disciplina de Matemática (395 linhas e 33 colunas), e o outro referente a Português (649 linhas e 33 atributos). Os atributos constantes nas bases de dados estão descritos na Figura 3.

Figura 3 – Atributos da Base de Dados

ID	ATRIBUTOS	DESCRIÇÃO	ID	ATRIBUTOS	DESCRIÇÃO
1	school	Escola do aluno (binário: 'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira)	18	paid	Aulas extras pagas dentro da disciplina (matemática ou português) (binário: sim ou não)
2	sex	Gênero do aluno (binário: 'F' - feminino ou 'M' - masculino)	19	activities	Atividades extracurriculares (binário: sim ou não)
3	age	Idade do aluno (numérico: de 15 a 22)	20	nursery	Cursou creche (binário: sim ou não)
4	address	Tipo de endereço residencial do aluno (binário: 'U' - urbano ou 'R' - rural)	21	higher	Deseja cursar o ensino superior (binário: sim ou não)
5	famsize	Tamanho da família (binário:' LE3 '- menor ou igual a 3 ou' GT3 '- maior que 3)	22	internet	Acesso à internet em casa (binário: sim ou não)
6	Pstatus	Status de coabitação dos pais (binário:' T '- morando junto ou' A '- à parte)	23	romantic	Com um relacionamento romântico (binário: sim ou não)
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)	24	famrel	Qualidade das relações familiares (numérico: de 1 - muito ruim a 5 - excelente)
8	Fedu	Escolaridade do pai (numérico: 0 - nenhuma, 1 - ensino primário (4° ano), 2 - 5° ao 9° ano, 3 - ensino secundário ou 4 - ensino superior)	25	freetime	Tempo livre depois da escola (numérico: de 1 - muito baixo a 5 - muito alto)
9	Mjob	Trabalho da mãe (nominal: 'professor', 'saúde' relacionado, 'serviços' civis (por exemplo, administrativo ou policial), 'em_casa' ou 'outro')	26	goout	Saindo com os amigos (numérico: de 1 - muito baixo a 5 - muito alto)
10	Fjob	Trabalho do pai (nominal: 'professor', 'saúde' relacionado, civil 'serviços' (por exemplo, administrativo ou policial), 'em_ casa' ou 'outro')	224	Dalc	Consumo de álcool durante o trabalho (numérico: de 1 - muito baixo a 5 - muito alto)
11	reason	Razão para escolher esta escola (nominal: perto de 'casa', escola 'reputação', 'curso' preferência ou 'outro')	28	Walc	Consumo de álcool no fim de semana (numérico: de 1 - muito baixo a 5 - muito alto)
12	guardian	Tutor do aluno (nominal: 'mãe', 'pai' ou 'outro')	29	health	Estado de saúde atual (numérico: de 1 - muito ruim a 5 - muito bom)
13	traveltime	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)	30	absences	Número de faltas na escola (numérico: de 0 a 93)
14	studytime	Tempo de estudo semanal (numérico: 1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas ou 4 -> 10 horas)	32	G1	Nota do primeiro período (numérico: de 0 a 20)
15	failures	Número de reprovações anteriores nas aulas (numérico: n se 1 <= n <3, senão 4)	32	G2	Nota do segundo período (numérico: de 0 a 20)
16	schoolsup	Suporte educacional extra (binário: sim ou não)	33	G3	Nota final (numérico: de 0 a 20, meta de saída)
17	famsup	Suporte educacional familiar (binário: sim ou não)			

Fonte: UCI – *Machine Learning*

e-ISSN 2237-9185

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 4 tarefas principais foram realizadas: (1) Formatação dos dados no formato de Data Frame – como a base de dados possui atributos numéricos e categóricos a única opção de formatação na linguagem R é o Data Frame; (2) Divisão da base de dados em treinamento e teste – em que 80% dos dados das bases foram definidos para treino, e 20% para teste, de forma aleatória; (3) Padronização dos atributos numéricos – realizado, por meio do método scale; e (3) Transformação do atributo a ser previsto (atributo meta), a nota final do aluno (G3) para Factor – para a aplicação de algoritmos de AM 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 (G3) que é prevista no processo de predição de desempenho, continua sendo numérica, mas assume essa configuração.

4.3 Terceira Etapa: Processamento Analítico e Algoritmos

No que tange a processamento analítico, neste estudo foram utilizados algoritmos de AM para regressão. Tanto a classificação como a regressão são tarefas de Aprendizagem de Máquina Supervisionada, nesse tipo de técnica a base de dados possui colunas categorizadas que servem para treinar o modelo, que deve, na próxima etapa, a de teste, identificar as categorias de cada linha, de forma independente, generalizando regras que foram aprendidas na fase de treinamento. 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, ou as notas de alunos em uma disciplina, como neste estudo, em que o atributo meta é um valor numérico (KUBAT, 2017).

Alguns dos algoritmos de AM para regressão mais conhecidos são: Árvores de Decisão, Regressão Linear, *Random Forest* (RF), *Suport Vector Machines* (SVM) e Redes Neurais Artificiais do tipo *Multilayer Perceptron* (MLP). Estes algoritmos foram empregados neste estudo para previsão do desempenho dos alunos, sobretudo por serem muito utilizados em tarefas de MDE, bem como pela sua grande aceitação por pesquisadores da área (SHAHIRI; HUSAIN; RASHID, 2015). Assim foi possível identificar quais desses algoritmos geram modelos mais eficazes. Para realização da aplicação desses algoritmos foram utilizadas bibliotecas específicas do R: Árvores de Decisão – biblioteca "*rpart*"; *Random Forest* – biblioteca "*randomForest*"; *Suport Vector Machine* – biblioteca "*e1071*"; a Regressão Linear é nativa da linguagem R, portanto não é necessária uma biblioteca específica apenas chama-se o método "*lm*"; Redes Neurais – biblioteca "*nnet*".

4.4 Quarta Etapa: Análise dos Resultados

Por fim, no intuito de complementar o processo proposto por Aggarwal (2015), acrescenta-se mais uma etapa ao 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 regressão (ou 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

e-ISSN 2237-9185

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 em torno de 80% das instâncias para treinamento e 20% 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 predições, em outras palavras quantificar o seu desempenho. Uma das métricas mais utilizadas no contexto de avaliação de modelos de Aprendizagem Supervisionada é a acurácia, esta corresponde a uma métrica de avaliação comum para problemas de regressão, ou classificação. É o número de previsões corretas feitas como uma proporção de todas as previsões realizadas sobre a base de testes. Devido a sua ampla utilização e aceitação (JAPKOWICZ; SHAH, 2014) está é a métrica utilizada neste estudo.

5. RESULTADOS

Os resultados desse estudo foram sistematizados de acordo com as três questões de pesquisa, e são apresentados na sequência.

5.1. QP1 – Qual o desempenho de modelos gerados a partir de algoritmos baseados em Regressão na previsão do desempenho dos alunos?

Foi gerado um modelo para cada algoritmo selecionado para esse estudo, para as duas bases de dados: Português e de Matemática, os resultados referentes ao método de avaliação Treinamento/Teste e a métrica da acurácia para a análise do desempenho dos modelos estão disponíveis na Tabela 1.

Tabela 1 – Acurácia dos Algoritmos

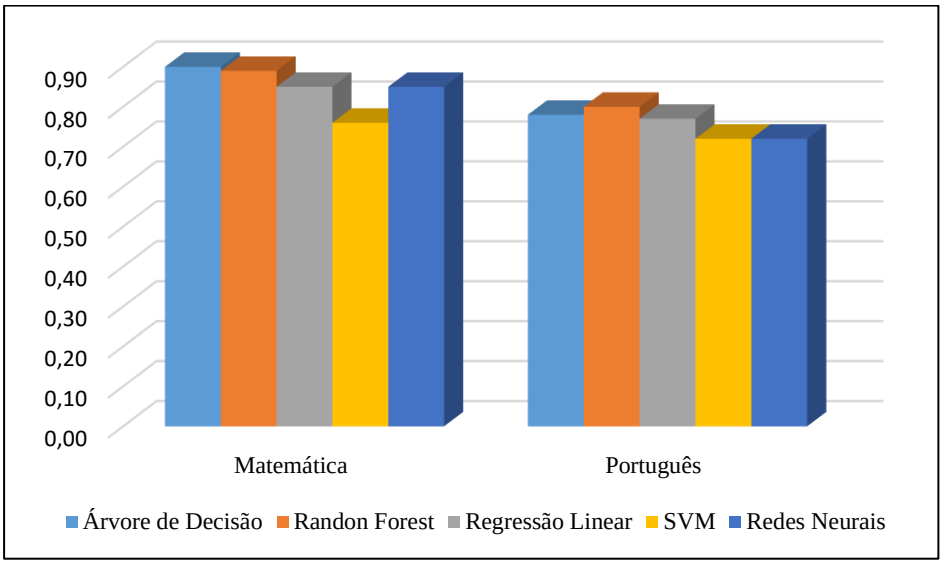
Disciplina	Árvore de Decisão	*Random Forest*	Regressão Linear	SVM	Redes Neurais (MLP)
Matemática	0,90	0,89	0,85	0,76	0,85
Português	0,78	0,8	0,77	0,72	0,72

Fonte: Autores (2021)

Com esses resultados percebe-se que os algoritmos de regressão aplicados sobre as bases, geraram modelos com precisão significativa, em torno 80% na maioria dos casos, exceto para o SVM, sendo boas opções para a previsão das notas finais dos alunos, em bases de dados similares a deste estudo. Para uma melhor visualização das análises, pode-se observar de forma comparativa os resultados da acurácia na Figura 4.

e-ISSN 2237-9185

Figura 4 – Gráfico da Acurácia dos Algoritmos

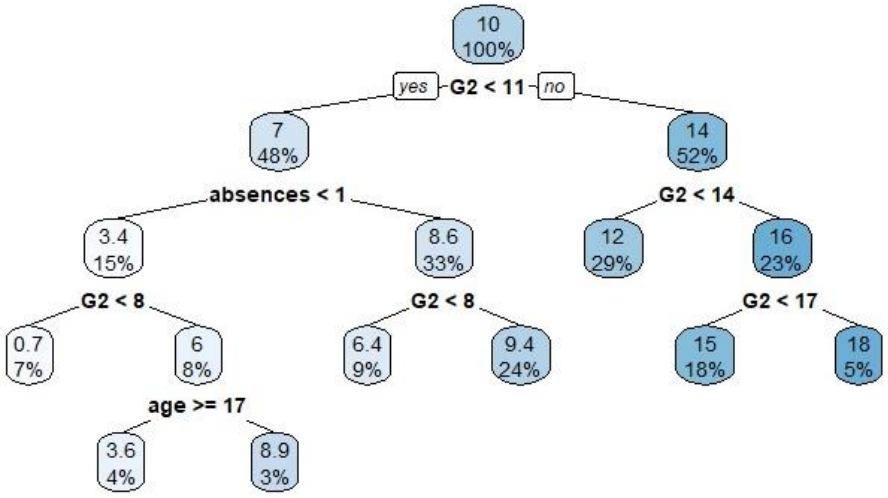


Fonte: Autores (2021)

5.2. QP2 – Qual dos algoritmos de Regressão gera o modelo com melhor desempenho?

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 referente a disciplina de Matemática, com 90% de acurácia, enquanto que para os registros relacionados a disciplina de Português o melhor desempenho foi apresentado pelo algoritmo *Random Forest* (também baseado em árvores de decisão), com 80% de precisão. Embora os algoritmos de Regressão Linear e Redes Neurais também tenham apresentado um bom desempenho, para as duas bases. Mas, um fato que chamou a atenção foi o modelo de Árvore de Decisão ter tido um desempenho tão satisfatório, visto que ele é um algoritmo simples, tanto na sua implementação como na sua interpretação, o que pode ser confirmado ao se analisar as Figuras 5 e 6. Nelas estão representadas os gráficos de Árvore de Decisão (gerados com a biblioteca *rpart.plot*) das duas bases de dados, em que podem ser percebidos quais os principais atributos considerados para realização da predição do desempenho pelo algoritmo de Árvore de Decisão.

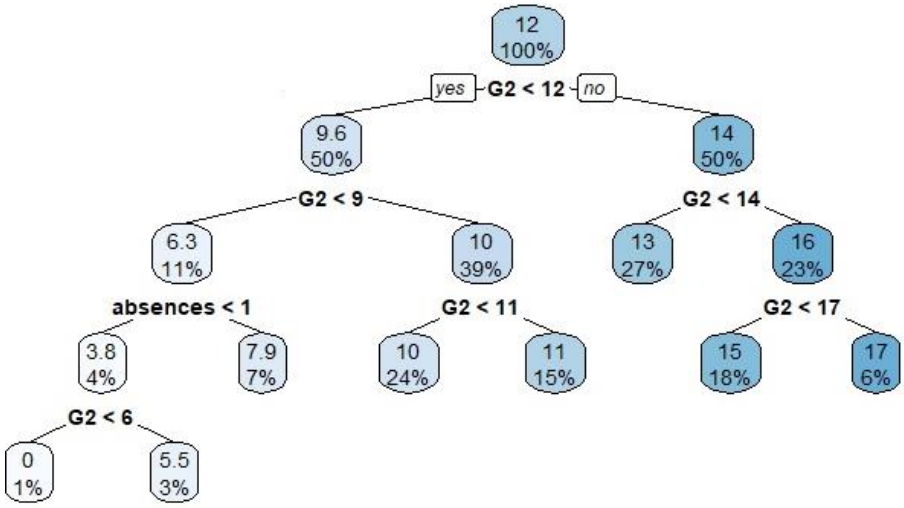
Figura 5 – Gráfico de Árvores de Decisão a disciplina de Matemática



Fonte: Autores (2021)

e-ISSN 2237-9185

Figura 6 – Gráfico de Árvores de Decisão da disciplina de Português



Fonte: Autores (2021)

O algoritmo de Árvore de decisão requer poucos recursos computacionais e tem tempo de processamento bem pequeno, devido a estes fatores este algoritmo seria a melhor escolha para realizar a predição nestes conjuntos de registros. Ainda que para a base de dados de Português, o algoritmo *Random Forest* tenha tido o melhor desempenho, devido a sua necessidade de mais recursos computacionais e também uma maior complexidade de implementação, acredita-se que o algoritmo de Árvore de Decisão seja uma boa opção, pois a diferença do valor da acurácia entre os modelos gerados por estes algoritmos foi pequena.

Outro fator a ser analisado foi o baixo desempenho das Redes Neurais e do SVM, considerados algoritmos bastante eficientes, acredita-se que devido à baixa quantidade de registros e atributos não foi possível extrair toda a eficiência destes algoritmos, impossibilitando a generalização das regras aprendidas de forma satisfatória, o que gerou um *overfitting*[4], visto que os algoritmos tiveram ótimos resultados no treinamento – para a base de dados de Português o SVM atingiu 91% de acurácia, enquanto as Redes Neurais 95%; e na base referente a Matemática o SVM teve 90% de precisão e as Redes Neuras alcançaram 98% – porém nos conjunto de registros de teste isso não aconteceu.

5.3. QP3 – Qual conjunto de atributos tem maior influência sobre as notas finais dos alunos?

Para a identificação dos atributos que mais influenciam na predição do desempenho dos alunos foram calculadas as correlações entre todos os atributos das bases de dados, para as duas disciplinas, e geradas as matrizes de correlação[5]. A função do coeficiente de correlação é determinar qual é a intensidade da relação que existe entre conjuntos de dados ou informações conhecidas. Para interpretar o coeficiente é preciso saber que 1 significa que a correlação entre as variáveis é perfeita positiva e -1 significa que é perfeita negativa. Se o coeficiente for igual a 0 significa que as variáveis

[4] Um cenário de *overfitting* ocorre quando, nos dados de treino, o modelo tem um desempenho excelente, porém quando utilizamos os dados de teste o resultado é ruim. Neste caso, o modelo aprendeu tão bem as relações existentes no treino, que acabou apenas decorando o que deveria ser feito, e ao receber as informações das variáveis nos dados de teste, o modelo tenta aplicar as mesmas regras decoradas, porém com dados diferentes esta regra não tem validade, e o desempenho é afetado.

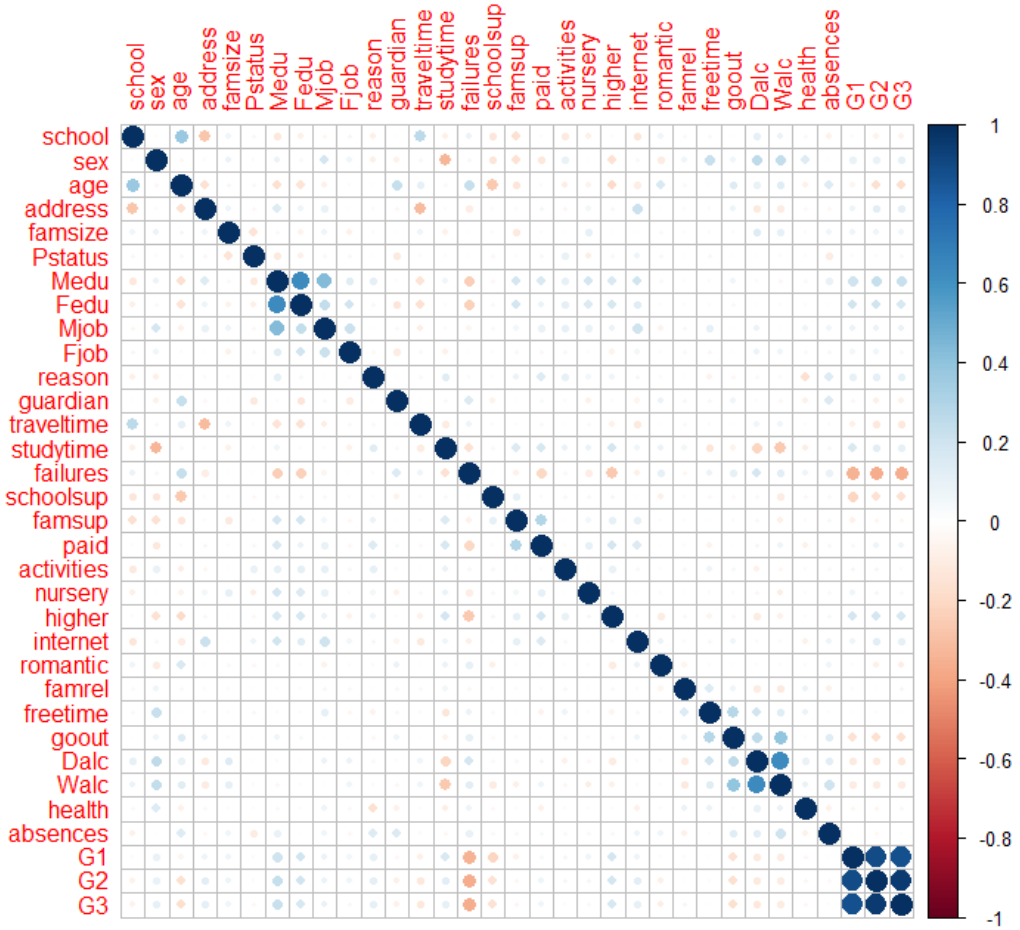
[5] Para acessar as matrizes de correlação completas, no formato planilha, utilizar os seguintes links: Matemática: <https://tinyurl.com/24hb3z83> e Português <https://tinyurl.com/3enbx9y2>.

e-ISSN 2237-9185

não dependem uma da outra. Neste estudo foi utilizado o coeficiente de correlação *Spearman*, que tem a função de medir a intensidade da relação entre duas variáveis, sendo elas lineares ou não.

Para a base de dados referente a disciplina de Matemática foram identificadas correlações bem altas, o que já era esperado, entre as notas anteriores G1 e G2 com o desempenho final (G3), em torno de 88% e 96% respectivamente. Todavia, outros atributos também influenciaram de forma considerável e devem analisados com cuidado no que se refere a performance dos alunos, como: a quantidade de reprovações anteriores, que possui uma correlação de -36% (correlação negativa, com forte influência) com as notas finais; a escolaridade da mãe com 22%; a intenção de realizar curso superior com 17%; o nível de escolaridade do pai também com 17%; e se estão saindo com os amigos -17% de correlação com desempenho final do aluno. Para melhor apresentar as correlações obtidas entre os atributos da base de dados, na Figura 7 pode-se visualizar o gráfico de correlação ou correlograma (gerado com a biblioteca *corrplot*), referente base de dados de Matemática. Neste gráfico, quanto maior o tamanho do círculo, e mais intensa a cor, maior o valor da correlação. As correlações perfeitas, valor 1, são indicadas pela cor azul escuro, enquanto que -1 é indicado pela cor vermelho escuro.

Figura 7 – Gráfico de Correlação da base de dados de Matemática



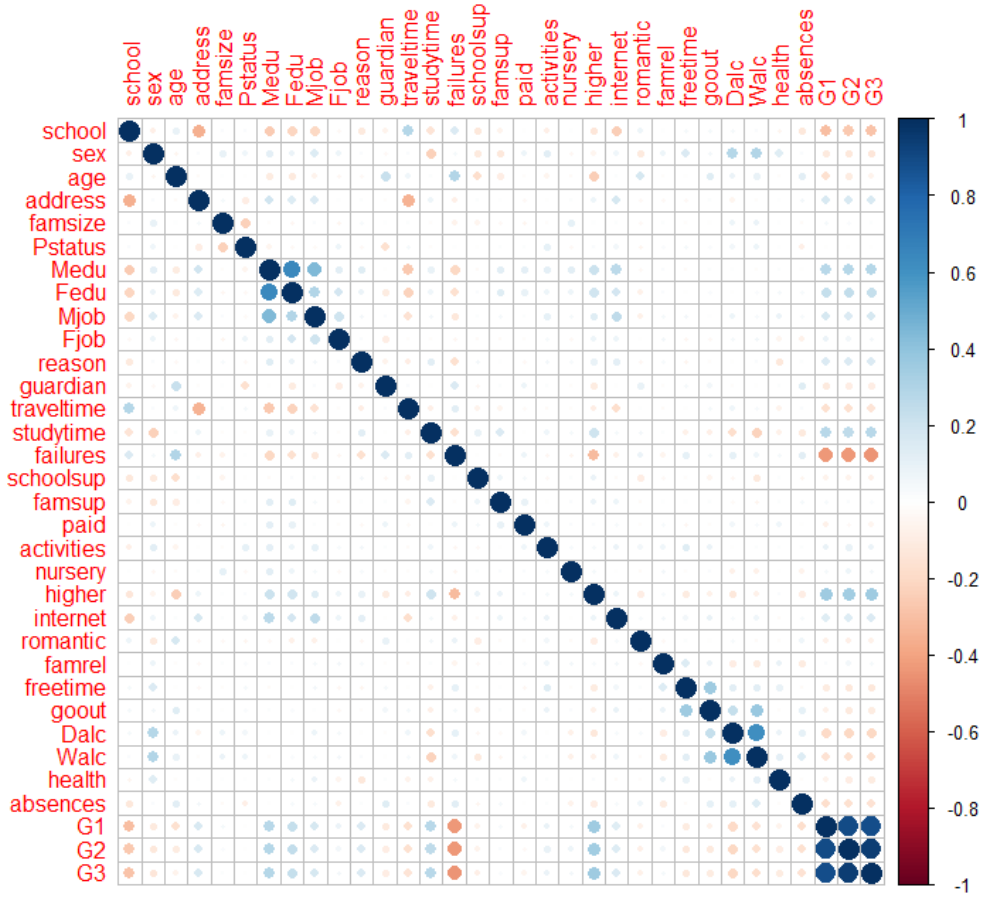
Fonte: Autores (2021)

Com relação da base de dados de Português também foram identificadas correlações bem altas, entre as notas anteriores com o desempenho final, em torno de 88% para G1 e 94% para G2. Além destes, foram identificados outros atributos que também influenciam de maneira significativa nas notas finais dos alunos, sendo as maiores correlações para: as reprovações anteriores com -45%; a vontade de realizar um curso superior com 35%; a escolaridade da mãe com 28%; o tempo de estudo extraclasse

e-ISSN 2237-9185

com 27%; o nível de escolaridade do pai com 23%; e por fim, um fator que chama bastante a atenção são os valores das correlações dos atributos associados ao consumo de álcool, durante o período de aula (*Dalc*), e nos finais de semana (*Walc*) que tiveram correlações de -20% e -17% respectivamente, para com as notas finais, o que suscita um sinal de alerta sobre o consumo de drogas, mesmo que licitas, por jovens estudantes. Para melhor ilustrar as correlações obtidas, também apresenta-se o gráfico de correlação para a base de dados da disciplina de Português (Figura 8).

Figura 8 – Gráfico de Correlação da base de dados de Português



Fonte: Autores (2021)

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 para regressão, processo que constatou que o modelo gerado pelo algoritmo Árvore de Decisão foi o mais preciso para o conjunto de dados da disciplina de Matemática alcançando uma acurácia de 90%, e o modelo gerado pelo algoritmo *Random Forest* foi o mais eficiente para os registros referentes a Português, com uma acurácia de 80%. Com o desenvolvimento deste processo de MDE foi possível verificar que uma base de dados composta por: atributos pessoais, atributos comportamentais, antecedentes acadêmicos e antecedentes econômicos é suficiente para realizar a geração de modelos eficazes na previsão do desempenho dos alunos. No geral os algoritmos utilizados originaram modelos com bons desempenhos. Na base de dados de Matemática a precisão dos modelos foram: *Random Forest* – 89%; Regressão Linear e Redes Neurais – 85%; e SVM – 76%. Para a base de dados da disciplina de Português os modelos atingiram acurácias nos seguintes valores: Árvores de Decisão – 78%; Regressão Linear – 77%; SVM – 72%; e Redes Neurais – 72%.

e-ISSN 2237-9185

Da mesma forma que na pesquisa desenvolvida por Malini e Kalpana (2021), que utilizaram a mesma base de dados pública deste estudo, os modelos gerados pelas Redes Neurais do tipo MLP não obtiveram, no geral, um bom desempenho, os autores relataram 72% de precisão, enquanto que no presente estudo, em média para as duas bases de dados, obtiveram 78,5%. Considerando apenas a base de Português, os mesmos 72% de precisão relatado pelos autores, confirmando que mesmo com ferramentas diferentes, pré-processamento de dados distintos, bem como a aplicação do processo de MDE, os resultados obtidos são semelhantes, possibilitando a percepção que o processo conduzido neste estudo foi adequado. Ademais, os autores obtiveram a melhor taxa de precisão para o modelo constituído a partir do método de Bagging, 88% de precisão, taxa muito próxima a alcançada pelos modelos gerados com os algoritmos de melhores desempenhos deste estudo, Árvores de Decisão com 90% para a base de Matemática, e *Random Forest* com 80% para a base de Português.

Além da geração de modelos para a previsão do desempenho, também foram analisados quais os principais atributos influenciam nas notas finais dos alunos. Como no estudo realizado por Dabhade *et al.* (2021), foi percebido que o desempenho passado recente é o fator mais importante para a previsão do desempenho futuro, visto que os atributos com maior correlação com as notas finais (G3) do terceiro período, foram as notas intermediárias do segundo período (G2), com 96% de correlação para a disciplina de Matemática e 94% para Português. Destaca-se também as notas do primeiro período (G1) que obtiveram também correlações bem altas, 88% para Matemática e Português. Essa relação também foi descrita por Riestra-González, Paule-Ruíz e Ortin (2021), estes salientaram em suas conclusões que a participação dos alunos nos cursos, ou seja, os registros relacionados com atividades realizadas mostram uma forte influência no desempenho dos alunos nas fases iniciais do curso, e no decorrer do curso as tarefas de avaliação tornam-se mais influentes. Pelo lado negativo, também há grande dependência do desempenho passado dos alunos, ficando o atributo reprovações anteriores como o terceiro mais influente sobre as notas finais, com correlação de -36% para Matemática e -45% para Português. Um último atributo relativo aos antecedentes acadêmicos e importante para a aprendizagem, é o tempo de estudo extraclasse, este apresenta uma correlação 27% com as notas finais dos alunos, para os registros de Português.

Ademais, não somente os fatores relacionados aos antecedentes acadêmicos influenciam no desempenho dos alunos, como salientado por Dabhade *et al.* (2021), existe uma forte relação entre as características de comportamento dos alunos e o desempenho acadêmico, isso foi evidenciado neste estudo quando se observa a intenção (comportamento de motivação dos alunos) em cursar a graduação, visto que este fator tem uma influência de 35% sobre as notas finais dos alunos na disciplina de Português e 17% na de Matemática. Outro fator comportamental que impactou na performance dos alunos foi o consumo de álcool, mais especificamente com relação a base de dados de Português, em que foram identificadas correlações de -20% e -17%, para o consumo durante o período de aula (*Dalc*) e nos finais de semana (*Walc*), respectivamente; ocorrência que implica em atenção sobre o consumo de drogas, mesmo que lícitas. Um último atributo comportamental, que também impacta de forma significativa a performance dos alunos é se os estudantes estão saindo com os amigos (*goout*), identificado nos registros de Matemática, este atributo possui uma correlação de -17% com o desempenho final, configurando uma interferência negativa, assim como o consumo de álcool.

Por fim, um aspecto muito importante a ser considerado, citado também por Malini e Kalpana (2021), são os atributos relacionados ao contexto econômico dos estudantes, os autores evidenciaram que

e-ISSN 2237-9185

estes desempenham um papel importante na vida do aluno e o afeta de várias maneiras, sobretudo no comportamento de aprendizagem. Neste estudo também foi possível perceber como alguns fatores relacionados aos antecedentes econômicos dos alunos têm uma forte influência sobre as notas finais, como o nível de escolaridade da mãe e do pai, que estão entre os atributos econômicos com maior impacto no aprendizado dos alunos, estes apresentaram correlações de 28% e 22% para o nível de escolaridade da mãe; e 23% e 17% para o nível de escolaridade do pai, nas bases de dados de Matemática e Português, respectivamente.

Em suma, com o desenvolvimento deste estudo foi possível perceber que é possível realizar a previsão do desempenho de alunos com uma base de dados formada por atributos pessoais e comportamentais, antecedentes acadêmicos e econômicos, mostrando que os modelos gerados com algoritmos de AM para Regressão são técnicas eficientes para esta tarefa de MDE. Além disso, esta pesquisa apresenta como uma das principais contribuições a apresentação do desenvolvimento do processo de MDE, em um conjunto de dados público, que pode ser replicado por outros pesquisadores, ademais os resultados das avaliações dos algoritmos exibidos podem dar suporte na escolha de métodos mais eficazes para a MDE em que o objetivo seja a previsão do desempenho de alunos. Outra contribuição foi detalhar os principais atributos que impactam no desempenho dos alunos, dessa forma estes podem ser monitorados de forma a auxiliar na diminuição de problemas como a evasão e o baixo desempenho de forma antecipada, apoiando ações preventivas, principalmente quando os alunos apresentam rendimentos aquém do esperado.

7. REFERÊNCIAS

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

ALMASRI, A.; CELEBI, E.; ALKHAWALDEH, R.S. (2019). EMT: ensemble meta-based tree model for predicting student performance. **Scientific Programming**, v. 19, p. 1-14. Disponível em: <https://doi.org/10.1155/2019/3610248>.

BAKER, R.; ISOTANI, S.; CARVALHO, A. (2011). Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 2, p. 3–13. Disponível em: <https://doi.org/10.5753/rbie.2011.19.02.03>.

BAKER, R. S.; INVENTADO, P. S. (2014). Educational Data Mining and Learning Analytics. In: J.A. Larusson and B. White (EDS.) (org.). **Learning Analytics: From Research to Practice.** 1. ed. New York, USA: Springer, 1–195. E-book. Disponível em: <https://doi.org/10.1007/978-1-4614-3305-7>.

BAKER, R. S. J. D. (2015). **Big data and education.** 2. ed. New York, USA: A Massive Online Open Textbook (MOOT) - Teachers College, Columbia University. Disponível em: <http://www.columbia.edu/~rsb2162/bigdataeducation.html>.

CORTEZ, P.; SILVA, A. (2008). Using Data Mining to Predict Secondary School Student Performance. **In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008).** Disponível em: <http://www3.dsi.uminho.pt/pcortez/student.pdf>.

DABHADE, P.; AGARWAL, R.; ALAMEEN, K. P.; FATHIMA, A. T.; SRIDHARAN, R.; GOPAKUMAR, G. (2021). Educational data mining for predicting students' academic performance using machine

learning algorithms. **Materials Today: Proceedings**. Disponível em: <https://doi.org/10.1016/j.matpr.2021.05.646>.

JAPKOWICZ, N.; SHAH, M. (2014). Evaluating Learning Algorithms: A Classification Perspective. 1a Ed. ed. Cambridge, E-book. Disponível em: <https://dl.acm.org/doi/book/10.5555/1964882>.

KUBAT, M. (2017). **An Introduction to Aprendizagem de Máquina.** 2. ed*. Coral Gables, FL, USA: Springer.* E-book. Disponível em: <https://doi.org/10.1007/978-3-319-63913-0>.

MALINI, J.; KALPANA, Y. (2021). Investigation of factors affecting student performance evaluation using education materials data mining technique. **Materials Today: Proceedings**. Disponível em: <https://doi.org/10.1016/j.matpr.2021.05.026>.

MDE. **Sociedade Internacional de Mineração de Dados Educacionais**. (2020). Disponível em: <http://educationaldatamining.org/>. Acesso em: 15 set. 2021.

RIESTRA-GONZÁLEZ, M.; PAULE-RUÍZ, M. DEL P.; ORTIN, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. **Computers & Education**, v. 163, p. 1-20. Disponível em: <https://doi.org/10.1016/j.compedu.2020.104108>.

ROMERO, C.; VENTURA, S. (2013). Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,** v. 3, n. 1, p. 12–27. Disponível em: <https://doi.org/10.1002/widm.1075>.

ROMERO, C.; VENTURA, S. (2020). Educational data mining and learning analytics: An updated survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,** v. 10, n. 3, p. 1–21. Disponível em: <https://doi.org/10.1002/widm.1355>.

SHAHIRI, A. M.; HUSAIN, W.; RASHID, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. **Procedia Computer Science,** v. 72, p. 414–422. Disponível em: <https://doi.org/10.1016/j.procs.2015.12.157>.

SINGH, R.; PAL, S. (2020). Machine learning algorithms and ensemble technique to improve prediction of students performance. **International Journal of Advanced Trends in Computer Science and Engineering**, v. 9, n. 3, p. 3970-3976. Disponível em: <https://doi.org/10.30534/ijatcse/2020/221932020>.

SOUZA, V. F.; PERRY, G. T. (2020). Tendências de Pesquisas em Mineração de Dados Educacionais em MOOCs: um Mapeamento Sistemático. **Revista Brasileira de Informática na Educação**, v. 28, p. 491-508. Disponível em: <http://dx.doi.org/10.5753/rbie.2020.28.0.491>.

TALAL, H.; SAEED, S. (2019) A study on adoption of data mining techniques to analyze academic performance. **ICIC Express Letters, Part B: Applications,** v. 10, n. 8, p. 681-687. Disponível em: <http://doi.org/10.24507/icicelb.10.08.681>.

YAACOB, W. F. W.; NASIR, S. A. M; YAACOB, W. F. W.; SOBRI, N. M. (2019). Supervised data mining approach for predicting student performance. **Indonesian Journal of Electrical Engineering and Computer Science**, v. 16, n. 3, p. 1584- 1592. Disponível em: <http://doi.org/10.11591/ijeecs.v16.i3.pp1584-1592>.

Submissão: 31/12/2021

Aceito: 18/02/2022