UNIVERSIDADE FEDERAL DO TRIÂNGULO MINEIRO - UFTM

MESTRADO PROFISSIONAL EM MATEMÁTICA EM REDE NACIONAL - PROFMAT

Dissertação de Mestrado

Mineração de Dados como Suporte Educacional

Larissa de Pádua Miranda

Uberaba - Minas Gerais

Dezembro de 2018

Mineração de Dados como Suporte Educacional

Larissa de Pádua Miranda

Dissertação de Mestrado apresentada à Comissão Acadêmica Institucional do PROFMAT-UFTM como requisito parcial para obtenção do título de Mestre em Matemática.

Orientador: Prof. Dr. Leandro Cruvinel Lemes

Uberaba - Minas Gerais

Dezembro de 2018

Catalogação na fonte: Biblioteca da Universidade Federal do Triângulo Mineiro

M644m

Miranda, Larissa de Pádua
Mineração de dados como suporte educacional / Larissa de Pádua Miranda. -- 2018.
52 f. : il., fig., graf., tab.

Dissertação (Mestrado Profissional em Matemática em Rede Nacional) -- Universidade Federal do Triângulo Mineiro, Uberaba, MG, 2018
Orientador: Prof. Dr. Leandro Cruvinel Lemes

1. Matemática - Estudo e ensino. 2. Mineração de dados (Computação). 3. Python (Linguagem de programação de computador). 4. Pesquisa educacional. I. Lemes, Leandro Cruvinel. II. Universidade Federal do Triângulo Mineiro. III. Título.

CDU 51(07)

LARISSA DE PÁDUA MIRANDA

Mineração de Dados como Suporte Educacional

Dissertação apresentada ao curso de Mestrado Profissional em Matemática em Rede Nacional-PROFMAT, da Universidade Federal do Triângulo Mineiro, como parte das atividades para obtenção do título de Mestre em Matemática.

17 de dezembro de 2018

Banca Examinadora

Prof. Dr. Leandro Cruvinel Lemes
Orientador
Universidade Federal do Triângulo Mineiro

Profa. Dra. Vanessa de Paula Cintra
Universidade Federal do Triângulo Mineiro

Prof. Dr. Jair Rocha do Prado
Universidade Federal de Uberlândia

A Deus, pois a Sua graça me alcançou.
Ao meu marido, pelo incentivo e compreensão.

Agradecimentos

Agradeço a Deus, pela sua imensa misericórdia em todas as etapas dessa jornada, por me sustentar e renovar a minha fé para crer que tudo seria possível.

À Universidade Federal do Triângulo Mineiro (UFTM), pela oportunidade de aprofundar os meus estudos de maneira efetiva e com qualidade; ao coordenador do curso de Mestrado Profissional em Matemática em Rede Nacional (PROFMAT), da UFTM, Prof. Dr. Rafael Rodrigo Ottoboni, e à secretária do curso, Fernanda, pela paciência e atenção desde o primeiro contato.

A CAPES, pelo apoio financeiro.

Ao meu esposo Fernando, pela compreensão dos momentos em que estive ausente para me dedicar às atividades do mestrado, por ter me apoiado a ingressar e acreditado em mim quando eu já não conseguia.

Aos professores, por dedicarem parte do seu dia e compartilharem seus conhecimentos, em especial ao Prof. Dr. Leandro Cruvinel, orientador deste trabalho, que, com suas propostas, me desafiaram e me fizeram evoluir. À minha mãe, por toda dedicação, sabedoria e compreensão que precisei.

Às minhas amigas Natalia e Larissa, por ouvirem meus desabafos, por me orientarem e me animarem nos momentos de angústia com sábios conselhos; e ao Edson e ao Henderson que, nos momentos de descontração, renovaram a minha energia.

Aos meus amigos "5K" que, mesmo nos poucos encontros e por diversas vezes apenas virtuais, me incentivaram nesse processo.

Ao João Lucas que, apesar de não conhecer pessoalmente, se tornou um amigo que, com sua paciência, me esclareceu dúvidas e me auxiliou no que diz respeito à programação.

Aos meus colegas de classe, em especial a Ananda, Silvana, André e Alex, por todos os almoços de sexta-feira que tornavam a minha obrigação mais leve e prazerosa. Muito obrigada por dividirem suas vidas, conhecimentos, anseios e vontades que, unidas às minhas, fizeram com que a turma PROFMAT 2016 tivesse mais representantes até o fim.

Às equipes gestoras da Escola Municipal do Bairro Shopping Park e à Escola Estadual Felisberto Alves Carrejo, pela compreensão e adaptação das minhas aulas, e a todos os meus amigos e colegas de trabalho que, mesmo nos corredores dessas instituições, dedicaram um minuto do seu tempo para me incentivar. Em especial a Franciele, Maria Helena, Lilian, Daniela, Sônia, Danielle, Mareni, Bruno e Daliane – bastava um olhar para virem me socorrer.

"Assim, ao Rei eterno, imortal, invisível, Deus único, honra e glória pelos séculos dos séculos. Amém!"

1 Timóteo 1:17

Resumo

Ao considerar as demandas educacionais, o processo de informatização do ambiente escolar e o baixo desempenho dos alunos na disciplina de matemática, propõe-se pesquisar, conhecer e adaptar materiais que viabilizem a classificação de estudantes a partir de perfis pautados na relação entre fatores sociais, escolares (avaliações bimestrais e somativas) e extraescolares que possam influenciar sobremaneira no rendimento deles ao final do ano letivo. Para isso, utilizou-se a Mineração de Dados Educacionais (do inglês *Educational Data Mining*, EDM) – especificamente os algoritmos *Random Forest*, *Gradient Boosting Classifier* e *KNearest Neighbor Algorithm (KNN)* – e a linguagem de programação *Python* para adaptar modelos por meio de dois conjuntos de dados disponíveis em um site gratuito para comparar a eficiência entre eles. Pretende-se identificar fatores sociais e extraescolares, além de construir perfis de alunos de determinada comunidade, com vistas a discutir metodologias que podem otimizar o processo de ensino e aprendizagem.

Palavras-chave: Mineração de Dados Educacionais. Práticas escolares. Fatores sociais.

Abstract

When considering the educational demands, the process of computerization of the school environment and the low performance of the students in the mathematics discipline, it is proposed to research, to know and to adapt materials that enable the classification of students from profiles based on the relation between social factors, (bimonthly and summative assessments) and extracurricular activities that may greatly influence their performance at the end of the school year. For this, Educational Data Mining (EDM) - specifically the Random Forest algorithms, Gradient Boosting Classifier and KNearest Neighbor Algorithm (KNN) - and the programming language Python to adapt models through two sets of data available on a free website to compare the efficiency between them. It aims to identify social and extra-school factors, as well as to build profiles of students from a particular community, with a view to discussing methodologies that can optimize the teaching and learning process.

Keywords: Educational Data Mining. School practices. Social factors.

Sumário

Lista de ilustrações

Lista de tabelas

INTRODUÇÃO

Com o intuito de dar sentido aos dados gerados pelas redes sociais, econômicas e nos ambientes escolares, transformando-os em informações que podem gerar conhecimento, a Mineração de Dados (do inglês *Data Mining*)surgiu como uma proposta no final da década de 1980 que ainda é tida como uma das tecnologias mais utilizadas nos últimos anos. De acordo com Witten e Frank (2005), Olson e Dellen (2008) e Bramer (2007), a Mineração de Dados é aplicada satisfatoriamente em diversas áreas, como no departamento de cobranças, para detecção de fraudes; em serviços de telemarketing, para acessar dados dos clientes; nos partidos eleitorais, para traçar o perfil de possíveis eleitores; em departamentos de recursos humanos, para identificar competências em currículos; nas tomadas de decisão, para filtrar informações relevantes e fornecer indicadores de probabilidade; e na educação, foco deste estudo.

Utilizou-se a área da Mineração de Dados Educacionais do inglês *Educational Data Minig, EDM* - qque permite a construção de métodos para investigar aspectos educacionais, com a intenção de conhecer as configurações em que os alunos aprendem, como uma ferramenta para o planejamento e a avaliação das atividades ministradas em sala de aula. Diante disso, visa-se identificar fatores sociais e extraescolares, construindo perfis de estudantes de determinada comunidade, com vistas a discutir as metodologias que podem otimizar o processo de ensino e aprendizagem.

No Capítulo 1, ressalta-se a importância dos fatores sociais e sua influência no processo de ascensão escolar. Na sequência, o Capítulo 2 discorre sobre a Mineração de Dados Educacionais, suas etapas para execução e os tipos de aprendizado empregados. O Capítulo 3 aborda os termos usados na investigação, o objetivo a ser alcançado e os trabalhos relacionados ao assunto. Enquanto isso, o Capítulo 4 consiste na descrição dos processos construídos por meio dos conjuntos de dados Student-mat e xAPI-Edu-Data, as ferramentas utilizadas e as primeiras análises. Por fim, o Capítulo 5 traz as conclusões obtidas nesta pesquisa, um breve relato de mediações pedagógicas para a sala de aula na disciplina de matemática e considerações relevantes.

1 DESEMPENHO ESCOLAR E OS FATORES SOCIAIS

Ao considerar os estudantes como seres sociais, dotados de conhecimento e vivências extra escolares que implicam nas relações professor-aluno e aluno-aluno, na captação e interiorização do conhecimento, acredita-se que a

> *avaliação das dificuldades de aprendizagem deve ser um processo amplo de coleta de dados, que contemple uma análise quantitativa e qualitativa, e que vise verificar o nível de execução das tarefas escolares, desenvolvimento anterior da criança, comportamentos em sala de aula, opinião do professor, ambiente familiar e social, aptidões, métodos de aprendizagem, métodos de avaliação da aprendizagem e outros aspectos referentes ao ambiente e ao aluno. (CAPELLINI et al., 2004, p.82)*

Assim se fizeram importantes a identificação e a seleção de características sociais relevantes no processo de ensino e aprendizagem e, por conseguinte, no desempenho escolar. De acordo com Siqueira e Gurgel-Giannetti (2011, p.80):

> *para uma aprendizagem de "sucesso" são necessárias várias habilidades cognitivas associadas a oportunidades adequadas. Ambientes enriquecidos de experiências sensoriais são fundamentais, sendo que a privação pode levar a prejuízos. Ambientes familiares pouco estimuladores e com pouca interação sociolinguística podem levar a criança ao não desenvolvimento de suas aptidões e habilidades. É bem estabelecido na literatura que condições desfavoráveis socioeconômico-culturais influenciam negativamente no desempenho cognitivo e acadêmico, ocasionam maior índice de mau desempenho e insucesso escolar.*

Com base nas ideias do autor, ressalta-se a importância de se conhecer o perfil dos estudantes com quem os docentes lidam a período letivo, para viabilizar reflexões e consequentes considerações que potencializem o aprendizado escolar.

De fato, todas as estruturas sociais do estudante, não apenas aquelas estabelecidas na comunidade escolar, podem ser determinantes para o seu desempenho. Nesse contexto se sobressaem as ideias de Alves et al. (2013, p.588) ao relatarem os dados de Soares e Collares (2006), que:

> *trabalharam com dados do Sistema de Avaliação da Educação Básica para avaliar a relação entre recursos familiares e desempenho cognitivo de alunos da 8ª série (atual 9º ano) do ensino fundamental. Na discussão dos resultados, destacaram o envolvimento dos pais (indicador que registra o grau de participação dos pais tanto na vida diária quanto no acompanhamento escolar dos filhos, incluindo itens como conversas com os filhos, refeições em comum etc.), atribuindo a ele um "papel de ativação dos recursos culturais" (p.636) que permitiria concretizar, no desempenho dos alunos, as vantagens econômicas e culturais da família.*

Macedo (2004) cita que o rendimento escolar se refere ao resultado da junção de fatores individuais (trajetória e motivação, por exemplo); familiares, que abrangem tanto as questões estruturais diretamente relacionadas aos pais tanto ao nível de escolaridade e à participação da vida escolar; e os fatores escolares proporcionados pelas relações entre os alunos, a escola e os professores.

Já de acordo com Miguel, Rijo e Lima (2012, p.134) o abandono e o baixo rendimento escolar se baseiam em seis importantes pontos, a saber:

> (1)características dos alunos - gênero (em que são os rapazes quem manifesta mais alienação), estatuto socioeconômico e aspirações acadêmicas dos alunos; (2) diminuição, durante a adolescência, da motivação dirigida à escola[...]; (3) experiência repetida de fracasso escolar conducente a baixas expectativas de autoeficácia escolar; (4) relações pobres com os professores; (5) contexto de aprendizagem pouco promotor de experiências positivas (e.g. experiências negativas em contexto de sala de aula); e (6) influência das relações interpessoais com pares e os pais - aspirações acadêmicas, valores, normas, atitudes negativas face à escola, experiências escolares negativas, baixos objetivos e valores escolares dos pais e dos pares influenciam de forma importante as expectativas dos alunos face à escola.

Diante disso, Macedo (2004, p.3) pondera que este estudo pretende "investigar como se dá em um determinado grupo amostral, a relação existente entre os fatores familiares e escolares e o rendimento escolar destes indivíduos". Há reflexões sobre suas características e as metodologias que podem ser implementadas, com vistas a um desempenho escolar satisfatório.

2 MINERAÇÃO DE DADOS EDUCACIONAIS

As formas de comunicação e interação em sociedade vêm sofrendo grandes modificações nas últimas décadas com o avanço de tecnologias que, para Gonçalves (2018, p.61):

> *que as tecnologias são criações humanas que vão surgindo e se modificando com o passar do tempo, de acordo com o contexto social, cultural, econômico e profissional em que estão inseridas, para facilitar e aprimorar as ações cotidianas do homem.*

Nesse entremeio, as tecnologias de informação são recursos alternativos para as práticas metodológicas, o que tem modificado as relações escolares na perspectiva de melhorar o ensino e tentar motivar os alunos na busca pelo conhecimento. Conforme Moran (2007, p.10):

> As mudanças que estão acontecendo são de tal magnitude que implicam reinventar a educação, em todos os níveis, de todas as formas. As mudanças são tais que afetam tudo e todos: gestores, professores, alunos, empresas, sociedade, metodologias, tecnologias, espaço e tempo.

Cm consonância com as ideias do autor, identificou-se na Mineração de Dados uma opção de ferramenta tecnológica para colaborar nas mediações pedagógicas, não apenas em sala de aula, como também na preparação de atividades. A Mineração de Dados Educacionais (*Educacional Data Mining*, EDM), "[...] vem, sobretudo, reforçar o papel do professor na preparação, condução e avaliação do processo de ensino e aprendizagem" (BRASIL, 1998, p.45). Ademais, ela viabiliza a identificação dos fatores sociais, econômicos e afetivos que podem causar maior impacto no rendimento escolar de estudantes de uma mesma comunidade.

> Favoráveis ou não, é chegado o momento em que nós, profissionais da educação, que temos o conhecimento e a informação como nossas matérias-primas, enfrentamos os desafios oriundos das novas tecnologias. Esses enfrentamentos não significam a adesão incondicional ou a oposição radical ao ambiente eletrônico, mas, ao contrário, significam criticamente conhecê-los para saber de suas vantagens e desvantagens, de seus riscos e possibilidades, para transformá-los em ferramentas e parceiros em alguns momentos e dispensá-los em outros instantes. (KENSKI, 1998, p.61)

Diante das ideias da autora, este trabalho se propôs como um possível desencadeador de discussão sobre a viabilidade de umas dessas tecnologias como ferramenta de suporte para a área educacional. O objetivo foi descrever uma experiência do uso da

EDM como suporte para o professor no processo de ensino e aprendizagem, sugerindo análises do contexto escolar e social do educando, além da viabilidade. Dessa maneira, minimiza-se a quantidade de procedimentos necessários para uma intervenção pedagógica individualizada, potencializando o compartilhamento de informações entre docentes e estudantes no ambiente escolar. Para isso, foram implementados algoritmos e analisados os resultados obtidos por eles.

2.1 Mineração de Dados Educacionais

Na busca por ferramentas computacionais que fossem capazes de facilitar e potencializar o processo de ensino e aprendizagem, viu-se na Mineração de Dados uma possibilidade de investigar e encontrar padrões entre o desempenho escolar e os fatores sociais. Nesse contexto, Simoudis (1996, p.26) conceitua a Mineração de Dados como "[...] o processo de extração de informações válidas, compreensíveis e úteis e previamente desconhecidas, a partir de grandes bases de dados e utilizá-las para a tomada de decisões cruciais". Já para Dunham (2003, p. 3), "[...] minerar dados é encontrar informações escondidas em um banco de dados".

A Mineração de Dados faz uso da *Exploratory Data Analysis* (EDA) que, segundo Kaski e Kohonen (1996, p.2, tradução nossa), "[...] tem por objetivo apresentar o conjunto de dados de maneira mais compreensível, ao mesmo tempo preservando o maior número de informações essenciais do conjunto de dados original, tanto quanto possível". Sendo assim, a EDA auxilia a Mineração de Dados na identificação de características relevantes dos atributos relacionados no conjunto de dados.

Do ponto de vista educacional, as ações de mediação para aprendizagem podem ser relevantes tanto para o desenvolvimento do aluno, quanto para do processo educacional. Para isso, o professor deve estar atento ao conjunto de ações do estudante, "sem pular" etapas da construção do conhecimento. Para Severo (2011, p.71), o uso da Mineração de Dados Educacionais, na perspectiva dos professores, está voltado:

> *para a obtenção de maior realimentação em relação a instruções, avaliação da estrutura e conteúdo de cursos, bem como, a efetividade do processo de aprendizagem. Outro foco de interesse desta aplicação é a classificação de estudantes em grupos, baseados nas necessidades de monitoramento e orientação, busca de padrões de aprendizagem regulares e irregulares, busca de erros frequentes, busca de atividades que são mais efetivas, busca de informações para adaptação e customização de cursos.*

A EDM, segundo Baker, Isotani e Carvalho (2011, p. 4), "[...] é definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjunto de dados coletados em ambientes educacionais". Essa ferramenta tem

sido frequentemente utilizada pelos cursos de Educação a Distância (EaD) e em programas com suporte computacional, ao captar informações geradas nos ambientes de interação virtual para melhoria no ensino (WITTEN; FRANK, 2005; OLSON; DELLEN, 2008; BRAMER, 2007). Há registros de seu uso na gestão administrativa de cursos de graduação, para criar ou reduzir vagas ofertadas, além de detectar fatores que influenciam na evasão e conclusão acadêmica.

O uso da Mineração de Dados na educação tem como um dos objetivos criar mecanismos para identificar padrões de agrupamento, seleção e classificação de uma grande quantidade de dados gerados nos ambientes de ensino e aprendizagem. A EDM, assim como as outras áreas de aplicação da Mineração de Dados, pode transformar as informações coletadas em algo relevante e, posteriormente, em conhecimento.

Vale ressaltar que, nesta dissertação, visou-se perceber a EDM como uma ferramenta facilitadora para os professores no processo de acompanhamento pedagógico individualizado, possibilitando detectar os métodos mais eficazes de aprendizagem e demais fatores envolvidos na construção do conhecimento.

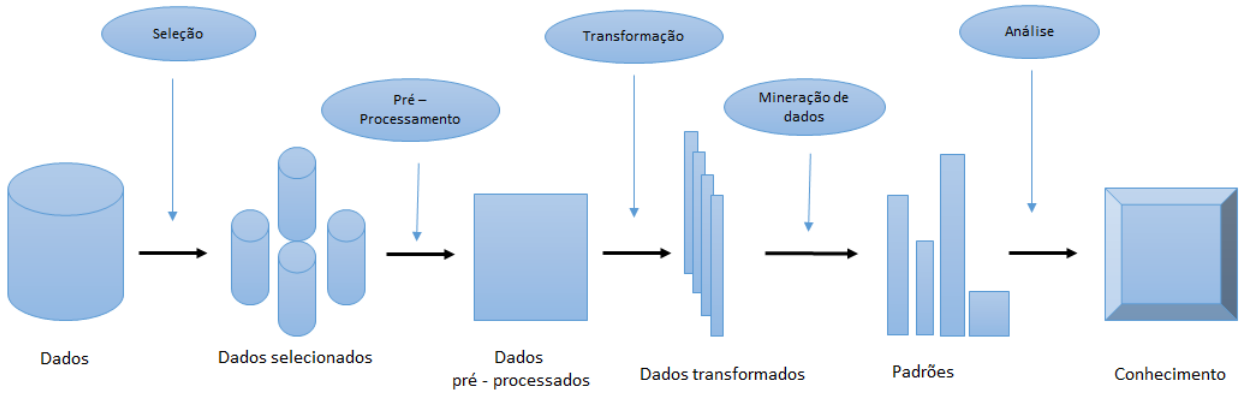
2.2 Conhecendo o KDD

O processo de informatização colabora para a produção cada vez mais acelerada de dados escolares em ambientes virtuais. Nesse contexto surge a extração de dados de conhecimento (do inglês *Knowledge Discovery in Databases*, KDD) que, segundo Fayyad (1996, p.40-41), é "[...] o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados". Sua principal função é, portanto, identificar e selecionar dados que propiciem a formação de saberes.

Ainda de acordo com Fayyad (1996), a execução desse processo é composta por algumas etapas: coleta e seleção de dados, pré-processamento, transformação, Mineração de Dados e análise, que estão representadas na Figura 1. A primeira etapa influencia diretamente o processo com a Mineração de Dados, visto que, a partir da coleta e seleção inicial de informações, se determina as mais relevantes para responder a questionamentos. Já o pré-processamento possibilita organizar os dados selecionados, corrigir a seleção de informações irrelevantes e preencher lacunas de dados incompletos. Por sua vez, na transformação, os dados se tornam numéricos e são estruturados por meio de técnicas de vetorização. Nessa fase, quando necessário, há uma adequação dos dados ao algoritmo de modelagem proposto, geralmente feita segundo processos de normalização. A partir desse estágio, aplica-se a Mineração de Dados que, com o auxílio do método e do algoritmo selecionados, agrupa-os de acordo com os padrões e atributos identificados. A análise

é a parte final, em que há a interpretação e avaliação dos processos executados – aqui é necessário que o responsável avalie o caminho percorrido, verificando cada fase para garantir que não houve falhas e verificar se as informações obtidas são relevantes.

Figura 1 – Etapas do KDD (adaptadas de Fayyad et al., 1996)



Fonte: Adaptado pela autora a partir de Fayyad et al. (1996)

Para maior eficiência no uso da Mineração de Dados, diferentes pesquisadores implementaram e avançaram em relação a modelos categorizados de acordo com os tipos de tarefa (classificação ou regressão) que realizam e os métodos executados. Quando há a necessidade de classificar unidades amostrais em n classes distintas, os algoritmos de classificação são mais utilizados e, ao considerar $n = 2$, diz-se que é uma tarefa de classificação binária. Podem-se citar como exemplos: a classificação de estudantes como "aprovados" ou "reprovados", os diagnósticos de doenças como "positivo" e "negativo", as ações em bolsas de valores como "vai ter alta" ou "vai ter baixa", os tripulantes de um transatlântico como "sobrevivente" ou "não sobrevivente", entre outros. Nesse modelo, após a análise do conjunto de dados fornecidos, é possível indicar classes de acordo com critérios selecionados e, após essa categorização, os novos dados inseridos são direcionados ao conjunto que mais se assemelham.

No ambiente educacional, esse método pode auxiliar na identificação dos casos de reprovação de um grupo de alunos, uma vez que os padrões que os associam se relacionam a características do baixo desempenho escolar.

2.3 Aprendizado supervisionado e não supervisionado

As tarefas em Mineração de Dados podem ser divididas em atividades de aprendizado supervisionado e não supervisionado, conforme o modo como os algoritmos são descritos. Segundo Monard (2003, p.40), "[...] no aprendizado supervisionado é fornecido ao algoritmo de aprendizado, ou indutor, um conjunto de exemplos de treinamento para os quais o

rótulo da classe associada é conhecido"; ainda de acordo com o autor, "[...] no aprendizado não supervisionado, o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos ou clusters".

Nesse ínterim, Matsubara (2004, p.11) cita que "[...] o processo de aprendizado supervisionado se dá pela apresentação de um conjunto de exemplos de treinamento rotulados a um indutor". Já no aprendizado não supervisionado, denominado por ele como "aprendizado por observação e descoberta" (p.12), os algoritmos são utilizados para descobrir um padrão a partir de uma característica.

Logo, no aprendizado supervisionado, a variável de interesse é previamente rotulada, ao passo que, no não supervisionado, pode haver ou não uma variável de interesse. Se isso ocorrer, ela não é previamente rotulada.

3 MÉTODOS E PROCEDIMENTOS

Para identificar características e fatores sociais que podem influenciar no rendimento escolar de uma amostra, visando à classificação dos estudantes no que tange à aprovação ou reprovação no conteúdo de matemática, importaram-se dois conjuntos de dados do repositório online <https://www.kaggle.com/datasets>. Esse site disponibiliza uma diversidade de dados para download gratuito, além de conter códigos de análises realizadas com diferentes algoritmos.

De acordo com a descrição do site, os arquivos selecionados foram obtidos de uma pesquisa com alunos de matemática do ensino médio. Assim, o trabalho foi iniciado com a preparação dos dados para as análises, com a consequente classificação das estruturas das informações como numéricas ou categóricas.

Nesses termos, para visualizar a distribuição dos dados, houve a análise univariada com o auxílio de gráficos e tabelas (ver APÊNDICE A) para as variáveis numéricas. Em relação aos valores únicos maiores ou iguais a cinco, optou-se pelo uso de boxplot, que possibilita visualizar os atributos conforme a disposição inicial, bem como a distribuição dos dados de acordo com as respostas obtidas. A análise bivariada permite verificar simultaneamente duas variáveis e o índice de relacionamento entre elas; logo, foi feita a correspondência dos atributos disponíveis com o target selecionado, apresentando os resultados dos atributos numéricos por meio do boxplot e dos categóricos não ordinais utilizando os gráficos de barras empilhadas (ver APÊNDICES A e B).

Em seguida, identificou-se a coluna que apresenta os dados de interesse, a qual foi modificada por meio de uma discretização binária, que consiste na classificação do atributo utilizando dois rótulos – nesse caso, uma das classes sempre é vista como positiva "[...] numericamente ou lexicograficamente" (PEDREGOSA et al., 2011). As variáveis categóricas foram transformadas para valores numéricos por meio de *dummies*, o que leva à subdivisão de cada coluna de acordo com as respostas apresentadas, utilizando as variáveis booleanas para associar cada resposta ao valor de saída '0' ou '1'. Cumpre dizer que "[...] uma variável Booleana é uma variável que assume valores binários B = falso, verdadeiro, ou 0, 1" (KLOCK; RIBAS; REIS, 2010, p.34).

Para os métodos de classificação, é viável a normalização dos dados, em que os valores são transformados normalmente para um intervalo entre 0,0 e 1,0. Dentre as técnicas conhecidas, a *Min–max normalization* foi utilizada e, de acordo com Jain e Bhandare (2011, p.48), ela:

> *mantém a distribuição original das pontuações, exceto para um fator de escala e transforma todas as pontuações em um ponto comum no intervalo*

> *[0, 1]. Os escores de distância podem ser transformados em pontuações de similaridade subtraindo-se o escore normalizado min-max de 1.*

Ainda de acordo com o autor, os valores d de cada atributo P do conjunto de dados são levados a d' pela seguinte fórmula (JAIN; BHANDARE, 2011, p.47, tradução nossa):

$$d' = \frac{[d - min(p)] * [novo\_max(p) - novo\_min(p)]}{max(p) - min(p)} + novo\_min(p)$$

Em que *min(p)* = valor mínimo atribuído e *max(p)* = valor máximo atribuído.

Nesta dissertação, os dados foram separados em treino e teste. Esse procedimento, segundo Kiralj e Ferreira (2009, p.772, tradução nossa), permite que o conjunto de dados seja "[...] dividido em um conjunto de treinamento (aprendizado) usado para construir o modelo e um conjunto de validação externa (também chamado de conjunto de previsão) que é empregado para testar o poder do modelo". Tal separação é feita para que o algoritmo "aprenda" relações existentes no conjunto de treino e, assim, apresente o critério de confiabilidade para o conjunto de teste após a implementação dos algoritmos.

As métricas de precisão se dividem em três modelos:

- *precision* - "[...]estima o valor preditivo de um rótulo, seja positivo ou negativo, dependendo da classe para a qual é calculado; em outras palavras, avalia o poder preditivo do algoritmo"(SOKOLOVA; JAPKOWICZ; SZPAKOWICZ, 2006, p.1018, tradução nossa);

- *recall* - de acordo com Basu et al. (1998) se assemelha à métrica anterior; contudo, suas estimativas são feitas para um valor superior à previsão correta;

- *f1-score* - "[...]é uma medida composta que beneficia os algoritmos com maior sensibilidade e contrariedade com uma região específica"(SOKOLOVA; JAPKOWICZ; SZPAKOWICZ, 2006, p.1018, tradução nossa).

Nesse sentido, tais técnicas são utilizadas para medir o desempenho do conjunto de teste e estão associadas à matriz de confusão que, de acordo com Carrijo (2004), oferece o quociente entre o número de classificações corretas e a quantidade de classificações que foram preditas para a classe a partir do conjunto obtido.

A implementação dos algoritmos deste trabalho foi feita com o suporte do *Jupyter Notebook* e com a linguagem de programação *Python* que, segundo Lutz e Ascher (2007) é orientada a objetos e pode ser utilizada em uma variedade de domínios tanto para programas independentes como em aplicações. Apesar de ela ser, de acordo com os termos computacionais, uma linguagem de programação de "alto nível", é dinâmica e necessita de poucas linhas de comando para ser executada. Nesse sentido, sua escolha se deve ao

fato de ser acessível ao maior número de profissionais da educação, em diferentes níveis de instrução tecnológica.

Para a modelagem dos dados serão utilizados três algoritmos. O primeiro deles é o Random Forest que, em sua tradução literal, representa uma "floresta de árvores" de decisão e possibilita a classificação a partir do conjunto de dados selecionado. Breiman (1998 apud PEDREGOSA, 2011, p.256, tradução nossa) discorre que:

> *cada árvore do conjunto é criada a partir de uma amostra desenhada com substituição (ou seja, uma amostra de inicialização) do conjunto de treinamento. Além disso, ao dividir um nó durante a construção da árvore, a divisão escolhida não é mais a melhor divisão entre todos os recursos.Ao invés disso, a divisão escolhida é a melhor divisão entre um subconjunto aleatório dos recursos. Como resultado dessa aleatoriedade, o viés da floresta em geral aumenta ligeiramente (se comparado ao viés de uma única árvore não aleatória), mas, devido a média sua variância também diminui, compensando o aumento, resultando asssim um modelo global melhor.*

As árvores de decisão são essenciais para as técnicas de mineração, uma vez que é possível estabelecer e fixar regras a serem reaplicadas em outros conjuntos de dados que considere os mesmos atributos e variáveis, além de expressar os resultados obtidos em uma linguagem popular e de fácil compreensão. O *Random Forest* permite visualizar a porcentagem de influência de cada atributo do conjunto de dados sobre a classe de interesse, destacando os fatores mais relevantes à classificação pretendida.

Já o segundo algoritmo de aprendizado supervisionado (*KNearest Neighbor*, KNN), utiliza o conjunto de treino para prever a categoria com maior probabilidade, selecionando o número mais próximo de vizinhos. E o *Gradient Boosting Classifier* realiza tarefas semelhantes ao anterior, mas, de acordo com Pedregosa et al. (2011), se diferencia pela capacidade de se ajustar automaticamente a novos modelos para fornecer uma estimativa mais precisa; manipulação de dados heterogênea; viabilização da escolha do número de estimativas; funções para corrigir os dados de saída; e um alto poder de precisão.

3.1 Trabalhos relacionados

Nesta seção se apresentam alguns trabalhos desenvolvidos na área de Mineração de Dados Educacionais.

Em Athani et al. (2017), a classificação dos alunos de uma escola de Portugal foi feita por meio de um questionário próprio da instituição e do desempenho na disciplina de matemática. Para isso, implementaram a *Multiclass Support Vector Machine*, que utiliza uma combinação binária das informações ao separá-las em conjuntos com dados semelhantes e validação cruzada do *Kfold*, obtendo uma precisão de 89% para classificar tais estudantes de acordo com as respostas dadas no questionário.

Enquanto isso, E-Calderon e Aranibar (2015) utilizaram a EDM conforme o algoritmo de classificação da Rede Neural Artificial (do inglês Artificial Neural Network), detectando fatores com maior influência na nota do semestre de alunos do ensino superior. O conjunto de dados considerava 39 atributos, como motivação do estudante, relacionamento com os pais/responsáveis e níveis de estresse. Identificaram, com acurácia de 84,86%, que os fatores mais relevantes são: idade, gênero, nota do exame de admissão nas respostas objetivas e discursivas, práticas para aprendizagem, aspectos mais significativos à vida, calma em relação a situações difíceis e grau de dificuldade em lidar com fatos desagradáveis.

No trabalho de Ferreira (2015), a autora utilizou os dados disponíveis no Censo Escolar da Educação Básica do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) de 2014, em se tratando da cidade de Porto Alegre, no Rio Grande do Sul, para pesquisar quais características, pessoais ou sociais, têm mais impacto na evasão escolar e, consequentemente, na não conclusão do Ensino Fundamental. Para isso, ela empregou a regra de Classificação J48 e o filtro CfsSubsetEval, ambos disponíveis no *software* Weka. Após a sua implementação e algumas seleções de filtros e grupos analisados, destacou-se com acurácia de 96,17% que os alunos do ensino privado, em razão dos recursos de internet, laboratório de ciências e auditório na escola, apresentam maior probabilidade de concluírem o ensino fundamental. Ao filtrarem outros atributos, constatou-se menor precisão, com aproximadamente 91,64%, de fatores relevantes para a conclusão do ensino fundamental – cor/raça branca, acesso às aulas de inglês, espanhol, arte e outras disciplinas não obrigatórias.

E o texto de Gottardo, Kaestner e Noronha (2012) utilizou as técnicas de Mineração de Dados para inferir sobre o desempenho de alunos de um curso de EaD, implementando os algoritmos de classificação *Random Forest* e *Multilayer Perceptron*. As informações foram divididas em três dimensões – perfil geral do uso do Ambiente Virtual de Aprendizagem (AVA), interação estudante-estudante e interação professor-estudante –, para as implementações determinarem com maior precisão os atributos que mais influenciam no desempenho dos educandos. As análises com *Random Forest* geraram uma precisão de 76,2%, e com o *Multilayer Perceptron*, a acurácia foi de 76,5%, classificando os discentes no nível de desempenho correspondente.

Tabela 1 – Síntese dos trabalhos relacionados

	Algoritmo	Resultado
Athani et al. (2017)	Multiclass Support Vector Machine	89,00%
E-Calderon e Aranibar (2015)	Artificial Neural Network	84,86%
Ferreira (2015)	J48	96,17%
Gottardo, Kaestner e Noronha (2012)	Random Forest	76,20%
	Multilayer Perceptron	76,50%

Fonte: Elaborado pelo autor

4 MATERIAIS E RESULTADOS

Muito se discute sobre o uso das tecnologias na sala de aula como recurso metodológico no processo de ensino aprendizagem de diversos conteúdos programáticos, em especial na matemática. Moran (1995, p.25) postula que:

> *é possível criar usos múltiplos e diferenciados para as tecnologias. Nisso está o seu encantamento, o seu poder de sedução. Os produtores pesquisam o que nos interessa e o criam, adaptam e distribuem para aproximá-lo de nós. A sociedade, aos poucos, parte do uso inicial, previsto, para outras utilizações inovadoras ou inesperadas.*

Nesse sentido, propõe-se a análise de conjuntos de dados disponíveis em um repositório online, com o escopo de traçar o perfil de estudantes de acordo com características sociais e econômicas, além do desempenho em avaliações.

4.1 Análise do conjunto de dados Student-mat

A primeira análise foi feita com o conjunto de dados Student-mat, constituído pelas respostas de estudantes do ensino médio a um questionário social, de gênero e de caráter estudantil, disponível no site <https://www.kaggle.com/uciml/student-alcohol-consumption>.

Em geral, o conjunto de dados utilizados para a implementação da EDM tem, por respostas aos seus atributos, variáveis que podem ser classificadas em dois tipos: numérico – valores discretos ou contínuos em intervalos de um conjunto numérico; categórico – conjunto de expressões, siglas, abreviações ou valores que não podem ser ordenados. Conforme essas definições, as variáveis ora tratadas estão relacionadas na Tabela 2:

Tabela 2 – Categorização dos atributos - Student-mat

Nome	Descrição	Tipo de variável	Possíveis respostas
Escola	Tipo de escola que frequenta	Categórico	'GP' , 'MS'
sexo	Sexo	Categórico	'F', 'M'
Idade	Idade	Numérico	Número inteiro positivo
Endereço	Endereço residencial	Categórico	'U', 'R'
Motivo	Motivo da escolha da escola que frequenta	Categórico	'course', 'other', 'home', 'reputation'
Tamanho da família	Número de pessoas com quem reside	Categórico	'GT3' , 'LE3'

Tabela 2 - Categorização dos atributos - Student-mat

Nome	Descrição	Tipo de variável	Possíveis respostas
Educação-mãe	Nível de escolaridade da mãe	Categórico	Número natural de 0 a 4
Educação-pai	Nível de escolaridade do pai	Categórico	Número natural de 0 4
Emprego-mãe	Ocupação da mãe	Categórico	'at_home', 'health', 'other', 'services', 'teacher'
Emprego-pai	Ocupação do pai	Categórico	'at_home', 'health', 'other', 'services', 'teacher'
Deslocamento	Tempo de deslocamento da residência à escola	Numérico	Número natural de 1 a 4
Responsável	Pessoa responsável pelo estudante	Categórico	'mother', 'father', 'other'
Tempo de estudo	Número de horas de estudo semanal	Numérico	Número natural de 1 a 4
Reprovações	Número de reprovações	Numérico	Número natural de 0 a 3
Aulas extra	Atividades extraturno	Categórico	'yes', 'no'
Suporte familiar	Suporte familiar na vida escolar	Categórico	'yes', 'no'
Aulas particulares	Frequenta aulas particulares	Categórico	'yes', 'no'
Atividades	Atividades extracurriculares	Categórico	'yes', 'no'
Pré-escola	Frequentou aulas em pré-escola	Categórico	'yes', 'no'
Ensino superior	Pretenção em fazer curso superior	Categórico	'yes', 'no'
Internet	Acesso à internet em casa	Categórico	'yes', 'no'
Relacionamento amoroso	Relacionamento amoroso	Categórico	'yes', 'no'
Relacionamento familiar	Qualidade do relacionamento familiar	Categórico	Número natural de 1 a 5
Tempo livre	Número de horas livres	Numérico	Número natural de 1 a 5
Relacionamento dos pais	Relacionamento dos pais (moram juntos ou separados)	Categórico	'A' - 'T'
Encontro com amigos	Qualidade de encontro com os amigos	Categórico	Número natural de 1 a 5
Álcool dias úteis	Consumo de álcool nos dias úteis	Categórico	Número natural de 1 a 5
Álcool fim de semana	Consumo de álcool nos finais de semana	Categórico	Número natural de 1 a 5
Saúde	Nível atual do estado de saúde	Categórico	Número natural de 1 a 5
Faltas	Número de ausências nas aulas	Numérico	Número inteiro postivo
G1	Nota na avaliação 1	Numérico	Número inteiro postivo
G2	Nota na avaliação 2	Numérico	Número inteiro postivo
G3	Nota na avaliação final	Numérico	Número inteiro postivo

Fonte: Elaborado pelo autor

Além das respostas à parte social, a pesquisa considerou como atributos o desempenho em três avaliações no conteúdo de matemática, denominadas de G1, G2 e G3. Optou-se por tomar G3 como target de interesse, pois representa a nota final dos alunos e é utilizada para predizer as relações com os demais itens do conjunto de dados Student-mat. A Tabela 3 apresenta o índice de correlação entre os atributos citados acima, em que se percebeu o quão relacionados eles estão, ao demonstrar a chance de o estudante repetir seu desempenho em todas as avaliações que forem feitas no período analisado.

Tabela 3 – Correlação entre os atributos G1, G2 e G3

	G1	G2	G3
G1	1,00	0,85	0,80
G2	0,85	1,00	0,90
G3	0,80	0,90	1,00

Fonte: Elaborado pelo autor

A análise univariada dos dados numéricos (age, absences, G1 e G2) apresentados na Figura 2 e no Apêndice A permitiu verificar características relevantes sobre o perfil dos estudantes, tais como:

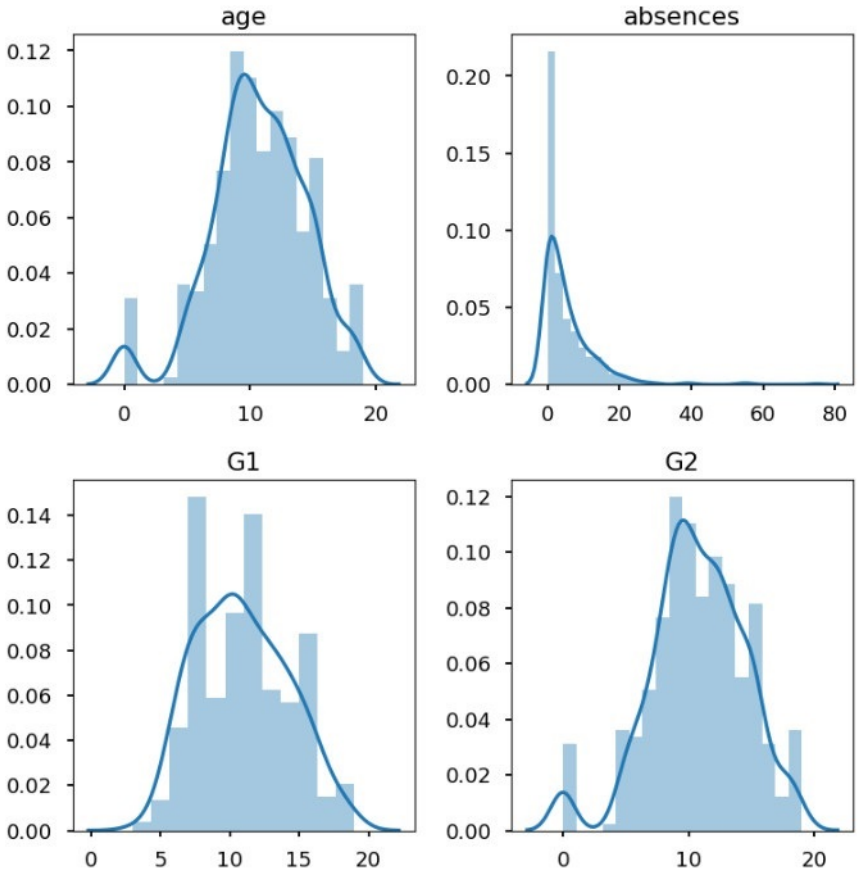
- a média e a moda das idades dos alunos são, respectivamente, 16,69 e 16 anos de idade;
- dos 395 entrevistados apenas 34 não tiveram nenhuma falta durante o semestre letivo e a média de faltas foi de aproximadamente 6;
- a média das notas de G1 e G2 foi de aproximadamente 10. No primeiro, a nota com maior frequência foi 10, e no segundo, 9.

O Apêndice B apresenta a distribuição das respostas referente aos atributos categóricos organizados conforme moda e frequência absoluta e relativa agrupadas – a Tabela B detalha as respostas em cada classe. Nesse caso, identificou-se que:

- cerca de 88,39% frequentam a mesma instituição escolar;
- 77,72% residem na área urbana;
- 281 alunos (aproximadamente 71,14%) têm até três membros na família
- 10,38% são filhos de pais separados.

As respostas sobre desempenho na avaliação G3, que é a variável de interesse, foram separadas em duas classes utilizando a discretização binária. Consideram-se as notas com

Figura 2 – Atributos numéricos



Fonte: Elaborado pela autora

valores maiores ou iguais a 10 como classe positiva, com o rótulo "aprovado", e os valores inferiores a 10 se referem à classe negativa, nomeada como "reprovado".

Os itens do questionário classificados aqui como dados do tipo categórico (ver Tabela 2), descritos na linguagem Python como dtype = object, representavam categorias. Ao serem apresentados com códigos numéricos ou não, foi necessária uma mudança de parâmetros por *dummies*. Após transformação, a técnica *min–max normalization* foi então aplicada ao conjunto de dados.

Nesse sentido, o conjunto foi dividido em conjunto de treino e teste para a implementação dos algoritmos *Random Forest*, *Gradient Boosting Classifier* e *KNN*, com a porcentagem de separação para validação na medida de 30% e 70%, respectivamente, com 20, 500, 1.000 e 2.000 rodagens, tendo como resultado a Tabela 4:

Tabela 4 – Resultados das implementações - Student-mat

	Algoritmo	Resultado
Análise (20 rodagens)	Random Forest	89,24% (±3,21%)
	Gradient Boosting	90,56% (±2,90%)
	KNN	69,58% (±4,06%)

Tabela 4 - Resultados das implementações - Student-mat

	Algoritmo	Resultado
Análise (500 rodagens)	Random Forest	88,49% (±3,11%)
	Gradient Boosting	91,32% (±2,54%)
	KNN	69,35% (±4,15%)
Análise (1000 rodagens)	Random Forest	88,42% (±3,14%)
	Gradient Boosting	91,32% (±2,55%)
	KNN	69,48% (±4,12%)
Análise (2000 rodagens)	Random Forest	88,38% (±3,18%)
	Gradient Boosting	91,26% (±2,57%)
	KNN	69,33% (±4,14%)

Fonte: Elaborado pela autora

Diante dos resultados apresentados, percebeu-se que o *Gradient Boosting Classifier* teve a melhor acurácia de 91,32%, para os casos de 500 e 1.000 rodagens, seguido pelo *Gradient Boosting Classifier*, com 89,24% (20 rodagens) e, por último, há o *KNN*, que obteve 69,58% com 20 rodagens.

As análises seguintes foram obtidas pela implementação do *Random Forest* que, diferentemente dos algoritmos supracitados, permite apresentar o índice de relevância exercido pelos atributos do conjunto sobre o target escolhido (considera-se aqui o atributo G3).

Após a implementação do algoritmo, obteve-se a porcentagem de importância sobre G3, em se tratando dos atributos de maior relevância no conjunto de dados:

Tabela 5 – Percentual de relevância sobre G3 - Student-mat

	Importância		Importância
G2	31,38%	G1	29,39%
Faltas	3,10%	Idade	2,74%
Reprovações	2,51%	Atividades_sim	1,89%
Relacionamento familiar_2	0,12%	Relacionamento familiar_3	0,26%
Relacionamento familiar_4	0,81%	Relacionamento familiar_5	1,66%
Álcool em dias úteis_2	1,37%	Álcool em dias úteis_3	0,78%
Álcool em dias úteis_4	0,00%	Álcool em dias úteis_5	0,21%
Álcool fim de semana_2	0,24%	Álcool fim de semana_3	0,33%
Álcool fim de semana_4	0.,09%	Álcool fim de semana_5	0,36%
Encontro com amigos_2	0,53%	Encontro com amigos_3	0,44%
Encontro com amigos_4	1,32%	Encontro com amigos_5	0,39%
Motivo_outro	1,24%	Motivo_reputação	0,50%
Motivo_residência	0,33%	Responsável_mãe	0,94%
Responsável_outro	0,07%	Educação–mãe_1	0,50%
Educação–mãe_2	0,35%	Educação–mãe_3	0,41%
Educação–mãe_4	0,87%	Educação–pai_1	0,33%
Educação–pai_2	0,31%	Educação–pai_3	0,12%
Educação–pai_4	0,55%	Emprego–mãe_outro	0,82%
Emprego–mãe_saúde	0,32%	Emprego–mãe_professor	0,25%
Emprego–mãe_serviços	0,23%	Emprego–pai_outro	0,41%
Emprego–pai_serviços	0,37%	Emprego–pai_professor	0,18%
Emprego-–pai_saúde	0,17%	Sáude_2	0,41%

Tabela 5 - Percentual de relevância sobre G3 - Student-mat

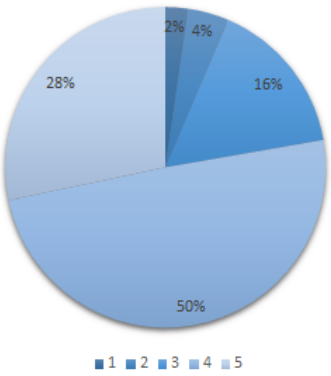
	Importância		Importância
Saúde_3	0,29%	Saúde_4	0,28%
Sáude_5	0,58%	Sexo_M	0,78%
Internet_sim	0,74%	Endereço_Urbano	0,60%
Pré_sim	0,48%	Suporte familiar_sim	0,47%
Relacionamento dos pais_juntos	0,33%	Relacionamento amoroso_sim	0,29%
Tempo livre	0,21%	Ensino Superior_sim	0,10%
Deslocamento	0,87%	Tempo de estudo	1,45%
Tamanho da família_menorouigual3	0,86%	Aulas particulares_sim	0,84%
Escola_MS	0,00%	Suporte educacional extra_sim	0,94%

Fonte: Elaborado pela autora

Mostrou-se relevante a qualidade do relacionamento familiar, em específico o atributo "Relacionamento familiar_5", com 1,66%, acompanhado pelo "Relacionamento dos pais_juntos", que indicou uma relevância de 0,33%.

A representação gráfica apresentada da Figura 3 possibilita uma melhor visualização do desempenho dos alunos e a influência do relacionamento dos pais. Percebeu-se que a maior parte dos entrevistados classificados como aprovados apontaram a qulidade do relacionamento familiar com nível 4 ou 5.

Figura 3 – Percentual de aprovação pela relação familiar



Fonte: Elaborado pela autora

Nesse sentido, Bastiani (1993 apud BHERING; SIRAJ-BLATCHFORD, 1999, p.192):

> *afirma que "o envolvimento de pais com a escola passou a ser considerado nos últimos anos como uma preocupação necessária e legítima e não pode ser mais uma opção extra"que as escolas poderiam ou não ter. Isso não parece ser diferente na nossa realidade brasileira, pois estudos nacionais salientam a importância do contato com os pais em outras situações que não sejam somente em situações extremas e problemáticas.*

Com base nas ideias da autora, o critério "Responsável_mãe", após a implementação do algoritmo, obteve relevância de 0,94%. Ao observar a Tabela 6 e a Figura 4, é possível constatar que, dos 395 estudantes questionados:

- 273 responderam que o responsável legal é a mãe;
- 90 citaram o pai;
- 32 indicaram outros parentes.

Tabela 6 – Resultado final e responsável legal do estudante

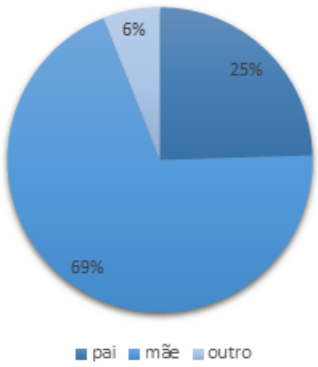
Responsável G3	pai	mãe	outro
0	8	25	5
4	0	1	0
5	1	6	0
6	4	11	0
7	1	7	1
8	5	21	6
9	6	18	4
10	16	37	3
11	8	35	4
12	9	20	2
13	8	19	4
14	8	19	0
15	6	26	11
16	4	11	1
17	2	4	0
18	4	7	1
19	0	5	0
20	0	1	0

Fonte: Elaborado pela autora

Como grande parte dos dados se referiu à opção "mãe", é esperado que apresente a maior concentração em todas as notas obtidas no resultado final; contudo, vale destacar que as notas mais altas se sobressaíram apenas nesse caso. Pode-se perceber que o nível de escolaridade da mãe ("Educação–mãe_4") é um atributo relevante na classificação do estudante, com a influência de 0,87%.

Outro critério que apresentou relevância foi "Faltas", que contabiliza o número de ausências de cada aluno durante o semestre da pesquisa. Obteve-se o resultado de 3,10%, fator que deve ser observado pelo profissional da educação e pode vir a ser determinante para a situação do educando. O gráfico para visualização desse atributo foi o boxplot, pois a diversidade de respostas recebidas nesse critério não possibilitou uma visualização adequada no gráfico de barras.

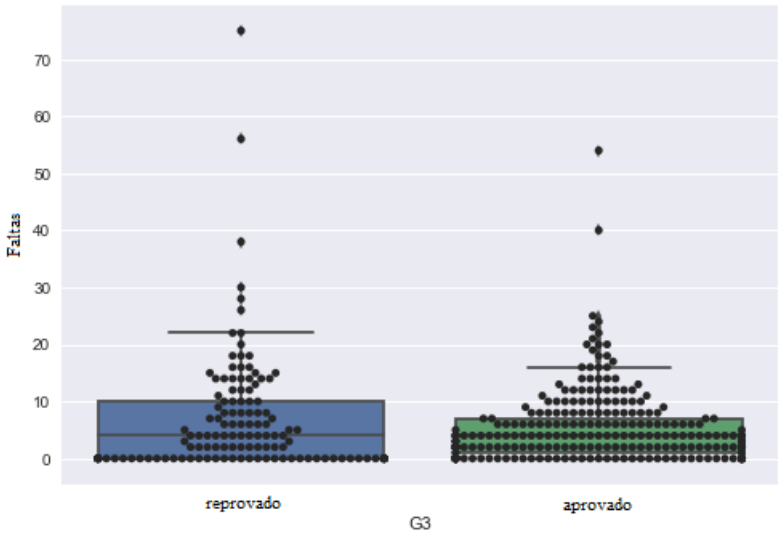
Figura 4 – Percentual de aprovação pelo tipo de responsável



Fonte: Elaborado pela autora

A Figura 5 mostra que os alunos classificados como "aprovados" faltaram a menos de 10 aulas, enquanto os discentes descritos como "reprovados" apresentaram uma quantidade de ausências mais elevada.

Figura 5 – Número de faltas



Fonte: Elaborado pela autora

Identificou-se que o quesito "Atividades" obteve 1,89% com a implementação do algoritmo Random Forest. Conforme a Tabela 7 e a Figura 6, foi possível identificar que, nesse item, houve melhor distribuição das respostas, com 201 estudantes que frequentavam aulas de reforço na disciplina e 194 que não tinham esse suporte.

De maneira análoga ao critério analisado anteriormente, na plotagem do gráfico (Figura 6), a quantidade de alunos classificados como "reprovados" é praticamente idêntica para as respostas do atributo: cerca de 32% dos alunos que frequentavam algum tipo de

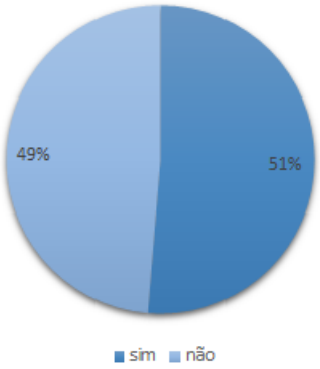
Tabela 7 – Relação dos atributos: Nota final e atividades extraescolares

Atividades	G3
Não	194
Sim	201

Fonte: Elaborado pela autora

reforço escolar foram reprovados, e, dos que não frequentavam atividades extraescolares, 34% tiveram esse desempenho.

Figura 6 – Percentual de aprovações por frequência em atividades extraescolares



Fonte: Elaborado pela autora

Corrobora-se com Miguel, Rijo e Lima (2012, p.33), ao afirmarem que:

> *o cumprimento de rotinas escolares, como o estudo para os testes ou realização de trabalhos de casa, contribui para o sucesso do aluno, na medida em que potencia melhores resultados escolares (Alexander, Entwisle, & Horsey, 1997). Estas tarefas desempenham um papel importante na melhoria dos resultados apresentados pelos alunos e promovem a sua autorregulação (uma vez que o aluno necessita de estar concentrado para executar com sucesso a tarefa) e sentimentos de maestria.*

O item "Reprovações", que concerne ao número de reprovações em semestres que antecederam a pesquisa, se destacou na análise do conjunto, com o resultado de 2,51%. A Tabela 8 mostra a relação dos 395 estudantes que responderam ao questionário, dos quais:

- Dos 312 que não tiveram reprovações, 234 foram classificados como "aprovados" e 78 como "reprovados";
- Dos 50 que reprovaram em apenas uma disciplina, 24 foram "aprovados" e 26 "reprovados";

- Dos 17 reprovados em duas matérias, três foram classificados como “aprovados” e 14 como “reprovado”;
- Dos 16 que tiveram reprovações em três disciplinas, quatro foram “aprovados” e 12 “reprovados”.

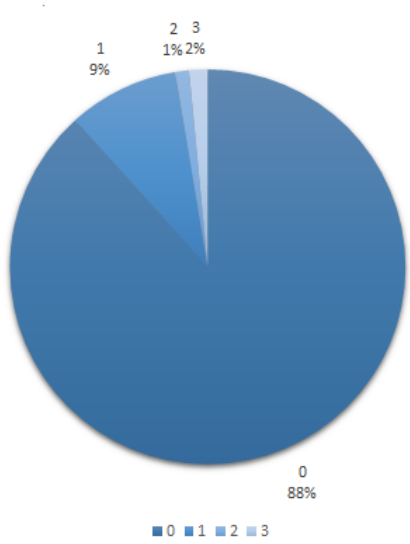
Conforme a Figura 7, ficou claro que, apesar de haver concentração das respostas no caso em que não houve reprovações em situações anteriores – isto é, essa resposta se apresentou como a mais frequente –, percebe-se que ocorreu a maior discrepância entre as classes de classificação.

Tabela 8 – Relação dos atributos: Números de reprovações e nota final

G3 Reprovações	Aprovado	Reprovado
0	234	78
1	24	26
2	3	14
3	4	12

Fonte: Elaborado pela autora

Figura 7 – Percentual de aprovação por número de reprovações anteriores



Fonte: Elaborado pela autora

4.1.1 Comparativo de resultados

Na literatura utilizada como embasamento teórico, foram encontrados diversos textos que sugerem análises semelhantes às realizadas nesta pesquisa. O estudo de caso

feito por Cortez e Silva (2008), por exemplo, utilizou o mesmo conjunto de dados e o algoritmo *Random Forest* para classificar os indivíduos envolvidos na investigação. Na Tabela 9 a seguir, sugere-se um breve comparativo utilizando a análise feita pelo autor e as acurácias obtidas nas implementações.

Para fins de comparação, calculou-se o F1-score obtendo os resultados da Tabela 9. Ressalta-se que este se diferencia do texto citado anteriormente por implementar os algoritmos *Random Forest*, *Gradient Boosting* e *KNN* simultaneamente, além do número de validações solicitadas (alternou-se o número de "rodagens" para 20, 500, 1.000 e 2.000). As modificações foram realizadas com o intuito de disponibilizar ao leitor outros modelos gratuitos de critérios de previsão de aprovação de alunos com diferentes níveis de acertabilidade e que possibilitassem a melhor visualização dos dados.

Tabela 9 – Comparativo - Dados retirados e adaptados de Cortez e Silva (2008)

	Conjunto de dados	Algoritmo	Tarefa	Acurácia	Métrica
$(CORTEZ; SILVA, 2008)$	Student-mat	Random Forest	Classificação	72,40% (±0,40%)	F1-score
Análise (20 rodagens)	Student-mat	Random Forest	Classificação	86,83% (±7,33%)	F1-score
	Student-mat	Gradient Boosting	Classificação	89,34% (±4,17%)	F1-score
	Student-mat	KNN	Classificação	60,97% (±20,29%)	F1-score
Análise (500 rodagens)	Student-mat	Random Forest	Classificação	87,02% (±5,86%)	F1-score
	Student-mat	Gradient Boosting	Classificação	90,13% (±4,53%)	F1-score
	Student-mat	KNN	Classificação	58,55% (±22,17%)	F1-score
Análise (1000 rodagens)	Student-mat	Random Forest	Classificação	87,02% (±5,80%)	F1-score
	Student-mat	Gradient Boosting	Classificação	90,15% (±4,59%)	F1-score
	Student-mat	KNN	Classificação	58,51% (±22,00%)	F1-score
Análise (2000 rodagens)	Student-mat	Random Forest	Classificação	86,82% (±5,96%)	F1-score
	Student-mat	Gradient Boosting	Classificação	90,11% (±4,66%)	F1-score
	Student-mat	KNN	Classificação	58,56% (±22,10%)	F1-score

Fonte: Elaborado pela autora

4.2 Análise do conjunto de dados xAPI-Edu-Data

A próxima análise foi feita com o conjunto de dados **xAPI-Edu-Data**, disponível no site <https://www.kaggle.com/dan195/classification-of-student-marks/data>. Esta pesquisa não se trata apenas de alunos que cursam disciplinas relacionadas à matemática, mas foi escolhida por poder mostrar a influência de critérios análogos para a classificação

dos estudantes. Além disso, esse conjunto de dados constata a nacionalidade dos sujeitos envolvidos, trabalhando com uma diversidade que ajuda a desmistificar padrões ou concepções que caracterizam uma população. Retirado do mesmo repositório online do Student-mat, esse arquivo resulta de uma investigação com 480 indivíduos que responderam a 16 atributos descritos e categorizados na Tabela 10:

Tabela 10 – Classificação dos atributos - xAPI-Edu-Data

Descrição	Tipo de variável	Possíveis respostas
Gênero	Categórico	'M', 'F'
Nacionalidade	Categórico	'KW', 'Lebanon', 'Egypt', 'SaudiArabia', 'USA', 'Jordan', 'Venezuela', 'Iran', 'Tunis', 'Morocco', 'Syria', 'Palestine','Iraq', 'Libya'
Nível de escolaridade	Categórico	'lowerlevel', 'MiddleSchool', 'HighSchool'
Local de nascimento	Categórico	'Kuwait', 'Lebanon', 'Egypt', 'SaudiArabia', 'USA', 'Jordan','Venezuela', 'Iran', 'Tunis', 'Morocco', 'Syria', 'Iraq','Palestine', 'Libya'
Estágio	Categórico	'G-04', 'G-07', 'G-08', 'G-06', 'G-05', 'G-09', 'G-12', 'G-11','G-10', 'G-02'
Classe matriculado	Categórico	'A', 'B', 'C'
Tipo de curso que frequenta	Categórico	'IT', 'Math', 'Arabic', 'Science', 'English', 'Quran', 'Spanish','French', 'History', 'Biology', 'Chemistry', 'Geology'
Semestre letivo	Categórico	'F', 'S'
Responsável pelo estudante	Categórico	'Father', 'Mum'
Número de participações	Numérico	Número inteiro positivo
Número de visitas no material do curso	Numérico	Número inteiro positivo
Verificações as novas notificações do curso	Numérico	Número inteiro positivo
Número de discussões	Numérico	Número inteiro positivo
Participação do responsável	Categórico	'Yes', 'No'
Satisfação do responsável	Categórico	'Good', 'Bad'
Número de ausências	Categórico	'Under-7', 'Above-7'
Classe	Categórico	'M', 'L', 'H'

Fonte: Elaborado pelo autor

De maneira análoga à primeira análise do trabalho, iniciaram-se as etapas do KDD, nas quais se confirmou que não havia dados nulos que prejudicariam os próximos passos da mineração. Os processos para seleção das variáveis – discretização, normalização, separação dos conjuntos de treino e teste e algoritmos – foram os mesmos implementados anteriormente. A Tabela 11 mostra os resultados obtidos:

Tabela 11 – Resultados das implementações - xAPI-Edu-Data

	Algoritmo	Acurácia
Análise (20 rodagens)	Random Forest	67,64% (±3,05%)
	Gradient Boosting	73,40% (±0,32%)
	KNN	62,50% (±0,00%)
Análise (500 rodagens)	Random Forest	66,60% (±2,81%)
	Gradient Boosting	73,46% (±0,29%)
	KNN	62,50% (±0,00%)
Análise (1.000 rodagens)	Random Forest	66,55% (±2,94%)
	Gradient Boosting	73,48% (±0,28%)
	KNN	62,50% (±0,00%)
Análise (2.000 rodagens)	Random Forest	66,62% (±2,87%)
	Gradient Boosting	73,48% (±0,29%)
	KNN	62,50% (±0,00%)

Fonte: Elaborado pelo autor

Conforme a Tabela 11, para a implementação de *KNN*, manteve-se o resultado, independentemente do número de rodagens solicitadas, com 62,50%; com Gradient Boosting, a acurácia foi de 73,48% para os casos de 1.000 e 2.000 rodagens; e com Random Forest, a melhor acurácia foi obtida para o caso de 20 rodagens, com 67,64%.

De modo análogo ao caso anterior, definiu-se a variável de interesse – nesse caso, o atributo "Classe" –, visando identificar as relações existentes entre este e os demais atributos do conjunto de dados. Novamente, fez-se uso do algoritmo *Random Forest*, em que se obtiveram os percentuais de influência descritos na Tabela 12:

Tabela 12 – Percentual de influência de cada atributo - xAPI-Edu-Data

Atributo	Resultado
Gênero	0,93%
Nacionalidade	0,88%
Local de nascimento	0,20%
Nível de escolaridade	1,06%
Estágio	0,28%
Curso matriculado	0,27%
Semestre	1,94%
Responsável	7,12%
Número de participações em sala	7,13%
Número de visitas ao material do curso	2,84%
Verificações as notificações do curso	3,71%
Número de discussões	1,04%
Participação do responsável	1,70%
Satisfação do responsável	6,74%
Número de ausências	62,42%

Fonte: Elaborado pela autora

Entende-se, portanto, que:

> *ainda ao nível comportamental, mas no que diz respeito ao percurso escolar do sujeito, os alunos desistentes são também aqueles que pos-*

> *suem uma maior taxa de absentismo, cujos estudos indicam menos 68% de probabilidade de atingirem um nível de formação média a superior (Balfanz, Herzog,& Maclver,2007) possuindo também maior número de reprovações, aos quais conduzem ao enfraquecimento da vinculação à escola. (MIGUEL; RIJO; LIMA, 2012, p.133)*

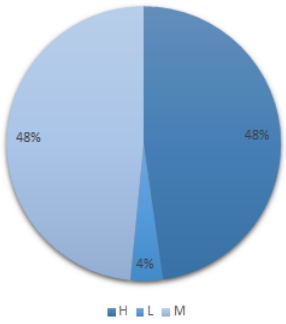
Conforme as ideias da autora e os resultados obtidos, constatou-se que o atributo "Número de ausências" classifica cada indivíduo em duas categorias, de acordo com a quantidade de faltas: inferior a sete ausências e superior a sete ausências – aqui, a influência na classificação final foi de 62,42%. Segundo a Figura 8 e a Tabela 13, que relaciona os mesmos atributos, os alunos mais assíduos estão nos níveis H e M, ao passo que a maior parte dos discentes do nível L tiveram o número de faltas superior a sete.

Tabela 13 – Número de faltas por nível de proficiência

	Superior-7	Inferior-7
H	4	138
M	116	11
L	71	40

Fonte: Elaborado pela autora

Figura 8 – Número de faltas em cada nível de proficiência



Fonte: Elaborado pela autora

Seguindo a análise, verificou-se que outro atributo relevante ao desempenho dos estudantes é "Responsável pelo estudante" que, com a implementação do Random Forest, obteve porcentagem de 7,12%. Conforme a Tabela 14 e a Figura 9, os alunos no nível mais avançado de estudo (H = Higher) responderam que seu responsável é a mãe (mum) – dos 142 estudantes que se encontram nesse nível, 100 citaram essa alternativa. Nos demais níveis (L – Low, M – Medium), os dados foram agrupados para outra opção de responsável (pai), em que esse membro familiar se responsabilizaria por eles.

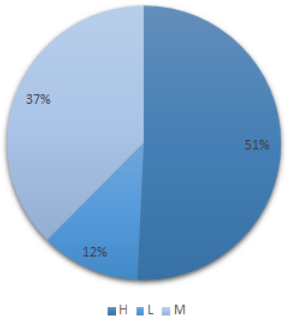
Outro atributo de porcentagem significativa na classificação foi "número de participações em sala", correspondendo a 7,13%. O gráfico da Figura 10, obtido a partir da

Tabela 14 – Nível de proficiência e responsável

	Pai	Mãe
H	42	100
M	104	23
L	137	74

Fonte: Elaborado pela autora

Figura 9 – Nível de proficiência e responsável

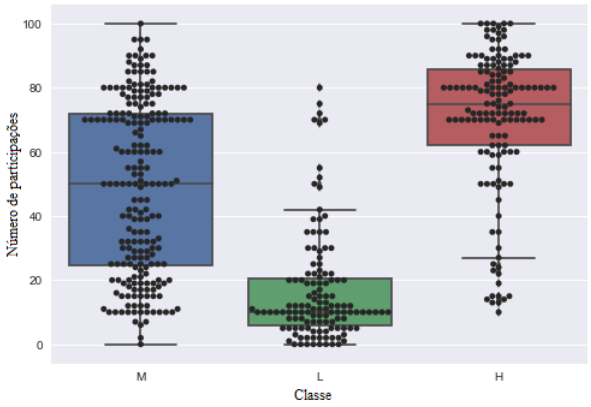


Fonte: Elaborado pela autora

comparação dos dois critérios, permitiu avaliar que as informações são consistentes – os alunos classificados em nível básico tiveram a participação média inferior a 20 vezes. Em contrapartida, discentes do nível avançado apresentaram maior incidência de participações, o que sugere que a maior parte dos estudantes se posicionou mais de 80 vezes durante o semestre.

Como as respostas a esse atributo foram diversas, o gráfico de barras não possibilitou uma visualização apropriada dos dados. Para isso, optou-se pela plotagem do gráfico do tipo boxplot que, com uma quantidade razoável de informações, possibilita verificar se a distribuição ocorre de forma simétrica ou se os dados estão dispersos.

Figura 10 – Número de participações por nível de proficiência



Fonte: Elaborado pela autora

5 CONSIDERAÇÕES FINAIS

Ao considerar que o foco deste estudo diz respeito à reflexão sobre as práticas que viabilizam o aprendizado significativo e efetivo dos estudantes:

> *pretende-se ainda que essa intervenção seja realizada o mais precocemente possível, não devendo ser necessário que o aluno manifeste um nível de ruptura significativa com o sistema escolar, ou que abandone a escola, para que seja identificado e alvo de intervenção. (MIGUEL; RIJO; LIMA, 2012, p.139)*

Nesse contexto foram adotadas, como conceitos do papel do professor, as ideias de Perrenoud (2000, p.139), ao considerar que tal profissional se concentra "[...] na criação, na gestão e na regulação das situações de aprendizagem". Desde a escolha do conjunto de dados até o momento das análises feitas após a aplicação da EDM, consideram-se imprescindíveis os fatores sociais e econômicos dos discentes, e de igual modo são determinantes no processo de ensino e aprendizagem. Nesse sentido, a EDM é uma ferramenta que auxilia o docente no gerenciamento das informações fornecidas e, ao mesmo tempo, omitidas pelo contexto escolar, uma vez que passam despercebidas no cotidiano.

Com acesso a linhas de comandos bem definidas e previamente adaptadas à realidade de atuação, o docente pode utilizar tais ferramentas para identificar características determinantes que viabilizem a aprendizagem. Nessa perspectiva, o professor teria a função de alimentar o conjunto de dados e solicitar, com apenas um comando no *Jupyter Notebook*, que eles sejam "rodados" novamente, visto que os algoritmos estabelecem padrões e regras a partir do conjunto de teste.

Nas palavras de Gil (2010 apud DAMIANI et al., 2013, p.132), um experimento "[...] consiste essencialmente em determinar um objeto de estudo, selecionar as variáveis capazes de influenciá-lo e definir as formas de controle e de observação dos efeitos que a variável produz no objeto" – é nessa perspectiva que o trabalho foi proposto. Experimentações feitas nos dados de Student-mat e xAPI-Edu-Data permitiram verificar fatores sociais com expressiva influência no resultado final dos estudantes desses conjuntos, para ressaltar a relevância destes para o desempenho escolar. Isso possibilita ao docente priorizar a identificação dessas situações em sala de aula para a mediação.

Se fossem consideradas apenas as experiências anteriores de professores e seus relatos, supor-se-ia que um dos fatores determinantes para o bom desempenho dos estudantes nas aulas seria o acompanhamento frequente da disciplina ministrada. Outro item que salientado é a relação familiar, em que os alunos com os melhores desempenhos responderam,

em grande medida, que são acompanhados pela mãe na vida escolar – a EDM possibilitou que deixassem os campos da suposição para confirmar por meio das análises estatísticas.

Assim, nos casos individuais, como estudantes que ingressam no decorrer do ano letivo, os resultados podem auxiliar no diagnóstico e na intervenção preventiva, em que não é preciso um longo período de convivência para identificar outros fatores relevantes. Diante das duas análises, ficou evidente que o número de faltas (ausências) é um fator com alta acurácia, ou seja, exerce grande influência no rendimento final do aluno; logo, com a consulta simples ao histórico escolar do discente (documento que registra as informações do estudante durante a sua vida escolar e é obrigatório para matrícula e transferência), o professor consegue identificar essa característica e intervir, caso haja necessidade.

Frente aos fatores identificados com o auxílio da EDM, metodologias podem ser implementadas para favorecer as mediações pedagógicas em sala de aula que, para Damiani et al. (2013, p.58), são "[...] investigações que envolvem o planejamento e a implementação de interferências (mudanças, inovações) destinadas a produzir avanços, melhorias, nos processos de aprendizagem dos sujeitos que delas participam". Assim, neste capítulo final serão discutidos conceitos relacionados à medicação pedagógica e à sua influência no processo de ensino e aprendizagem, com destaque a ferramentas e recursos tecnológicos utilizados para esse fim.

Prado (2006, p.103) pontua que:

> O fato de o professor observar e entender como o aluno aprende - suas fragilidades conceituais, potencialidades e estratégias de resolução - lhe dá condições para ensinar por meio da criação de situações de aprendizagem que possam ser significativas para o aluno. A criação destas situações de aprendizagem demanda do professor (antes e durante a sua ação pedagógica) o desenvolvimento de estratégias envolvendo os materiais, as atividades e as interações, mas não de forma isolada e/ou sequencial. Neste caso, como o foco centra-se na articulação entre o ensino e aprendizagem, os elementos da mediação vão se entrelaçando na ação, expressando, com isto, a integração dos aspectos relacionados às necessidades e interesses dos alunos, bem como aqueles relacionados à intencionalidade pedagógica do professor.

Corrobora-se com Gervai (2007, p.29-30) que, embasada nas ideias de Vygotsky, pondera que:

> A construção das funções psíquicas está ligada à apropriação da cultura humana, que acontece, segundo o autor, através das relações interpessoais dentro da sociedade à qual o indivíduo pertence. Vygotsky considera que essa apropriação ocorre através da educação e do ensino com mediação de adultos e/ou pares mais experientes.

Nesse sentido, é necessário que os profissionais da educação busquem ferramentas que viabilizem o processo de seleção e tratamento de informações relevantes à mediação pedagógica, propiciando avanços na construção dos conhecimentos que não aconteceriam

sem a intervenção do professor. Ainda de acordo com a autora, a mediação permite a criação de "[...] um programa de trabalho pedagógico para um grupo de pessoas com um histórico de vida, com conhecimentos e habilidades parecidas" (p.35).

Ressalta-se que:

> A mediação pedagógica requer preocupação com a organização didático - metodológica dos saberes a serem ensinados, contemplando as ações em torno do conhecimento contextualizado em função da linguagem, das práticas socioculturais e dos avanços científico-tecnológicos. (MALLMANN, 2010, p.135)

Com respaldo nas ideias de Mallmann (2010) citadas anteriormente, indica-se que a mediação pedagógica se inicia na elaboração do plano de aula, uma vez que o docente deve considerar diversos aspectos que viabilizam ou não o aprendizado do aluno, analisar o contexto sociocultural em que a sala de aula está inserida, diagnosticar o conhecimento prévio dos estudantes sobre o conteúdo a ser abordado e verificar se há os recursos necessários para a construção desse conhecimento para, depois disso, optar pela metodologia que melhor se adapta à situação em questão.

De acordo com a pesquisa bibliográfica realizada neste trabalho, foram encontradas metodologias de ensino que viabilizam atividades em sala de aula e exploram as múltiplas habilidades dos estudantes na construção do conhecimento, permitindo mediações pedagógicas diversas. Uma delas se refere aos jogos, mas:

> Quando nos referimos à utilização de jogos nas aulas de matemática como um suporte metodológico, consideramos que tenha utilidade em todos os níveis de ensino. O importante é que os objetivos com o jogo estejam claros, a metodologia a ser utilizada seja adequada ao nível em que se está trabalhando e, principalmente, que represente uma atividade desafiadora ao aluno para o desencadeamento do processo. (GRANDO, 2004, p.25,26)

Usualmente utilizada para introduzir um conteúdo, reforça-se que essa metodologia possibilita ao aluno se desvincular do "rigor" matemático, bem como auxilia no processo de abstração e elaboração de estratégias para solucionar problemas.

Marco (2004, p.35), com base em Chateau (1987), define os jogos nas aulas de matemática como:

> (...) uma atividade dinâmica e de prazer, desencadeada por um movimento próprio, desafiando e motivando os jogadores à ação, permitindo, possivelmente, uma ponte para algum conhecimento. O jogo está em quem joga e não em quem assiste ao jogo, ou seja, o jogo está na ação e é ação.

É notório que os jogos levam à construção de situações variadas para as mediações pedagógicas, seja por meio dos questionamentos sobre as escolhas feitas pelo jogador ou do

esgotamento de possibilidades para obter o mesmo resultado de maneira diferente ou mais eficaz. Assim, o jogo se torna "[...] um ambiente de aprendizagem e (re)criação conceitual e não apenas de reprodução mecânica do conceito, como ocorre na resolução de uma lista de exercícios denominados problemas" (MARCO, 2004, p.37).

A investigação em sala de aula tem como uns dos principais percussores João Pedro da Ponte que, juntamente com outros pesquisadores, definiu que: "Investigar é procurar conhecer o que não se sabe" (PONTE; BROCARDO; OLIVEIRA, 2003, p.13). Nesse sentido, as atividades investigativas devem ser propostas em sala de aula para os alunos trilharem os próprios caminhos no processo de aprendizagem, serem efetivamente construtores do seu conhecimento e "[...] chamados a agir como um matemático" (PONTE; BROCARDO; OLIVEIRA, 2003, p.23), elaborando hipóteses e provas, além de arguirem com professores e colegas sobre os resultados obtidos.

A metodologia de ensino mais apropriada será a que melhor se adequar ao perfil estudantil trabalhado, algo definido a partir dos fatores sociais de maior incidência na comunidade em questão. Sendo assim, a preparação das aulas até a sua execução percorrem o caminho que objetiva à ascensão escolar do maior número de alunos.

Ressalta-se que as linhas de programação usadas em Student-mat foram as mesmas implementadas em xAPI-Edu-Data. Assim, os mesmos comandos podem ser aplicados em quaisquer conjuntos de dados gerados por questionários que considerem atributos semelhantes ou que modifiquem a nomenclatura de cada item de acordo com o conjunto a ser analisado.

Destarte, tais considerações reiteram a eficácia da EDM em diferentes culturas educacionais, cujo uso possibilita destacar fatores relevantes e características em comum de uma comunidade com o mesmo perfil estudantil. De fato, isso pode facilitar a criação de estratégias para a mediação pedagógica, com o escopo de consolidar e progredir o conhecimento voltado ao maior número de indivíduos.

Referências

ALVES, M. T. G. et al. Fatores familiares e desempenho escolar: uma abordagem multidimensional. *DADOS-Revista de Ciências Sociais*, Universidade do Estado do Rio de Janeiro, v. 56, n. 3, 2013.

AMRIEH, E. A.; HAMTINI, T.; ALJARAH, I. Preprocessing and analyzing educational data set using x-api for improving student's performance. *Applied Electrical Engineering and Computing Technologies (AEECT)*, 2015 IEEE Jordan Conference on (pp. 1-5). IEEE, 2015.

AMRIEH, E. A.; HAMTINI, T.; ALJARAH, I. Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, n. 9(8), p. 119–136, 2016.

ATHANI, S. S. et al. Student performance predictor using multiclass support vector classification algorithm. *International Conference on Signal Processing and Communication*, p. 341–346, 2017.

BASU, C. et al. Recommendation as classification: Using social and content-based information in recommendation. In: *Aaai/iaai*. [S.l.: s.n.], 1998. p. 714–720.

BHERING, E.; SIRAJ-BLATCHFORD, I. A relaçãoo escola-familia - um modelo de trocas e colaboração. 1999.

BRAMER, M. Undergraduate topicsin computer science – principles of data mining. *Springer*, 2007.

BRASIL, M. Secretaria de educação. *Parâmetros Curriculares Nacionais: terceiro e quarto ciclos do ensino fundamental - Matemática*, Brasília: MEC/SEF, 1998.

BREIMAN, L. Arcing classifiers. *Annals of Statistics*, 1998.

CAPELLINI, S. A. et al. Medidas de desempenho escolar: avaliação formal e opinião de professores. *Estudos de Psicologia (Campinas)*, SciELO Brasil, 2004.

CARRIJO, I. B. *Extração de regras operacionais ótimas de sistemas de distrubuição de água através de algoritmos genéticos multiobjetivo e aprendizado de máquina*. Tese (Doutorado) — Universidade de São Paulo, 2004.

CORTEZ, P.; SILVA, A. M. G. Using data mining to predict secondary school student performance. EUROSIS-ETI, p. 5–12, 2008.

DAMIANI, M. F. et al. Discutindo pesquisas do tipo intervenção pedagógica. *Cadernos de Educação*, v. 45, p. 57–67, 2013.

DUNHAM, M. H. Data mining:introductory and advanced topics. Upper Saddle River, p. 1–6, 2003.

E.-CALDERON, O. A.; B.-ARANIBAR, D. Optimal selection of factors using genetic algorithms and neural networks for the prediction of student's academic performance. *Latin America Congress on Computational Intelligence (LA-CCI)*, p. 341–346, 2015.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37–54, 1996.

FERREIRA, G. S. Investigação acerca dos fatores determinantes para a conclusão do ensino fundamental utilizando mineração de dados educacionais no censo escolar da educação básica do inep 2014. *Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação(CBIE)*, p. 1034–1043, 2015. Disponível em: <http://dx.doi.org/10.5753/cbie.wcbie.2015.1034>.

GERVAI, S. M. S. A mediação pedagógica em contexto de aprendizagem online. 2007. Disponível em: <https://tede2.pucsp.br/handle/handle/13857>.

GONÇALVES, E. H. A utilização de tecnologias digitais no curso de licenciatura em matemática parfor/ead da universidade federal de uberlândia. Universidade Federal de Uberlândia, 2018.

GOTTARDO, E.; KAESTNER, C.; NORONHA, R. V. Previsão de desempenho de estudantes em cursos ead utilizando mineração de dados: uma estratégia baseada em séries temporais. *Anais do 23º Simpósio Brasileiro de Informática na Educação*, 2012. Disponível em: <http://dx.doi.org/10.5753/cbie.sbie.2012.\%25p>.

GRANDO, R. C. O jogo e a matemática no contexto da sala de aula. *Paulus*, Campinas, 2004.

JAIN, Y. K.; BHANDARE, S. K. Min max normalization based data perturbation method for privacy protection. *International Journal of Computer & Communication Technology*, v. 2, n. 8, p. 45–50, 2011.

KASKI, S.; KOHONEN, T. Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. In: CITESEER. *Neural networks in financial engineering. Proceedings of the third international conference on neural networks in the capital markets*. [S.l.], 1996.

KENSKI, V. M. Novas tecnologias. *Revista Brasileira de Educação*, n. 8, p. 58–71, 1998.

KIRALJ, R.; FERREIRA, M. Basic validation procedures for regression models in qsar and qspr studies: theory and application. *Journal of the Brazilian Chemical Society*, SciELO Brasil, v. 20, n. 4, p. 770–787, 2009.

KLOCK, C. E.; RIBAS, R. P.; REIS, A. I. Karma: um ambiente para o aprendizado de síntese de funções booleanas. *Brazilian Journal of Computers in Education*, v. 18, n. 02, p. 33, 2010.

LUTZ, M.; ASCHER, D. Aprendendo python. Tradução TORTELLO, p. 568, 2007.

MACEDO, G. A. Fatores associados os rendimento escolar de alunos da 5ª série(2000) – uma abordagem do valor adicionado. *XIV Encontro Nacional de Estudos Populacionais - ABEP*, 2004.

MALLMANN, E. M. *Redes e Mediação: Princípios Epistemológicos de Teoria de Rede de Mediadores da Educação*. [S.l.: s.n.], 2010. v. 54. 221–241 p.

MANHÃES, L. M. B.; AL et. Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. *XXII Simpósio Brasileiro de Informática na Educação - XVII Workshop de Informática na Escola*, p. 150–159, 2011.

MARCO, F. F. d. Estudo dos processos de resolução de problema mediante a construção de jogos computacionais de matemática no ensino fundamental. [sn], 2004. Disponível em: <http://repositorio.unicamp.br/bitstream/REPOSIP/253205/1/Marco_FabianaFiorezide_M.pdf>.

MATSUBARA, E. T. O algoritmo de aprendizado semi-supervisionado co-training e sua aplicação na rotulação de documentos. 2004. Acesso em: 25/07/2018. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-19082004-092311/en.php>.

MIGUEL, R. T.; RIJO, D.; LIMA, L. N. Fatores de risco para o insucesso escolar: A relevância da variáveis psicológicas e comportamentais do aluno. *Revista Portuguesa de Pedagogia*, v. 46, p. 89–114, 2012.

MONARD, M. C.; BARANAUSKAS, J. A. *Conceitos sobre Aprendizado de Máquina*. [S.l.]: Editora Manole, 2003. 89–114 p.

MORAN, J. M. *A Educação que desejamos: novos desafios e como chegar lá*. Campinas, SP: [s.n.], 2007.

NUNES, F. B.; VOSS, G. B.; CAZELLA, S. C. Mineração de dados educacionais e mundos virtuais: um estudo exploratório no opensim. *Anais dos Workshops do IV Congresso Brasileiro de Informática na Educação(CBIE)*, p. 1044–1053, 2015. Disponível em: <http://dx.doi.org/10.5753/cbie.wcbie.2015.1044>.

OLSON, D. L.; DELLEN, D. Advanced data mining techniques. *Springer*, 2008.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PERRENOUD, P. Dez competências para ensinar. *Artes Médicas Sul.*, 2000.

PONTE, J. P. da; BROCARDO, J.; OLIVEIRA, H. *Investigações matemáticas na sala de aula*. Belo Horizonte: [s.n.], 2003.

PRADO, M. E. B. B. A mediação pedagógica: suas relações e interdependências. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2006. v. 1, n. 1, p. 101–110.

SEVERO, C. E. P. et al. *Mediação Pedagógica em Ambientes Virtuais de Ensino Aprendizagem Através de Agentes de Mineração de Dados Educacionais*. 2011. Acesso em: 28/08/2017. Disponível em: <http://www.seer.ufrgs.br/index.php/InfEducTeoriaPratica/article/view/14223/16840>.

SIMOUDIS, E. Reality check for data mining. IEEE Expert, p. 26–33, 1996.

SIQUEIRA, C. M.; GURGEL-GIANNETTI, J. Mau desempenho escolar: uma visão atual. *Revista da Associação Médica Brasileira*, Elsevier, v. 57, n. 1, p. 78–87, 2011.

SOKOLOVA, M.; JAPKOWICZ, N.; SZPAKOWICZ, S. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: SPRINGER. *Australasian joint conference on artificial intelligence*. [S.l.], 2006. p. 1015–1021.

WERSTCH, J. V. *Vygotsky y la formación social de la mente*. [S.l.]: Paiadós, 1988.

WITTEN, I. H.; FRANK, E. Data mining - practical machine learning tools and techniques. *Elsevier*, 2005.

A Atributos numéricos

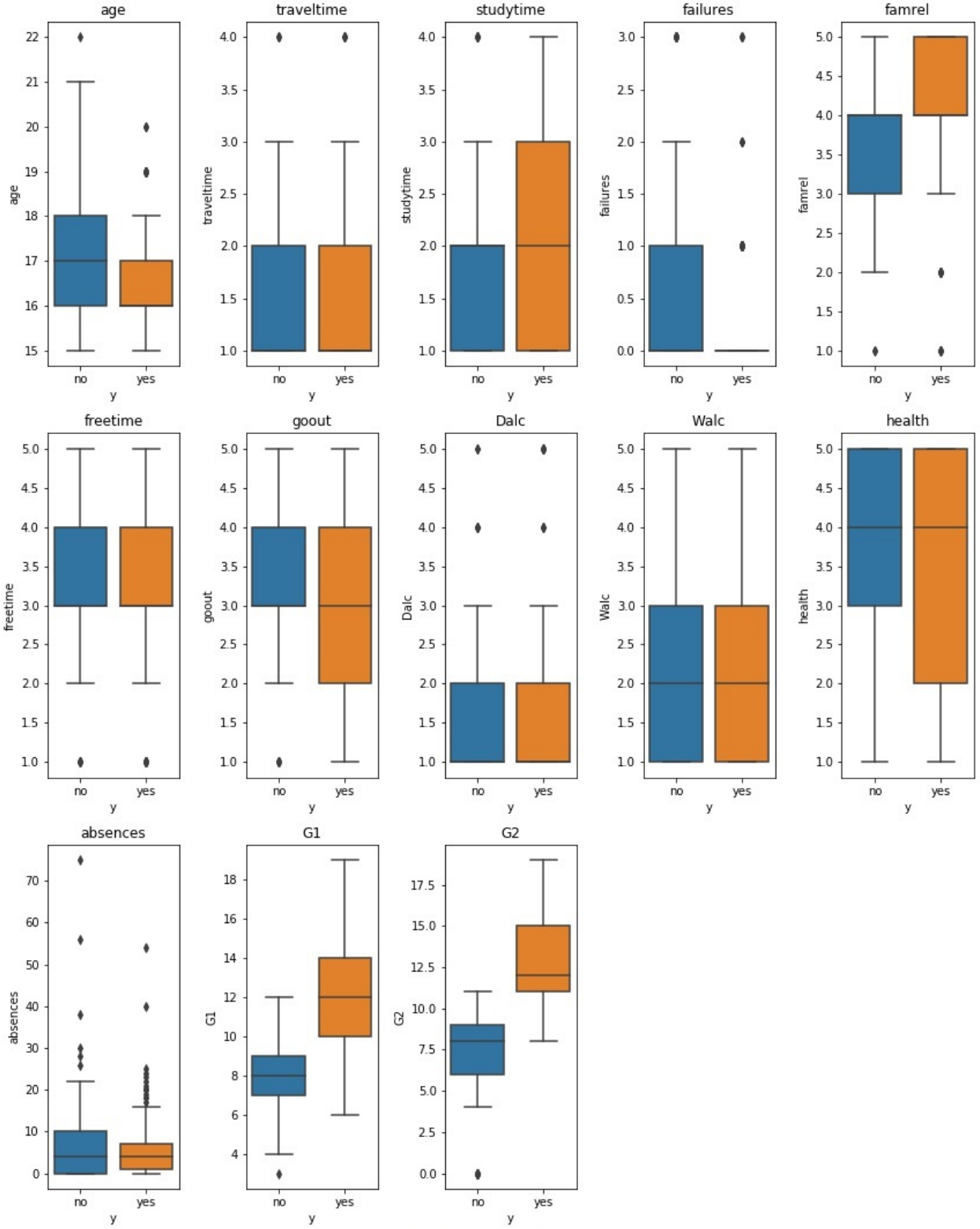
Análise univariada dos atributos numéricos:

Tabela 15 – Atributos numéricos - Student-mat

	Cont.	Média	Des padrão	Min.	1°Q	2°Q	3°Q	Máx.	Únicos	Moda	Freq. rel. moda	Ampl.	Skew	z-score
idade	395	16,69	1,27	15	16	17	18	22	8	16	26,33%	7	0,46	3,66
tempo de deslocamento	395	1,44	0,69	1	1	1	2	4	4	1	65,06%	3	1,60	9,72
tempo de estudo	395	2,03	0,83	1	1	2	2	4	4	2	50,13%	3	0,63	4,81
reprovações	395	0,33	0,74	0	0	0	0	3	4	0	78,99%	3	2,38	12,24
relacionamento familiar	395	3,94	0,89	1	4	4	5	5	5	4	49,37%	4	-0,95	-6,73
tempo livre	395	3,23	0,99	1	3	3	4	5	5	3	39,75%	4	-0,16	-1,34
encontro com os amigos	395	3,10	1,11	1	2	3	4	5	5	3	32.,91%	4	0.,12	0,96
Consumo de álcool dias úteis	395	1,48	0,89	1	1	1	2	5	5	1	69.,87%	4	2,18	11,68
Consumo de álcool finais de semana	395	2,29	1,28	1	1	2	3	5	5	1	38,23%	4	0,61	4,67
saúdede	395	3,55	1,39	1	3	4	5	5	5	5	36,96%	4	-0,49	-3,87
ausências	395	5,70	8,00	0	0	4	8	75	34	0	29,11%	75	3,66	15,11
G1	395	10.,90	3.,31	3	8	11	13	19	17	10	12,91%	16	0,24	1,95
G2	395	10.71	3,76	0	9	11	13	19	17	9	12,66%	19	-0,43	-3,41

Fonte: Elaborado pela autora

Gráficos da distribuição dos atributos numéricos de acordo com o target:



Fonte: Elaborado pela autora

B Tabela de atributos categóricos

Tabela 16 – Atributos categóricos - Student-mat

	Contagem	Únicos	Classes	Moda	Freq. abs. moda	Freq. rel. moda
scholl	395	2	GP, MS	MP	349	8,35%
sex	395	2	F, M	F	208	52,66%
address	395	2	U, R	U	307	77,72%
famsize	395	2	GT3, LE3	GT3	281	71,14%
Pstatus	395	2	A, T	T	354	89,62%
Medu	395	5	0, 1, 2, 3, 4	4	131	33,16%
Fedu	395	5	0, 1, 2, 3, 4	2	115	29,11%
Mjob	395	5	at_home, health, other, services, teacher	other	141	35,70%
Fjob	395	5	at_home, health, other, services, teacher	other	217	54,94%
reason	395	4	course, other, home, reputation	course	145	36,71%
guardian	395	3	mother, father, other	mother	273	69,11%
schoolsup	395	2	yes, no	no	344	87,09%
famsup	395	2	yes, no	yes	241	61,27%
paid	395	2	yes, no	no	214	54,18%
activities	395	2	yes, no	yes	201	50,89%
nursey	395	2	yes, no	yes	314	79,49%
higher	395	2	yes, no	yes	375	94,94%
internet	395	2	yes, no	yes	329	83,29%
romantic	395	2	yes, no	no	263	66,58%

Fonte: Elaborado pela autora

Frequência das resposta em cada atributo categórico:

Tabela 17 – Frequência dos atributos categóricos

Coluna	Classe	freq. abs.	freq. rel.
school	GP	349	88,35%
	MS	46	11,65%
sex	F	208	52,66%
	M	187	47,34%
address	U	307	77,72%
	R	88	22,28%
famsize	GT3	281	71,14%
	LE3	114	28,86%
Pstatus	A	41	10,38%
	T	354	89,62%
Medu	4	131	33,16%
	1	59	14,94%
	3	99	25,06%
	2	103	26,08%
	0	3	0,76%
Fedu	4	96	24,30%
	1	82	20,76%
	2	115	29,11%
	3	100	25,32%
	0	2	0,51%
Mjob	at_home	59	14,94%
	health	34	8,61%
	other	141	35,70%
	services	103	26,08%
	teacher	58	14,68%
Fjob		29	7,34%
	other	217	54,94%
	services	111	28,1%
	health	18	4,56%
	at_home	20	5,06%
reason	course	145	36,71%
	other	36	9,11%
	home	109	27,59%
	reputation	105	26,58%
guardian	mother	273	69,11%
	father	90	22,78%
	other	32	8,10%
schoolsup	yes	51	12,91%
	no	344	87,09%
famsup		153	38,73%
	yes	242	61,27%
paid	no	214	54,18%
	yes	181	45,82%

Coluna	Classe	freq. abs.	freq. rel.
activities	no	194	49,11%
	yes	201	50,89%
nursery		314	79,49%
	no	81	20,51%
higher	yes	375	94,94%
	no	20	5,06%
internet		66	16,71%
	yes	329	83,29%
romantic	no	263	66,58%
	yes	132	33,42%
famrel	4	195	49,37%
	5	106	26,84%
	3	68	17,22%
	1	8	2,03%
	2	18	4,56%
goout	4	86	21,77%
	3	130	32,91%
	2	103	26,08%
	1	23	5,82%
	5	53	13,42%
Dalc	1	276	69,87%
	2	75	18,99%
	5	9	2,28%
	3	26	6,58%
	4	9	2,28%
Walc	1	151	38,23%
	3	80	20,25%
	2	85	21,52%
	4	51	12,91%
	5	28	7,09%
health	3	91	23,04%
	5	146	36,96%
	1	47	11,90%
	2	45	11,39%
	4	66	16,71%

Fonte: Elaborado pela autora