UNIVERSIDADE FEDERAL DO CEARÁ
INSTITUTO UNIVERSIDADE VIRTUAL
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA EDUCACIONAL

HÉLDER ANTERO AMARAL NUNES

MINERAÇÃO DE DADOS SOCIOECONÔMICOS E EDUCACIONAIS DE DISCENTES PARA PREDIÇÃO DE EVASÃO E RETENÇÃO ESCOLAR

FORTALEZA
2023

HÉLDER ANTERO AMARAL NUNES

MINERAÇÃO DE DADOS SOCIOECONÔMICOS E EDUCACIONAIS DE DISCENTES PARA PREDIÇÃO DE EVASÃO E RETENÇÃO ESCOLAR

Dissertação apresentada ao Curso de Mestrado Profissional em Tecnologia Educacional do Programa de Pós-Graduação em Tecnologia Educacional do Instituto Universidade Virtual da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Tecnologia Educacional. Área de Concentração: Tecnologia Educacional

Orientador: Prof. Dr. Leonardo Oliveira Moreira

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

N925m Nunes, Hélder Antero Amaral.
Mineração de Dados Socioeconômicos e Educacionais de Discentes para Predição de Evasão e Retenção Escolar / Hélder Antero Amaral Nunes. – 2023.
93 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Instituto UFC Virtual, Programa de Pós-Graduação em Tecnologia Educacional, Fortaleza, 2023.
Orientação: Prof. Dr. Leonardo Oliveira Moreira.

1. Mineração de dados educacionais. 2. Evasão. 3. Retenção. 4. Predição. I. Título.

CDD 371.33

HÉLDER ANTERO AMARAL NUNES

MINERAÇÃO DE DADOS SOCIOECONÔMICOS E EDUCACIONAIS DE DISCENTES PARA PREDIÇÃO DE EVASÃO E RETENÇÃO ESCOLAR

Dissertação apresentada ao Curso de Mestrado Profissional em Tecnologia Educacional do Programa de Pós-Graduação em Tecnologia Educacional do Instituto Universidade Virtual da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Tecnologia Educacional. Área de Concentração: Tecnologia Educacional

Aprovada em: 27/10/2023

BANCA EXAMINADORA

Prof. Dr. Leonardo Oliveira Moreira (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Emanuel Ferreira Coutinho
Universidade Federal do Ceará (UFC)

Prof. Dr. José Gilvan Rodrigues Maia
Universidade Federal do Ceará (UFC)

Prof. Dr. Thiago Iachiley Araújo de Souza
Centro Universitário Christus (Unichristus)

Aos meus filhos Miguel (Gueguel) e Daniel (Diel), desde que me tornei pai, tudo o que faço é por vocês.

AGRADECIMENTOS

Ao Prof. Dr. Leonardo Moreira, por me orientar durante toda a minha jornada no mestrado. Muito obrigado pela sua imensa paciência!

Ao Prof. Dr. Edgar Marçal, coordenador do programa PPGTE, pelo seu esforço em estabelecer o curso de mestrado.

Aos meus filhos, Miguel e Daniel, que nos momentos em que estive ausente, focado nos estudos, sempre entenderam quando eu deixava de brincar para estudar, codificar ou escrever a dissertação. Além de agradecer, quero deixar registrado que tudo o que faço é sempre pensando em vocês!

À minha esposa, que sempre me incentiva e motiva a me dedicar cada vez mais!

Aos meus pais, irmã, tios e sogros, que nos momentos em que estive ausente, focado nos estudos, sempre me fizeram compreender que o futuro é construído com dedicação constante no presente! Agradeço por cada oração feita!

Agradeço a todos os professores por me proporcionarem o conhecimento não apenas racional, mas também por manifestarem caráter e afetividade na educação durante o meu processo de formação. Agradeço a todos que se dedicaram a mim, não apenas por me ensinarem, mas por me ajudarem a aprender.

"O sonho é que leva a gente para frente. Se a gente for seguir a razão, fica aquietado, acomodado."

(Ariano Suassuna)

RESUMO

A evasão e retenção escolar sempre foram temas abordados na educação brasileira. Esses problemas afetam não apenas a vida dos alunos, de suas famílias e da sociedade em que vivem, mas também o orçamento das instituições de ensino. Isso é evidenciado pelo fato de que a alta taxa de evasão e retenção representa um desperdício de recursos públicos. Diante disso, as instituições de ensino devem desempenhar seu papel como educadoras e buscar abordagens inovadoras para aplicar seus recursos no combate à evasão e retenção. A Mineração de Dados Educacionais possibilita o conhecimento de fatores que podem melhorar a proposta educacional, bem como a previsão do desempenho dos alunos e dos fatores que influenciam o aprendizado. Para uma compreensão mais aprofundada do tópico e do estado atual da arte, foi conduzido uma Revisão Sistemática da Literatura. Essa revisão permitiu identificar os algoritmos de Inteligência Artificial mais amplamente empregados, bem como os dados associados a eles. Como resultado, essa RSL desempenhou um papel fundamental ao aprimorar a compreensão e ao definir os requisitos para o software proposto. Com base nessas características, o objetivo deste trabalho é desenvolver uma ferramenta que faça uso da mineração de dados educacionais e socioeconômicos. Através da aplicação de técnicas de classificação, a ferramenta visa auxiliar gestores educacionais no enfrentamento da evasão e retenção escolar, desde o momento da matrícula até o início das aulas. Essa ferramenta foi desenvolvida em Java, fazendo uso da biblioteca Weka. No processo de validação do experimento, foram utilizadas duas bases de dados distintas, resultando em uma impressionante taxa de acurácia de mais de 97% em ambas as bases. Para a validação da usabilidade do software, foi aplicado o questionário Sistema de Avaliação da Usabilidade, ao qual foram adicionadas duas perguntas adicionais com o intuito de compreender melhor as eventuais dificuldades na utilização do software pela comunidade escolar. A aplicação do questionário foi conduzida por profissionais da área de educação em escolas situadas na região do sertão pernambucano. Os resultados desse processo de validação oferecem uma visão valiosa sobre o desempenho e a usabilidade da ferramenta, contribuindo para sua avaliação e aprimoramento contínuo.

Palavras-chave: mineração de dados educacionais; evasão; retenção; predição.

ABSTRACT

Dropout and school retention have always been topics addressed in Brazilian education. These issues impact not only students' lives, their families, and the society they live in but also the budget of educational institutions. This is evidenced by the fact that the high dropout and retention rates represent a waste of public resources. In light of this, educational institutions must fulfill their role as educators and seek innovative approaches to allocate their resources effectively in combating dropout and retention. Educational Data Mining enables the understanding of factors that can enhance the educational proposal, as well as the prediction of students' performance and the factors influencing learning. For a deeper understanding of the topic and the current state of the art, a Systematic Literature Review was conducted. This review allowed for the identification of the most widely used Artificial Intelligence algorithms, along with the associated data. As a result, this SLR played a fundamental role in improving the understanding and defining the requirements for the proposed software. Based on these characteristics, the goal of this work is to develop a tool that utilizes educational and socioeconomic data mining. Through the application of classification techniques, the tool aims to assist educational managers in addressing dropout and retention from the moment of enrollment until the start of classes. This tool was developed in Java, using the Weka library. In the experiment validation process, two distinct databases were used, resulting in an impressive accuracy rate of over 97% in both databases. To assess the software's usability, the System Usability Scale questionnaire was administered, with two additional questions added to better understand potential difficulties in using the software within the school community. The questionnaire was conducted by education professionals in schools located in the region of the Pernambuco hinterlands. The results of this validation process provide valuable insights into the tool's performance and usability, contributing to its ongoing evaluation and improvement.

Keywords: educational data mining; evasion; retention; prediction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

BI	*Business Intelligence*
CPF	Cadastro de Pessoas Físicas
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSV	*Comma-Separated Values*
DW	*Data Warehouse*
EAD	Ensino a Distância
GPL	*General Public License*
IA	Inteligência Artificial
IE	Instituição de Ensino
KNN	*K-Nearest Neighbor*
LGPD	Lei Geral de Proteção de Dados
MDE	Mineração de Dados Educacionais
MLP	*Multi Layer Perceptron*
ODS	*OpenDocument Spreadsheet Document*
OLAP	*Online Analytical Processing*
PDF	*Portable Document Format*
QP	Questões de Pesquisa
RNA	Rede Neural Artificial
RNAs	Redes Neurais Artificiais
RSL	Revisão Sistemática da Literatura
SUS	*System Usability Scale*
SVM	*Support Vector Machine*
TCLE	Termo de Consentimento Livre e Esclarecido
TI	Tecnologia da Informação
UFC	Universidade Federal do Ceará
UFPI	Universidade Federal do Piauí
UnB	Universidade de Brasília
XLS	Microsoft Excel 97-2003 *Workbook*
XLSX	Microsoft Excel *Workbook*

LISTA DE SÍMBOLOS

k	Parâmetro que direcionara a quantidade de vizinhos no algoritmo K-Nearest Neighbor

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo, é apresentada a motivação e a contextualização desta pesquisa, que tem como finalidade a construção de um software para mineração de dados escolares, com o intuito de realizar a predição de evasão e retenção de alunos. Na seção 1.1, é discutida a contextualização e a motivação desta dissertação. Já na seção 1.3, são apresentados os objetivos e as questões de pesquisa. Na seção 1.4, por sua vez, são apresentados os principais resultados e contribuições da pesquisa. Por fim, a seção 1.5 detalha a estrutura desta dissertação e os capítulos seguintes.

1.1 Motivação e Contextualização

O sucesso de um estudante em qualquer nível de formação desempenha um papel fundamental na sociedade e nas instituições de ensino. Nesse contexto, muitas instituições de ensino têm adotado abordagens abrangentes para aprimorar a experiência educacional, incluindo a implementação de equipes multidisciplinares compostas por psicólogos, pedagogos e assistentes sociais. Além disso, têm explorado inovações tecnológicas e metodologias de ensino, todas voltadas para a redução das taxas de retenção e evasão (SUPERBE; SILVA, 2018). Essas ações são uma resposta direta aos desafios enfrentados pelos alunos, como a gravidez na adolescência, a pressão decorrente da baixa renda familiar, que muitas vezes os leva a trabalhar para contribuir com o sustento de suas casas, as dificuldades de aprendizado e a falta de interesse, tanto por parte dos alunos quanto das próprias instituições de ensino (BROCK; SCHWARTZMAN, 2005; ANDRADE, 2016).

Uma das inovações tecnológicas que pode auxiliar os gestores escolares em reduzir a taxa de insucesso dos alunos é a Mineração de Dados Educacionais (MDE). A MDE é uma área em constante crescimento, que se preocupa em explorar os dados de desempenho acadêmico e, com base nas informações coletadas, encontrar as falhas e até mesmo prever evasão ou retenção de alunos. Por consequência, os gestores acadêmicos podem planejar as melhores soluções para a tomada de decisões das instituições de ensino e, por fim, reduzir o número de evasões e retenções (MARTINHO, 2014; COUTO, 2017).

No contexto da evasão escolar, Calixto *et al.* (2017) realizaram um trabalho que consiste na identificação de variáveis relacionadas a este indicador educacional, utilizando os dados do censo escolar no âmbito dos estados de Ceará e Sergipe. As análises se deram por meio

de técnicas de indução de regras e regressão logística.Calixto *et al.* (2017) teve como conclusão que a idade, a etapa e a modalidade de ensino, a existência de laboratórios e localização da escola se destacaram como variáveis influentes na evasão escolar. Ainda, Marques *et al.* (2019) fizeram um estudo bibliográfico com o objetivo de identificar os trabalhos que abordam o problema da evasão escolar utilizando técnicas de mineração de dados. O estudo permitiu identificar que as principais técnicas utilizadas são Redes Neurais Artificiais (RNAs) (HAYKIN, 2001), classificações, equações estruturais, regressão, análise de sobrevivência, análise fatorial, árvores de decisão e algoritmos de agrupamento (NORVIG; RUSSELL, 2013).

Enquanto no âmbito do desempenho escolar, o trabalho de Laisa e Nunes (2015) analisaram a aprovação e a reprovação utilizando base de dados de alunos do ensino médio através do algoritmo J48[1]. Cechinel *et al.* (2015) realizaram investigações sobre a reprovação no cenário acadêmico do Ensino a Distância (EAD). Utiliza-se como atributo as contagens de interações com ambiente virtual e demonstrou que as Redes Bayesianas se mostraram o modelo mais adequado de predição. Alguns outros, como Couto (2017), Martinho (2014) e Superbe e Silva (2018) realizaram predições tanto de evasão quanto de retenção utilizando RNAs e Redes Bayesianas.

Outro trabalho que é importante citar é o de Dharmawan *et al.* (2018) que utilizou dados demográficos, financeiros, de interação social e da personalidade do aluno, para realizar a predição. Pesquisas como essas se mostram importantes pois logo após a matrícula do aluno, ou seja, antes de iniciar suas aulas, e consequentemente, antes do aluno sentir dificuldades nas disciplinas, assim como, perceber as dificuldades para frequentar as aulas, o algoritmo de árvore de decisão e SVM, que obtiveram melhores resultados, já podem realizar a predição do insucesso escolar. Sendo assim, a gestão escolar pode agir e tomar suas decisões, antes mesmo do início das aulas.

No entanto, após realizar uma Revisão Sistemática da Literatura (RSL), que será explicada posteriormente, não foram encontradas pesquisas sobre o desenvolvimento de software de MDE para auxiliar a gestão escolar na predição de evasão e retenção escolar. Sendo assim, este trabalho busca desenvolver um software de fácil uso pelas equipes multidisciplinares da instituição de ensino, para prever os casos de insucesso escolar através de dados coletados na matrícula ou coletados nos sistemas de gestão escolar, utilizando MDE para prever os casos de evasão e retenção escolar. Tendo como hipótese que diferentes bases de dados possuem

[1] Árvore de Decisão baseada em um conjunto de dados rotulados.

diferentes ótimos-algoritmos, e assim, um software pode fazer uso de diversos algoritmos para a mesma base de dados, sem que isso tenha um custo computacional que venha a inviabilizar o software.

1.2 Definição do Problema de Pesquisa e Hipótese

Segundo Gil (2002), algumas regras práticas para formulação de problemas científicos são: a) deve ser estruturada como uma pergunta; b) deve ser a mais específica possível; e c) utilizar terminologias claras com significativo preciso. Assim, pode-se formular o problema científico que permeia este trabalho com a seguinte questão: "Como soluções computacionais podem identificar situações de evasão e retenção escolar?". Além da questão que reflete o problema científico desta pesquisa, é formulada a seguinte hipótese: "Soluções no âmbito da mineração de dados educacionais podem indicar situações de evasão e retenção escolar."

1.3 Objetivos

Neste contexto, este trabalho propõe um software que faz uso de MDE, com a finalidade de fornecer uma predição de evasão e retenção, através de métodos de classificação, para os gestores educacionais. Para entender o estado da arte da literatura, foi conduzido um mapeamento sistemático de literatura sobre os termos: MDE, evasão e retenção escolar. Com isso, o objetivo geral deste trabalho é desenvolver um software para que toda a equipe multidisciplinar das escolas possa, através da MDE, adquirir mais informações para as tomadas de decisões que possam causar impacto na permanência e êxito escolar de seus alunos. Segundo Marconi e Lakatos (2003) os objetivos específicos visam, de um lado, atingir o objetivo geral e, de outro, aplicá-los a situações particulares. Assim, para alcançar o objetivo geral, os seguintes objetivos específicos foram elencados:

a) realizar um levantamento do estado da arte em busca de soluções voltadas para a redução da evasão e retenção escolar;
b) analisar as soluções identificadas no objetivo anterior, destacando suas contribuições científicas e as técnicas que empregam;
c) desenvolver um software que permita aos gestores, com base nos dados de sua própria instituição e alunos, realizar previsões de evasão e retenção escolar;
d) realizar experimentos com o software proposto para avaliar sua eficácia; e

e) conduzir uma análise de usabilidade do software junto ao público-alvo, incluindo professores, gestores e equipes multidisciplinares de instituições de ensino.

O foco principal da pesquisa, assim como o seu produto, é construir um software de fácil uso, para que a equipe escolar possa realizar uso de MDE, mesmo sem ter conhecimento técnico na área, e assim, este trabalho consiga alcançar respostas para as seguintes questões:

1) É possível fazer uso de MDE para auxiliar na permanência e êxito escolar?
2) Quais são os principais problemas para a utilização de técnicas de MDE nas instituições de ensino?

1.4 Principais Resultados e Contribuições

Os principais resultados obtidos e contribuições desta dissertação são:

a) uma RSL com o objetivo de identificar o uso de MDE para ajudar as instituições de ensino a garantirem que o discente conclua seu itinerário formativo;
b) um software de MDE para auxiliar gestores educacionais no combate à evasão e retenção escolar;
c) uma avaliação com duas bases de dados educacionais; e
d) uma avaliação de usabilidade por meio do questionário SUS.

Após a elaboração da RSL e a identificação dos principais algoritmos usados na predição de evasão e retenção escolar, bem como o desenvolvimento do software correspondente, criou-se um aplicativo que permite aos usuários inserir seus dados por meio de arquivos CSV ou XLS e realizar aprendizado de máquina. Durante o processo de aprendizado, o usuário pode especificar quais dados deseja utilizar para previsão por meio da classificação. O software realiza o aprendizado com base em cinco algoritmos distintos: Árvore de Decisão, Floresta Aleatória, KNN, *Multi Layer Perceptron* (MLP) e Naive Bayes. A aprendizagem e os testes são conduzidos com validação cruzada de 10 *folds*, identificando o algoritmo com o melhor desempenho para realizar previsões na segunda base de dados. Outra informação obtida após realizar a RSL é que não foi encontrado nenhum outro trabalho que trate de um software genérico e de fácil uso para profissionais da educação realizarem mineração de dados escolares.

Para avaliar a precisão do software, utilizamos duas bases de dados públicas: uma contendo dados reais de escolas portuguesas e outra com dados fictícios. Em ambos os casos, a precisão foi superior a 97%.

A usabilidade foi avaliada por meio de um questionário SUS, ao qual foram adici-

onadas duas perguntas extras, com o objetivo de identificar possíveis desafios que os usuários possam enfrentar ao utilizar o software proposto. O resultado do questionário resultou em uma pontuação SUS de 91,125, com uma classificação “A”.

1.5 Estrutura da Dissertação

Esta dissertação está organizada em seis capítulos. O Capítulo 1 descreve a introdução, destacando a contextualização, motivação, questões de pesquisa, hipótese, objetivos, metodologia e organização do texto. Já o Capítulo 2 aborda todos os conceitos teóricos necessários para uma melhor compreensão do trabalho. Os aspectos metodológicos utilizados no trabalho e suas respectivas etapas são detalhados no Capítulo 3. O Capítulo 4 apresenta detalhadamente todo o planejamento e o protocolo adotado para a realização da RSL com o intuito de proporcionar uma visão global do que é pesquisado sobre a colaboração da MDE para ajudar as instituições de ensino a garantirem que o discente conclua seu itinerário formativo. Já o Capítulo 5 exibe e detalha o produto educacional proposto, abordagem descrita nesta dissertação, para indicar situações de evasão e retenção escolar por meio da área de conhecimento da Inteligência Artificial, em particular, Mineração de Dados Educacionais. O Capítulo 6 apresenta a avaliação proposta, além de analisar e discutir os resultados da avaliação para fins de validação do produto educacional. Por fim, o Capítulo 7 conclui o trabalho e apresenta os trabalhos futuros que podem dar continuidade ao presente trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo reúne todo o arcabouço teórico e conceitual necessário para um melhor entendimento e compreensão deste trabalho. Portanto, são apresentados conceitos básicos sobre a evasão e retenção escolar, mineração de dados, ciência de dados, mineração de dados educacionais, inteligência artificial e árvore de decisão. Por fim, será apresentada a ferramenta Weka.

2.1 Evasão e Retenção Escolar

O êxito de um estudante em qualquer nível educacional reveste-se de suma importância, não apenas para as instituições de ensino, mas também para o tecido social. Por conseguinte, inúmeras instituições têm buscado métodos variados, que vão desde a constituição de equipes multidisciplinares, compostas por profissionais como psicólogos, pedagogos e assistentes sociais, até a adoção de inovações tecnológicas e abordagens pedagógicas inovadoras. O propósito dessas iniciativas é aprimorar a qualidade da experiência educacional, com o intuito de minimizar as taxas de retenção e evasão (SUPERBE; SILVA, 2018).

Para elucidar o conceito de evasão no contexto do ensino superior, uma comissão especial incumbida de analisar a evasão (MEC, 1997) definiu evasão de cursos de graduação como a interrupção permanente da trajetória acadêmica de um estudante, antes de sua conclusão. Além disso, a comissão estabeleceu uma diferenciação entre diferentes formas de evasão: a) evasão do curso, que engloba situações como abandono (falta de matrícula), desistência oficial, transferência ou mudança de curso e exclusão institucional; b) evasão da instituição, quando o estudante se desvincula da instituição na qual estava matriculado; e c) evasão do sistema, na qual o estudante deixa o ensino superior de forma temporária ou definitiva. A comissão também propôs várias abordagens para calcular a evasão.

No entanto, enfrentar desafios na aquisição e análise dos dados necessários para identificar suas causas e consequências é uma questão crítica relacionada à evasão. Frequentemente, as instituições de ensino possuem os dados necessários para investigar a evasão, como registros de matrículas, desistências, trancamentos de matrícula e taxas de reprovação em turmas. No entanto, esses dados tendem a ser desorganizados, redundantes, desatualizados e, muitas vezes, de difícil acesso ou de leitura complicada.

Dentre os fatores primordiais associados à evasão escolar, destacam-se situações

como a gravidez na adolescência, a influência da baixa renda familiar, que muitas vezes obriga os alunos a trabalhar para contribuir com as despesas domésticas, bem como as dificuldades de aprendizado e o desinteresse, tanto por parte dos próprios estudantes quanto das instituições de ensino (BROCK; SCHWARTZMAN, 2005; ANDRADE, 2016). É importante ressaltar que a problemática da evasão e da retenção manifesta-se em todos os níveis e modalidades de ensino.

A pesquisa de Chiquitto e Baida (2020) revela que tanto a evasão quanto a retenção são problemas existentes no ensino médio técnico no Instituto Federal do Mato Grosso do Sul e em diversas outras instituições. Em seu trabalho, foram utilizados dados da plataforma Nilo Peçanha e do Atlas do Desenvolvimento Humano do Brasil com a intenção de encontrar associações fortes entre a evasão escolar e outras variáveis, a fim de fornecer evidências das causas da evasão escolar. No entanto, concluíram que os motivos da evasão escolar são fenômenos complexos e envolvem diversos fatores agravantes, o que dificulta a análise dos dados individualmente

No entanto, Silva (2018), em sua pesquisa sobre a evasão escolar em cursos de ensino médio noturnos no município de Paranavaí – PR, chegou à conclusão de que alguns fatores levaram à desistência do aluno. São eles:

- o trabalho do aluno, devido à necessidade de renda para auxiliar no sustento familiar;
- o desinteresse do aluno nos estudos, pois eles não têm perspectivas ou não acreditam que o resultado escolar vá impactar na vida do estudante;
- o apoio da família do aluno, pois alguns já possuem filhos e muitas vezes não têm onde deixar os filhos no turno da noite; e
- fatores intraescolares, como a realidade da sala de aula, com metodologias de ensino não apropriadas para alunos que chegam já estando fora da sala de aula há alguns anos e como estão em trabalhos cansativos, chegam para o momento da aula com fadiga do seu trabalho.

No contexto da evasão escolar, é relevante mencionar o estudo realizado por Silva *et al.* (2012). Nessa pesquisa, foram analisadas as taxas de conclusão e evasão no curso de Educação Física da Universidade Federal do Piauí (UFPI) no ano de referência de 2005, revelando taxas de 28,8% para conclusão e 48,5% para evasão. Vale destacar que os outros 22,7% dos estudantes permaneceram retidos, potencialmente contribuindo para taxas mais elevadas de evasão subsequente.

Uma investigação mais recente, conduzida por Coutinho *et al.* (2018), abordou a evasão no curso de graduação da Universidade Federal do Ceará (UFC). Os resultados dessa pesquisa indicam que as taxas de evasão foram de 9,78% no primeiro semestre de 2015 e 12,10%

no segundo semestre do mesmo ano. Esses dados destacam a importância de compreender e abordar a evasão, pois ela pode ter um impacto significativo no sucesso acadêmico dos estudantes e nas instituições de ensino.

No contexto da retenção escolar, este estudo aborda a situação em que um estudante não obtém sucesso em uma disciplina ou módulo específico, resultando na sua retenção e necessidade de retomar o estudo dessa matéria no próximo semestre. É importante ressaltar que a retenção frequentemente está associada a um impacto significativo na evasão escolar. O trabalho conduzido por Lima *et al.* (2019) apresenta dados sobre a retenção na Universidade de Brasília (UnB) no período entre 2002 e 2008. Essa pesquisa revela que, durante esse período, a taxa de retenção variou entre 54,6% e 66,3%. Esses números destacam a relevância desses desafios, tanto a evasão quanto a retenção, e como eles podem afetar as instituições de ensino de forma substancial.

Uma das inovações tecnológicas que pode aumentar a taxa de sucesso dos alunos é a MDE. A MDE é uma área em constante crescimento, que se preocupa em explorar os dados de desempenho acadêmico e, com base nas informações coletadas, encontrar as falhas e até mesmo prever a evasão ou retenção de alunos. Como consequência, os gestores acadêmicos podem planejar as melhores soluções para a tomada de decisões das instituições de ensino e, por fim, reduzir o número de evasões e retenções (MARTINHO, 2014; COUTO, 2017).

2.2 Mineração de Dados

Historicamente, o ser humano tem a necessidade de armazenar informações para serem utilizadas posteriormente. As pinturas rupestres, ou um pouco mais recentemente, a escrita cuneiforme e os hieróglifos, foram utilizados no passado para que as informações fossem transmitidas de forma íntegra. Devido ao armazenamento dessas informações, hoje, depois de mais de dois milênios, conseguimos obter diversas informações sobre a vida e cultura desses povos. Com o passar do tempo e o avanço das tecnologias, como as máquinas de escrever, a imprensa e os computadores digitais, conseguimos armazenar e produzir cada vez mais informações. Com os computadores digitais, ocorreu um aumento no armazenamento e processamento de dados.

A capacidade dos computadores em armazenar grandes quantidades de dados de forma segura, íntegra e confiável, bem como o poder de processamento para manipular esses dados, fez com que fossem desenvolvidos cada vez mais softwares para auxiliar empresas e orga-

nizações. No entanto, os softwares que apenas manipulavam alguns poucos dados numéricos e textuais começaram a se tornar obsoletos, intensificando o desenvolvimento de softwares capazes de lidar com um grande volume de dados, muitas vezes distribuídos em locais geograficamente distintos. Assim, novas tecnologias surgiram para atender a essa demanda de processamento de dados com o objetivo de obter informações mais precisas.

Nessa perspectiva, a imensa quantidade de dados disponíveis por meio de softwares, aplicativos para computadores e *smartphones*, bem como a Internet, criou um desafio para a ciência da computação, dando origem a uma ramificação chamada de ciência de dados.

2.2.1 Ciência de Dados

Para contextualizar a ciência de dados, inicialmente deve-se compreender o conceito sobre o seu alicerce, os dados. Segundo Elmasri e Navathe (2019), dados podem ser definidos como um conjunto de elementos que descrevem um fato, seja ele real ou abstrato, que possuem significado implícito e que podem ser armazenados de maneira persistente. Desse modo, algo mais tangível, como por exemplo, um produto, ou algo mais abstrato, como as informações de uma venda, ambos são considerados dados. Outro exemplo similar no âmbito escolar são os dados de um aluno ou a frequência de uma determinada turma.

Devido a existir um custo para armazenamento, similar aos arquivos de papel que torna necessário comprar mais armários ou alocar um espaço físico para esse ambiente, na maioria das vezes, esses dados digitais são armazenados objetivando economia de espaço. Assim, se faz necessário utilizar processamentos para realizar consultas, associações e combinações diversas, para que esses dados, que inicialmente estavam armazenados separadamente para evitar redundância, possam trazer um novo significado, muito mais relevante e decisivo para que os gestores possam tomar suas decisões.

Na ciência de dados, quando se processa os dados com um determinado objetivo em ter respostas, se obtém as informações (DATE, 2004). Portanto, um mesmo conjunto de dados, ou base de dados, pode ser utilizado para extrair diversas informações bem distintas, dependendo dos atributos que estão sendo analisados, assim como da perspectiva que se deseja alcançar.

Nos primórdios da computação, os custos elevadíssimos eram um dos maiores problemas para sua propagação. Pensar no projeto e construção de equipamentos, assim como o desenvolvimento de aplicações, tinha um alto custo financeiro. A questão sobre o armazenamento de dados não era diferente. Discos rígidos eram tão grandes que tinham que ser transportados por

mais de uma pessoa, ou as fitas magnéticas, além de serem caras, armazenavam poucos volumes de dados. Em 1980, o custo para se armazenar cada gigabyte de dados estava estimado em cerca de cento e noventa e três dólares (TANENBAUM; AUSTIN, 2013), enquanto que, realizando uma rápida consulta em lojas virtuais no início do ano de 2022, o custo caiu para pouco mais de 4 centavos de dólar.

Sendo assim, o armazenamento de dados digitais já não era um problema tão grande quando ocorreu o surgimento da internet, que se transformou no conector que reduziu as distâncias do mundo, e assim, o número de dados aumentou de forma exponencial. Informações que para alguns poderiam ser tratadas como desnecessárias, para outros, se tornaram sua maior fonte de renda. Gigantes empresas como Google ou Facebook ganham dinheiro com as informações deixadas por seus usuários ao navegarem, e essas informações são processadas pelas gigantes e vendidas em forma de propaganda direcionada ao seu público-alvo.

Nos dias de hoje, não existem preocupações com o armazenamento de dados; pelo contrário, quanto mais dados, melhor. Agora, se faz necessário saber como abstrair conhecimento sobre algo em um universo de informações. Assim, novos conceitos sobre dados foram surgindo, como, por exemplo, *Business Intelligence* (BI), *Data Warehouse* (DW) e *Online Analytical Processing* (OLAP), assim como *Big Data* e *Data Mining*, passaram a ser palavras do dia a dia da ciência de dados. Ao mesmo tempo que foram surgindo áreas específicas, como, por exemplo, a MDE.

2.2.2 Mineração de Dados Educacionais

Assim como a internet e outros setores comerciais, o setor escolar tem uma grande capacidade de gerar dados. Além do grande volume de informações no momento da matrícula do aluno, diariamente são gerados mais dados sobre ele, desde a frequência escolar até informações sobre atividades, provas, testes, monitorias, atividade de educação física, projetos de extensão, entre outros. Isso ocorre não apenas nas escolas tradicionais; no caso de cursos online, o número e o volume de dados são ainda maiores. Utilizar esses dados escolares para inferir conhecimento é o cerne da Mineração de Dados Educacionais (MARTINHO, 2014) (COUTO, 2017).

Devido ao insucesso escolar ser um problema que afeta tanto instituições de ensino privadas quanto públicas, pois diminui o lucro ou a receita e a visibilidade social, é ainda mais prejudicial para o aluno. Além de perder a segurança que o ambiente escolar proporciona, o aluno terá grandes chances de não conseguir se qualificar profissionalmente (SUPERBE; SILVA,

2018).

Dessa forma, a MDE tem sido utilizada com o objetivo de identificar possíveis alunos que possam, infelizmente, alcançar o insucesso escolar. Caso esse aluno seja identificado previamente, a instituição de ensino, juntamente com sua equipe responsável, pode tomar medidas para reduzir esse problema (CALIXTO *et al.*, 2017).

Os pesquisadores normalmente utilizam dados escolares, o que significa que o aluno já deve estar estudando para que esses dados estejam disponíveis. Isso implica que o aluno já está sofrendo e sendo penalizado, seja por questões financeiras, pessoais, ou como mencionado por Calixto *et al.* (2017), por falta de bagagem escolar. Isso dificulta a gestão em tomar decisões como concessão de bolsas, apoio psicológico ou abertura de turmas de nivelamento antes de os alunos terem tido impactos negativos no ensino. As pesquisas fazem uso de diversos tipos diferentes de dados, incluindo dados pessoais, como demonstrado por Pertiwi *et al.* (2017) e Pereira e Zambrano (2017), que além de dados pessoais e educacionais também utilizaram o algoritmo de árvore de decisão, assim como Silveira *et al.* (2019), que utilizou o algoritmo de Floresta Aleatória.

No entanto, o trabalho de Dharmawan *et al.* (2018) quebra esse paradigma de esperar que o aluno comece a estudar para inferir algum conhecimento a partir dos dados gerados durante o curso. Nesse caso, os autores não utilizaram dados escolares, apenas dados socioeconômicos para realizar uma predição de insucesso escolar. Eles conduziram experimentos com três algoritmos diferentes: árvores de decisão, SVM e KNN.

2.3 Inteligência Artificial

O computador surgiu com o propósito de realizar cálculos balísticos, uma atividade repetitiva e demorada para os seres humanos. Com o avanço e o aumento do poder computacional, os computadores tornaram-se capazes de resolver problemas complexos em um curto período de tempo. No entanto, ainda existem problemas cuja solução é tão complexa que mesmo um supercomputador levaria anos para resolvê-los. Esses problemas são conhecidos como problemas NP-complexos ou NP-difíceis (CORMEN, 2013). Portanto, algoritmos de inteligência artificial foram desenvolvidos para tentar encontrar soluções viáveis em um tempo consideravelmente curto, embora não haja garantia de que essas soluções sejam as melhores.

Os algoritmos de inteligência artificial são frequentemente divididos em duas categorias: os supervisionados e os não supervisionados. Os algoritmos supervisionados de alguma

forma contam com a orientação de um supervisor durante o processo de aprendizagem, permitindo que o algoritmo faça ajustes automáticos ou seja punido com base nos dados. Por exemplo, se o objetivo for ensinar o algoritmo a diferenciar um humano de um cachorro, é necessário fornecer exemplos, ou seja, instâncias de cachorros e humanos para que o algoritmo aprenda.

A outra categoria engloba os algoritmos não supervisionados, nos quais o algoritmo não possui informações prévias sobre a classificação das instâncias. Ele agrupa as instâncias com base em características que identifica nas mesmas. Essa categoria também é conhecida como agrupamento (NORVIG; RUSSELL, 2013).

Durante a pesquisa para este trabalho, uma RSL foi realizada, e uma das respostas esperadas era identificar quais algoritmos foram mais utilizados nos últimos anos em problemas de mineração de dados educacionais. Portanto, a seguir, serão apresentados alguns desses algoritmos.

2.3.1 Árvore de Decisão

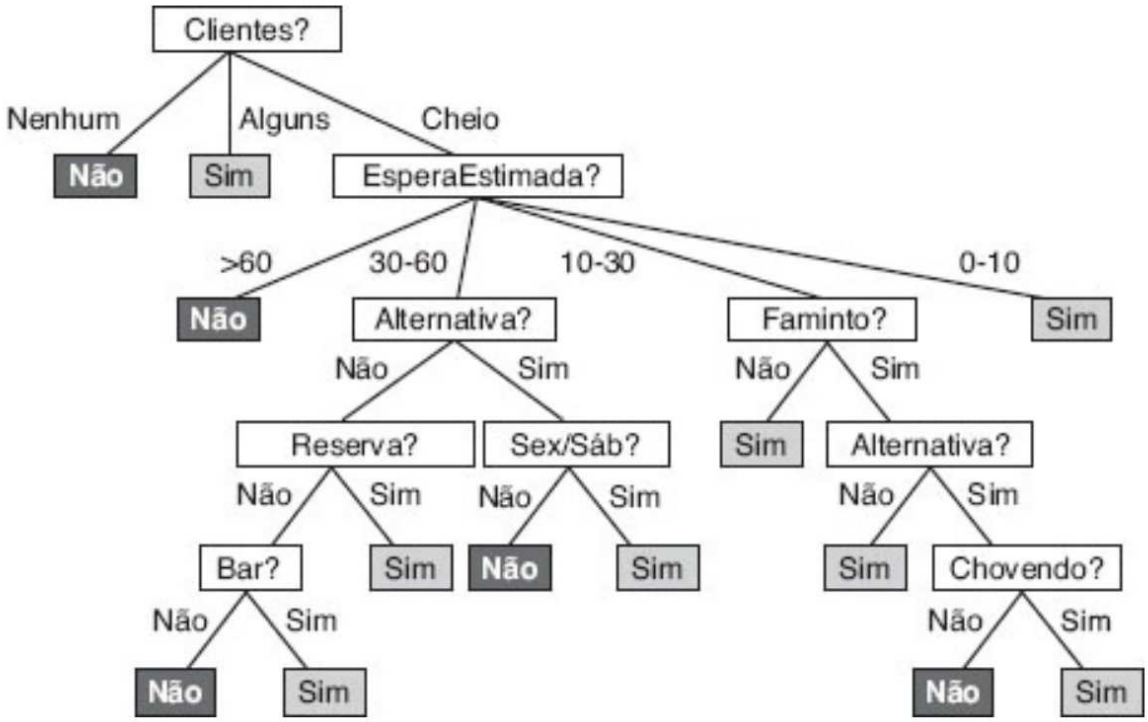
As árvores de decisão, segundo Norvig e Russell (2013), são uma das formas mais simples e bem-sucedidas de aprendizado de máquina. O funcionamento de uma árvore de decisão ocorre a partir de uma entrada com um vetor ou conjunto de valores e produz uma única saída, que representa a decisão tomada pelo algoritmo.

Uma árvore de decisão obtém a sua decisão após a execução de uma sequência de testes. Cada nó interno da árvore corresponde a uma verificação do valor de um determinado atributo informado na entrada. Considere-se que os nós internos sejam representados por A_i, e as ramificações dos nós são classificadas com os valores possíveis do atributo, $A_i = valor_{ik}$. Enquanto o algoritmo vai realizando testes e percorrendo o caminho da árvore até chegar ao nó folha, sendo este o valor a ser retornado pela função, ou seja, o valor “decisão”.

A representação gráfica de um árvores de decisão é bem natural e facilita muito o seu entendimento. Sendo assim, a Figura 1 representa uma árvore de decisão com a intenção de decidir se ocorrerá uma espera por uma mesa em um restaurante. Importante destacar que esta árvore possui 9 atributos, que na Figura 1, estão representados por retângulos com fundo na cor branca. Dependendo do valor para cada atributo o algoritmo irá percorrendo a árvore até chegar em um nó folha, que estão representados por retângulos nas cores cinza claro, para decisões positivas, e cinza escuro, para decisões negativas. É importante, para este trabalho, esclarecer que árvores de decisões podem apresentar mais de dois valores de decisões, e não apenas dois

tipos como foi mostrado neste exemplo.

Figura 1 – Exemplo de árvore de decisão.



Fonte: Norvig e Russell (2013).

No entanto, o grande problema é definir qual a melhor árvore para solucionar o problema. Visto que obtendo os atributos pode-se ser desenhado diversas árvores diferentes, sendo este um problema exponencial, exemplificando com um problema de 10 atributos com valores booleanos, teríamos uma tabela-verdade com 2^{10} linhas, ou seja 1.024 linhas e se considerarmos apenas dois valores para a decisão da árvore, teríamos 2^{1024} diferentes árvores. Sendo assim, fazemos uso de indução para obter, dentre tantas possíveis, a árvore que pode alcançar os melhores resultados.

Uma das formas de realizar essa indução é através de um algoritmo que adota uma estratégia gulosa de dividir para conquistar: sempre testar o atributo mais importante em primeiro lugar. Após encontrar o melhor atributo, o teste divide o problema em subproblemas menores que podem então ser resolvidos de forma recursiva. O atributo definido como mais importante é aquele que possui a maior diferença na classificação dos atributos, ou seja, a maior entropia. Isso permite obter uma classificação correta com um número menor de testes, resultando em uma árvore com caminhos mais curtos e menos profundidade.

2.3.2 Floresta Aleatória

A Floresta Aleatória, também conhecida como *Random Forest*, é um algoritmo de aprendizado de máquina supervisionado amplamente utilizado devido à sua flexibilidade e facilidade de uso. Este algoritmo opera através da criação de uma floresta de decisões de maneira aleatória, o que o torna altamente eficaz e robusto (HASTIE *et al.*, 2009).

A construção de uma Floresta Aleatória envolve a criação de um conjunto de árvores de decisão, utilizando uma técnica conhecida como *Bagging*, que é uma combinação das palavras "*bootstrap aggregating*". No contexto estatístico, o *bootstrap* é um método de amostragem que envolve a seleção de objetos com reposição, mantendo o mesmo número de objetos que o conjunto de dados original. Geralmente, cerca de dois terços dos objetos são escolhidos aleatoriamente para criar uma amostra maior, que é usada para treinar as árvores de decisão. A terça parte restante é reservada para fins de teste.

Ao final do processo, as árvores de decisão individuais são combinadas por meio de uma votação simples, utilizando o erro médio de todas as amostras. Essa abordagem de "*bootstrap aggregating*" tem um impacto significativo na redução do erro médio quadrático e na diminuição da variância do classificador treinado. Isso resulta em um modelo que não varia muito em diferentes amostras de dados, tornando-o menos suscetível ao *overfitting*, que é o ajuste excessivo aos dados de treinamento (BASTOS *et al.*, 2013).

A capacidade da Floresta Aleatória de mitigar o *overfitting*, ao mesmo tempo em que mantém uma alta capacidade de generalização, é uma das razões pelas quais é amplamente adotada em uma variedade de problemas de aprendizado de máquina. Sua versatilidade o torna adequado para tarefas de classificação, regressão e até mesmo detecção de anomalias. Compreender a dinâmica por trás da criação dessa "floresta" de árvores de decisão é essencial para aproveitar todo o potencial do algoritmo *Random Forest* em suas aplicações práticas.

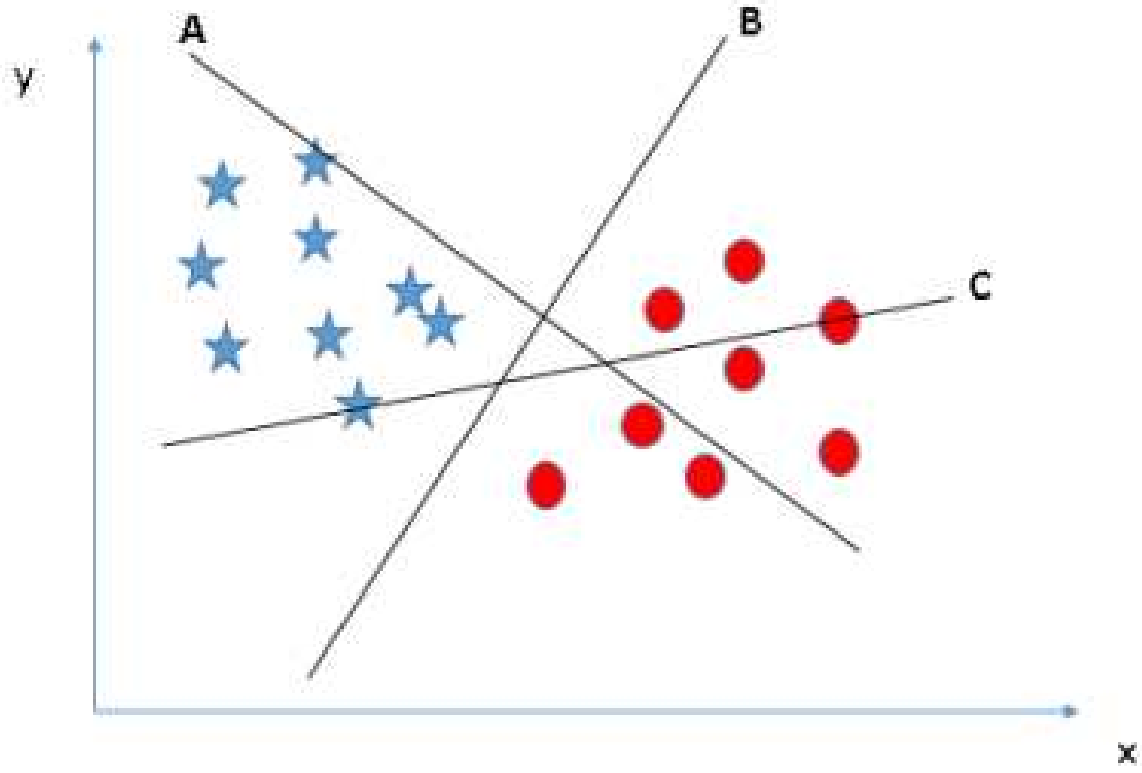
2.3.3 Support Vector Machine

O algoritmo SVM é uma técnica de aprendizado de máquina amplamente adotada, conhecida por sua eficácia em tarefas de classificação e regressão. Este algoritmo supervisionado tem como principal objetivo treinar modelos capazes de classificar dados de acordo com suas características. Uma característica distintiva do SVM é a representação das instâncias como pontos em um espaço n-dimensional, onde "n" corresponde ao número de atributos (NORVIG;

RUSSELL, 2013). Cada valor de atributo é mapeado para uma coordenada no espaço, permitindo ao SVM encontrar o hiperplano ótimo para separar as diferentes classes.

A Figura 2 ilustra o conceito de hiperplanos no contexto do SVM. Em um problema de classificação com duas classes, representadas por círculos e estrelas, são apresentados três hiperplanos: A, B e C. Notavelmente, apenas o hiperplano B é capaz de realizar uma separação eficaz entre as classes. Isso destaca a capacidade do SVM em encontrar o hiperplano ideal, aquele que maximiza a margem de separação entre as classes.

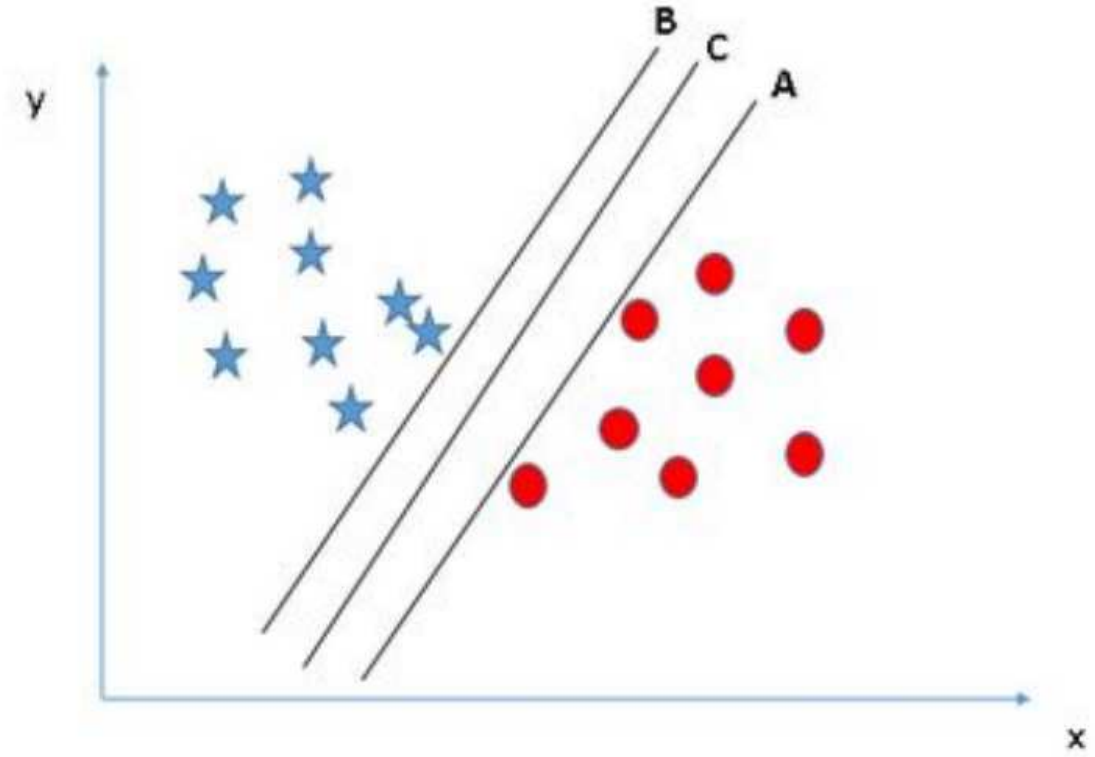
Figura 2 – Demonstração de possíveis hiperplanos.



Fonte: Google Imagens. Disponível em: <https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcSXOB4TQ4_aRHE6BnReSsaITGu_2-HKEQIVeg&usqp=CAU>. Acesso em: 14 de março de 2022.

Durante o processo de treinamento do SVM, vários hiperplanos candidatos são gerados para separar as classes, conforme ilustrado na Figura 3. A principal estratégia do SVM é selecionar o hiperplano que está equidistante das instâncias das diferentes classes, proporcionando a máxima margem de separação. No exemplo da Figura 3, o hiperplano C é escolhido por manter uma distância igual das instâncias de ambas as classes. Essa decisão estratégica contribui para a robustez e eficácia do SVM em problemas de classificação, mesmo em espaços de alta dimensionalidade.

Figura 3 – Margem dos hiperplanos em um SVM.



Fonte: Google Imagens. Disponível em: <https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_4.png>. Acesso em: 14 de março de 2022.

Além de sua notável capacidade de classificação, o SVM oferece flexibilidade ao permitir o uso de diferentes funções de *kernel*, como o kernel linear, polinomial e radial, para lidar com dados que podem não ser linearmente separáveis. Essa capacidade de adaptação a diferentes cenários torna o SVM um dos algoritmos mais versáteis em aprendizado de máquina (NORVIG; RUSSELL, 2013).

Algumas vantagens do uso do SVM são que ele apresenta bons resultados em problemas com margem de separação clara e é eficaz quando o número de dimensões é maior do que o número de instâncias.

2.3.4 K-Nearest Neighbors

O algoritmo KNN é uma técnica de aprendizado de máquina que se destaca por sua abordagem flexível e adaptativa, onde a estrutura do modelo é determinada diretamente pelo conjunto de dados em uso. Diferentemente de muitos outros algoritmos, o KNN não requer um processo de treinamento prévio com os dados. Em vez disso, ele mantém os exemplos do conjunto de treinamento no próprio modelo, tornando-o um algoritmo do tipo "*lazy*" ou

preguiçoso (GUO *et al.*, 2003).

A característica fundamental do KNN é a variável "*k*", que representa o número de vizinhos mais próximos a serem considerados durante a classificação ou regressão de um ponto de dados desconhecido. A escolha adequada de "*k*" desempenha um papel crítico no desempenho do algoritmo, pois um valor baixo de "*k*", como 1, resulta em uma classificação sensível ao ruído nos dados, enquanto um valor muito alto de "*k*" pode suavizar as fronteiras de decisão, tornando o modelo menos sensível a padrões sutis.

Para realizar a classificação, o KNN utiliza a distância entre pontos como métrica de proximidade. Isso envolve a representação dos dados como pontos em um espaço multidimensional, onde cada característica corresponde a uma coordenada. O algoritmo calcula a distância entre o ponto a ser classificado e todos os pontos de treinamento, selecionando os "*k*" pontos mais próximos. A classe ou valor alvo mais comum entre esses vizinhos é atribuído ao ponto a ser classificado.

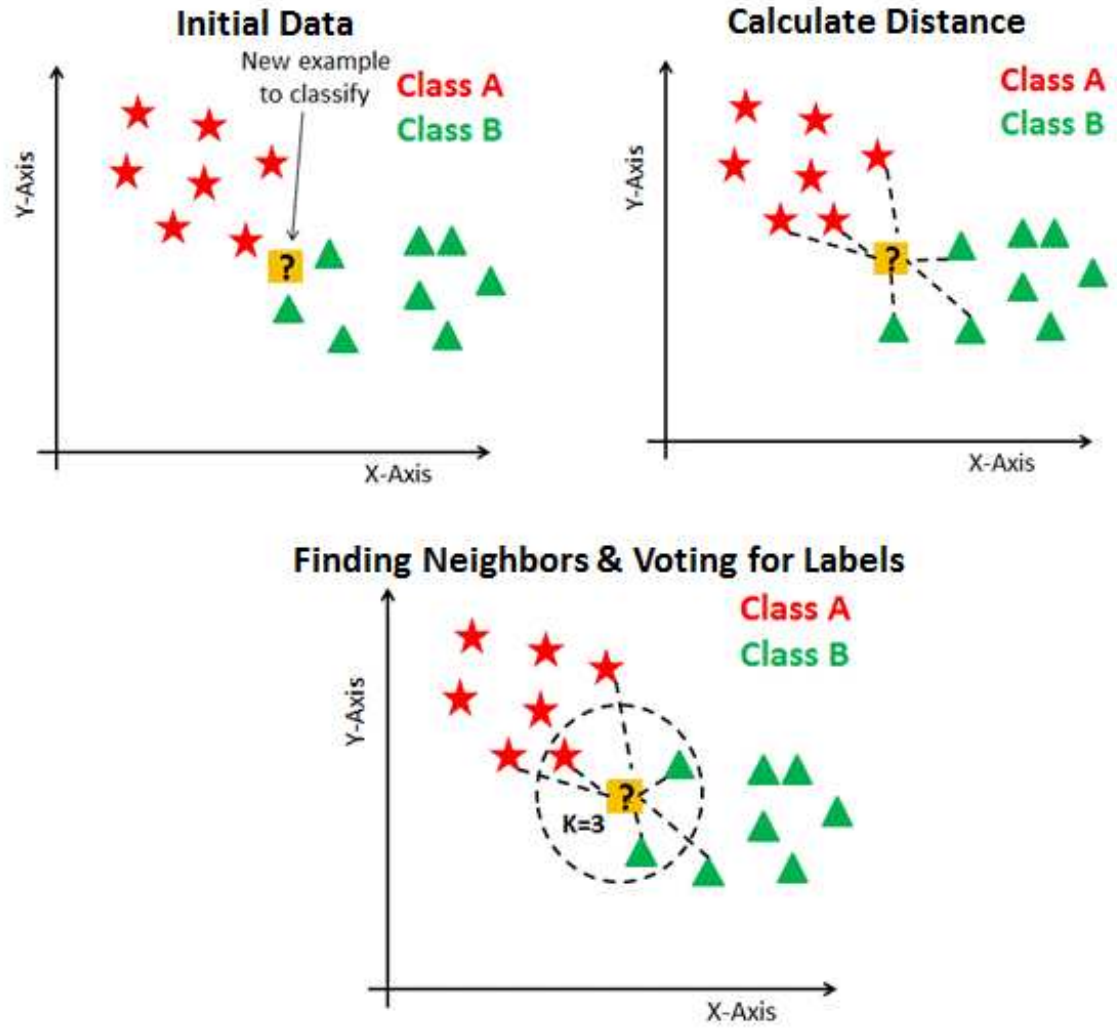
A Figura 4 ilustra esse processo, onde um objeto representado por um quadrado amarelo deve ser classificado. Se "*k*" for definido como 1, o algoritmo identificará o vizinho mais próximo, classificando o objeto como uma estrela. No entanto, se "*k*" for definido como 3, ele considerará o vizinho mais próximo (uma estrela) e os dois vizinhos seguintes (triângulos), classificando o objeto como um triângulo com base na maioria dos vizinhos.

Além desses aspectos essenciais, é importante destacar que o KNN pode ser afetado pela escolha da função de distância, que pode variar de acordo com o problema. Também é fundamental discutir estratégias de normalização de dados para garantir que todas as características contribuam igualmente para o cálculo da distância. Por fim, vale mencionar que o KNN é aplicável em uma variedade de contextos, desde classificação de textos até sistemas de recomendação e detecção de anomalias, tornando-o uma ferramenta versátil no campo da aprendizagem de máquina.

2.3.5 Multi Layer Perceptrons

O cérebro humano é constituído, principalmente, por dois tipos de células: as glias e os neurônios. As glias estão presentes em cerca de dez vezes o número de neurônios, e sua principal função é dar sustentação ao cérebro. Já os neurônios são basicamente as unidades de processamento dos sinais e estímulos que recebem. A forma como os neurônios se comportam serviu de inspiração para a elaboração do algoritmo de redes neurais artificiais. O neurônio é

Figura 4 – Classificação utilizando algoritmo KNN.



Fonte: Google Imagens. Disponível em: <http://res.cloudinary.com/dyd911kmh/image/upload/f_auto,q_auto:best/v1531424125/KNN_final1_ibdm8a.png>. Acesso em: 14 de março de 2022.

subdividido em axônio, dendritos e corpo celular. O axônio é o prolongamento do neurônio por onde os impulsos nervosos são levados a outro neurônio ou outro tipo de célula, enquanto que os dendritos são as ramificações presentes no corpo celular que recebem os impulsos nervosos. Através dos neurônios e de suas interconexões ocorrem as sinapses, o que possibilita a ocorrência de processos como o pensamento, a emoção, a cognição, o movimento, entre outros (HAYKIN, 2001).

As Rede Neural Artificials (RNAs), constituídas por neurônios artificiais seguindo o conceito dos neurônios naturais, são modelos estatístico-matemáticos. As RNAs são modelos adaptáveis, assim como os neurônios naturais. Essa adaptação ocorre através da variação de alguns parâmetros de controle nos neurônios artificiais, permitindo que as RNAs realizem aprendizagem de máquina e, com isso, realizem classificação ou reconhecimento de padrões. Segundo

Haykin (2001) e Kovács (2002), os modelos de RNAs surgiram junto com os computadores na década de 1940, com o propósito de resolver problemas extremamente complexos.

A ideia original das RNAs era ser o mais semelhante possível ao funcionamento dos sistemas neurológicos humanos. No entanto, mesmo com o avanço da tecnologia e, consequentemente, do poder computacional, essa ideia original foi deixada de lado devido ao aumento do conhecimento sobre a fisiologia complexa do neurônio biológico. No entanto, as pesquisas desenvolvidas sobre as redes neurais artificiais mostraram-se muito importantes na resolução de diversos tipos de problemas. Isso ocorre devido a uma característica importante das RNAs, que é a tolerância a falhas. Portanto, mesmo quando a informação é parcial, a rede pode escolher um padrão ou informação de saída que a rede considera o mais próximo do desejado (AGUIAR, 2000).

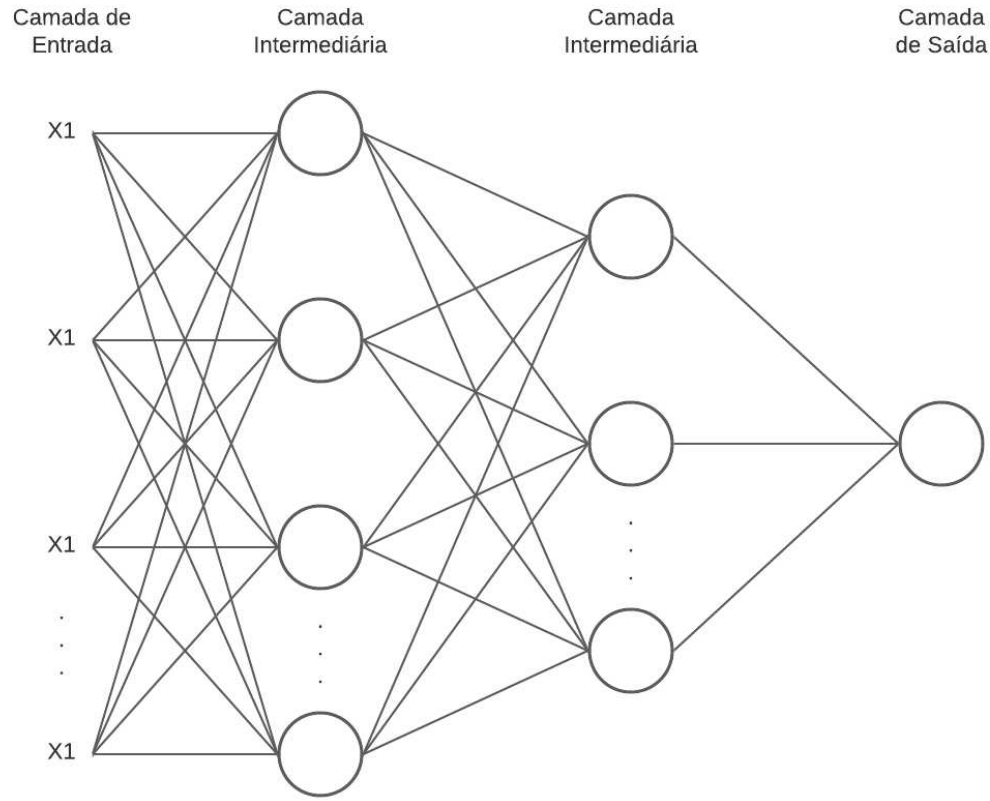
As RNAs são constituídas por unidades ou elementos de processamento, que por sua vez são organizados em pelo menos três camadas. A primeira camada, chamada de camada de entrada, recebe os dados ou padrões de entrada. A camada de saída fornece as respostas da rede ou os padrões de saída. Entre elas, podem ser incluídas uma ou mais camadas chamadas de intermediárias, ocultas ou também internas, conforme podemos visualizar na Figura 5. Ainda sobre a Figura 5, é possível observar que nas RNAs, o número de atributos de entrada pode variar, assim como o número de camadas intermediárias, que nesta figura é de duas camadas.

No sistema nervoso humano, as transmissões sinápticas são o processo em que a informação gerada ou processada é transmitida de um neurônio para o outro ou até a célula efetora. Sendo assim, na RNA, funciona de forma similar, na qual cada elemento de processamento é associado com um peso sináptico. Esse peso é o que faz uma analogia com a realidade das sinapses nas redes neurais biológicas. Portanto, a sequência da sinapse ocorrerá e seguirá adiante caso o valor seja superior ao peso do elemento de processamento. A aprendizagem da rede neural ocorre com as mudanças e adequações dos pesos sinápticos de ligação para cada problema.

Uma rede neural que tem o seu fluxo apenas em um sentido, da esquerda para direita em uma RNA, conforme a Figura 5, recebe o nome de rede direta ou *feedforward*, ou ainda, de acordo com Haykin (2001), de rede alimentada para frente. Outra arquitetura muito utilizada é a MLP.

A MLP ou rede *perceptron* de múltiplas camadas, conforme Haykin (2001). A vantagem desta arquitetura é a regra de aprendizado delta generalizada, também conhecida como *backpropagation* ou regra de retro propagação do erro, que, segundo Abdi *et al.* (1999), é uma

Figura 5 – Modelo de uma Rede Neural Artificial.



Fonte: Elaborado pelo autor.

das mais poderosas regras de aprendizado das redes neurais. A MLP faz uso do algoritmo de treinamento *backpropagation*, desenvolvido por Rumelhart e McClelland (1987), que, graças ao algoritmo, foi possível treinar eficientemente redes com camadas intermediárias, resultando no modelo de Redes Neurais Artificiais mais utilizado atualmente.

Sendo assim, cada camada da MLP tem uma função específica. A camada de saída recebe os estímulos da camada intermediária, e assim, é inferido o padrão que resultará na resposta do problema. Enquanto que as camadas intermediárias funcionam como extratoras de características, os pesos dos neurônios nesta camada são uma codificação das características informadas na camada de entrada, e assim, a rede consegue criar sua própria representação com maiores detalhes sobre o problema. Portanto, um fator crucial em uma RNA é encontrar o número de neurônios para a camada de entrada e intermediária na qual a rede consiga inferir a solução do problema.

Para realizar o treinamento de uma rede com o algoritmo *backpropagation*, deve-se seguir dois passos. No primeiro passo, o algoritmo informa um padrão para a camada de entrada da rede, permitindo que a atividade resultante flua através da rede, seguindo o fluxo das camadas,

até que a resposta seja produzida pela camada de saída. No segundo passo, o resultado obtido na camada de saída é comparado com a saída desejada para essa instância do padrão informado no primeiro passo. Se o valor obtido na RNA não for o mesmo do padrão, o erro é calculado. Sendo assim, o erro é propagado no sentido inverso, ou seja, da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo ajustados de acordo com o erro, sendo este retro-propagado, explicando assim a origem do seu nome. Devido a isso, o treinamento das redes MLP com *backpropagation* pode levar um tempo consideravelmente longo.

Por fim, após a conclusão do treinamento da rede e a taxa de erro estiver em um nível satisfatório, a rede neural poderá ser utilizada como uma ferramenta para a classificação de novas instâncias. Neste caso, a rede não utilizará o *backpropagation*, visto que seu uso é apenas na aprendizagem de máquina, sendo utilizado apenas o modo progressivo. Ou seja, novas instâncias serão informadas na camada de entrada, processadas nas camadas intermediárias e, por fim, os resultados poderão ser obtidos na camada de saída, que será a interpretação da rede para aquela instância.

Como dito anteriormente, uma MLP possui a desvantagem de ter um tempo de treinamento consideravelmente maior que outros algoritmos, sendo esse tempo proporcional à quantidade de neurônios e camadas. No entanto, caso o algoritmo possua um número suficiente de camadas e neurônios, ele pode solucionar problemas que não sejam separados linearmente, superando algoritmos como o *Naive Bayes*.

2.3.5.1 Deep Learning

Uma importante técnica baseada em redes neurais que vem a cada dia ganha mais destaque é a *Deep Learning* e mesmo não sendo utilizada neste trabalho, vale a pena citar pois pode ser utilziada em trabalhos futuros. Ao contrário do MLP, o *Deep Learning* representa uma abordagem mais sofisticada para a resolução de problemas complexos de aprendizado de máquina. O conceito-chave que distingue o *Deep Learning* é a presença de redes neurais profundas, caracterizadas por uma arquitetura com várias camadas. Essas redes profundas permitem a aprendizagem de representações hierárquicas complexas dos dados, capacitando-as a discernir padrões e características abstratas (LECUN *et al.*, 2015).

As aplicações do *Deep Learning* são vastas e incluem domínios como reconhecimento de imagem, processamento de linguagem natural, reconhecimento de voz e jogos. Em

comparação com o MLP, o *Deep Learning* destaca-se em tarefas que exigem uma compreensão mais profunda e contextual dos dados. Pode citar como exemplo as Redes neurais convolucionais que são amplamente empregadas para tarefas de visão computacional, como também as redes neurais recorrentes são eficazes para lidar com dados sequenciais (GOODFELLOW *et al.*, 2016).

O treinamento em *Deep Learning*, embora compartilhe semelhanças com o MLP, muitas vezes envolve arquiteturas mais complexas e pode demandar recursos computacionais substanciais. Em resumo, enquanto o MLP representa um passo inicial no campo das redes neurais, o *Deep Learning* expande consideravelmente essa capacidade, proporcionando soluções mais avançadas para problemas complexos de aprendizado de máquina.

2.3.6 Naive Bayes

O *Naive Bayes* é um classificador probabilístico baseado no teorema de Bayes, criado por Thomas Bayes, que realiza uma suposição de independência entre os preditores. Em outras palavras, o algoritmo de classificação *Naive Bayes* assume que as características particulares de um determinado objeto ou classe não estão relacionadas com a presença de qualquer outro atributo. Sendo assim, o classificador considera que um fruto pode ser considerado como um limão se possuir a cor verde, formato redondo e tiver entre 4 e 7 centímetros de diâmetro. Mesmo que os atributos dependam uns dos outros ou mesmo da existência de outros atributos não utilizados, todas essas características influenciam de forma independente na probabilidade de que este fruto seja um limão, e é por isso que o algoritmo é conhecido como "*Naive*", ou ingênuo em português (BERRAR, 2019).

O teorema de Bayes, princípio deste algoritmo, é dado pela equação:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

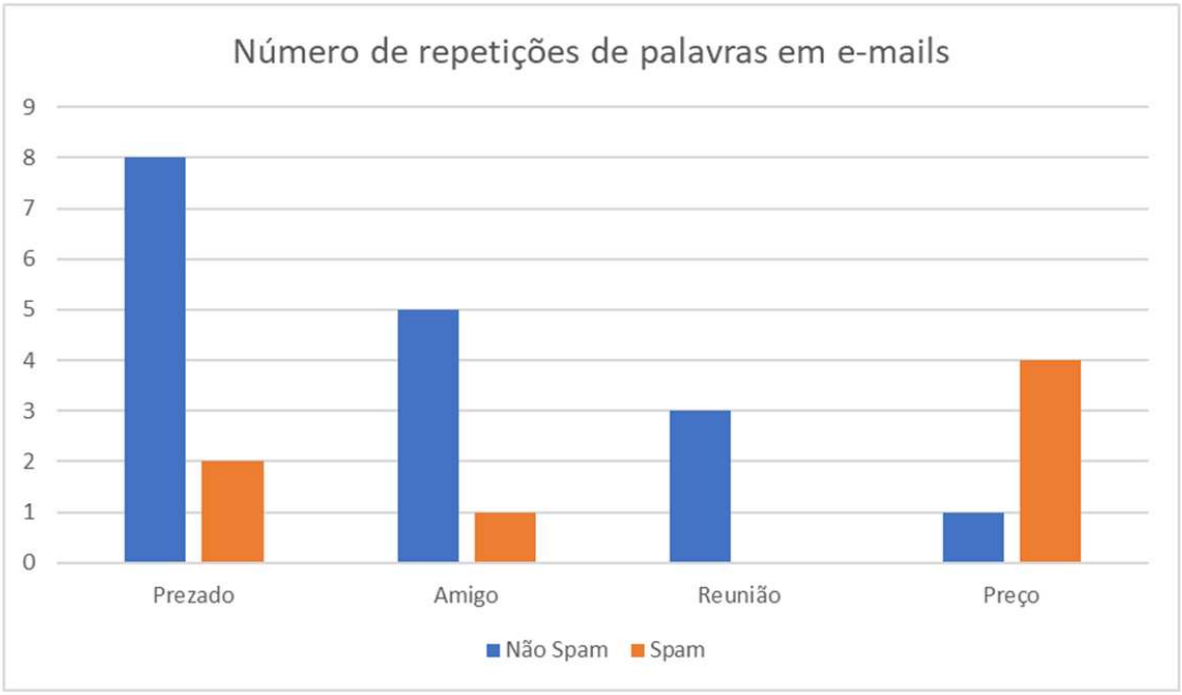
Sendo que:

- P(A|B) é uma probabilidade condicional, que é a probabilidade do evento A ocorrer dado que B seja verdade;
- P(B|A) também é uma probabilidade condicional, sendo que, é a probabilidade do evento B ocorrer quando A for verdade; e
- P(A) e P(B) são as probabilidades de A ou B ocorrerem sem quaisquer condições.

Este algoritmo de classificação pode ser utilizado para diversos fins, sendo um

deles a análise textual, como por exemplo, análise de *spam* em e-mails. O algoritmo deverá inicialmente identificar o número de repetições das palavras contidas no e-mail, o que é melhor identificado humanamente através de um gráfico. Para melhor compreensão do algoritmo, iremos considerar que nos e-mails utilizados para aprendizagem de máquina foram identificadas as palavras contidas no gráfico da Figura 6.

Figura 6 – Número de repetições de palavras em e-mails.



Fonte: Elaborado pelo autor.

Sendo assim, pode-se considerar que a probabilidade da palavra “prezado” em e-mails que não são spam é de $P(\text{prezado}|\text{nao\_spam}) = 8/17$, ou seja, aproximadamente 0,47. Sendo 8 o número de repetições da palavra “prezado” nos e-mails que não foram considerados *spam* e 17 o número total de palavras nestes e-mails nao_spam. Seguindo a mesma lógica, a probabilidade da palavra “prezado” estar contida em um e-mail considerado spam é de $P(\text{prezado}|\text{spam}) = 2/7$, ou seja, aproximadamente 0,29. O Quadro 1 apresenta o resultado do cálculo das probabilidades das demais palavras da Figura 6:

O próximo passo do algoritmo é verificar a probabilidade de os e-mails serem *spam* ou não. Neste exemplo, a base conta com 12 e-mails, sendo 4 spams e 8 e-mails não_spam. Tendo a probabilidade de $P(\text{spam}) = 0,33$ e $P(\text{nao\_spam}) = 0,67$, sendo esta última chamada de prioridade a priori. O próximo passo é definir a probabilidade de um e-mail não_spam conter as palavras “prezado amigo”. Sendo assim:

Tabela 1 – Probabilidade das palavras contidas no e-mail.

Não Spam	Spam
P(prezado | nao_spam) = 0,47	P(prezado | spam) = 0,29
P(amigo | nao_spam) = 0,29	P(amigo | spam) = 0,14
P(reuniao | nao_spam) = 0,18	P(reuniao | spam) = 0
P(preco | nao_spam) = 0,06 = 0,18	P(preco | spam) = 0,57

Fonte: Elaborado pelo autor.

$$P(nao\_spam|prezado\_amigo) = P(nao\_spam) * P(prezado|nao\_spam) * P(amigo|nao\_spam) \quad (2.2)$$
$$P(nao\_spam|prezado\_amigo) = 0,67 * 0,47 * 0,29 = 0,09$$

Enquanto que para a probabilidade de o e-mail spam conter o texto prezado amigo é de:

$$P(spam|prezado\_amigo) = P(spam) * P(prezado|spam) * P(amigo|spam) \quad (2.3)$$
$$P(spam|prezado\_amigo) = 0,33 * 0,29 * 0,14 = 0,01$$

Portanto, o algoritmo de classificação irá classificar o novo e-mail como um nao_spam, devido à probabilidade de 0,09 ser maior.

Assim como o KNN, o *Naive Bayes* e outros algoritmos baseados na teoria de Bayes obtêm resultados melhores do que as árvores de decisão quando se trata de ocorrências raras, como por exemplo em diagnósticos de doenças raras (EHSANI-MOGHADDAM *et al.*, 2018).

2.3.7 Weka

A ferramenta Weka é um pacote de software desenvolvido na linguagem de programação Java pela Universidade de Waikato, Nova Zelândia. Tendo sua licença de uso *General Public License* (GPL), ou seja, o seu código-fonte é livre para ser estudado e alterado. Ou seja, é amplamente estudado e atualizado por pessoas em diferentes locais do planeta. Esta ferramenta tem como objetivo agregar algoritmos de inteligência artificial dos mais diferentes paradigmas e abordagens com o intuito de auxiliar no estudo de aprendizagem de máquina (PERTIWI *et al.*, 2017).

Além da ferramenta já disponibilizar diversos algoritmos de inteligência artificial, como arvores de decisão, KNN, SVM, redes neurais, florestas aleatórias, *Naive Bayes*, regres-

sões lineares e logísticas entre outros, também é possível realizar o download de *plugins* com ferramentas desenvolvidas pelos colaboradores do projeto (WITTEN *et al.*, 2016).

Além de já disponibilizar diversos algoritmos de inteligência artificial, como árvores de decisão, KNN, SVM, redes neurais, florestas aleatórias, *Naive Bayes*, regressões lineares e logísticas, entre outros, também é possível realizar o download de *plugins* com ferramentas desenvolvidas pelos colaboradores do projeto (WITTEN *et al.*, 2016).

Diversos trabalhos sobre evasão e retenção utilizaram a ferramenta *Weka*, como o trabalho de Manhães *et al.* (2012), Santos e Fernandes (2022) e Araujo (2018). Além disso, o fato de a ferramenta estar disponibilizada em arquivo de distribuição .jar e possuir uma ampla documentação, facilita a integração com outros softwares. Sendo assim, reitera-se que o desenvolvedor possui a disponibilidade de todos os algoritmos de aprendizagem de máquina que o *Weka* dispõe, aliado ao frequente uso da ferramenta em trabalhos semelhantes.

2.4 Conclusão do Capítulo

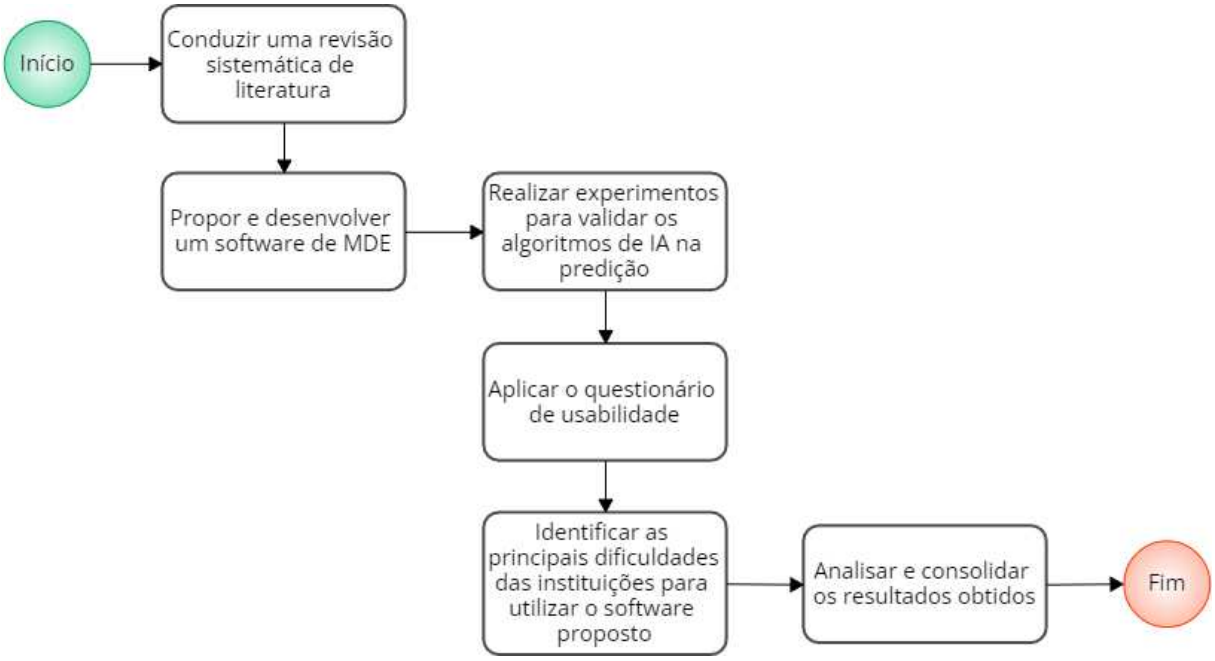
Diante desse cenário, torna-se evidente que a evasão e retenção escolar representam desafios de magnitude considerável para nossa sociedade, demandando uma abordagem eficaz para atenuar essas problemáticas. A utilização da MDE surge como uma oportunidade promissora para abordar essas questões de maneira mais proativa e eficiente. Através da aplicação de algoritmos de aprendizagem de máquina, é possível realizar análises preditivas utilizando técnicas de classificação, o que proporciona uma visão aprofundada sobre padrões de comportamento e fatores de risco associados à evasão e retenção escolar.

Essa capacidade preditiva não apenas oferece uma compreensão mais profunda dos desafios educacionais enfrentados por alunos, mas também capacita a equipe multiprofissional das instituições de ensino a tomar medidas preventivas e personalizadas. Ao antecipar potenciais casos de evasão ou retenção, a equipe pode implementar estratégias específicas e intervenções direcionadas a alunos específicos, proporcionando um suporte personalizado que aborda as necessidades individuais de cada estudante. Assim, a aplicação estratégica da MDE não apenas identifica riscos iminentes, mas também instrumentaliza as instituições de ensino para desenvolver iniciativas proativas que contribuam significativamente para a redução efetiva dos índices de evasão e retenção.

3 METODOLOGIA

Este trabalho foi subdividido em algumas etapas com intuito de cumprir os objetivos elencados no Capítulo 1. Inicialmente algumas atividades comuns de pesquisa foram executadas para compreender melhor os conceitos, aplicações e problemas da literatura. Estas atividades foram: um levantamento bibliográfico e um estudo por meio de um mapeamento sistemático de literatura para os trabalhos relacionados. O fluxograma apresentado na Figura 7 demonstra a sequência da execução das atividades deste trabalho.

Figura 7 – Fluxo metodológico para a execução deste trabalho



Fonte: Elaborado pelo autor.

Os passos estão detalhados a seguir.

Conduzir uma Revisão Sistemática de Literatura

Para compreender o estado da arte da literatura e possibilitar embasamento teórico para este trabalho, foi realizado, como primeiro passo, uma RSL. Assim, é possível abstrair fatores que contribuem de maneira positiva ou negativa à proposta deste trabalho. A RSL é um tipo de estudo secundário utilizado para buscar de forma abrangente trabalhos primários relacionados com uma questão específica de pesquisa (SILVA, 2015). Para esta revisão foi utilizada o método de Kitchenham *et al.* (2009), este método aborda o protocolo como a parte do trabalho que especifica as questões de pesquisa, estratégias que serão utilizadas para condução

da revisão, critérios de inclusão e exclusão, extração e síntese dos dados.

Sendo assim, definiu-se os seguintes critérios de inclusão e de exclusão: a) 1ª seleção: leitura de títulos e de resumos para identificação dos artigos que interessam à pesquisa; b) 2ª seleção: leitura da introdução e das considerações finais para filtragem mais apurada dos artigos escolhidos na 1ª seleção. Na primeira etapa de seleção, foram excluídos os trabalhos que se limitavam a ser uma RSL ou que abordavam o tema da Mineração de Processos Educacionais sem focalizar especificamente o combate à evasão e retenção escolar. Além disso, foram desconsiderados os trabalhos redigidos em idiomas diferentes do português ou inglês.

Na segunda fase de avaliação, os mesmos critérios da primeira seleção foram aplicados, com a adição da necessidade de fornecer informações detalhadas sobre a base de dados utilizada e os algoritmos empregados. Desta forma, buscamos realizar um estudo abrangente que permitisse compreender as contribuições da Inteligência Artificial (IA) na predição de indicadores de abandono escolar. Para alcançar esse objetivo, formulamos um conjunto de Questões de Pesquisa (QP) que nortearam a investigação.

Este conjunto com 4 QP, enumeradas de QP1 a QP4, podem ser visualizadas no Quadro 1 da Seção 4.1.

Foram pesquisados artigos científicos, publicados em conferências e periódicos, nas bibliotecas digitais *ACM Digital Library*, *Google Scholar* e *IEEE Xplorer*. Mais detalhes sobre o procedimento e protocolo da RSL, será descrito no Capítulo 4.

Propor e Desenvolver um Software de MDE

Na inferência, que posteriormente será validada através de questionário, de que parte das instituições de ensino não possuem técnicos na área de Tecnologia da Informação (TI) a proposta de software deve ser de fácil uso e principalmente de fácil instalação. Sendo assim, a proposta é de que o software seja do tipo *portable*, ou seja, dispensa de instalação, sendo apenas necessário executar. A linguagem de programação escolhida para o desenvolvimento é o Java, pode-se citar alguns dos principais motivos da seleção desta linguagem:

a) facilidade de integração com o Weka (WITTEN *et al.*, 2016), que também é escrito na mesma linguagem;
b) Ser uma linguagem mundialmente difundida, e assim, facilitar o compartilhamento, evolução e manutenção do código por diversas pessoas;

c) Domínio da linguagem pelo autor do trabalho;
d) Ser uma linguagem que funciona em diversos sistemas operacionais;
e) Fazendo uso da biblioteca de interface gráfica *Java Swing* faz com que não seja necessário execução de servidor, abrir navegador nem outras ações que poderia dificultar o uso, ou seja, para utilizar o software basta executar o programa com um clique duplo.

O desenvolvimento do software ocorre utilizando a metodologia de desenvolvimento ágil SCRUM (SOMMERVILLE, 2019), O SCRUM é um *framework* de gerenciamento de projetos ágil que se concentra na entrega contínua de valor ao cliente. Ele é amplamente utilizado na indústria de software, mas também encontra aplicação em diversos outros setores. Uma das características mais distintas do SCRUM é a abordagem iterativa e incremental, na qual o trabalho é dividido em períodos de tempo chamados "*sprints*", geralmente com duração de 2 a 4 semanas. Cada *sprint* resulta em um incremento potencialmente entregável do produto, permitindo que as equipes respondam rapidamente a mudanças e prioridades em constante evolução. No SCRUM, as equipes são auto-organizadas e multidisciplinares, o que significa que têm a responsabilidade de decidir como realizar o trabalho e de quais tarefas são capazes.

A transparência é um dos princípios centrais do SCRUM, garantindo que todos os envolvidos no projeto tenham visibilidade sobre o progresso, obstáculos e qualidade do trabalho. As reuniões regulares, como a reunião diária, a revisão da *sprint* e a retrospectiva, promovem a comunicação eficaz e permitem que a equipe se ajuste continuamente para melhorar seu desempenho. Com essa abordagem adaptativa e foco na entrega de valor, o SCRUM é altamente eficaz em ambientes em que os requisitos não estão totalmente definidos e as mudanças são comuns.

Quanto a metodologia de mineração de dados foi empregada a metodologia Cross-Industry Standard Process for Data Mining (CRISP-DM) (CHAPMAN *et al.*, 2000). O CRISP-DM é um modelo padrão de processo usado para guiar projetos de mineração de dados. Ele é composto por várias fases interconectadas que ajudam a organizar e planejar o desenvolvimento de projetos de mineração de dados. Aqui estão as principais fases do CRISP-DM: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem , Avaliação , Implantação E Documentação.

Realizar Experimentos para Validar os Algoritmos de IA na Predição

Neste estudo foram realizados dois experimentos com diferentes bases de dados com o software proposto. No experimento será analisado o tempo gasto para execução da aprendizagem de máquina, sua acurácia e erro médio absoluto. O experimento foi realizado em um computador com processador i7-3537U CPU 2.00GHz e memória RAM de 8GB. A aprendizagem de maquina foi realizada através de validação cruzada com 10 *folds*. Foram verificados a acurácia, tempo de treinamento e erro médio absoluto de 5 diferentes algoritmos, são eles: Árvore de Decisão, Floresta Aleatória, KNN, MLP e Naive Bayes.

Ambas as bases de dados possuem dados públicos, uma sendo base fictícia e outra real. Mais detalhes sobre a base, número de instâncias, atributos e os resultados obtidos nos experimentos, podem ser visto na Seção 6.1 do Capítulo 6 que trata dos resultados deste trabalho.

Aplicar o Questionário de Usabilidade

O método de validação de usabilidade utilizado neste trabalho é o SUS, proposto por Brooke (1986). Este método é uma escala que mede a usabilidade de uma interface, com ele, além de obter impressões sobre a interface, pode-se analisar se as tarefas propostas pelo software foram cumpridas com facilidade e eficiência.

Ainda segundo Brooke (1996), a usabilidade de um software é um fator muito particular, e por isto, deve ser adequado ao contexto em que se encontra, assim, torna-se muito difícil de mensurá-la de maneira direta.

Identificar as Principais Dificuldades das Instituições para Utilizar o Software Proposto

Além de avaliar a usabilidade do software proposto, o questionário aplicado incluía as dez perguntas do SUS, juntamente com mais duas perguntas adicionais. Essas duas perguntas tinham como objetivo identificar possíveis problemas que a instituição poderia enfrentar ao tentar utilizar o software. Como o software de MDE necessita dos dados para realizar a mineração, durante o desenvolvimento do software levantou-se a dúvida se as instituições de ensino possuem técnicos responsáveis pelo setor de informática na instituição.

Outra questão que foi levantada é se a instituição de ensino faz uso de algum sistema de gestão escolar digital. Pois caso a gestão dos dados escolar ainda seja tratado de forma física em formulários e pastas de papel, muito provavelmente se tornará inviável o uso da ferramenta

proposta neste trabalho.

Analisar e Consolidar os Resultados Obtidos

Por fim, após a execução das etapas anteriores, as informações foram consolidadas para avaliar os resultados obtidos. O intuito é verificar se o software proposto possui uma boa acurácia na predição de insucesso escolar. Assim como, a qualidade de sua usabilidade.

Para expor estas informações foi consolidado os dados através de gráficos e relatórios em forma de tabelas, afim de, analisar as informações adquiridas, e em seguida relacionar com os objetivos e hipótese deste trabalho.

4 REVISÃO SISTEMÁTICA DA LITERATURA

Neste capítulo será apresentado uma RSL realizada e citada no Capítulo 3. Essa RSL buscou trabalhos acadêmicos nas bibliotecas digitais *ACM Digital Library*, *IEEE Xplore* e *Google Scholar* para reunir produções que pudessem colaborar com a pesquisa sobre o uso de MDE no intuito de abordar indicadores que propiciam ao insucesso estudantil. Assim, o principal objetivo desta RSL é proporcionar uma visão global do que é pesquisado sobre a colaboração da MDE para ajudar as instituições de ensino a garantirem que o discente conclua seu itinerário formativo. Como objetivos secundários, este capítulo visa contribuir com pesquisadores sobre o tema da MDE com a finalidade de diminuir o número de insucesso escolar, na intenção de identificar os atributos, algoritmos e tipo de algoritmos mais utilizados.

4.1 Metodologia da RSL

O método de revisão selecionado para realizar o estudo é o sugerido por Kitchenham (2004). Este método segue algumas etapas bem definidas para a execução de uma RSL. A etapa inicial é a definição do protocolo de pesquisa. Kitchenham aborda o protocolo como a parte do trabalho que especifica as questões de pesquisa, estratégias que serão utilizadas para condução da revisão, critérios de inclusão e exclusão, extração e síntese dos dados. Sendo assim, definiu-se os seguintes critérios de inclusão e de exclusão:

a) 1ª seleção: leitura de títulos e de resumos para identificação dos artigos que interessam à pesquisa;
b) 2ª seleção: leitura da introdução e das considerações finais para filtragem mais apurada dos artigos escolhidos na 1ª seleção;

Foram considerados excluídos na primeira seleção os trabalhos que eram uma RSL, que tratavam sobre o tema de Mineração de Processos Educacionais, que não tivesse como problemática o combate da evasão e retenção escolar, que não estavam escritos na língua portuguesa ou inglesa. Na segunda sessão, foram avaliados os mesmos critérios da primeira seleção, acrescentando a necessidade de informar os dados da base de dados como também os algoritmos utilizados. Portanto, buscou-se produzir um estudo que possibilitasse compreender as contribuições da IA para predição de indicadores do abandono escolar, perante isso, utilizou-se um conjunto de Questões Pesquisas. Este conjunto com 4 QP, enumeradas de QP1 a QP4, podem ser visualizadas no Quadro 1.

Quadro 1 – Questões de Pesquisa (QP) da RSL

Nº	Questão de Pesquisa
QP1	Qual o tipo de algoritmo utilizado?
QP2	Qual o algoritmo utilizado?
QP3	Qual a modalidade de ensino da base de dados utilizada?
QP4	Quais os dados contidos na base de dados (socioeconômicos, educacionais, localidades geográficas, entre outros)?

Fonte: Elaborado pelo autor.

Dessa maneira, decidiu-se buscar artigos científicos nas bibliotecas digitais *ACM Digital Library*, *Google Scholar* e *IEEE Xplorer* fazendo das strings de buscas expostas no Quadro 2.

Quadro 2 – Strings de buscas nas plataformas digitais.

Biblioteca	Idioma dos artigos	String de busca
ACM Digital Library	Língua inglesa	((Data mining AND education AND (dropout OR retention)) OR (Educational data mining OR EDM) AND (dropout OR retention))
Google Scholar	Língua portuguesa	((Mineração de dados AND educação AND (evasão OR retenção)) OR (Mineração de dados educacionais OR MDE) AND (evasão OR retenção))
IEEE Xplore	Língua Inglesa e Portuguesa	("All Metadata":Data mining education) AND ("All Metadata":dropout)

Fonte: Elaborado pelo autor.

Diante do exposto, esclarece-se que a RSL envolveu a indicação de uma string de busca que proporcionasse o agrupamento de trabalhos da área do estudo desta revisão, logo, foi baseada nas QPs e de modo a atender quais os principais métodos de predição utilizados e ao mesmo tempo verificar se existe uma relação entre a modalidade, os algoritmos e seus tipos, como também, nos dados utilizados para a aprendizagem de máquina. Ademais, os termos OR e AND contribuem para variar e aproximar os termos da pesquisa, todavia, foi empregada na língua inglesa na plataforma *ACM Digital Library*, em língua portuguesa na biblioteca *Google Scholar* e em português e inglês na *IEEE Xplorer* com o objetivo de recolher produções nas duas línguas e, assim, enriquecer a elaboração desta RSL.

4.2 Resultados e Discussão

Após a realização das buscas nas bibliotecas através das strings de buscas, dispostas no Quadro 2, obteve-se o resultado exposto no Quadro 3. No quadro também é possível observar

o número de artigos que foram excluídos e incluídos em cada avaliação, os métodos destas avaliações foram descritos na sessão de Materiais e Métodos.

Quadro 3 – Quantidade de produções acadêmicas encontradas nas plataformas digitais.

	ACM Digital Library	**Google Scholar**	**IEEE Xplore**
Artigos encontrados	1.664	599	62
1ª exclusão e inclusão	Excluídos: 1.661 Incluídos: 3	Excluídos: 492 Incluídos: 107	Excluídos: 15 Incluídos: 47
2ª exclusão e inclusão	Excluídos: 0 Incluídos: 3	Excluídos: 88 Incluídos: 19	Excluídos: 23 Incluídos: 24

Fonte: Elaborado pelo autor.

Dito isto, após a filtragem e apuração dos trabalhos apresentados pelas plataformas, partiu-se para estudos de cada artigo selecionado com o intuito de verificar sua adequação às QPs, portanto, procedeu-se à análise, realizando a leitura integralmente dos documentos, e discussão dos dados obtidos a partir do diálogo entre os 46 artigos escolhidos e a proposta da pesquisa: analisar a contribuição do uso de IA para predição do insucesso estudantil, considerando indicadores de evasão e retenção escolar.

Após a análise dos 46 artigos, pode-se obter as respostas do objetivo deste trabalho. Os resultados serão apresentados nas sequências das perguntas das QP do Quadro 1.

A QP1 questiona sobre o tipo de algoritmo mais utilizado, conforme pode ser observado na Figura 8, nesta RSL os artigos não tratavam sobre os grupos de algoritmos, como também era muito comum utilizarem diferentes tipos na mesma base de dados. No entanto, nos artigos da RSL que apresentaram qualquer descrição sobre o algoritmo utilizado, sempre citavam o uso da ferramenta Weka (WITTEN *et al.*, 1999), sendo assim, este trabalho faz o agrupamento dos algoritmos da mesma forma que a ferramenta Weka. Foram encontrados 8 tipos diferentes de algoritmos, sendo que um deles, o Amazon Machine Learning, não se encontra na ferramenta Weka. Conforme pode ser visto na Figura 8, os algoritmos baseados em árvores foram os mais frequentemente utilizado, sendo utilizado em 36 dos trabalhos, ou seja, um pouco mais de 78% dos trabalhos. Tendo sequência, na frequência de uso, os algoritmos do grupo de funções, com 35 utilizações nos trabalhos incluídos. Importante esclarecer que alguns artigos estão sendo contabilizados em mais de um tipo de algoritmo devido utilizar mais de um algoritmo no seu experimento.

A resposta da QP2 – Qual o algoritmo mais utilizado, vêm como reflexo da resposta da QP1, tendo o algoritmo de árvore de decisão J48 e a Floresta aleatória como os algoritmos mais utilizados, tendo 23 e 13 utilizações. Foram utilizados um total de 19 algoritmos. O número

Figura 8 – Representação gráfica das respostas a “QP 1 - Qual tipo de algoritmo mais utilizado?”.

Fonte: Elaborado pelo autor.

de trabalhos em cada algoritmo pode ser observado na Figura 9.

Figura 9 – Representação gráfica das respostas a “QP 2 - Qual algoritmo mais utilizado?”.

Fonte: Elaborado pelo autor.

A resposta da QP3 será dividida em duas partes, primeiro respondendo se o curso ocorre de forma presencial ou à distância. Tendo um total de 7 trabalhos com curso distância e os demais 39 são cursos de forma presencial.

A segunda parte, que é a questão da modalidade de ensino. Quatro trabalhos abordaram a evasão do ensino médio, enquanto que os trabalhos com base de dados de alunos da modalidade técnico/profissionalizante foram 6, os demais 36 artigos possuíam dados de alunos do ensino superior. Outro ponto interessante observado nesta RSL, é que o curso mais frequentemente utilizado nos experimentos eram na área de computação, 21 dos 36 citaram que eram da área de computação.

Por fim, a QP4 trata sobre a base de dados utilizada, realizando uma intersecção

entre os grupos citados nos artigos de Dharmawan *et al.* (2018) e Tenpipat e Akkarajitsakul (2020). Sendo assim, este trabalho utiliza os seguintes grupos de dados: demográfico ou sociais, educacionais, econômicos, interações sociais, personalidade do estudante, informações sobre os docentes da instituição de ensino, do município da instituição de ensino. A Figura 10 representa, em forma de gráfico, os tipos de dados mais utilizados nos artigos da revisão. Importante dizer, que apenas o artigo de Dharmawan *et al.* (2018) não utilizou dados educacionais em sua pesquisa.

Figura 10 – Representação gráfica das respostas a "QP 4 - Grupo de dados utilizado nos artigos?".

Fonte: Elaborado pelo autor.

No entanto, foi realizada uma pesquisa maior nos dados para identificar os mais utilizados. Após essa ação, foi possível identificar 191 diferentes atributos utilizados nos 46 artigos desta RSL. Devido a grande quantidade de atributos encontrados, o Quadro 4 apresenta os atributos mais utilizados.

Alguns dos atributos não tiveram uma frequência tão alta, no entanto, chamaram a atenção dos autores devido a sua importância nos resultados ou o fato de ser incomum, mas mesmo assim ter impactados nos trabalhos que o utilizaram. A etnia foi um dos atributos mais utilizados, no entanto, Mehra *et al.* (2019) em seu trabalho utilizou um outro atributo social Indiano muito impactante, a casta. Outros atributos como os hábitos do aluno sobre sua frequência em leitura, fumar ou consumo de bebidas alcoólicas nos finais de semana, relacionamento com a família e colegas (DHARMAWAN *et al.*, 2018). Sobre os dados da instituição o tempo na qual os professores estão na instituição, que em outras palavras, pode ser dito como a rotatividade dos professores na escola, a adequação do professor e disciplina lecionada, formação do professor como suas pós-graduação ou se é licenciado ou bacharel (NASCIMENTO *et al.*, 2018).

Por fim, foi analisado os países dos autores dos artigos, na intenção de identificar os países que tentam solucionar esses problemas com técnicas de mineração de dados educacionais. Sendo que, 24 artigos são de autores brasileiros, 5 da Índia, 3 da Indonésia, Colômbia,

Quadro 4 – Atributos mais utilizados nos artigos da RSL.

Atributo	**Repetições**	**Atributo**	**Repetições**
Gênero	31 Artigos	Profissão do pai e da mãe	5 Artigos
Rede de origem (pública, privada)	14 artigos	Média de horas-aulas diárias	4 artigos
Área do curso (Saúde, Eng. E etc)	13 artigos	Quant. de disciplinas por semestre	4 artigos
Data de nascimento ou idade	13 artigos	Quant. de repro. em disc. básicas	4 artigos
Idade no início do curso	12 artigos	Já possui curso superior	4 artigos
Ano de ingresso	11 artigos	Núm./porcent. de disc. concluídas	4 artigos
Estado civil	11 artigos	Número de reprovações	4 artigos
Renda per capita familiar	9 artigos	Distância ou tempo entre escola e lar	4 artigos
Cidade de nascimento	8 artigos	Está trabalhando	4 artigos
Curso matriculado	7 artigos	Campus	4 artigos
Etnia	7 artigos	Recebe alguma bolsa estudantil	4 artigos
Forma de ingresso	6 artigos	País de origem	4 artigos
Turno	6 artigos	UF de origem	3 artigos
Número de membros na família	6 artigos	tempo da conclusão do ensino médio	3 artigos
Escolaridade do pai e da mãe	6 artigos	Já evadiu de outro curso	3 artigos
Nota do vestibular/Enem	5 artigos	Meio de transporte	3 artigos
Média e/ou notas das disc. aprovadas	5 artigos	Possui internet em casa	3 artigos
Semestre do ano de ingresso	5 artigos	Tempo gasto por dia nas redes sociais	3 artigos
Zona do lar (rural ou urbana)	5 artigos	Taxa de evasão/retenção da instituição	2 artigos

Fonte: Elaborado pelo autor.

Espanha, Itália e Tailândia tiveram 2 autores. Tiveram apenas um autor representando África do Sul, Bangladesh, Chile, China, Fiji, Paquistão, Polônia, Portugal e Romênia. Dois trabalhos apresentavam pesquisas com autores de países diferentes, sendo este o motivo do número de países não ser o mesmo número de artigos selecionados.

4.3 Conclusão da RSL

Após a conclusão da pesquisa de revisão pode-se observar que os algoritmos mais utilizados são as árvores de decisão e florestas aleatórias, no entanto os algoritmos de regressões, MLP e SVM também foram utilizados com uma certa frequência assim como em alguns trabalhos foram os que obtiveram o maior sucesso, sendo que a grande maioria dos trabalhos fizeram uso da ferramenta Weka. Também foi possível chegar à conclusão que vários dados podem ser utilizados na mineração de dados educacionais para prever a evasão e retenção, incluindo dados que não são do âmbito escolar.

Graças à RSL, foi possível constatar que a MDE já se consolidou como uma abordagem amplamente difundida na previsão de evasão e retenção escolar. Essa constatação confere maior robustez e validade à presente pesquisa. Além disso, a análise da RSL possibilitou a identificação dos algoritmos e ferramentas mais amplamente empregados nesse contexto, fundamentando assim o desenvolvimento do software proposto, que se utiliza desses algoritmos.

Outro aspecto que contribuiu de maneira significativa para o aprimoramento do software em questão é a compreensão de que diferentes conjuntos de dados podem ser empregados para alimentar os modelos de aprendizado de máquina. Consequentemente, o software proposto foi concebido para aceitar diversos tipos de dados como entrada, proporcionando flexibilidade e adaptabilidade na implementação de técnicas de aprendizado de máquina. Essa característica visa garantir que a ferramenta seja aplicável em contextos educacionais diversos, possibilitando a personalização e eficácia na previsão de evasão e retenção escolar.

5 PRODUTO EDUCACIONAL

A proposta do trabalho se deu na construção de um software para mineração de dados de maneira abrangente, permitindo sua aplicação em diversas bases de dados e com a utilização de diferentes algoritmos de aprendizagem de máquina. Além disso, foi priorizada a acessibilidade aos usuários, não exigindo conhecimentos técnicos para sua utilização. Para garantir a facilidade de uso, a usabilidade do software foi avaliada por meio do método SUS.

5.1 Requisitos do software proposto

Diversos membros das instituições de ensino podem realizar ações que possam impactar a vida escolar dos estudantes, e com isso, ocasionar na diminuição dos índices de insucesso escolar. Monitoria (SAMPAIO; SILVA, 2019), bolsas de assistência estudantil (CARRANO *et al.*, 2018), cursos de auxiliares ou disciplinas auxiliares (ANDRADE *et al.*, 2019) são exemplos de ações que auxiliam no combate a evasão. No entanto, nem sempre as instituições contam com todos os recursos financeiros para oferecer bolsas e monitorias ou com infraestrutura e recurso humano para os cursos extras. Sendo assim, as instituições e seus gestores tem que, de certo modo, identificar e incluir aos programas os alunos mais indicados, para que assim, possam minimizar a evasão e/ou retenção escolar. Uma tecnologia que vem sendo trabalhada com esse objetivo é a mineração de dados educacionais.

A MDE é uma subárea da mineração de dados, que tem como objetivo realizar processos para encontrar anomalias, padrões e correlações em grandes conjuntos de dados, e assim, poder prever resultados (HAN *et al.*, 2011). Sendo assim, a mineração de dados educacionais, faz uso dos dados educacionais para realizar esta mineração. Frequentemente, nas pesquisas, esses dados são separados em subconjuntos de dados, sendo estes os mais comuns subconjuntos:

- dados da instituição de ensino (COLPANI, 2018), como por exemplo notas de qualidade da instituição, localidade, cursos que oferta, titulação de docentes ou infraestrutura da instituição;
- dados socioeconômicos (DHARMAWAN *et al.*, 2018), na qual são informações frequentemente solicitadas pelas instituições durante a matrícula, por exemplo, local de residência, tipo de residência, renda familiar, escolaridade dos pais, escola antecessora, possui plano de saúde particular; e

- dados educacionais (COUTO, 2017) (SUPERBE; SILVA, 2018), podendo ser citados os atributos de frequência escolar, notas, monitorias, cursos de extensão, curso matriculado e disciplinas matriculadas.

Sendo assim, tarefas de predição consiste na análise de um conjunto de dados que estão descritos por atributos e seus respectivos valores associados a um rótulo. Podendo ser exemplificado com um conjunto de dados dos alunos matriculados na disciplina de Cálculo I, que possui diversos atributos como as notas, frequência, número de vezes que está estudando a disciplina, e seus respectivos valores para cada aluno no conjunto de dados. Por fim, um possível rótulo pode ser a descrição se o aluno foi reprovado ou aprovado. Com isto, na entrada de novas informações de alunos recém matriculados, a mineração de dados pode fazer uma predição sobre o seu rótulos, provável aprovado ou reprovado, através de padrões de informações já observadas.

Ainda sobre a predição, segundo Silva *et al.* (2016) a análise preditiva necessita de técnicas ou algoritmos para ser executadas. Existindo diversos algoritmos de inteligência artificial com funcionamento de formas diferentes, desde algoritmos estatísticos, árvores de decisões ou redes neurais artificiais. Sendo que todos eles possuem vantagens e desvantagens, na qual alguns podem alcançar um melhor resultado para conjuntos de dados diferentes ou um resultado similar, porém com um tempo de processamento menor.

Isto posto, O Quadro 5 pode-se definir os principais requisitos do software proposto, sendo eles:

Quadro 5 – Requisitos do software proposto

Requisito	**Tipo**	**Motivação ou Descrição**
Usabilidade	Não Funcional	Devido ter usuários de diverso níveis de computação e mineração de dados
Permitir base de dados genérico	Funcional	Possibilitar que qualquer instituição utilize sua base de dados emitida pelo seu sistema de gestão acadêmica
Permitir a remoção de atributos	Funcional	O usuário possa escolher qual dos atributos da base irá utilizar para a aprendizagem de máquina, pois assim, o usuário não precisa alterar o arquivo da base emitido pelo sistema acadêmico
Realizar a predição com diferentes algoritmos	Funcional	Devido diferentes bases obterem resultados melhores com diferentes algoritmos, o software proposto deve encontrar qual o melhor algoritmo para este problema
Emitir relatório	Funcional	O relatório irá auxiliar na tomada de decisão dos gestores escolares
Salvar o treinamento	Funcional	A mesma base é utilizada com frequência no mesmo semestre, não alterando o seu conteúdo, podendo ser reutilizado as mesmas configurações de treinamento

Fonte: Elaborado pelo autor.

O ultimo requisito exposto no Quadro 5, de salvar o treinamento, foi identificado durante a entrevista de usabilidade, visto que um dos algoritmos de aprendizagem de máquina, o MLP, possui um tempo elevado de treinamento. Os entrevistados citaram que frequentemente

utilizariam a mesma base de dados durante boa parte do semestre. Sendo assim, caso fosse possível salvar o treinamento, este trabalho seria realizado apenas uma única vez no semestre.

5.2 O Software Proposto

5.2.1 Desenvolvimento do software

O software proposto foi desenvolvido para plataforma *desktop* e na linguagem de programação Java. O motivo da linguagem escolhida se explica devido a facilidade da integração com o pacote de software Weka (WITTEN *et al.*, 2016), que também está escrito em Java, facilitando sua integração. Além do fato de posteriormente poder reutilizar parte do código fonte para que o software seja convertido para uma versão web, outro fator determinado é devido o domínio da linguagem pelo autor deste trabalho. Enquanto que o motivo de utilização do Weka, foi considerado devido ao número de utilizações que foi analisada através da RSL exposta no Capítulo 4, assim como, ao tipo de licença do software de GPL sendo portanto possível estudar e alterar o respectivo código-fonte. Assim, fazendo com que o software proposto também possa seguir com a mesma licença e poder ter contribuições de código-fonte de terceiros.

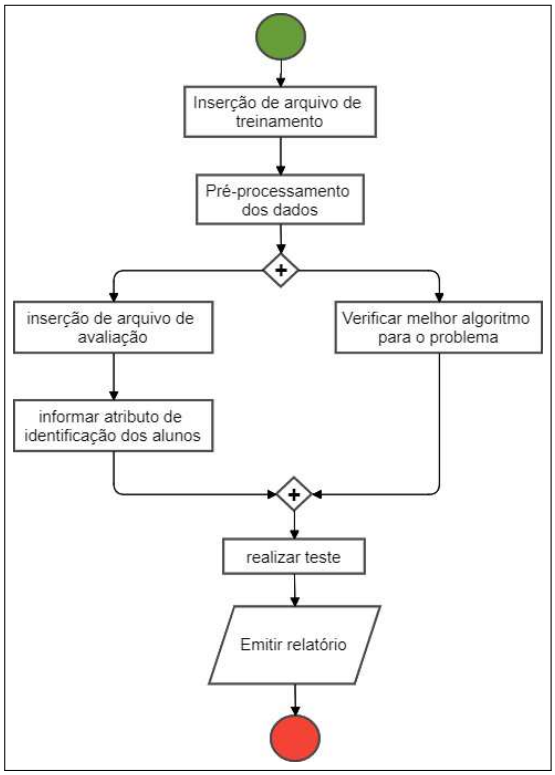
Enquanto a plataforma *desktop* se explica devido o software poder ser executado de forma que nem seja necessário instalação, sendo conhecido como *software portable* ou aplicativo portátil. Alguns pesquisas comprovam que existe uma falta de recurso humano no setor de TI nas Instituição de Ensino (IE)s como citam os trabalhos de Freire *et al.* (2021), Carvalho e Oliveira (2019). Por outro lado, para que o software proposto seja na forma de aplicação web, seria necessário um pouco mais de conhecimento para a instalação de um servidor web e a configuração do *firewall* para possibilitar o seu funcionamento na intranet da instituição educacional.

O trabalho de aprendizagem de máquina geralmente envolve alto custo computacional e, dependendo da quantidade de dados de treinamento, pode levar alguns minutos. Portanto, após receber os dados de treinamento e realizar o pré-processamento e validação, o software inicia a aprendizagem de máquina com diversos algoritmos, buscando encontrar o que proporciona os melhores resultados para aquela base de dados específica.

No entanto, como essa ação pode demorar algum tempo, ela funciona em concorrência com outras ações do usuário, com o objetivo de minimizar o tempo de resposta final, como mostrado na Figura 11. Além disso, no fluxograma da Figura 11, enquanto a aprendizagem está

em processo, o software continua apto a receber informações da base de dados para realizar a verificação, conhecida como base de teste. Essa base deve ter apenas um detalhe específico: deve ser informado um atributo para identificar o aluno, que pode ou não ser utilizado na avaliação.

Figura 11 – Fluxograma do funcionamento do software proposto.



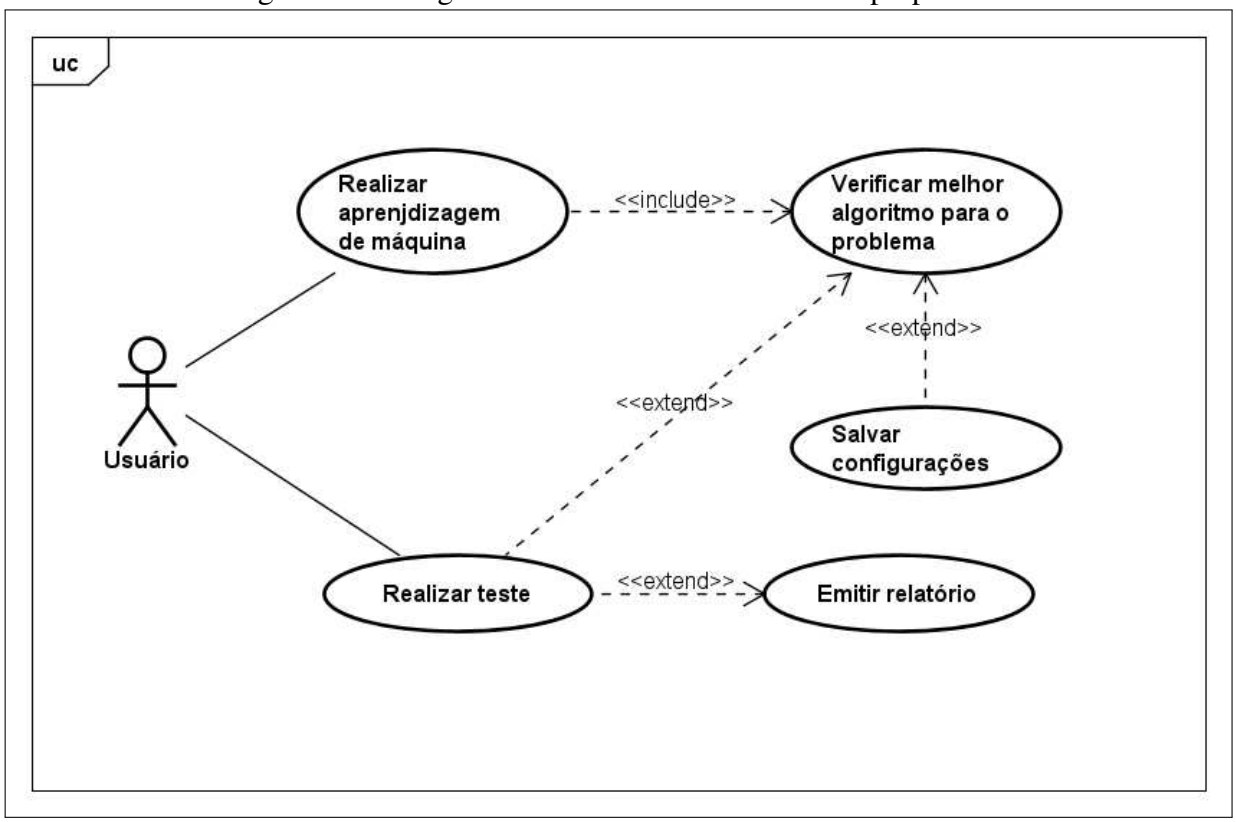
Fonte: Elaborado pelo autor.

A Figura 12 apresenta o Diagrama de Caso de Uso do software proposto, descrevendo as funcionalidades gerais do programa. Durante a etapa de aprendizagem de máquina, o software identificará qual classificador obteve o melhor resultado. Após essa identificação e treinamento do algoritmo selecionado, o usuário poderá salvar as configurações do programa em um arquivo. Essa funcionalidade será especialmente útil, uma vez que o procedimento é computacionalmente custoso e, consequentemente, mais longo.

Dessa forma, o usuário poderá utilizar o mesmo arquivo de treinamento para diversas predições diferentes, desde que não ocorram mudanças na base de dados. Ao final da classificação, o software fornecerá um relatório informando quais alunos foram classificados como evadidos, retidos ou alcançaram a conclusão no tempo correto. Esse relatório poderá ser impresso para ser utilizado como apoio nas tomadas de decisões.

A maneira de inserir os dados na base de dados do software é por meio de arquivos

Figura 12 – Diagrama de Caso de Uso do software proposto.



Fonte: Elaborado pelo autor.

em formatos *Comma-Separated Values* (CSV), *OpenDocument Spreadsheet Document* (ODS), Microsoft Excel 97-2003 *Workbook* (XLS) ou Microsoft Excel *Workbook* (XLSX), por duas razões.

A primeira razão é facilitar para os usuários do software, uma vez que esses são os principais formatos de transferência de dados, os quais provavelmente podem ser exportados por um software de gestão escolar ou formulários *online*.

A segunda razão é evitar que os dados sejam informados manualmente por meio de formulários, o que poderia levar a erros de digitação ou outras falhas humanas, especialmente considerando o possível volume de dados.

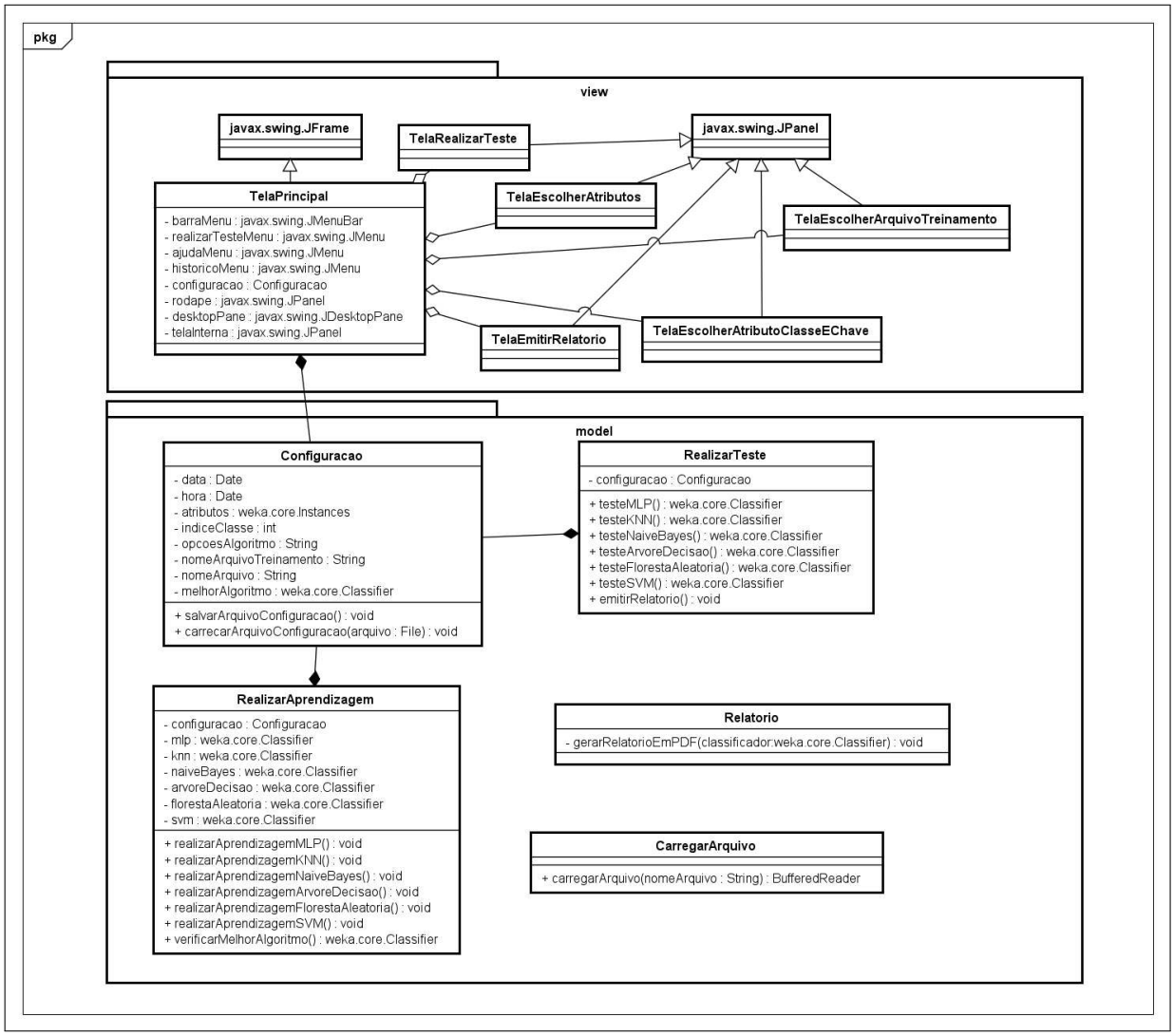
A Figura 13 representa o diagrama de classe do software proposto, que foi dividido em dois pacotes: o pacote de visão e o pacote de modelo. Na imagem, podemos observar que a interface gráfica será desenvolvida utilizando o *toolkit Java Swing*. A tela principal herda da classe `JFrame` e outras classes do Java Swing, como `JItemMenu`, `JButtom`, `JTextField`, `JLabel` e outras relacionadas a formulários, tiveram seus atributos retirados do diagrama para evitar excesso de detalhes.

Além disso, os métodos de controle e encapsulamento (*gets* e *sets*) também foram

simplificados para tornar o diagrama mais legível. As telas, com exceção da tela principal, realizam herança com a classe `JPanel` para generalizar o código ao instanciar novas telas e adicioná-las ao atributo `desktopPane` da classe `TelaPrincipal`.

Por meio do `desktopPane`, que é um objeto da classe `JDesktopPane`, é possível inserir o painel de uma nova tela, como um objeto da classe `TelaEscolherAtributos`, dentro de um objeto da classe `JInternalFrame`. Dessa forma, as telas ficam abertas na janela principal do programa, evitando que o usuário final abra diversas telas e as perca na barra de status do sistema operacional.

Figura 13 – Diagrama de Classe do software proposto.



Fonte: Elaborado pelo autor.

No pacote modelo da Figura 13, estão representadas cinco classes que desempenham funções essenciais nas principais regras de negócio do software. A única classe que possui relacionamento direto com as classes do pacote de visão é a classe `Configuracao`. Essa classe

é responsável por armazenar informações já processadas, como os dados inseridos e o algoritmo que obteve os melhores resultados para a classificação.

Dessa forma, as classes `Aprendizagem` e `RealizarTeste`, que são responsáveis pela aprendizagem de máquina e pela classificação, fazem parte da composição do objeto `Configuracao`. Por essa razão, a classe `Configuracao` é implementada como um objeto `Singleton`, garantindo que haverá apenas uma instância dela durante todo o ciclo de vida da execução do software.

As informações armazenadas na classe `Configuracao` serão salvas em arquivos, por meio de um método específico da classe, e esses arquivos conterão os valores contidos nesse único objeto.

Assim como nas classes do pacote de visão, os métodos de encapsulamento foram omitidos para facilitar a visualização do diagrama. Além disso, a classe `Singleton` e seus atributos privados e estáticos também foram representados sem seus métodos de encapsulamento.

5.2.2 Apresentação do software

Nesta subseção, apresentaremos o software desenvolvido e sua forma de utilização. Na Figura 14, é possível visualizar a tela principal do software. A usabilidade é facilitada pela barra de menu localizada no topo da janela, e o status do uso pode ser verificado no painel na parte inferior.

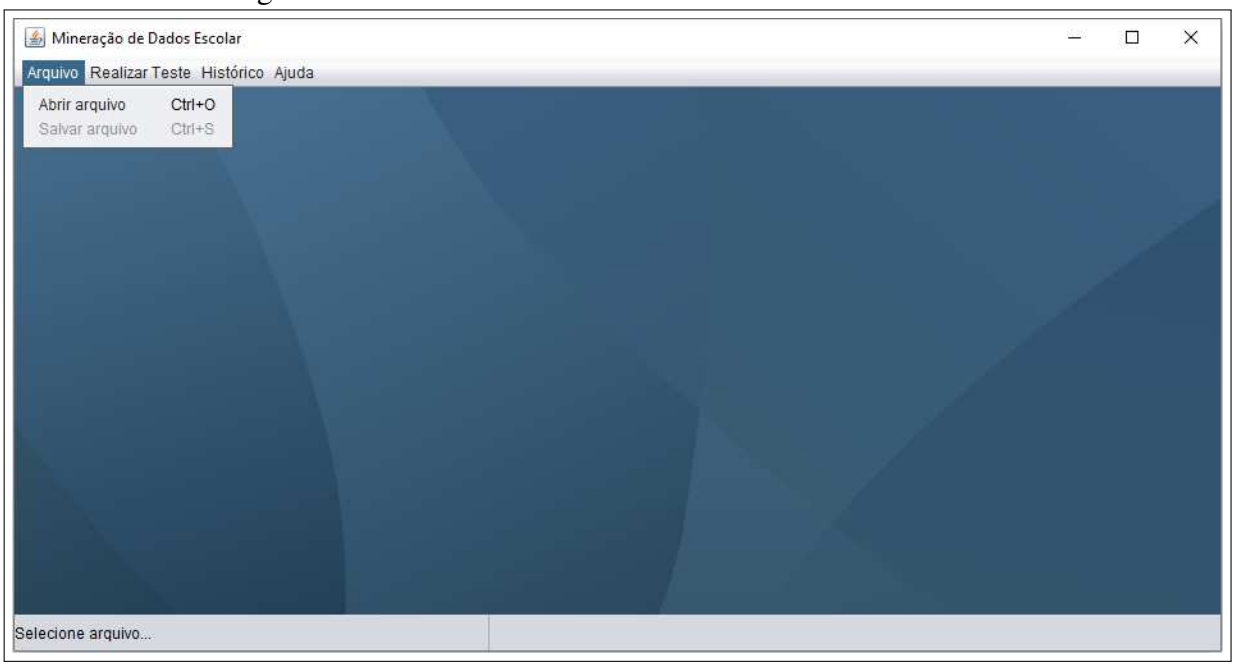
Ao utilizar os itens do menu, as janelas internas são abertas dentro do painel de trabalho, que, na Figura, possui um plano de fundo azul. Na Figura 14, o menu “Arquivo” está aberto, expondo as opções para abrir um arquivo de treinamento e salvar os dados. Neste momento, como o software ainda não realizou nenhuma aprendizagem de máquina, a opção de salvar o arquivo de treinamento está desativada.

Afim de facilitar o entendimento do uso do software, os dois processos do uso do software foi dividido em dois tópicos, sendo o primeiro realizando a MDE através da base de dados, enquanto que o segundo, será feito a partir do arquivo de treinamento salvo pelo software.

5.2.2.1 Realizando MDE por meio do arquivo de base de dados.

Para realizar a mineração de dados usando arquivos da base de dados extraídos do software de gestão acadêmica, é necessário acessar o menu “Realizar Teste”, conforme mostrado na Figura 15. O menu “Realizar Teste” possui quatro itens: “Realizar Aprendizagem”, “Atributos

Figura 14 – Tela inicial.

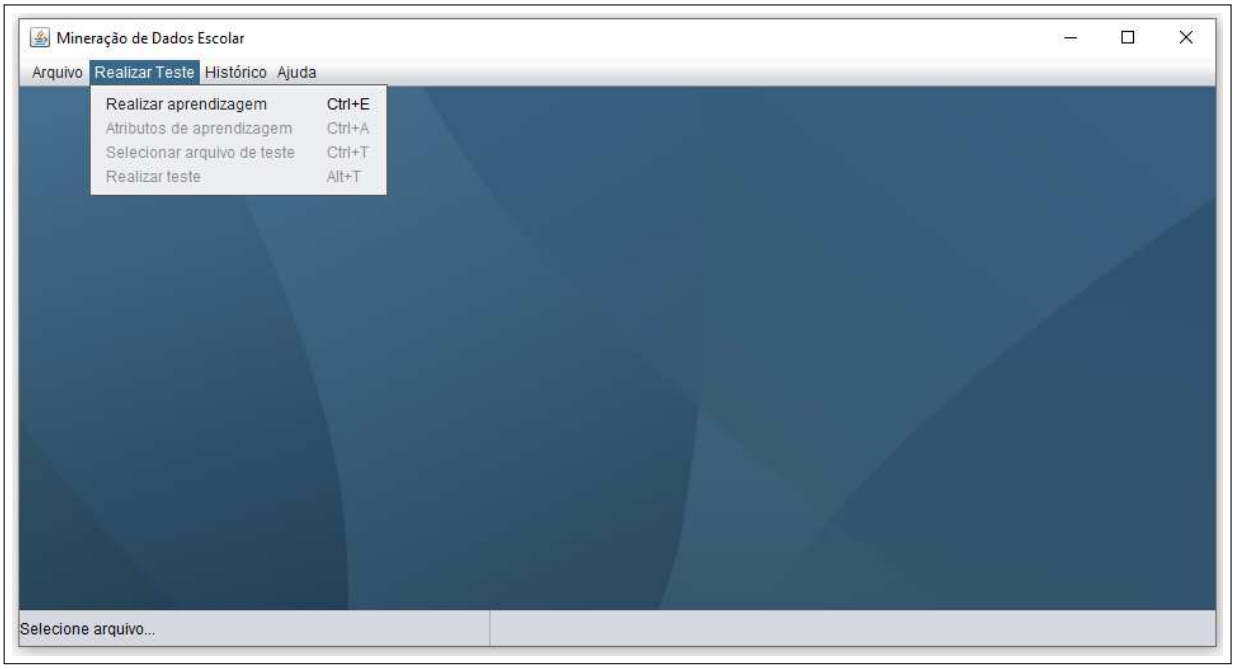


Fonte: Elaborado pelo autor.

de Aprendizagem", "Selecionar Arquivo de Teste" e "Realizar Teste".

No entanto, até o momento, o software ainda não realizou a aprendizagem de máquina, portanto, somente a opção "Realizar Aprendizagem" está disponível.

Figura 15 – Tela Inicial - Menu de Realizar Teste.

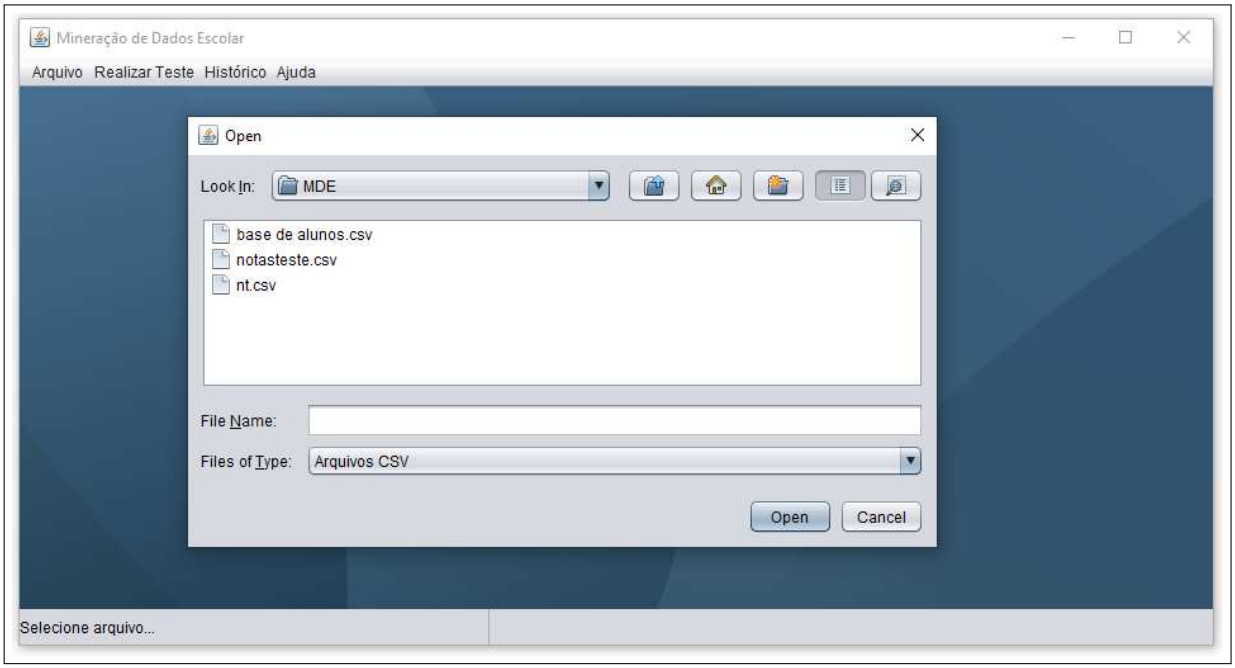


Fonte: Elaborado pelo autor.

Ao clicar no item de menu "Realizar aprendizagem" ou utilizar o atalho "Ctrl + E" irá carregar uma tela para informar o arquivo de aprendizagem de máquina. Este arquivo será

o da base de dados da instituição de ensino, como por exemplo, as informações retirados do software de gestão acadêmico da instituição. A Figura 16 apresenta esta tela.

Figura 16 – Selecionar arquivo de dados para realizar aprendizagem de máquina.

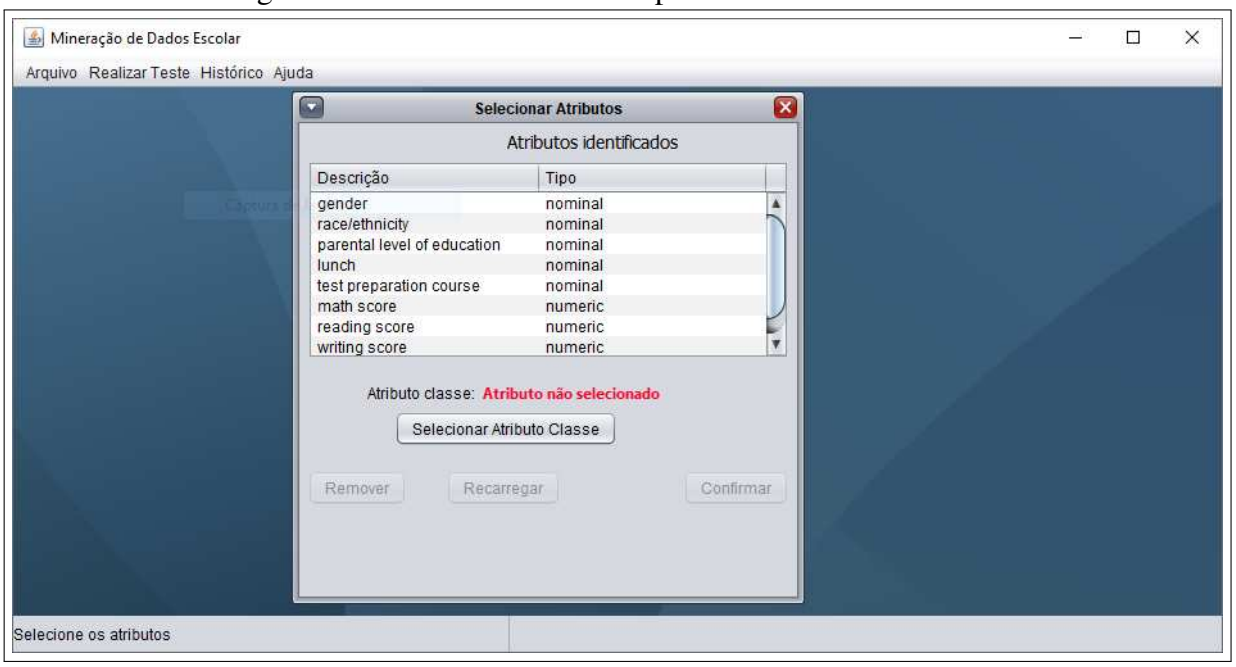


Fonte: Elaborado pelo autor.

Após selecionar o arquivo de treinamento os dados serão carregados e caso não contenha erros, os atributos e seus respectivos tipos serão expostos na próxima janela. Conforme pode ser observado na Figura 17. Através da Figura, pode-se perceber uma tabela com os atributos identificados no arquivo informado. Através desta tabela é possível selecionar qual atributo será utilizado para ser o atributo classe do experimento, como também, remover atributos da aprendizagem de máquina, ou seja, realizar a preparação dos dados para treinar os algoritmos. Para realizar alguma ação, basta clicar na linha da tabela e em seguida no botão da ação. O botão de "recarregar" irá recarregar a tabela com todas as informações iniciais, assim, o usuário pode refazer o processo caso aconteça algum erro ao remover um atributo acidentalmente. Somente após selecionar o atributo classe, o botão de confirmar ficará clicável e ao clicar iniciará o processo de aprendizagem de máquina.

Após confirmar os atributos e clicar em continuar, o processo de aprendizagem de máquina iniciará. O usuário poderá acompanhar na barra de status o andamento do treinamento do software em cada algoritmo. Como o processo de aprendizagem de máquina pode levar algum tempo, o usuário poderá informar ao software o arquivo para teste. A Figura 18 apresenta o

Figura 17 – Selecionar atributos para realizar treinamento.



Fonte: Elaborado pelo autor.

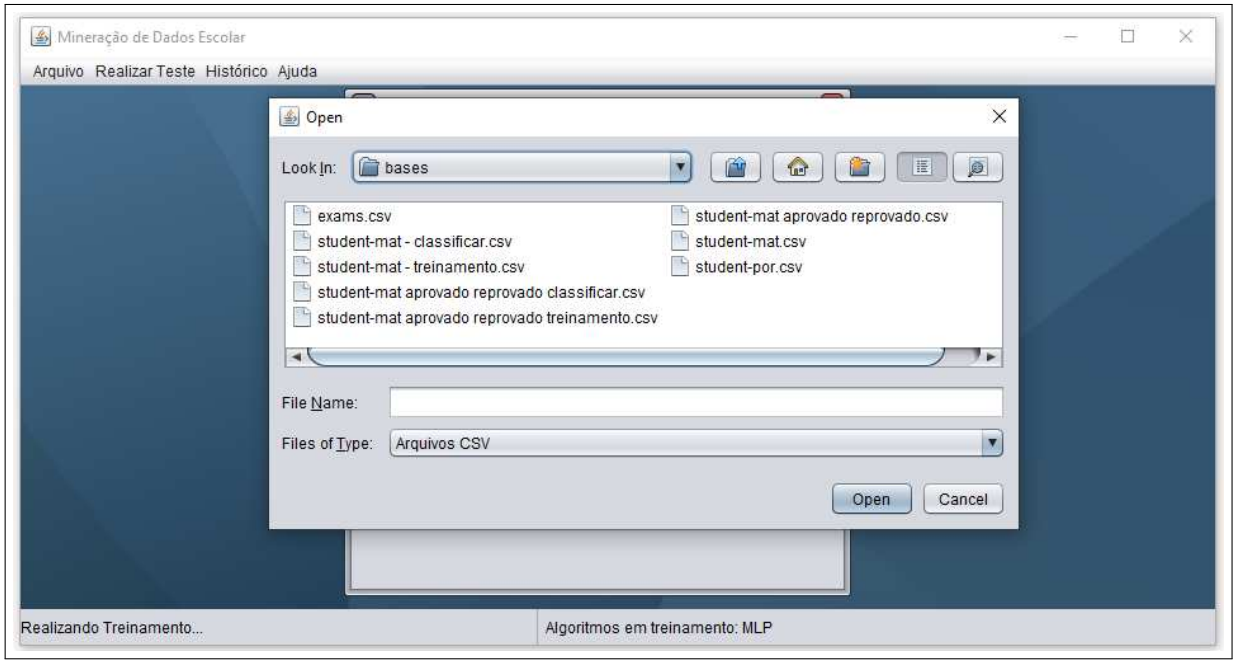
momento em que a janela da escolha de arquivo de classificação enquanto realiza a aprendizagem de máquina, perceba que no momento em que a imagem foi capturada, apenas o algoritmo de MLP continua em treinamento. Os demais algoritmos, Árvore de decisão, Floresta Aleatória, KNN, *Naive Bayes* e SVM já foram concluídos. Após o treinamento da base de dados, o texto de treinamento completo irá aparecer na barra de status, independentemente de qual tela está sendo apresentada. Também ficará disponível para utilização o item de menu “Salvar”, que está presente no menu “Arquivo”, este item de menu está exposto de forma desativada na Figura 14.

A Figura 19 apresenta a tela de selecionar atributos de classificação. Está janela é similar a que foi apresentada na Figura 17, no entanto, esta janela possui o campo para selecionar o atributo que identifica o objetivo a ser classificado, como por exemplo nome, matrícula, Cadastro de Pessoas Físicas (CPF) ou outro atributo, visto que, ao fim do processo, será informado a classificação do conjunto de dados e assim, auxiliar na tomada de decisões da gestão escolar. Após informar o atributo classe e o atributo de identificação o botão de classificar ficará disponível.

Após realizar a classificação, o usuário será apresentado a um resultado detalhado em um arquivo em formato PDF, como mostrado na Figura 20. Esse arquivo contém informações valiosas sobre a classificação dos dados, incluindo a probabilidade de confiança na qual o melhor algoritmo classificou cada instância.

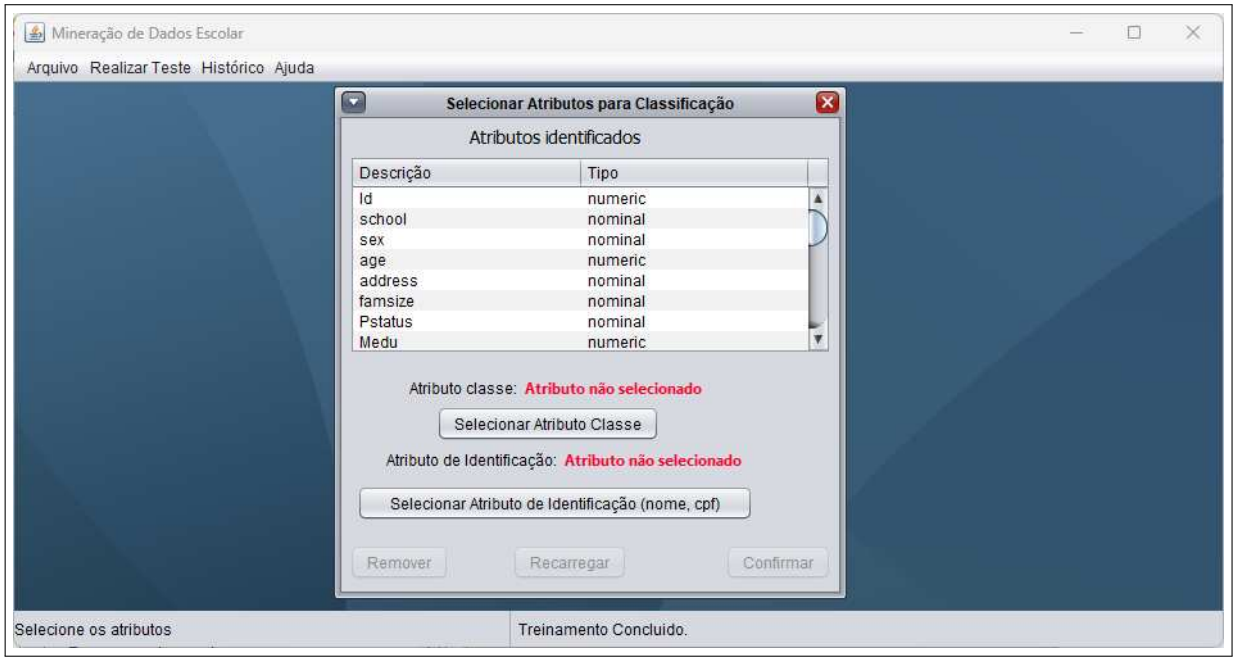
No arquivo *Portable Document Format* (PDF), os resultados são organizados de

Figura 18 – Selecionar arquivo para realizar teste enquanto realiza aprendizagem de máquina.



Fonte: Elaborado pelo autor.

Figura 19 – Selecionar atributos para realizar teste enquanto realiza aprendizagem de máquina.

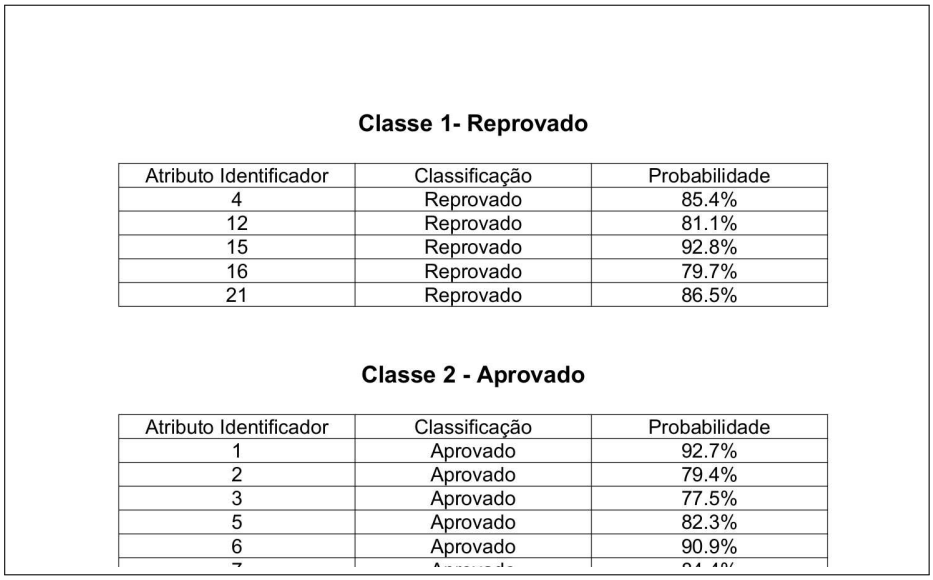


Fonte: Elaborado pelo autor.

forma clara e fácil de compreender, com os dados agrupados por classe. Para cada classe, o usuário pode verificar os registros classificados e suas respectivas probabilidades. Essa apresentação facilita a análise dos resultados e fornece *insights* importantes sobre a performance do algoritmo de aprendizagem de máquina em relação a cada classe.

A saída em formato PDF torna o processo de avaliação mais conveniente e permite que os resultados sejam compartilhados facilmente com outras partes interessadas, como gestores e equipe acadêmica. Dessa forma, a análise dos resultados da classificação torna-se mais eficiente e contribui para uma gestão escolar mais informada e direcionada no combate a evasão e retenção escolar.

Figura 20 – Resultado da classificação.



Classe 1- Reprovado

Atributo Identificador	Classificação	Probabilidade
4	Reprovado	85.4%
12	Reprovado	81.1%
15	Reprovado	92.8%
16	Reprovado	79.7%
21	Reprovado	86.5%

Classe 2 - Aprovado

Atributo Identificador	Classificação	Probabilidade
1	Aprovado	92.7%
2	Aprovado	79.4%
3	Aprovado	77.5%
5	Aprovado	82.3%
6	Aprovado	90.9%

Fonte: Elaborado pelo autor.

A Figura 20 Apresenta apenas parte do relatório final, O relatório completo pode ser visualizado no Apêndice B.

6 AVALIAÇÃO PROPOSTA

A proposta do trabalho consistiu na criação de um software genérico para a mineração de dados escolares. Esse software foi projetado para ser compatível com diversas bases de dados e algoritmos de aprendizagem de máquina diferentes. Para avaliar sua versatilidade, foram conduzidos experimentos com duas bases de dados públicas distintas. Além disso, o software foi desenvolvido de forma a ser acessível a usuários sem conhecimentos técnicos específicos. Portanto, o software foi projetado com ênfase na facilidade de uso, e essa usabilidade foi avaliada utilizando o método SUS. Neste capítulo, abordaremos tanto os experimentos realizados como a avaliação de usabilidade.

6.1 Experimentos

Neste estudo, foram conduzidos dois experimentos utilizando o software proposto, cada um com bases de dados diferentes. O primeiro experimento utilizou a base de dados pública empregada por Cortez e Silva (CORTEZ; SILVA, 2008). Essa base de dados consiste em um conjunto de informações de alunos de escolas portuguesas, com classificações referentes às notas nas disciplinas de matemática e português. Ao longo deste estudo, essa base será mencionada como “Cortez e Silva”.

Por outro lado, o segundo experimento utilizou uma base de dados[1] também pública, porém fictícia. Essa base de dados fictícia foi criada com o propósito específico de conduzir experimentos de aprendizagem de máquina sobre as notas dos alunos. Para os fins deste estudo, essa base será denominada como “Exams”.

Com essas duas bases de dados, o software proposto foi testado e avaliado quanto à sua eficiência na realização da aprendizagem de máquina e classificação dos dados. A utilização de diferentes bases de dados proporcionou uma análise abrangente da capacidade do software em lidar com diferentes contextos e conjuntos de informações, permitindo uma compreensão mais completa de seu desempenho e aplicabilidade em cenários variados.

Podemos obter uma compreensão mais detalhada das diferenças entre as bases de dados ao examinar a Tabela 2. Nela, são apresentadas a origem dos dados, o número de atributos, a lista desses atributos e, por último, o número de instâncias da base. Como discutido com maior profundidade no Capítulo 4, um aspecto crucial do processo foi a sistemática divisão dos

[1] Exam Scores - A fictional dataset and should only be used for data science training purposes. <http://roycekimmons.com/tools/generated_data/exams>. Acessado em: 02 de abril de 2022.

atributos em grupos distintos. Essa divisão foi realizada com o objetivo principal de aprimorar a compreensão abrangente da base de dados, proporcionando uma visão mais estruturada e significativa. As bases de dados utilizadas são compostas por atributos agrupados nas seguintes categorias: Educação, Demografia/Social e Econômico.

Tabela 2 – Comparação das bases de dados utilizadas no experimento.

	Cortez e Silva	Exams
Origem dos dados	Instituições de ensino Portuguesa	Dados fictícios
Número de Atributos	32/33	8
Atributos	colégio, sexo, idade, endereço, número de membros na família, mora com a família, escolaridade da mãe, escolaridade do pai, trabalho da mãe, trabalho do pai, motivo de escolher esta escola, responsável do estudante, tempo de translado para escola, tempo de estudo por semana, número de reprovações, suporte extra escolar, suporte dos pais para estudar, faz reforço/monitoria escolar, realiza atividades extracurricular, frequentou creche, deseja fazer ensino superior, possui acesso à internet em casa, está em relacionamento romântico, qualidade do relacionamento familiar, tempo livre depois da escola, sai com os amigos, consome álcool em dias de semana, consome álcool em finais de semana, situação de saúde, número de faltas, nota do primeiro semestre, nota do segundo semestre, nota final.	gênero, etnia, nível de graduação dos pais, almoça na escola, curso preparatório para teste, pontuação matemática, pontuação de leitura, pontuação de redação
Número de Instâncias	649	1000

Fonte: Elaborado pelo autor.

Esses grupos não apenas organizam os atributos de maneira coerente, mas também desempenham um papel de destaque na mineração de dados educacionais. O grupo “Educação” abrange aspectos intrinsecamente relacionados ao ambiente de aprendizado e ao desempenho acadêmico dos indivíduos. Através desses atributos, é possível capturar informações cruciais sobre os percursos educacionais, identificar padrões de aprendizado e prever possíveis desafios, como evasão e retenção.

Por sua vez, o grupo “Demografia/Social” compreende uma ampla gama de dados

que oferecem informações fundamentais sobre o contexto social e familiar dos estudantes. Essa perspectiva é vital para entender como fatores externos, como a composição familiar, localização geográfica e acesso a recursos sociais, podem influenciar diretamente o sucesso educacional.

Já o grupo "Econômico" lança luz sobre a influência dos fatores econômicos no cenário educacional e familiar dos estudantes. Os dados econômicos podem abranger desde níveis de renda até indicadores macroeconômicos que impactam as oportunidades educacionais disponíveis para os indivíduos. A compreensão das interconexões entre esses aspectos econômicos e os resultados educacionais pode enriquecer significativamente a análise de dados.

Em síntese, a estruturação em grupos dos atributos de dados demonstra ser um componente-chave na exploração da mineração de dados educacionais. Esses grupos fornecem uma abordagem organizada e holística para entender as complexidades envolvidas nos processos educacionais, considerando não apenas o aspecto pedagógico, mas também as influências sociais e econômicas. Nesse sentido, a Tabela 3 apresenta um resumo dos dados em cada grupo na respectiva base de dados correspondente.

O experimento foi conduzido em um computador equipado com um processador i7-3537U CPU de 2.00GHz e 8GB de memória RAM. Conforme ilustrado no fluxograma da Figura 11, no decorrer deste experimento, o treinamento ocorre de maneira simultânea à interface gráfica. No entanto, foram realizados dois experimentos com cada base de dados. No primeiro, o treinamento ocorre de forma sequencial, ou seja, o segundo algoritmo só inicia após a conclusão do primeiro, e assim por diante. Enquanto no segundo experimento, a aprendizagem acontece de forma concorrente, permitindo que todos os algoritmos sejam treinados simultaneamente. Em todos os experimentos, a aprendizagem e os testes são realizados com o uso da validação cruzada, utilizando 10 subconjuntos.

6.1.1 Experimento com base de dados Cortez e Silva

Utilizando a base de dados de Cortez e Silva (CORTEZ; SILVA, 2008), foi possível realizar análises que resultaram nos valores apresentados na Tabela 4. Esta tabela exibe os resultados obtidos pela aplicação de diversos algoritmos de mineração de dados, juntamente com as métricas de desempenho correspondentes.

Na Tabela 4, é notável que os resultados variam entre os algoritmos utilizados. A métrica de acurácia, que reflete a precisão dos modelos, revela que a Floresta Aleatória alcançou o melhor desempenho, atingindo uma notável taxa de 97,4684%. Essa métrica é crucial, pois

Tabela 3 – Comparação das bases de dados utilizadas no experimento.

	Cortez e Silva	Exams
Educacional	colégio, tempo de estudo por semana, número de reprovações, suporte extra escolar, suporte dos pais para estudar, faz reforço/monitoria escolar, realiza atividades extracurricular, frequentou creche, número de faltas, nota do primeiro semestre, nota do segundo semestre, nota final.	curso preparatório para teste, pontuação matemática, pontuação de leitura, pontuação de redação
Demográfico/Social	sexo, idade, endereço, mora com a família, escolaridade da mãe, escolaridade do pai, motivo de escolher esta escola, responsável do estudante, tempo de translado para escola, deseja fazer ensino superior, possui acesso à internet em casa, está em relacionamento romântico, qualidade do relacionamento familiar, tempo livre depois da escola, sai com os amigos, consome álcool em dias de semana, consome álcool em finais de semana, situação de saúde,	gênero,etnia, nível de graduação dos pais, almoça na escola,
Econômico	trabalho da mãe, trabalho do pai,	

Fonte: Elaborado pelo autor.

avalia a capacidade dos modelos de fazer previsões precisas em relação aos dados de teste.

Além disso, a tabela destaca o tempo de treinamento de cada algoritmo. Nesse aspecto, observa-se que a Árvore de Decisão e o KNN se destacam por terem tempos de treinamento muito rápidos, indicando eficiência computacional. A Árvore de Decisão é ágil devido ao seu processo direto de divisão de dados, que envolve menos cálculos complexos. O KNN, por sua vez, tem um treinamento rápido porque memoriza o conjunto de treinamento, e o processo de teste envolve principalmente a identificação de vizinhos próximos, exigindo menos processamento.

Por outro lado, algoritmos como MLP e *Naive Bayes* apresentam tempos de treinamento mais longos. O MLP, com suas redes neurais de múltiplas camadas, exige cálculos iterativos e ajustes de parâmetros, tornando-o computacionalmente mais intensivo. Da mesma forma, embora o *Naive Bayes* seja considerado um algoritmo mais simples, ele ainda requer cálculos repetitivos de probabilidades condicionais, o que pode contribuir para tempos de treinamento

Tabela 4 – Resultado alcançado com a base de dados de Cortez e Silva.

Algoritmos	Métricas	Valores
Árvore de Decisão	Acurácia	96,962 %
	Tempo de treinamento	0,01 seg
	Erro médio absoluto	0,0456
Floresta Aleatória	**Acurácia**	**97,4684 %**
	Tempo de treinamento	**0,06 seg**
	Erro médio absoluto	**0,0961**
KNN	Acurácia	76,962 %
	Tempo de treinamento	0,00 seg
	Erro médio absoluto	0,02319
MLP	Acurácia	91,1392 %
	Tempo de treinamento	3,38 seg
	Erro médio absoluto	0,0947
Naive Bayes	Acurácia	92,4051 %
	Tempo de treinamento	0,00 seg
	Erro médio absoluto	0,0838

Fonte: Elaborado pelo autor.

mais prolongados.

Essa diferença de desempenho em termos de tempo de treinamento destaca a importância de considerar a eficiência computacional ao selecionar algoritmos para cenários onde o tempo de treinamento é um fator crucial.

O erro médio absoluto também é uma métrica relevante, pois mede a diferença média entre as previsões do modelo e os valores reais. Através da Tabela 4, é possível perceber que o KNN obteve o menor erro médio absoluto, indicando uma maior precisão em suas previsões.

Uma observação relevante é que o experimento com a base de dados de Cortez e Silva revelou que o algoritmo de Floresta Aleatória alcançou um desempenho superior em relação aos demais, tanto em termos de acurácia quanto de tempo de treinamento. Essa constatação tem implicações significativas para a escolha de algoritmos em cenários semelhantes, ressaltando a importância de considerar tanto a precisão quanto a eficiência computacional.

Além disso, é interessante notar que a utilização da lógica de aprendizado concorrente também impactou os resultados, resultando em um tempo ligeiramente menor em relação à abordagem anterior. Esse resultado sugere que, com bases de dados maiores ou mais complexas, essa estratégia de aprendizado concorrente pode oferecer vantagens ainda mais significativas.

Portanto, a análise dos resultados apresentados na Tabela 4 oferece uma visão abrangente das performances dos algoritmos utilizados, destacando a eficácia da Floresta Aleatória devido à sua natureza de conjunto e diversidade. A abordagem da Floresta Aleatória, que combina múltiplas árvores de decisão independentes e suas previsões, permite a captura de nuances e

padrões complexos nos dados. Essa metodologia reduz o risco de *overfitting* e melhora a generalização do modelo para novos dados, resultando em um desempenho notavelmente superior. Além disso, sua capacidade de selecionar características e amostras de maneira aleatória em cada árvore contribui para reduzir a variância e aumentar a estabilidade do modelo.

6.1.2 Experimento com base de dados Exams

Utilizando a base de dados Exams, foram obtidos os resultados apresentados na Tabela 5.

Tabela 5 – Resultado alcançado com a base de dados Exams.

Algoritmos	Métricas	Valores
Árvore de Decisão	Acurácia	97,9 %
	Tempo de treinamento	0,03 seg
	Erro médio absoluto	0,0247
Floresta Aleatória	**Acurácia**	**99 %**
	Tempo de treinamento	**0,32 seg**
	Erro médio absoluto	**0,0383**
KNN	Acurácia	88,7 %
	Tempo de treinamento	0,00 seg
	Erro médio absoluto	0,1139
MLP	**Acurácia**	**99 %**
	Tempo de treinamento	**1,59 seg**
	Erro médio absoluto	**0,0153**
Naive Bayes	Acurácia	97,6 %
	Tempo de treinamento	0,01 seg
	Erro médio absoluto	0,0496

Fonte: Elaborado pelo autor.

Na análise dos resultados da base de dados “Exams”, torna-se evidente que dois algoritmos se destacam ao atingir a mais alta acurácia. Tanto a Floresta Aleatória quanto o MLP obtiveram uma acurácia de 99%. No entanto, é notável que o MLP exigiu um tempo quase cinco vezes maior em comparação com a Floresta Aleatória. Importante destacar que os tempos apresentados referem-se à execução sequencial dos algoritmos, totalizando 1,95 segundos para a execução de todos os algoritmos. Realizando a aprendizagem dos algoritmos de forma simultânea, o tempo de execução foi reduzido para 1,61 segundos, resultando em uma economia de 0,34 segundos.

A análise dos resultados das duas bases de dados, conforme apresentado nas Tabelas 4 e 5, revela que diferentes algoritmos se destacam para cada conjunto de dados. Os algoritmos com resultados mais notáveis estão indicados em negrito. Além disso, observa-se que o tempo de treinamento para diferentes algoritmos não impõe um custo excessivamente alto, permitindo

que o software proposto explore a seleção do melhor algoritmo para a base de dados escolhida. Mesmo na base de dados que demandou mais tempo de treinamento, o processo foi concluído em 3,46 segundos.

6.2 Avaliação do SUS

Este trabalho tem como um dos objetivos específicos avaliar a usabilidade do software proposto por meio de um método simples e eficaz: o questionário SUS. A escolha desse método se deve à sua ampla utilização e comprovada eficácia na avaliação da usabilidade de sistemas e softwares.

O questionário SUS consiste em 10 perguntas cuidadosamente elaboradas para avaliar a facilidade de uso e a satisfação do usuário com o software em questão. Cada pergunta é avaliada em uma escala de concordância, variando de “discordo totalmente” a “concordo totalmente”. A pontuação total obtida a partir das respostas do usuário não é uma simples média aritmética, mas sim um cálculo ponderado que leva em consideração as características específicas do SUS. O questionário foi respondido por professores, coordenadores, responsáveis pelo setor de matrícula e psicologo que atuam na educação em escolas de três diferentes municípios do sertão pernambucano, Afogados da Ingazeira, Salgueiro e Mirandiba.

Além das 10 perguntas padrões do SUS, este estudo acrescentou duas questões adicionais que não dizem respeito diretamente à usabilidade do software, mas sim à dificuldade enfrentada pela gestão escolar ao obter os dados escolares. Essas questões foram incluídas para obter uma visão mais abrangente do contexto em que o software será utilizado e entender eventuais dificuldades externas que possam afetar a percepção da usabilidade pelos usuários.

A Tabela 6 apresenta as doze perguntas utilizadas no questionário. As 10 primeiras são as perguntas padrões do SUS, enquanto as duas últimas são as questões adicionais sobre a obtenção de dados escolares. Essas perguntas foram cuidadosamente selecionadas para fornecer informações valiosas sobre a usabilidade do software e sua integração com a gestão escolar.

Por meio desse questionário, busca-se obter informações significativos sobre a experiência do usuário com o software proposto e identificar possíveis pontos de melhoria na usabilidade. Ao avaliar a satisfação e a facilidade de uso do software.

Esta avaliação compreendeu um estudo tanto presencial como remoto, em que os entrevistados foram convidados a responder a um formulário digital nas pesquisas remotas e a um questionário em papel nas pesquisas presenciais. O público-alvo selecionado para o

Tabela 6 – Lista de perguntas utilizadas no questionário.

Número	Perguntas	Tipo de resposta
1	Eu acho que gostaria de usar essa aplicação com frequência.	1 até 5
2	Eu acho o sistema desnecessariamente complexo.	1 até 5
3	Eu achei o sistema fácil de usar.	1 até 5
4	Eu acho que precisaria de ajuda de uma pessoa com conhecimento técnicos para usar o sistema.	1 até 5
5	Eu acho que as várias funções do sistema estão muito bem integradas.	1 até 5
6	Eu acho que o sistema apresenta muitas inconsistências.	1 até 5
7	Eu imagino que as pessoas aprenderão a usar esse sistema rapidamente.	1 até 5
8	Eu achei o sistema complicado de usar.	1 até 5
9	Eu me senti confiante ao usar o sistema.	1 até 5
10	Eu precisei aprender várias coisas novas antes de usar o sistema.	1 até 5
11	Eu tenho acesso para extrair informações da base de dados do sistema de gestão escolar.	Sim ou não
12	A minha escola possui um técnico ou responsável no setor de Tecnologia da Informação.	Sim ou não

Fonte: Elaborado pelo autor.

questionário foi composto por professores, coordenadores de curso, coordenadores de registro acadêmico/matrícula e diretores de instituições educacionais.

Entretanto, durante as primeiras pesquisas, identificou-se uma dificuldade enfrentada por alguns entrevistados, que não possuíam acesso ou não sabiam como o sistema de gestão escolar disponibilizava os dados de aprendizagem de máquina. Diante disso, para facilitar o teste do software, uma base de dados fictícia foi disponibilizada em conjunto com o software, permitindo aos entrevistados a realização da avaliação mesmo sem acesso aos dados reais.

Os entrevistados tiveram a oportunidade de utilizar o software desenvolvido após acordarem com o Termo de Consentimento Livre e Esclarecido (TCLE) descrito no Apêndice A e, em seguida, responderam a um questionário contendo as doze questões mencionadas anteriormente. O período de coleta de dados através do questionário ocorreu ao longo de uma semana e não houve acompanhamento por parte do autor ou instruções específicas sobre como utilizar o software, uma vez que o objetivo do questionário era avaliar a facilidade de uso do sistema de forma autônoma e realista.

Após cada teste realizado, os usuários tiveram a liberdade de expressar suas apreciações, críticas e sugestões sobre o software, seja por meio de comentários escritos no papel ou por meio do formulário de resposta do questionário. Essa abertura para *feedbacks* proporcionou uma visão mais completa das percepções dos usuários em relação ao software e permitiu que o autor

do estudo coletasse informações e sugestões para melhorias no software, como foi o exemplo da funcionalidade de salvar o treinamento, citado no capítulo anterior.

O modelo de medição de usabilidade de sistemas SUS é uma proposta de Brooke (1986), consiste em avaliar efetividade, eficiência e satisfação do usuário, esta medição disponibiliza 10 questões fechadas para avaliar determinadas ações relacionadas ao sistema. Cada questão tem cinco possíveis respostas para o usuário, sendo uma variação de respostas de 1 a 5. Sendo que, o valor 1 está discordando completamente e o valor 5 concorda completamente. Para obter o resultado desta medição para uma única resposta, deve-se seguir as seguintes regras:

- Para todas as questões ímpares, ou seja, 1, 3, 5, 7 e 9, deve-se subtrair 1 da resposta do usuário;
- As perguntas pares, ou seja, 2, 4, 6, 8 e 10, deve-se subtrair a resposta do usuário de 5;
- Deve-se somar todas as respostas das 10 questões e multiplicar por 2,5.

Durante a pesquisa, foram realizadas entrevistas com 31 indivíduos provenientes de 6 instituições de ensino. Os resultados obtidos foram apresentados de forma gráfica para facilitar a compreensão e análise dos dados. A Figura 21 exibe o número de respostas referentes à pergunta 1: “Eu acho que gostaria de usar essa aplicação com frequência”.

Dos trinta e um entrevistados, dezesseis responderam com a nota máxima (5), concordando plenamente com a afirmação. Outras quatorze pessoas atribuíram a nota 4, demonstrando alto grau de concordância, e uma pessoa respondeu com a nota 3. Realizando os cálculos de pontuação: 0 + 0 + (3-1) + ((4-1)*14) + ((5-1)*16), chegamos a um total de 108 pontos para esta pergunta.

Após dividir o total de pontos pelo número de entrevistados (31), o *score* SUS para esta pergunta ficou em 3,48. Esse valor reflete a percepção geral dos entrevistados em relação à afirmação apresentada na pergunta 1. Vale ressaltar que quanto maior o *score* SUS, maior é a facilidade de uso percebida pelos usuários em relação à aplicação em questão. Dessa forma, o valor de 3,48 sugere que a maioria dos entrevistados avaliou positivamente a possibilidade de utilizar o software com frequência, indicando uma percepção favorável da usabilidade da aplicação.

A Figura 22 representa um gráfico das respostas referente a pergunta 2 “Eu acho o sistema desnecessariamente complexo”. Das trinta e uma pessoas pessoas entrevistadas, vinte e oito atribuíram a nota 1, divergindo com a pergunta e três pessoas responderam com a nota 2. Sendo assim, esta pergunta obteve 121 pontos através dos cálculos: ((5-1)*28) + ((5-2)*3) + 0 +

Figura 21 – Gráfico das respostas da Pergunta 1.

Fonte: Elaborado pelo autor.

0 . Sendo o seu *score* SUS de 3,90.

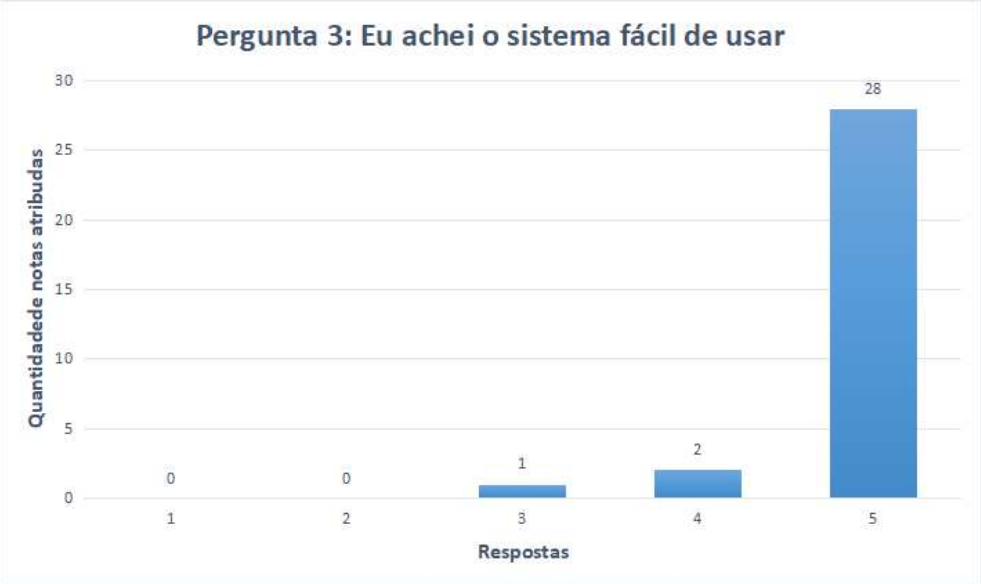
Figura 22 – Gráfico das respostas da Pergunta 2.

Fonte: Elaborado pelo autor.

O resultado da terceira pergunta “Eu achei o sistema fácil de usar” pode ser visualizado no gráfico da Figura 23. Nesta pergunta vinte e oito pessoas marcaram a resposta 5, duas pessoas marcaram a resposta 4 e uma pessoa a resposta 3. Sendo assim, a pontuação alcançada neste critério foi de 120. A maneira de alcançar este valor foi através do calculo 0 + 0 + (3-1) + ((4-1)*2) + ((5-1)*28). Nesta pergunta o *score* SUS é de 3,87 calculando.

A pergunta que trata do uso de pessoa com conhecimentos técnicos, a pergunta 4

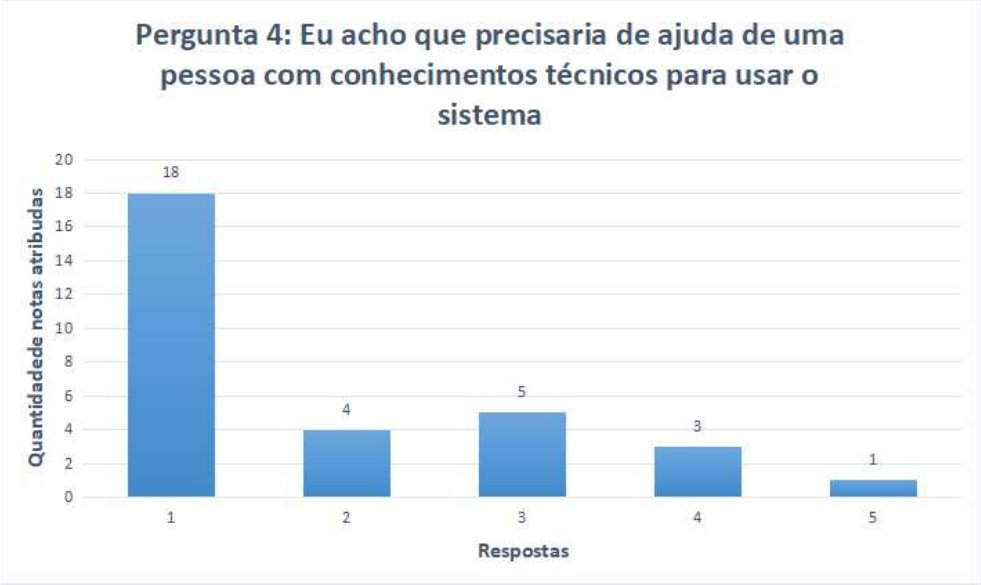
Figura 23 – Gráfico das respostas da Pergunta 3.



Fonte: Elaborado pelo autor.

”Eu acho que precisaria de ajuda de uma pessoa com conhecimento técnicos para usar o sistema”, foi a segunda pior nota. A distribuição de respostas foi de dezoito para a resposta 1, quatro para a resposta 2, cinco para a resposta 3 e uma para a resposta 5. Sendo o calculo da pontuação, ((5-1)*18) + ((5-2)*4) + ((5-3)*5) + ((5-4)*3) + 0, totalizando 97 pontos e o *score* SUS de 3,13.

Figura 24 – Gráfico das respostas da Pergunta 4.



Fonte: Elaborado pelo autor.

A Figura 25 expõe o número de respostas referente a pergunta 5 “Eu acho que as várias funções do sistema estão muito bem integradas.”. Vinte e nove deram a nota 5 e duas pessoas responderam com a nota 4. Realizando os cálculos 0 + 0 + 0 + ((4-1)*2) + ((5-1)*29),

esta pergunta alcançou a nota de 122 pontos. Após dividir pelo número de entrevistados, o *score* SUS para esta pergunta ficou de 3,93.

Figura 25 – Gráfico das respostas da Pergunta 5.

Fonte: Elaborado pelo autor.

Na Figura 26 é possível observar o gráfico das respostas referente a pergunta 6 “Eu acho que o sistema apresenta muitas inconsistências”. Esta foi a pergunta que alcançou o melhor resultado. Todos os entrevistados marcaram a resposta 1. Sendo assim, esta pergunta obteve 124 pontos através dos cálculos: ((5-1)*31) + 0 + 0 + 0 + 0 . Sendo o seu *score* SUS de 4.

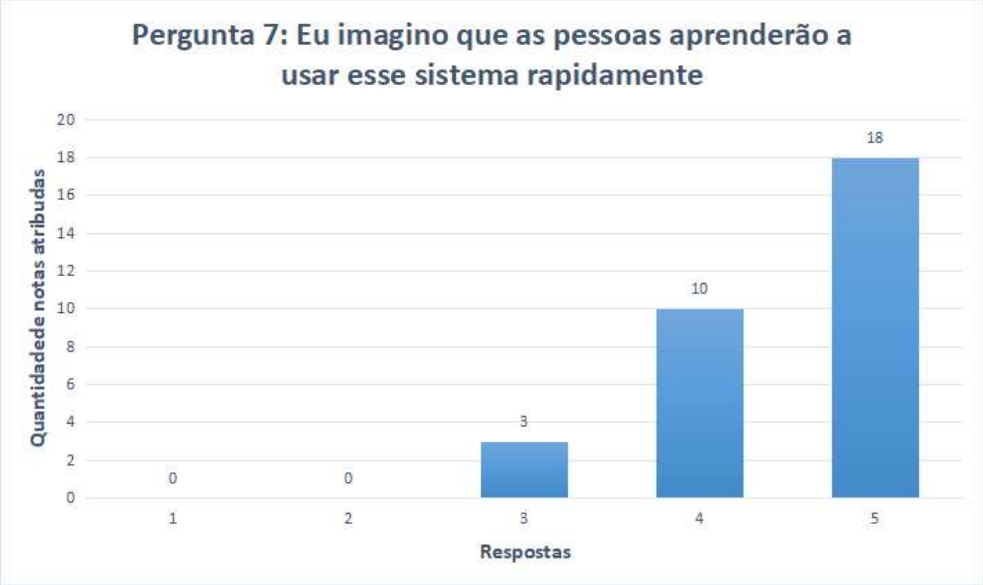
Figura 26 – Gráfico das respostas da Pergunta 6.

Fonte: Elaborado pelo autor.

O resultado da sétima pergunta “Eu imagino que as pessoas aprenderão a usar esse

sistema rapidamente" pode ser visualizado no gráfico da Figura 27. Nesta questão dezoito pessoas marcaram a resposta 5, dez pessoas marcaram a resposta 4 e três pessoas a resposta 3. Sendo assim, a pontuação alcançada neste critério foi de 108. A maneira de alcançar este valor foi através do calculo 0 + 0 + ((3-1)*3) + ((4-1)*10) + ((5-1)*18). Nesta pergunta o *score* SUS é de 3,48 calculando.

Figura 27 – Gráfico das respostas da Pergunta 7.



Fonte: Elaborado pelo autor.

A resposta da pergunta 8 "Eu achei o sistema complicado de usar", alcançou o segundo melhor resultado, o gráfico da Figura 29 apresenta as respostas dos entrevistados. Com um *score* SUS de 3,96. Nesta pergunta, trinta entrevistados marcaram a opção 1 e uma pessoa a opção 2. A pontuação total foi calculada da seguinte forma: ((5-1)*30) + (4-1) + 0 + 0 + 0 = 123.

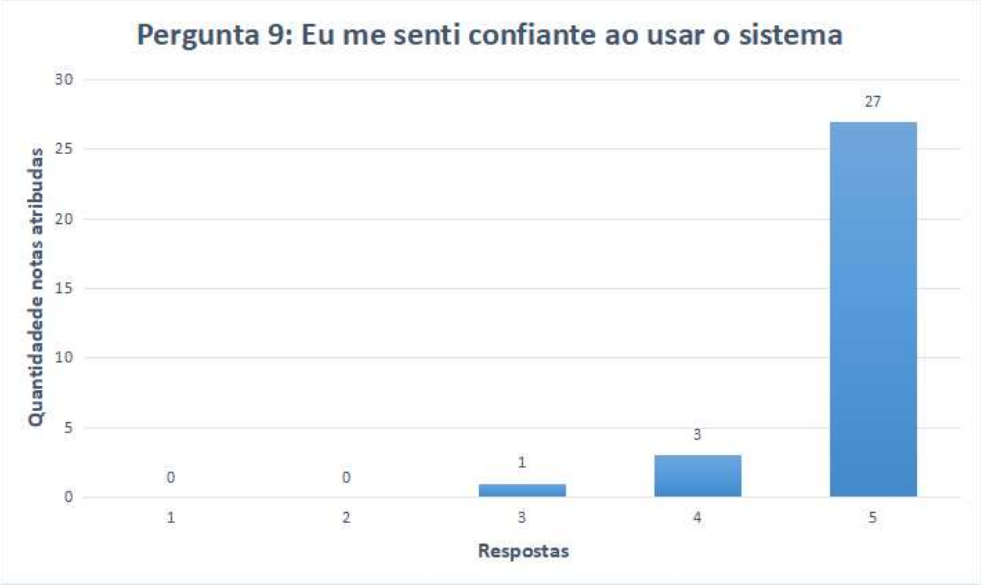
A Figura 29 expõe o número de respostas referente a pergunta 9 "Eu me senti confiante ao usar o sistema". Vinte e sete deram a nota 5, três atribuíram a nota 4 e um entrevistado a nota 3. Realizando os cálculos 0 + 0 + (3-1) + ((4-1)*3) + ((5-1)*27), esta pergunta alcançou a nota de 119 pontos. Após dividir pelo número de entrevistados, o *score* SUS para esta pergunta ficou de 3,83.

A ultima pergunta do formulário SUS foi a que alcançou o menor resultado. A pergunta "Eu precisei aprender várias coisas novas antes de usar o sistema" obteve um 88 pontos, sendo distribuído em dezesseis entrevistados para a resposta 1, sete entrevistados para a resposta 2, três entrevistados para a resposta 4 e cinco entrevistados para resposta 5. Seguindo o calculo de questões de número par: ((5-1)*16) + ((5-2)*7) + 0 + ((5-4)*3) + 0. O *score* SUS foi de 2,84.

Figura 28 – Gráfico das respostas da Pergunta 8.

Fonte: Elaborado pelo autor.

Figura 29 – Gráfico das respostas da Pergunta 9.

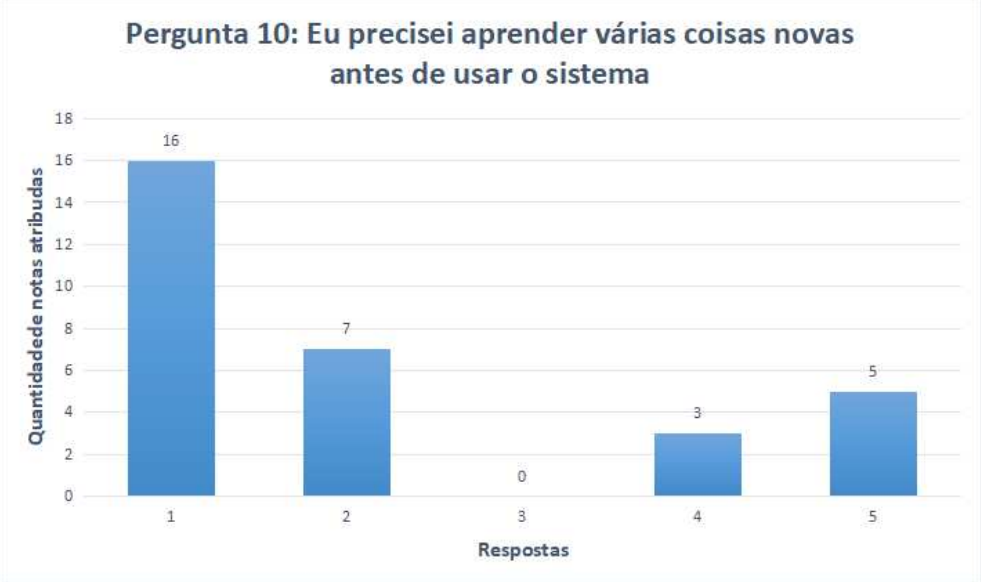


Fonte: Elaborado pelo autor.

Em dois formulários possui uma observações sobre a falta de conhecimento sobre inteligência artificial, o que provavelmente pode ser o motivo *score* baixo. O gráfico de respostas pode ser visto na Figura 30.

Ainda sobre a avaliação SUS, a Tabela 7 expõe uma síntese para auxiliar no entendimento e comparações de cada pergunta do formulário. A forma de calcular a pontuação foi de realizar o somatório da multiplicação da pontuação da resposta pelo número de repetições de resposta neste item. Posteriormente é realizado a soma da pontuação de todas as questões e dividido pelo número de respostas. Para alcançar o *score* SUS é realizado a multiplicação por

Figura 30 – Gráfico das respostas da Pergunta 10.



Fonte: Elaborado pelo autor.

2,5. Assim, o *score* SUS deste produto de software é de 91,125, por fim, segundo a proposta de Brooke (1986) o software possui classificação SUS de A. Demonstrando através da avaliação do método SUS exposta na Tabela 7 que o software possui uma boa usabilidade.

Tabela 7 – Tabela com o resultado da pesquisa SUS

	1	2	3	4	5	Calculo da Pontuação
P1	0	0	1	14	16	0 + 0 + (3-1)+ ((4-1)*14) + ((5-1)*16) = 108
P2	28	3	0	0	0	((5-1)*28) + ((5-2)*3) + 0 + 0 = 121
P3	0	0	1	2	28	0 + 0 + (3-1) + ((4-1)*2) + ((5-1)*28) = 120
P4	18	4	5	3	1	((5-1)*18) + ((5-2)*4) + ((5-3)*5) + ((5-4)*3) + 0 = 97
P5	0	0	0	2	29	0 + 0 + 0 + ((4-1)*2) + ((5-1)*29) = 122
P6	31	0	0	0	0	((5-1)*31) + 0 + 0 + 0 + 0 = 124
P7	0	0	3	10	18	0 + 0 + ((3-1)*3) + ((4-1)*10) + ((5-1)*18) = 108
P8	30	1	0	0	0	((5-1)*30) + (4-1) + 0 + 0 + 0 = 123
P9	0	0	1	3	27	0 + 0 + (3-1) + ((4-1)*3) + ((5-1)*27) = 119
P10	16	7	0	3	5	((5-1)*16) + ((5-2)*7) + 0 + ((5-4)*3) + 0 = 88
Total						(108 + 121 + 120 + 97 + 122 + 124 + 108 + 123 + 119 + 88)/31 = 36,45
Score SUS						36,45 * 2,5 = 91,125
Classificação SUS						A

Fonte: Elaborado pelo autor.

Como simples fator de comparação entre valores do *score* do SUS, o software de prevenção de tromboembolismo venoso proposto por Toledo *et al.* (2022) obteve *score* do SUS de 79,5 pontos. Já o software proposto por Nascimento *et al.* (2022) que auxilia no inclusão entre surdos e ouvintes obteve um *score* do SUS de 74,2.

Além das dez perguntas padrão do SUS, foram incluídas mais duas perguntas adicionais, apresentadas como pergunta 11 e pergunta 12 na Tabela 6. O propósito dessas duas

perguntas adicionais é verificar a possibilidade dos profissionais da área da educação terem o conhecimento e acesso necessários para extrair a base de dados do seu sistema de gestão escolar. Essa capacidade é fundamental para que eles possam utilizar seus próprios dados locais para realizar a aprendizagem de máquina com o software proposto.

A primeira pergunta visa identificar se os profissionais têm o conhecimento técnico e acesso adequado para extrair a base de dados do sistema de gestão escolar. Essa informação é relevante para compreender se o software será viável e útil para a instituição, uma vez que a utilização de dados locais é uma premissa importante para a aplicação efetiva da aprendizagem de máquina em contexto educacional.

Já a segunda pergunta tem o objetivo de identificar a possibilidade da instituição possuir algum técnico ou profissional de TI que possa auxiliar no uso da ferramenta proposta e também na extração dos dados do sistema de gestão escolar caso o usuário não tenha permissão. Essa pergunta destaca a importância de o software ser de fácil uso e de fácil instalação, considerando que nem todos os usuários têm habilidades avançadas em tecnologia.

Sendo assim, a Figura 31 mostra o gráfico com as respostas referente as perguntas 11 e 12. Referente a pergunta 11 “Eu tenho acesso para extrair informações da base de dados do sistema de gestão escolar”, pode-se perceber que 58% dos entrevistados conseguem pegar os dados dos alunos. Ainda sobre essa questão, vale destacar dois pontos escritos no campo de observação do formulário: A escola do entrevistado não possui sistema de gestão escolar digital, sendo que esta gestão é feita manualmente através de formulários físicos; O sistema de gestão escolar não possui a funcionalidade de exportar os dados dos alunos. Sendo este software uma ferramenta privada que até o momento da entrevista o gestor escolar não possuía acesso a base de dados. A partir destas observações dos entrevistados pode-se destacar dois pontos: algumas funcionalidades futuras podem ser desenvolvidas, como por exemplo acessar diretamente a base de dados do sistema de gestão escolar; a falta de informações nas instituições tanto pública como privadas de softwares livres de gestão escolar, como por exemplo o I-Educar, na qual Júnior (2019) e Speck *et al.* (2018) realizaram o trabalho de implantar e treinarem os usuários.

Já a pergunta 12 “A minha escola possui um técnico ou responsável no setor de Tecnologia da Informação”, a intenção desta pergunta é de como o software pode ser distribuído, visto que, nem todas as pessoas dominam as técnicas de TI. Na Figura 31 pode-se ver que 83% dos entrevistados trabalham em instituições que não possuem profissionais de TI, sendo assim, o software deve ser o mais simples possível de instalar e utilizar. Assim, o tipo de software

Figura 31 – Gráfico das respostas das Perguntas 11 e 12.

Fonte: Elaborado pelo autor.

desenvolvido que não tem necessidade de instalação, facilitando para usuários que não dominam esses conhecimentos. Portanto, é importante reforçar a necessidade de que o software seja fácil de usar e de instalar.

7 CONCLUSÃO

Neste trabalho, buscamos atingir uma série de objetivos específicos, que orientaram nossa pesquisa e análise ao longo dos capítulos. Recapitulando esses objetivos e indicando onde cada um deles foi abordado:

a) realizar um levantamento do estado da arte em busca de soluções voltadas para a redução da evasão e retenção escolar;
b) analisar as soluções identificadas no objetivo anterior, destacando suas contribuições científicas e as técnicas que empregam;
c) desenvolver um software que permita aos gestores, com base nos dados de sua própria instituição e alunos, realizar previsões de evasão e retenção escolar;
d) realizar experimentos com o software proposto para avaliar sua eficácia; e
e) conduzir uma análise de usabilidade do software junto ao público-alvo, incluindo professores, gestores e equipes multidisciplinares de instituições de ensino.

Sobre o objetivo a) e o objetivo b), no Capítulo 4, foi apresentada uma RSL, na qual foi utilizada para investigar a situação do estado da arte sobre o tema deste trabalho. O software proposto no objetivo específico c) foi apresentado no Capítulo 5. Quanto ao objetivo específico d), na Seção 6.1, foram apresentados os resultados do experimento com duas diferentes bases de dados. O objetivo específico e) foi alcançado e apresentado na Seção 6.2 com os resultados do teste de usabilidade realizado por meio do método SUS. Por fim, a hipótese apresentada na Seção 1.2, “Soluções no âmbito da mineração de dados educacionais podem indicar situações de evasão e retenção escolar”, pode ser confirmada por meio dos resultados do experimento e da usabilidade do produto de software apresentados anteriormente.

Quanto as perguntas expostas na Seção 1.3 do Capítulo 1:

1) É possível fazer uso de MDE para auxiliar na permanência e êxito escolar?
2) Quais são os principais problemas para a utilização de técnicas de MDE nas instituições de ensino?

Após a conclusão deste trabalho, alcançamos as respostas para as duas perguntas citadas anteriormente. A pergunta “1)” pode ser respondida pelo produto de software, juntamente com as validações do experimento e da usabilidade através do método SUS. Enquanto a resposta à pergunta “2)” foi obtida graças às duas questões utilizadas no formulário, em conjunto com o SUS. Sendo assim, as maiores dificuldades observadas são a falta de profissionais técnicos na área de TI nas instituições de ensino, o que dificulta a extração de dados do sistema de gestão

escolar digital. Em outros casos, a instituição de ensino não possui um sistema de gerenciamento digital, continuando a utilizar métodos de registro acadêmico baseados em formulários físicos.

Em resumo, este trabalho apresentou a construção de um software para mineração de dados educacionais, visando a predição de evasão e retenção escolar. Através de uma RSL, foi possível identificar as principais abordagens e algoritmos utilizados em estudos relacionados ao tema.

O software proposto foi desenvolvido utilizando a linguagem de programação Java, com foco na facilidade de uso e acesso para equipes multidisciplinares nas instituições de ensino. Os experimentos realizados com duas bases de dados distintas demonstraram que o software alcançou resultados promissores, com uma boa acurácia na predição dos casos de evasão e retenção.

A avaliação de usabilidade com o método SUS confirmou que o software obteve uma boa aceitação pelos usuários, indicando que a interface gráfica e a experiência do usuário foram bem-sucedidas. No entanto, durante o processo de avaliação, foi observado que algumas instituições podem enfrentar dificuldades para extrair informações da base de dados. Essas dificuldades podem ser atribuídas à falta de conhecimento para realizar a extração dos dados do sistema de gestão escolar ou à ausência de técnicos em Tecnologia da Informação para prestar auxílio nesse procedimento. Em situações mais desafiadoras, algumas instituições ainda não possuem um sistema de gestão escolar digital estabelecido.

Diante do exposto, uma possível proposta de pesquisa futura seria a realização de um estudo que analisasse e oferecesse suporte à implantação de softwares livres de gestão escolar em escolas privadas e públicas. Além de trazer benefícios para a modernização e eficiência do ambiente educacional, a adoção de um software público poderia ser vantajosa para a gestão e também para os cofres públicos. Ao optar por um software livre, as instituições educacionais eliminariam a dependência de empresas terceirizadas, reduzindo custos com licenciamento e suporte técnico. Além disso, a centralização em um servidor público para o funcionamento do sistema tornaria mais viável a sua manutenção e garantiria maior autonomia sobre os dados sensíveis das escolas. Essa transição para o software livre poderia ser planejada de forma estratégica, considerando as necessidades específicas de cada escola ou município, resultando em uma gestão mais eficiente e econômica.

É importante ressaltar que este trabalho apresenta algumas limitações, como a restrição a bases de dados específicas e a dependência do uso da plataforma Weka para a

aplicação dos algoritmos de mineração de dados. Além disso, a disponibilidade de dados e o acesso a sistemas de gestão escolar podem variar entre as instituições.

Como sugestões para pesquisas futuras, destacamos a ampliação do escopo do software para suportar diferentes plataformas de mineração de dados, a inclusão de mais algoritmos de aprendizado de máquina e a criação de uma interface para integração direta com diferentes sistemas de gestão escolar. Além disso, recomenda-se a realização de estudos adicionais para analisar o impacto do uso do software nas decisões e políticas educacionais das instituições de ensino.

Na conclusão desta pesquisa é fundamental reconhecer algumas limitações que afetaram as descobertas e análises. Primeiramente, as limitações dos dados desempenharam um papel crítico, com a qualidade e disponibilidade dos dados podendo impactar a precisão do software proposto. Outra consideração importante diz respeito ao viés nos dados, o que pode levar a resultados enviesados e questões éticas relacionadas à privacidade dos alunos, também importante citar a Lei Geral de Proteção de Dados (LGPD) que trata da proteção de dados pessoais e da privacidade dos indivíduos.

Também devemos ter em mente que a escolha dos algoritmos e hiperparâmetros pode influenciar significativamente os resultados, e a interpretabilidade de modelos complexos de IA pode ser um desafio. A inferência de causalidade a partir de correlações também é uma preocupação, e é importante evitar conclusões precipitadas nesse sentido. Como pesquisa futura, pode-se realizar um ajuste no software para que o mesmo possa fazer busca alterando os hiperparâmetros e encontrando a melhor solução para o problema.

Em suma, este trabalho representa um passo inicial na aplicação da mineração de dados educacionais para predição de evasão e retenção escolar. Esperamos que o software proposto possa contribuir para melhorar a experiência educacional e a tomada de decisões nas instituições de ensino, visando a redução dos índices de evasão e retenção de alunos.

REFERÊNCIAS

ABDI, H.; VALENTIN, D.; EDELMAN, B. **Neural Networks**. Thousand Oaks, CA: Sage Publications, Inc., 1999. ISBN 9780761914402.

AGUIAR, M. A. M. Complexidade e caos, por h. moysés nussenszveig (organizador). **Revista Brasileira de Ensino de Física (RBEF)**, v. 22, n. 2, p. 148–148, jun. 2000. ISSN 1806-9126.

ANDRADE, F.; ESQUINCALHA, A.; OLIVEIRA, A. T. O pré-cálculo nas licenciaturas em matemática das instituições públicas do rio de janeiro: o prescrito. **Revista Vidya**, v. 39, n. 1, p. 131–151, 2019.

ANDRADE, M. O. Mestrado em Gestão de Políticas Públicas e Segurança Social. **Evasão Escolar na Educação de Jovens e Adultos: um estudo a partir da Escola Monsenhor Gilberto Vaz Sampaio I - Varzedo/BA**. Cruz das Almas, BA: [s.n.], 2016.

ARAUJO, E. O. **Sistema de Mineração de Dados para Apoiar a Tomada de Decisão em uma Instituição de Ensino Superior: o problema da evasão escolar no IFTM**. 97 f. Dissertação (Mestrado em Assesoria de Administração) — Instituto Superior de Contabilidade e Administração do Porto, Instituto Politécnico do Porto, Porto, 2018.

BASTOS, D.; NASCIMENTO, P.; LAURETTO, M. Proposta e análise de desempenho de dois métodos de seleção de características para random forests. SBC, Porto Alegre, RS, Brasil, p. 49–60, 2013. Disponível em: <https://sol.sbc.org.br/index.php/sbsi/article/view/5675>.

BERRAR, D. Bayes' theorem and naive bayes classifier. In: RANGANATHAN, S.; GRIBSKOV, M.; NAKAI, K.; SCHÖNBACH, C. (Ed.). **Encyclopedia of Bioinformatics and Computational Biology**. Oxford: Academic Press, 2019. p. 403–412. ISBN 978-0-12-811432-2. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128096338204731>.

BROCK, C.; SCHWARTZMAN, S. **Os desafios da educação no Brasil**. Rio de Janeiro: Nova Fronteira, 2005. ISBN 978-8520917053.

BROOKE, J. System usability scale (sus): a quick-and-dirty method of system evaluation user information. **Reading, UK: Digital equipment co ltd**, v. 43, p. 1–7, 1986.

BROOKE, J. Sus-a quick and dirty usability scale. **Usability evaluation in industry**, London, England, v. 189, n. 194, p. 4–7, 1996.

CALIXTO, K.; SEGUNDO, C.; GUSMÃO, R. Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. **Simpósio Brasileiro de Informática na Educação (SBIE)**, v. 28, n. 1, p. 1447–1456, out. 2017. ISSN 2316-6533. Disponível em: <http://ojs.sector3.com.br/index.php/sbie/article/view/7674>.

CARRANO, D. P.; BERTASSI, A. L.; MELO-SILVA, G. Efetividade do pnaes enquanto política pública do estado para o combate à evasão universitária na ufsj. **Educação Online**, v. 13, n. 28, p. 1–19, aug. 2018. Disponível em: <https://doi.org/10.36556/eol.v13i28.417>.

CARVALHO, M.; OLIVEIRA, T. Infraestrutura de redes e dos laboratórios de informática de escolas públicas de conselheiro lafaiete. **Revista UFG (RUFG)**, v. 19, dez. 2019. Disponível em: <https://revistas.ufg.br/revistaufg/article/view/60605>.

CECHINEL, C.; ARAUJO, R. M.; DETONI, D. Modelagem e predição de reprovação de acadêmicos de cursos de educação a distância a partir da contagem de interações. **Revista Brasileira de Informática na Educação**, v. 23, n. 03, p. 1, 2015.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0: Step-by-step Data Mining Guide**. [S.l.]: SPSS, 2000.

CHIQUITTO, A. G.; BAIDA, A. C. Análise quantitativa das causas da evasão escolar dos cursos técnicos de nível médio integrado dos institutos federais de educação, ciência e tecnologia. **Encontro Internacional de Gestão, Desenvolvimento e Inovação (EIGEDIN)**, v. 4, n. 1, out. 2020. Disponível em: <https://periodicos.ufms.br/index.php/EIGEDIN/article/view/10934>.

COLPANI, R. Mineração de dados educacionais: um estudo da evasão no ensino médio com base nos indicadores do censo escolar. **Informática na educação: teoria & prática**, v. 21, n. 3, dez. 2018. Disponível em: <https://seer.ufrgs.br/index.php/InfEducTeoriaPratica/article/view/87880>.

CORMEN, T. H. **Desmistificando algoritmos**. 1a. ed. Rio de Janeiro: Elsevier, 2013. ISBN 978-8535271775.

CORTEZ, P.; SILVA, A. M. G. Using data mining to predict secondary school student performance. In: **5th Annual Future Business Technology Conference**. Porto: EUROSIS, 2008. p. 5–12. ISBN 978-9077381-39-7.

COUTINHO, E.; BEZERRA, J.; BEZERRA, C.; MOREIRA, L. Uma análise da evasão em cursos de graduação apoiado por métricas e visualização de dados. In: **Anais do XXIV Workshop de Informática na Escola**. Porto Alegre, RS, Brasil: SBC, 2018. p. 31–40. Disponível em: <https://sol.sbc.org.br/index.php/wie/article/view/14314>.

COUTO, D. C. Mestrado em Engenharia Elétrica. **Mineração de dados educacionais aplicada à busca de perfis de alunos em casos de evasão ou retenção: uma abordagem através de Redes Bayesianas**. Belém, PA: [s.n.], 2017. Disponível em: <http://repositorio.ufpa.br:8080/jspui/handle/2011/9463>.

DATE, C. J. **Introdução a Sistemas de Bancos de Dados**. 8a. ed. Rio de Janeiro: Elsevier, 2004. ISBN 978-8535212730.

DHARMAWAN, T.; GINARDI, H.; MUNIF, A. Dropout detection using non-academic data. In: **2018 4th International Conference on Science and Technology (ICST)**. Yogyakarta, Indonesia: IEEE, 2018. p. 1–4. ISBN 978-1-5386-5813-0.

EHSANI-MOGHADDAM, B.; QUEENAN, J. A.; MACKENZIE, J.; BIRTWHISTLE, R. V. Mucopolysaccharidosis type ii detection by naïve bayes classifier: An example of patient classification for a rare disease using electronic medical records from the canadian primary care sentinel surveillance network. **PLOS ONE**, Public Library of Science, v. 13, n. 12, p. 1–17, dec. 2018. Disponível em: <https://doi.org/10.1371/journal.pone.0209018>.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de Banco de Dados**. 7a. ed. São Paulo: Pearson Education do Brasil, 2019. ISBN 9788543025001.

FREIRE, R. R. B.; SILVA, E. V.; SOUZA, R. A. L.; VIEIRA, S. C. A realidade dos laboratórios de informática nas escolas públicas de maués: um estudo de caso. **Brazilian Journal of Development**, v. 7, n. 1, p. 3847–3858, 2021. ISSN 2525-8761.

GIL, A. C. **Como Elaborar Projetos de Pesquisa**. 4a. ed. São Paulo: Atlas, 2002. ISBN 85-224-3169-8.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: The MIT press, 2016. 800 p. ISBN 978-0262035613.

GUO, G.; WANG, H.; BELL, D.; BI, Y.; GREER, K. Knn model-based approach in classification. Springer Berlin Heidelberg, Berlin, Heidelberg, p. 986–996, 2003.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd. ed. New York: Morgan kaufmann, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Random forests. In: ____. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. New York, NY: Springer New York, 2009. p. 587–604. ISBN 978-0-387-84858-7. Disponível em: <https://doi.org/10.1007/978-0-387-84858-7_15>.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. 2a. ed. [S.l.]: Bookman, 2001. ISBN 9788577800865.

JÚNIOR, J. B. L. **Implantação e treinamento para o uso do sistema de gestão escolar i-Educar**. 27 f. Monografia (Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação)) — Unidade Acadêmica de Garanhuns, Universidade Federal Rural de Pernambuco, Garanhuns, 2019.

KITCHENHAM, B. **Procedures for Performing Systematic Reviews**. Department of Computer Science, Keele University, UK, 2004.

KITCHENHAM, B.; BRERETON, O. P.; BUDGEN, D.; TURNER, M.; BAILEY, J.; LINKMAN, S. Systematic literature reviews in software engineering–a systematic literature review. **Information and software technology**, Elsevier, v. 51, n. 1, p. 7–15, 2009.

KOVÁCS, Z. L. **Redes neurais artificiais**. [S.l.]: Editora Livraria da Fisica, 2002.

LAISA, J.; NUNES, I. Mineração de dados educacionais como apoio para a classificação de alunos do ensino médio. v. 26, n. 1, p. 1112, 2015.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Springer Nature Limited, v. 521, p. 436–444, 2015. Disponível em: <https://doi.org/10.1038/nature14539>.

LIMA, P.; BISINOTO, C.; MELO, N. S. d.; RABELO, M. Taxas longitudinais de retenção e evasão: uma metodologia para estudo da trajetória dos estudantes na educação superior. **Ensaio: Avaliação e Políticas Públicas em Educação**, Fundação CESGRANRIO, v. 27, n. 102, p. 157–178, Jan 2019. ISSN 0104-4036. Disponível em: <https://doi.org/10.1590/S0104-40362018002701431>.

MANHÃES, L.; CRUZ, S. da; COSTA, R. M.; ZAVALETA, J.; ZIMBRÃO, G. Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, v. 1, n. 1, 2012. ISSN 2316-6533. Disponível em: <http://milanesa.ime.usp.br/rbie/index.php/sbie/article/view/1585>.

MARCONI, M. d. A.; LAKATOS, E. M. **Fundamentos de Metodologia Científica**. 5a. ed. São Paulo: Atlas, 2003. ISBN 85-224-3397-6.

MARQUES, L. T.; CASTRO, A. F. D.; MARQUES, B. T.; SILVA, J. C. P.; QUEIROZ, P. G. G. Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um mapeamento sistemático da literatura. **Revista Novas Tecnologias na Educação (RENOTE)**, v. 17, n. 3, p. 194–203, dez. 2019. ISSN 1679-1916. Disponível em: <https://seer.ufrgs.br/index.php/renote/article/view/99470>.

MARTINHO, V. R. de C. Doutorado em Engenharia Elétrica. **Sistema inteligente para a predição de grupo de risco de evasão discente**. Ilha Solteira, SP: [s.n.], 2014. Disponível em: <https://hdl.handle.net/11449/100340>.

MEC. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. In: **Ministério da Educação - Comissão Especial de Estudos sobrea Evasão nas Universidades Públicas Brasileiras**. Brasília, DF, Brasil: [s.n.], 1997.

MEHRA, M.; KALBANDE, D. R.; MANKAR, S.; MUTSADDI, S. Data mining in educational systems for effective student mentoring. In: **2019 International Conference on Advances in Computing, Communication and Control (ICAC3)**. Mumbai, India: IEEE, 2019. p. 1–5.

NASCIMENTO, G.; VIEIRA, A.; MOURÃO, A.; FERNANDES, L. C.; MARTINS, J.; VIEIRA, N. A.; SCHULTZ, H. Pensamento computacional na concepção de estratégias para recurso de inclusão para comunicação entre surdos e ouvintes. In: **Anais do I Workshop de Pensamento Computacional e Inclusão**. Porto Alegre, RS, Brasil: SBC, 2022. p. 107–116. Disponível em: <https://sol.sbc.org.br/index.php/wpci/article/view/22564>.

NASCIMENTO, R. L. S. do; JUNIOR, G. G. da C.; FAGUNDES, R. A. de A. Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. **RENOTE**, v. 16, n. 1, 2018.

NORVIG, P.; RUSSELL, S. **Inteligência artificial: Tradução da 3a Edição**. 3a. ed. Rio de Janeiro: Elsevier, 2013. ISBN 978-85-352-3701-6.

PEREIRA, R. T.; ZAMBRANO, J. C. Application of decision trees for detection of student dropout profiles. Cancun, Mexico, p. 528–531, 2017.

PERTIWI, A. G.; WIDYANINGTYAS, T.; PUJIANTO, U. Classification of province based on dropout rate using c4.5 algorithm. In: **2017 International Conference on Sustainable Information Engineering and Technology (SIET)**. Malang, Indonesia: IEEE, 2017. p. 410–413.

RUMELHART, D. E.; MCCLELLAND, J. L. **Learning Internal Representations by Error Propagation**. [S.l.]: MIT Press, 1987. 318-362 p. ISBN 9780262291408.

SAMPAIO, J. C.; SILVA, K. S. P. d. Evasão na licenciatura em matemática: desafios e ações. **Brazilian Journal of Development**, v. 5, n. 12, p. 31096–31106, dez. 2019. Disponível em: <https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/5442>.

SANTOS, I. M.; FERNANDES, A. S. **EVASÃO UNIVERSITÁRIA: APLICANDO A MINERAÇÃO DE DADOS PARA A PREDIÇÃO DE UM PERFIL DE RISCO**. 21 f. Monografia (Bacharelado em Engenharia de Software) — Universidade Evangélica de Goiás (UniEVANGÉLICA), Anápolis, 2022.

SILVA, D. V. S. **Análise da qualidade de revisões sistemáticas em engenharia de software empírica**. 102 f. Dissertação (Mestrado em Ciência de Computação) — Pós-Graduação em Ciência da Computação, Centro de Informática, Universidade Federal de Pernambuco, Recife, 2015.

SILVA, F. A. **Um estudo sobre as causas da evasão escolar no ensino médio noturno de um colégio localizado no município de Paranavaí**. 45 f. Monografia (Especialização em Educação: Métodos e Técnicas de Ensino) — Universidade Tecnológica Federal do Paraná, Medianeira, 2018.

SILVA, F. I. C. d.; RODRIGUES, J. D. P.; BRITO, A. K. A.; FRANçA, N. M. d. Evasão escolar no curso de educação física da universidade federal do piauí. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, Publicação da Rede de Avaliação Institucional da Educação Superior (RAIES), da Universidade Estadual de Campinas (UNICAMP) e da Universidade de Sorocaba (UNISO), v. 17, n. 2, p. 391–404, Jul 2012. ISSN 1414-4077. Disponível em: <https://doi.org/10.1590/S1414-40772012000200006>.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados - Com Aplicações em R**. 1a.. ed. [S.l.]: GEN LTC, 2016. ISBN 9788535284478.

SILVEIRA, R. da F.; HOLANDA, M.; VICTORINO, M. de C.; LADEIRA, M. Educational data mining: Analysis of drop out of engineering majors at the unb - brazil. In: **2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)**. Boca Raton, FL, USA: IEEE, 2019. p. 259–262.

SOMMERVILLE, I. **Engenharia De Software**. 10a. ed. [S.l.]: Pearson Universidades, 2019. ISBN 9788543024974.

SPECK, R. A.; SCHREINER, M. A.; SOARES, J. P. R. d. S.; SILVA, L. B. d.; LENHART, G. A. A gestão educacional com o i-educar: análise da viabilidade de implantação no município de palotina – paraná. **Semina: Ciências Sociais e Humanas**, PePSIC, Londrina, v. 39, p. 65–74, jun. 2018. ISSN 1676-5443.

SUPERBE, A. R.; SILVA, R. I. A. Uso de redes neurais artificiais na predição de risco de evasão acadêmica. **Revista Terra & Cultura: Cadernos de Ensino e Pesquisa**, v. 34, n. esp., p. 160–166, 2018. ISSN 2596-2809. Disponível em: <http://periodicos.unifil.br/index.php/Revistateste/article/view/318>.

TANENBAUM, A. S.; AUSTIN, T. **Organização estruturada de computadores**. 6a. ed. São Paulo: Pearson Prentice Hall, 2013.

TENPIPAT, W.; AKKARAJITSAKUL, K. Student dropout prediction: A kmutt case study. In: **2020 1st International Conference on Big Data Analytics and Practices (IBDAP)**. Bangkok, Thailand: IEEE, 2020. p. 1–5.

TOLEDO, T. R. d. O.; PERES, A. L.; BARROS, P. E. S.; RUSSO, R. C.; CARVALHO, L. W. T. d. Prevtev: construção e validação de aplicativo móvel para orientações sobre tromboembolismo venoso. **Revista Brasileira de Educação Médica**, Associação Brasileira de Educação Médica, v. 46, n. 1, p. e032, 2022. ISSN 0100-5502. Disponível em: <https://doi.org/10.1590/1981-5271v46.1-20210405>.

WITTEN, I.; FRANK, E.; TRIGG, L.; HALL, M.; HOLMES, G.; CUNNINGHAM, S. Weka: Practical machine learning tools and techniques with java implementations. Hamilton, New Zealand, p. 1–4, 1999. ISSN 1170-487X.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data Mining: Practical Machine Learning Tools and Techniques**. 4a. ed. New York: Morgan Kaufmann, 2016. ISBN 9780128042915.

APÊNDICE A – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

Você está sendo convidado por Hélder Antero Amaral Nunes, aluno (a) do Mestrado do Programa de Pós-Graduação em Tecnologia Educacional (PPGTE) da Universidade Federal do Ceará (UFC), para participar de uma pesquisa. Leia atentamente as informações abaixo e tire suas dúvidas, para que todos os procedimentos possam ser esclarecidos.

A pesquisa com título "Mineração de dados socioeconômicos e educacionais de discentes para predição de evasão e retenção escolar" tem como objetivo Propor software de mineração de dados para prever evasão e retenção escolar. Dessa forma, a sua participação poderá trazer como benefícios validar a usabilidade do software proposto.

Para a sua realização, preciso que professores, técnicos administrativos escolares e gestores respondam a este questionário, ressaltando-se que a sua colaboração é de caráter voluntário e não implica em remuneração. Há o risco de você sentir-se constrangido com alguma pergunta, e caso isto ocorra, poderá a qualquer momento interromper a pesquisa e se for de sua vontade encerrar sua participação.

O questionário possui perguntas simples e deve tomar aproximadamente 20 minutos (o tempo de aplicação dependerá da pesquisa) do seu tempo. Os seguintes procedimentos serão respeitados:

1. Seus dados pessoais e outras informações que possam identificar você serão mantidos em segredo;
2. Você está livre para interromper a qualquer momento sua participação na pesquisa sem sofrer qualquer forma de retaliação ou danos; e
3. Os resultados gerais da pesquisa serão utilizados apenas para alcançar os objetivos e podem ser publicados em congresso ou em revista científica especializada.

Endereço do(s) responsável(is) pela pesquisa:

Pesquisador Responsável: Hélder Antero Amaral Nunes

Instituição: Universidade Federal do Ceará (UFC) / Programa de Pós-Graduação em Tecnologia Educacional (PPGTE)

Endereço: Bloco Acadêmico do Instituto UFC Virtual, Campus do PICI - CEP 60440-554- Fortaleza, Ceará, Brasil

Telefones para contato: (87) 99675-5112

E-mail: haanunes@gmail.com

ATENÇÃO: Se você tiver alguma consideração ou dúvida sobre a sua participação na pesquisa entre em contato com o pesquisador responsável.

O abaixo assinado _____, _____ anos, RG: _____ declara que é de livre e espontânea vontade que está participando da pesquisa. Eu declaro que li cuidadosamente este Termo de Consentimento Livre e Esclarecido (TCLE) e que, após sua leitura tive a oportunidade de fazer perguntas sobre o seu conteúdo, como também sobre a pesquisa e recebi explicações que responderam por completo minhas dúvidas. E declaro ainda estar recebendo uma cópia assinada deste termo e que minha participação é de caráter voluntário e não serei remunerado.

Pesquisador: Hélder Antero Amaral Nunes

Responsável:_____

Data: __/__/__

Participante: _____

Data: __/__/__

APÊNDICE B – RELATÓRIO DA CLASSIFICAÇÃO DO SOFTWARE PROPOSTO

Classe 1- Reprovado

Atributo Identificador	Classificação	Probabilidade
4	Reprovado	85.4%
12	Reprovado	81.1%
15	Reprovado	92.8%
16	Reprovado	79.7%
21	Reprovado	86.5%

Classe 2 - Aprovado

Atributo Identificador	Classificação	Probabilidade
1	Aprovado	92.7%
2	Aprovado	79.4%
3	Aprovado	77.5%
5	Aprovado	82.3%
6	Aprovado	90.9%
7	Aprovado	84.4%
8	Aprovado	91.3%
9	Aprovado	87.4%
10	Aprovado	93.8%
11	Aprovado	93.5%
13	Aprovado	70.6%
14	Aprovado	77.3%
17	Aprovado	91.4%
18	Aprovado	75.2%
19	Aprovado	86.2%
20	Aprovado	83.1%

Nome do arquivo de teste: Teste 21.csv
Melhor algoritmo: Floresta Aleatória
Tempo de Predição: 0 segundos