UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO DE JOINVILLE
CURSO DE ENGENHARIA MECATRÔNICA

BERNARDO DALFOVO DE SOUZA

ESTUDO DE CASO COM ANÁLISE DE DADOS PARA A DETECÇÃO DA DESISTÊNCIA DE ESTUDANTES EM DISCIPLINAS OFERTADAS COM APOIO DO AMBIENTE MOODLE

Joinville
2023

BERNARDO DALFOVO DE SOUZA

ESTUDO DE CASO COM ANÁLISE DE DADOS PARA A DETECÇÃO DA DESISTÊNCIA DE ESTUDANTES EM DISCIPLINAS OFERTADAS COM APOIO DO AMBIENTE MOODLE

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do título de bacharel em Engenharia Mecatrônica no curso de Engenharia Mecatrônica, da Universidade Federal de Santa Catarina, Centro Tecnológico de Joinville.

Orientador: Dr. Benjamin Grando Moreira

Joinville
2023

AGRADECIMENTOS

Agradeço aos meus pais, Rafael e Sheila, por me encorajarem em toda a minha trajetória estudantil. Seu apoio incondicional foi essencial para a conclusão desta etapa, e o desejo de orgulhá-los me deu forças para ir além.

À minha irmã, Ana Clara, expresso minha gratidão pelo carinho e celebração de cada conquista. Saber que você estava ao meu lado motivou-me a ser um modelo de pessoa e estudante a ser seguido.

À minha namorada, Lynda, agradeço pelo companheirismo durante os três semestres de desenvolvimento deste trabalho. Com o seu apoio, qualquer desafio foi facilmente superado.

Não poderia deixar de agradecer meu orientador, o professor Benjamin Grando Moreira, pelo conhecimento compartilhado na elaboração deste trabalho, além da preparação de quatro artigos para publicação, nos quais pude aplicar e aprimorar a sabedoria adquirida.

Felizmente, são muitos os amigos a quem gostaria de agradecer pela paciência e parceria durante o desenvolvimento deste trabalho e ao longo da minha jornada acadêmica. No entanto, não posso deixar de mencionar alguns que não apenas me apoiaram durante esse percurso, mas também o vivenciaram: Allan, Arthur, Eryk, Felipe, Gabriel, Lara Maria, Pedro Henrique, Vinícius e Vitor José. Muito obrigado a todos vocês; a experiência universitária foi infinitamente mais divertida com a presença de cada um.

A felicidade só é real quando compartilhada, e sou muito feliz por compartilhá-la com vocês.

A melhor coisa de ser um estatístico é que você pode brincar no quintal de todo mundo (Tukey).

RESUMO

A evasão escolar é um problema de âmbito nacional que causa problemas sociais, econômicos e pessoais aos alunos, incluindo consequências psicológicas, físicas, escolares e interpessoais. Uma das causas que podem levar à problemática evasão escolar é a desistência de disciplinas durante a graduação. Neste trabalho apresenta-se o desenvolvimento de um software que analisa dados obtidos a partir do MOODLE, identificando fatores que apontem a possibilidade de desistência de alunos em determinada disciplina. Duas análises foram realizadas, uma a partir de uma análise direta dos dados, e outra que utiliza algoritmos de aprendizado de máquina. Procurou-se definir modelos eficientes para analisar os dados pré-processados, de forma que a solução possa ser generalizada para diferentes disciplinas. A partir do uso do modelo Naive Bayes com seleção dos dez atributos mais relevantes pelo algoritmo chi-quadrado, foi possível atingir 95,31% de sensibilidade e 92,67% de acurácia com apenas 25% dos dados do semestre.

Palavras-chave: Previsão de desistência; modelos estatísticos; análise explícita e implícita.

ABSTRACT

School dropout is a national problem that causes social, economic, and personal issues for students, including psychological, physical, educational, and interpersonal consequences. One of the causes that can lead to school dropout is the dropping out of subjects during undergraduate studies. This work presents the development of a software that analyzes data obtained from MOODLE, identifying factors that indicate the possibility of student dropout in a specific subject. Two analyses were conducted, one from a direct analysis of the data, and another using machine learning algorithms. Efforts were made to define efficient models to analyze the pre-processed data, so that the solution can be generalized for different subjects. By using the Naive Bayes model with the selection of the ten most relevant attributes by the chi-square algorithm, it was possible to achieve 95.31% sensitivity and 92.67% accuracy with just 25% of the semester's data.

Keywords: Dropout prediction; statistical models; explicit and implicit analysis.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

σ	Variância
β	Relação entre os pesos das métricas de sensibilidade e precisão
Σ	Somatório

LISTA DE ABREVIATURAS E SIGLAS

AD	Árvore de Decisão
ANOVA	Análise de variância
ANOVA-5	Análise de variância com os cinco melhores atributos
ANOVA-10	Análise de variância com os dez melhores atributos
AVA	Ambiente Virtual de Aprendizagem
BN	Bayes Net
Chi2	Chi-quadrado
Chi2-5	Chi-quadrado com os cinco melhores atributos
Chi2-10	Chi-quadrado com os dez melhores atributos
COBENGE	Congresso Brasileiro de Educação em Engenharia
COVID-19	Coronavirus disease
CDE	Ciência de Dados Educacionais
CDI-1	Cálculo Diferencial e Integral I
CTJ	Centro Tecnológico de Joinville
DT	Decision Table
EaD	Ensino a Distância
Enem	Exame Nacional do Ensino Médio
ET	Extra Trees
FP	Falsos positivos
FN	Falsos negativos
GB	Gradient Boosting
GNB	Gaussian Naive Bayes
GI	Ganho de informação

GI-5	Ganho de informação com os cinco melhores atributos
HO	Holdout
IBK	Instance Based Learner
kNN	K-Nearest Neighbors
MEC	Ministério da Educação
MLP	Multilayer Perceptron
MOODLE	Modular Object-Oriented Dynamic Learning Environment
MDE	Mineração de Dados Educacionais
NB	Naive Bayes
OneR	One Rule
PESFAM	Probabilistic Ensemble Simplified Fuzzy Adaptive Resonance Theory Mapping
RG	Relação de ganho
RG-5	Relação de ganho com os cinco melhores atributos
RF	Random Forest
RL	Regressão logística
SciELO	Scientific Electronic Library Online
SA	Subamostragem aleatória
SC	Simple Cart
SL	Simple Logistic
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
UFSC	Universidade Federal de Santa Catarina
VC	Validação cruzada
VP	Verdadeiros positivos
VN	Verdadeiros negativos
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

Durante a realização de um curso de graduação, percebe-se que diversos alunos não conseguem completar o cronograma de determinadas disciplinas por vários motivos, como falta de motivação e dificuldade em compreender o conteúdo. A não conclusão da matéria, conhecida como *desistência*, caracteriza um tipo de fracasso escolar e configura um prejuízo pessoal e profissional para o aluno, além de impactar os resultados sobre a produtividade das universidades e sociedade (Nagai; Cardoso, 2017).

O Ministério da Educação (MEC) divide a definição de evasão educacional em três tipos: evasão de curso, que ocorre quando um estudante abandona um curso específico de graduação; evasão institucional, que se refere ao abandono da instituição de ensino pelo aluno; e evasão do sistema, quando o aluno deixa de participar do ensino superior como um todo (Brasil, 1996). Enquanto a literatura da área utiliza o conceito de evasão relativa aos cursos de graduação, ao realizar trabalhos de análise de dados para previsão da evasão, o presente trabalho avalia a não-conclusão de uma matéria de um curso de graduação, tratando desse conceito como desistência ou abandono (Manhães *et al.*, 2011; Morais, 2018; Queiroga *et al.*, 2018).

A evasão escolar, portanto, acarreta consequências significativas tanto para o indivíduo/aluno em um contexto psicológico quanto para o âmbito econômico e social, que são expectativas inerentes à instituição universitária. De acordo com as descobertas de Mallada (2011), quatro domínios podem ser impactados no que diz respeito ao estudante:

1. **Aspecto Psicológico:** Esta esfera pode manifestar-se por meio de sintomas que incluem depressão, ansiedade e até mesmo a consideração de ideias suicidas;
2. **Aspecto Físico:** A evasão escolar também pode ter efeitos físicos adversos, tais como alterações no padrão de sono e o desenvolvimento de quadros de hipertensão, correlacionados ao estresse e à pressão associados à não conclusão dos estudos;
3. **Aspecto Escolar:** Estudantes que enfrentam dificuldades na conclusão de seus cursos podem desenvolver atitudes negativas em relação às tarefas acadêmicas e experimentar uma diminuição no seu rendimento acadêmico;
4. **Aspecto Interpessoal:** A decisão de abandonar o ensino superior pode levar à desmotivação, à redução da qualidade de vida e até mesmo ao surgimento de estados de irritabilidade, influenciando negativamente o bem-estar global do aluno.

Essas implicações enfatizam a importância de abordar a questão da evasão

escolar de forma holística, considerando não apenas seus impactos acadêmicos, mas também as ramificações significativas no âmbito psicológico e social dos estudantes (Mallada, 2011).

Do ponto de vista econômico, a falta de retorno monetário a partir do investimento realizado pela universidade no aluno, diminui a rentabilização desse processo, inviabilizando o crescimento no âmbito local até o nacional (Fialho; Prestes, 2014). A sociedade é prejudicada pelo desperdício do investimento, visto que as vagas, uma vez ocupadas pelos alunos evadidos, não necessariamente são preenchidas (Gaioso, 2005 *apud* Adachi, 2009).

Durante a pandemia da Coronavirus disease (COVID-19), os Ambientes Virtuais de Aprendizagem (AVA) tornaram-se ferramentas essenciais para a realização de cursos no formato de Ensino a Distância (EaD), como pode-se perceber na Universidade Federal de Santa Catarina (UFSC), por exemplo, onde as aulas e avaliações foram ministradas integralmente pela plataforma Modular Object Oriented Dynamic Learning Environment (MOODLE), também utilizada pela UFSC durante o ensino presencial. Enquanto o ambiente virtual se mostrou eficiente para a disponibilização de vídeos, realização de aulas ao vivo e provas, limitou as interações entre aluno e professor, prejudicando o reconhecimento, mesmo que subjetivo, pelo docente, do risco de evasão dos discentes, por outro lado, o uso das plataformas digitais permitiu obter dados sobre a participação dos estudantes.

A busca por trabalhos acadêmicos produzidos sobre a análise de dados do MOODLE, com o objetivo de prever a desistência de alunos em determinadas matérias, foi empreendida nas plataformas Scientific Electronic Library Online (SciELO) e Base de Teses e Dissertações (BDTD), Anais publicados no Congresso Brasileiro de Educação em Engenharia (COBENGE), ScienceDirect, IEEE Xplore e Google Scholar. Os resultados apontam que é possível identificar alunos que apresentem os fatores que se atrelam à evasão dos cursos de graduação a partir das plataformas virtuais de ensino (Burgos *et al*., 2018; Garcia *et al*., 2022; Gottardo; Kaestner; Noronha, 2014; Manhães *et al*., 2011; Morais, 2018; Queiroga *et al*., 2018; Viana; Santana; Rabêlo, 2022), mas não contemplam a desistência de matérias individualmente, justificando, assim, o esforço de preencher essa lacuna em alguma medida.

Zarpelon (2016) utiliza seis variáveis quantitativas (nota na prova de Matemática do Exame Nacional do Ensino Médio (Enem), pesos para as provas do Enem, período de ingresso no curso, carga horária semanal de aulas, desempenho no teste diagnóstico e desempenho nos testes semanais), e uma variável qualitativa (comprometimento acadêmico), para analisar o desempenho dos alunos. Nagai e Cardoso (2017) utilizam 29 assertivas dentro de oito diferentes fatores (Estrutura do Curso, Escolha do Curso, Cidade, Conciliar estudo e trabalho, Estrutura da Instituição, Pessoal, Mercado de Trabalho e Aprendizado), demonstrando a complexidade e subjetividade desse assunto.

Dessa forma, ao considerar que o escopo deste trabalho leva em consideração a análise dos dados obtidos a partir do MOODLE, reconhece-se que os resultados alcançados representam apenas um dos fatores que podem levar ao fracasso escolar.

O presente estudo se concentra, de forma específica, na análise da não conclusão de disciplinas em um curso de graduação de modalidade presencial. Com o propósito de identificar indicadores que possam antecipar o potencial abandono de uma disciplina acadêmica, foi conduzido um estudo de caso. Segundo Manhães *et al*. (2011), utilizar a automatização para identificar discentes com risco de evasão, não só remove a subjetividade da análise por parte do docente, que depende diretamente do engajamento e experiências prévias do mesmo, como também os orienta a diversificar as atividades pedagógicas.

É relevante enfatizar que a coleta de dados para o presente trabalho foi realizada diretamente no ambiente MOODLE, pois várias instituições acadêmicas optam por evitar a instalação de complementos (plugins) em suas instâncias do sistema, visando preservar a integridade dos registros. O trabalho trata de questões relacionadas às melhores práticas de avaliação, incluindo a aplicação de métricas de avaliação, e diferentes métodos de classificação desenvolvidos para abordar esse tipo específico de problema. Destacam-se dois aspectos que diferenciam este estudo em relação ao que é convencional em trabalhos similares:

1. **Enfoque na desistência em disciplinas:** Concentra-se na análise da desistência em disciplinas específicas, diferenciando-se de abordagens que objetivam prever a evasão de curso ou reprovação em disciplinas em geral;
2. **Utilização exclusiva de dados da disciplina:** Utilizam-se exclusivamente dados da disciplina sob análise, o que representa uma diferença em relação a abordagens que podem considerar informações mais amplas;
3. **Elaboração de uma análise explícita:** Elaborou-se uma análise que não se baseia apenas em técnicas de aprendizado de máquina, mas também em uma observação dos dados diretamente.

A fim de identificar fatores que indiquem a possibilidade de abandono de disciplinas dos cursos de Engenharia do Centro Tecnológico de Joinville (CTJ), da Universidade Federal de Santa Catarina (UFSC), realizou-se um estudo de caso com dados extraídos dos registros de uso do AVA MOODLE das disciplinas de Programação 1 e Modelagem de Sistemas, ambos relativos ao primeiro semestre de 2023. Foram utilizados dados de frequência, notas e conclusão de atividades para desenvolver a análise, o que torna essa uma abordagem específica para cursos que disponibilizem tais recursos aos alunos.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Identificar, a partir da análise de dados extraídos do AVA MOODLE, alunos que apresentem possibilidade de desistência da disciplina e de maneira que haja tempo suficiente para uma intervenção por parte do professor.

1.1.2 Objetivos Específicos

- Gerar indicadores sobre o comportamento de desistência;
- Elaborar modelos para classificar alunos desistentes;
- Antecipar a identificação de alunos desistentes.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, é realizada uma revisão teórica dos conceitos essenciais para a condução deste estudo. Inicialmente, são abordados tópicos como o impacto financeiro e contábil da evasão, seguido pela análise de dados, pré-processamento dos dados e a seleção de atributos de maior relevância. Também são discutidas as métricas e as técnicas de divisão de dados empregadas na avaliação dos resultados alcançados. Por fim, este capítulo culmina em uma análise da literatura pertinente ao tema.

2.1 PREJUÍZO CONTÁBIL E ECONÔMICO

De acordo com Pinto (2021), apenas no ano de 2019 a quantidade de alunos que evadiram cursos de graduação, em universidades públicas chegou a 203.784, resultando em um custo contábil de R$ 8.974.795.155,90 às instituições públicas de ensino. Além do prejuízo contábil às universidades causado pela evasão, o discente também arca com um custo econômico, muitas vezes sem perceber. A autora descreve esse custo econômico como os potenciais salários renunciados pelos alunos evadidos, e resulta em uma média brasileira de R$ 21.277,59 para cada evadido, representando cerca de um ano de trabalho, de acordo com o salário mínimo da época (Pinto, 2021).

2.2 CIÊNCIA DE DADOS EDUCACIONAIS

A Ciência de Dados Educacionais (CDE) aplica análise de dados à educação e integra habilidades técnicas e sociais para entender práticas educacionais em diferentes contextos de aprendizagem. Originária dos anos 2000, a CDE evoluiu a partir de conferências sobre Mineração de Dados Educacionais (MDE) e Analítica de Aprendizagem, fortalecida por estudos em Inteligência Artificial na Educação (Filatro, 2020).

Romero e Ventura (2012) destacam a necessidade de métodos computadorizados para gerir o crescente volume de dados educacionais. A MDE, eficaz em sistemas de informação, permite analisar grandes quantidades de dados e identificar características de ambientes educacionais eficientes. O interesse em MDE é evidenciado pelo aumento significativo de citações em plataformas como Google Scholar e SciVerse Scopus. Esses dados incluem desempenho acadêmico, informações demográficas e dados de sistemas de gerenciamento de aprendizado. A MDE é aplicada em vários níveis educacionais e ambientes online, abrangendo previsão de desempenho estudantil, identificação de alunos em risco, personalização de aprendizagem e aprimoramento de cursos.

2.3 PRÉ-PROCESSAMENTO DE DADOS

No âmbito educacional, o pré-processamento de dados é crucial e complexo, frequentemente ocupando mais da metade do tempo dedicado a projetos de mineração de dados. Essa complexidade surge da natureza inicialmente inadequada dos dados educacionais para questões específicas, devido à sua diversidade e estrutura hierárquica. A seleção e transformação dos dados em um formato apropriado dependem do problema em questão, exigindo coleta meticulosa e alinhamento com os objetivos propostos. Ambientes educativos geram grandes volumes de dados de múltiplas fontes, o que requer integração com níveis de granularidade adequados. Simplificar tabelas ao reduzir variáveis e converter atributos numéricos em categorias aumenta a clareza da análise. A interpretação dos resultados e a compreensão dos limites dos modelos utilizados exigem a consideração do contexto. Para preservar a confidencialidade, os dados são anonimizados, eliminando informações pessoais irrelevantes para a mineração, como nomes e e-mails (Romero; Ventura, 2012).

Bases de dados que compreendem informações geradas naturalmente estão propensas a conter dados imprecisos, faltantes ou inconsistentes, devido ao seu grande volume e à provável coleta de múltiplas fontes heterogêneas. Quando submetidos a técnicas de mineração de dados, repositórios de baixa qualidade podem levar a resultados insatisfatórios. Entretanto, a aplicação de técnicas de pré-processamento pode melhorar substancialmente a qualidade dos dados e a eficiência do processo de mineração. A Figura 1 ilustra as quatro principais etapas desse processo: limpeza, integração, redução e transformação (Han; Pei; Tong, 2012).

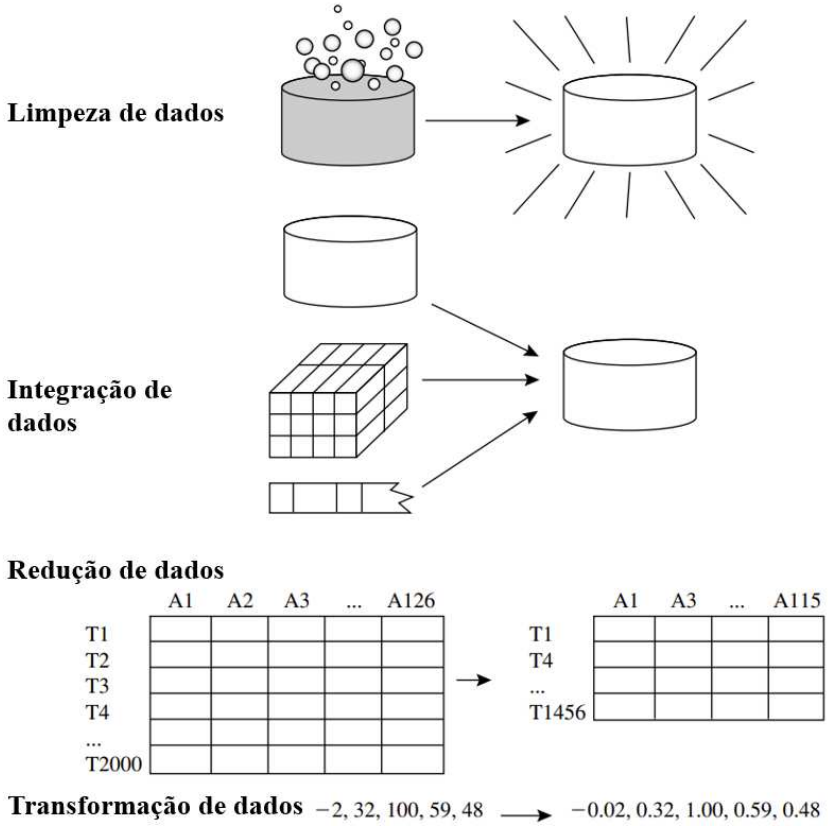
2.3.1 Limpeza de dados

A limpeza de dados consiste em aprimorar a qualidade dos dados ao tratar as questões relativas a informações incorretas, corrompidas, mal formatadas, duplicadas ou incompletas em um conjunto de dados. Esse processo não apenas preenche valores ausentes e suaviza dados ruidosos, mas também identifica e elimina discrepâncias e resolve inconsistências. Tal limpeza meticulosa é crucial; se os usuários perceberem os dados como *sujos*, a confiança nos resultados de minerações de dados subsequentes será prejudicada. Dados imprecisos também correm o risco de confundir os procedimentos de mineração, produzindo assim resultados pouco confiáveis (Han; Pei; Tong, 2012; Kaensar; Wongnin, 2023).

Han, Pei e Tong (2012) descrevem que a limpeza de dados pode ser necessária a partir da detecção de dados faltantes ou ruidosos. São dadas como soluções para dados faltantes: ignorar a tupla[1]; preencher a tupla manualmente; usar uma constante

[1] Entende-se como tupla um conjunto de pares atributo-valor que correspondem a uma unidade de dados.

Figura 1 – Quatro etapas do pré-processamento



Fonte: Han, Pei e Tong (2012), adaptado pelo autor (2023).

global para preencher a tupla; usar uma medida de tendência central (média e mediana, por exemplo) da tupla; usar uma medida de tendência central de entradas da mesma classe em que se está classificando os dados (em uma loja de televisores, no caso de um preço faltante, utiliza-se a média de preços de televisores de mesmo tamanho, por exemplo); determinar estatisticamente o valor mais provável para preencher a tupla, a partir dos métodos de regressão ou árvore de decisão, por exemplo.

2.3.2 Integração de dados

A integração de dados se faz necessária quando se extrai dados de diferentes fontes e tem como objetivo reduzir redundâncias e inconsistências na base de dados resultante, o que pode aprimorar a acurácia e velocidade do processo de mineração de dados. São descritos como desafios relacionados à integração de dados: identificação de entidade; análise de redundância e correlação; duplicação de tuplas; resolução e detecção de conflitos entre valores de dados. O primeiro desafio é encontrado quando existem valores equivalentes em duas bases de dados com estruturas diferentes (“nome_usuario” e “username”, por exemplo). O segundo desafio, que compreende a redundância de atributos, pode ser encontrado em uma base de dados quando uma variável é derivada de outra (faturamento anual e faturamento mensal, por exemplo).

Essa redundância pode ser detectada a partir de métodos de correlação de dados, na qual, dados dois atributos, é medida a influência de um sobre o outro (Han; Pei; Tong, 2012).

O terceiro desafio na integração de dados é a duplicação de tuplas, frequentemente resultante de erros na atualização parcial ou entrada incorreta de informações durante a fusão de diferentes bases. Por fim, os conflitos entre valores de dados emergem quando registros equivalentes apresentam divergências nas bases a serem integradas, atribuíveis a variações em representação, escala ou codificação. Um exemplo ilustrativo desse conflito é a discrepância nas unidades de medida, como a representação de distância em metros em uma base de dados, contrastando com milhas em outra (Han; Pei; Tong, 2012).

2.3.3 Redução de dados

Devido à possível complexidade e altos tempos de execuções atrelados à análise e mineração de dados, técnicas de redução de dados podem ser aplicadas para obter uma representação da base de dados em um volume inferior ao total, enquanto se mantém a integridade do conjunto original. Estratégias de redução de dados incluem a redução por dimensão, redução por número e compressão de dados. Métodos como a transformada wavelet, análise de componentes principais e seleção de atributos exemplificam abordagens adotadas na redução por dimensão, que visa reduzir a quantidade de variáveis aleatórias em consideração (Han; Pei; Tong, 2012).

Hall (1999) descreve a seleção de atributos como o processo de identificar e remover informações redundantes e irrelevantes. O que reduz a dimensionalidade dos dados e pode fazer com que os algoritmos de aprendizado operem mais eficiente e rapidamente, até mesmo aprimorando a acurácia de classificação. Exemplos de medidas de seleção de atributos incluem: ganho de informação (GI), relação de ganho (RG), chi-quadrado (Chi2) e análise de variância (ANOVA) (Han; Pei; Tong, 2012).

A medição de ganho de informação, descrita pela Equação 1, se baseia na avaliação da variação da entropia ao dividir um conjunto de dados com base em um atributo, que tem o valor de sua impureza medida. Caso essa divisão resulte em subconjuntos mais puros que o conjunto original, significa que a entropia é diminuída e o ganho de informação aumenta. Cada atributo tem seu ganho de informação calculado individualmente, e o atributo com maior ganho é escolhido para a divisão do conjunto. Esse atributo minimiza a informação necessária para classificar as tuplas e reflete a menor impureza da divisão. O uso do ganho de informação minimiza o número de testes necessários para classificar uma tupla e garante que uma árvore de decisão simples seja encontrada (Han; Pei; Tong, 2012).

$$\text{Ganho}(A) = \text{Info}(D) - \text{Info}_A(D) \tag{1}$$

Nessa equação, Ganho(A) representa o ganho de informação a partir da divisão do conjunto com um atributo A, Info(D) equivale à informação necessária para classificar uma tupla no conjunto de dados D, também conhecida como a entropia de D antes da divisão, e Info$_A(D)$ é a quantidade de informação necessária para classificar uma tupla em D após a divisão do conjunto por um atributo A, que equivale a entropia em D após a divisão por A (Han; Pei; Tong, 2012).

Porém, o ganho de informação tende a selecionar os atributos que tenham a maior quantidade de valores. Por exemplo, um atributo que representa a identificação de produtos em uma loja possui um valor para cada produto, ou seja, seu ganho de informação possivelmente será o mais alto do conjunto, visto que irá gerar divisões na mesma medida que há produtos, mesmo que tais divisões não representem análise relevante. Desta forma, a relação de ganho, representada pela Equação 2, leva em consideração a relação entre o número de tuplas geradas nas divisões e o total de tuplas do conjunto, de forma que mesmo que a seleção de um atributo gere a maior quantidade de divisões, o número de tuplas nessas divisões será considerado (Han; Pei; Tong, 2012).

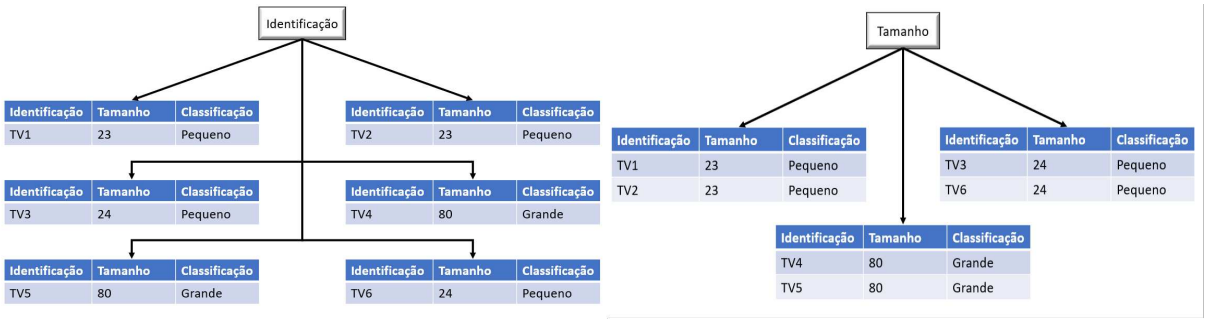
$$\text{RelGanho}(A) = \frac{\text{Ganho}(A)}{\text{DivInfo}_A(D)} \tag{2}$$

Na equação, RelGanho(A) representa a relação de ganho para o uso de um atributo A para a divisão, Ganho(A) é o ganho de informação, calculado a partir da Equação 1, e DivInfo$_A(D)$ equivale ao potencial ganho de informação gerado pela divisão da base de dados com o atributo A, que leva em consideração a quantidade de tuplas geradas em relação ao total de tuplas no conjunto D. A Figura 2 ilustra as diferenças entre o uso do ganho de informação e da relação de ganho para uma base de dados de uma loja de televisores de exemplo que possui cinco entradas de dados com os atributos de identificação do produto, diferentes para cada televisor, e tamanho do televisor, que pode ter os valores de 23, 24 ou 80 polegadas. Deseja-se identificar quais televisores são considerados como grandes a partir dos dados.

Percebe-se que, mesmo realizando mais divisões, as divisões geradas pelo ganho de informação não são tão relevantes à classificação quanto as divisões geradas pela relação de ganho, visto que no primeiro caso cada divisão possui apenas uma tupla.

O algoritmo Chi2 é baseado na estatística X^2 e propõe uma maneira de selecionar atributos diretamente de valores numéricos em paralelo com a discretização, em contraste com outros algoritmos que requerem a discretização antes da seleção. Para a primeira parte do algoritmo, é definido aos atributos numéricos um alto nível de significância para discretização. Após organizar as características de acordo com seus valores, é realizado o cálculo de X^2 para cada par de intervalos para então verificar se as frequências relativas de intervalos adjacentes são similares o suficiente para

Figura 2 – Representação dos resultados de a partir de ganho de informação (esquerda) e relação de ganho (direita)



Fonte: Elaborado pelo autor (2023).

justificar sua mescla. A repetição do processo de mescla é controlado por um limite de X^2, automaticamente definido pelo algoritmo a partir do parâmetro de significância. Ao atingir tal limite, verifica-se se um atributo foi mesclado a apenas um valor, o que indica que tal atributo não é relevante para a representação dos dados originais. Desta forma, ao fim da discretização a seleção de atributos também está concluída (Liu; Setiono, 1995; Hall, 1999).

A ANOVA é um método estatístico usado para determinar se existem diferenças significativas entre as médias de diferentes grupos de dados. O objetivo é avaliar se as variações observadas entre os grupos são maiores do que as variações dentro de cada grupo, o que sugere dependências significativas entre si. Este método envolve quatro etapas: coleta de amostras; cálculo das variâncias entre (σ_e^2) e dentro (σ_d^2) das amostras; cálculo da pontuação F, razão entre σ_e^2 e σ_d^2; e interpretação da pontuação F. Para a última etapa, se a pontuação F for menor ou apenas ligeiramente maior que um, indica que a variância entre as amostras não é significativamente maior do que a variância dentro delas, o que sugere que as amostras provavelmente vêm de populações com médias semelhantes. Por outro lado, se a pontuação F for significativamente maior que um, isso implica que a maior parte da variância na amostra total é devido às diferenças entre os grupos, portanto, é provável que as amostras venham de populações com médias diferentes (Mahbobi; Tiemann, 2010).

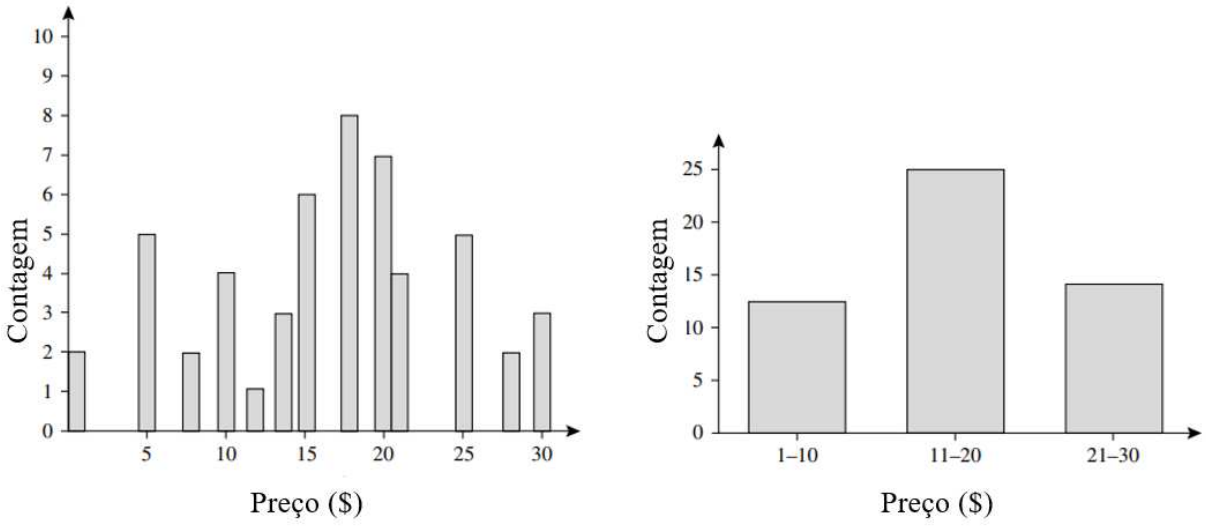
2.3.4 Transformação de dados

A transformação tem como objetivo tornar o processo de mineração mais eficiente e facilitar a identificação dos padrões presentes na base de dados. Estratégias para esse procedimento incluem: suavização; construção de atributos; agregação; normalização; discretização; e geração de uma hierarquia de conceitos. A normalização consiste em evitar a dependência da análise a uma escolha de unidades de medição, como quilogramas ao invés de libras, pela transformação dos dados para se encaixarem

em intervalos menores, como entre -1 e 1 ou entre 0 e 1 (Liu; Setiono, 1995; Hall, 1999).

Também considerado como um método de redução de dados, o processo de substituir os valores de atributos numéricos por intervalos ou conceitos, é conhecido como discretização. Os valores discretos podem então ser organizados recursivamente em representações de mais alto nível, possibilitando outra forma de transformação de dados, a hierarquia de conceitos. Uma forma de se discretizar dados é a partir de um histograma dos dados que particiona os valores de um atributo em intervalos que, idealmente, tenham a mesma quantidade de tuplas (frequências equivalentes), ou a partir da largura do intervalo (largura equivalente), para que as partições tenham a mesma largura, independente da quantidade de tuplas em si (Liu; Setiono, 1995; Hall, 1999). A Figura 3 ilustra o processo de transformação de dados de preço para intervalos em um histograma.

Figura 3 – Processo de transformação de dados de preço (esquerda) a partir de análise do histograma (direita)



Fonte: Han, Pei e Tong (2012), adaptado pelo autor (2023).

No caso da Figura 3, percebe-se que o histograma segue uma partição de dados para largura equivalente de 10$. Tal partição faz com que as análises sejam feitas a partir de três intervalos de dados, ao invés de 13 instâncias separadas.

2.4 ANÁLISE DE DADOS

A análise de dados é um processo multifacetado que envolve a aplicação de técnicas e modelos para extrair informações úteis, categorizado em análise exploratória, análise explícita e análise implícita. Entre os procedimentos para auxiliar

na compreensão de dados, a CDE faz uso de modelos matemáticos construídos para comunicar insights a educadores, alunos, gestores, designers instrucionais e outras partes interessadas (Filatro, 2020).

Na análise explícita, as informações desejadas são diretamente acessíveis nos dados, mas requerem operações específicas para sua extração e destaque. O objetivo desta análise é concreto e direcionado, contrastando com a análise exploratória, que busca uma compreensão mais ampla dos dados. Técnicas de análise explícita incluem a elaboração de predicados lógicos para filtrar subconjuntos de dados, identificação e tratamento de outliers que podem distorcer os resultados analíticos, e estratificação para resumir intervalos de variáveis numéricas (Amaral, 2016).

Por outro lado, a análise implícita aborda objetivos como classificação, regressão, agrupamento e associação. Esses objetivos são comumente alcançados com o auxílio de técnicas de aprendizado de máquina, incluindo algoritmos como árvores de decisão e redes neurais artificiais. Essas abordagens são fundamentais para interpretar grandes volumes de dados e identificar padrões não evidentes (Manhães *et al.*, 2011; Burgos *et al.*, 2018).

2.4.1 Classificação de dados

A classificação é uma forma de análise de dados em que um modelo é construído para prever rótulos de classificação, que são definidos por cada base de dados que se deseja classificar. Por exemplo, para a análise de dados médicos em que se deseja prever qual o melhor tratamento, os rótulos podem ser os nomes de três remédios. É possível também rotular em classes mais objetivas, como positivo e negativo. A classificação de dados pode ser dividida em etapas de aprendizado (ou treinamento), na qual o modelo é construído, e classificação, onde o modelo gerado é usado para prever os rótulos em dados de teste (Han; Pei; Tong, 2012).

No primeira etapa, o algoritmo de classificação constrói o classificador a partir de uma base de dados de treinamento, composta de tuplas (amostras, objetos, exemplos, instâncias ou dados) e suas respectivas classificações. Uma tupla é representada por um vetor de atributos de dimensão n, no qual estão contidos n valores de n atributos, e está contida em uma classe pré-determinada pelos atributos de rótulo de classe, os quais devem ser valores discretos. Como espera-se que os rótulos de classificação da base de dados de treinamento sejam providenciados ao modelo, sua construção é caracterizada como aprendizado supervisionado (Han; Pei; Tong, 2012).

Para o segundo passo, o modelo gerado é utilizado para a classificação, o que necessita o uso de uma base de dados de teste, visto que os classificadores tendem a realizar um sobreajuste (overfit) dos dados, fazendo com que anomalias e características específicas dos dados de treinamento sejam incorporadas no modelo,

mesmo que não representem os dados como um todo. Dessa forma, no mesmo formato que o primeiro passo, tuplas de teste completamente independentes do treinamento são utilizadas (Han; Pei; Tong, 2012). Após a construção de um modelo de classificação e a categorização de dados de teste, é necessário estimar a qualidade dos resultados encontrados a partir de técnicas de avaliação.

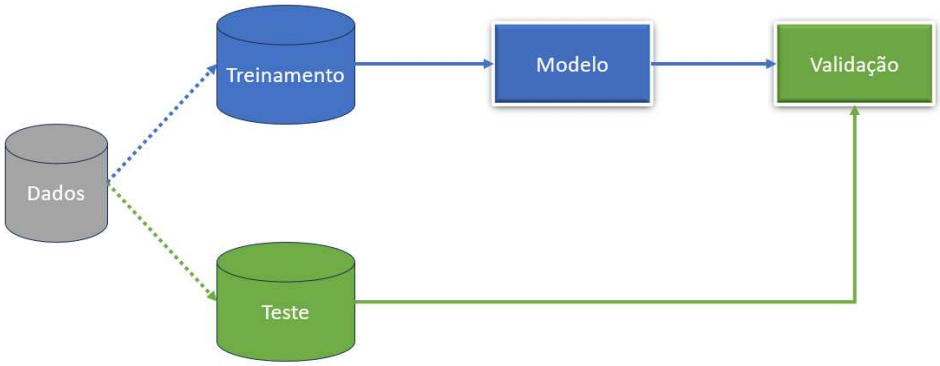
2.5 AVALIAÇÃO DE RESULTADOS

Nesta seção, serão descritos diferentes métodos de divisão de dados em treino e teste, além de uma descrição de métricas que podem ser utilizadas para mensurar a performance dos resultados e permitir uma comparação entre diferentes modelos, a fim de escolher um que melhor classifique os dados.

2.5.1 Método holdout

Também conhecido como validação simples nesse método, ilustrado na Figura 4, os dados são divididos aleatoriamente em treinamento e teste. Ressalta-se a importância de utilizar o conjunto de testes (dados novos e não vistos pelo modelo) apenas para a avaliação de métricas, não durante o treinamento, para evitar a inserção de um viés na classificação (Han; Pei; Tong, 2012).

Figura 4 – Validação de modelos a partir do método holdout



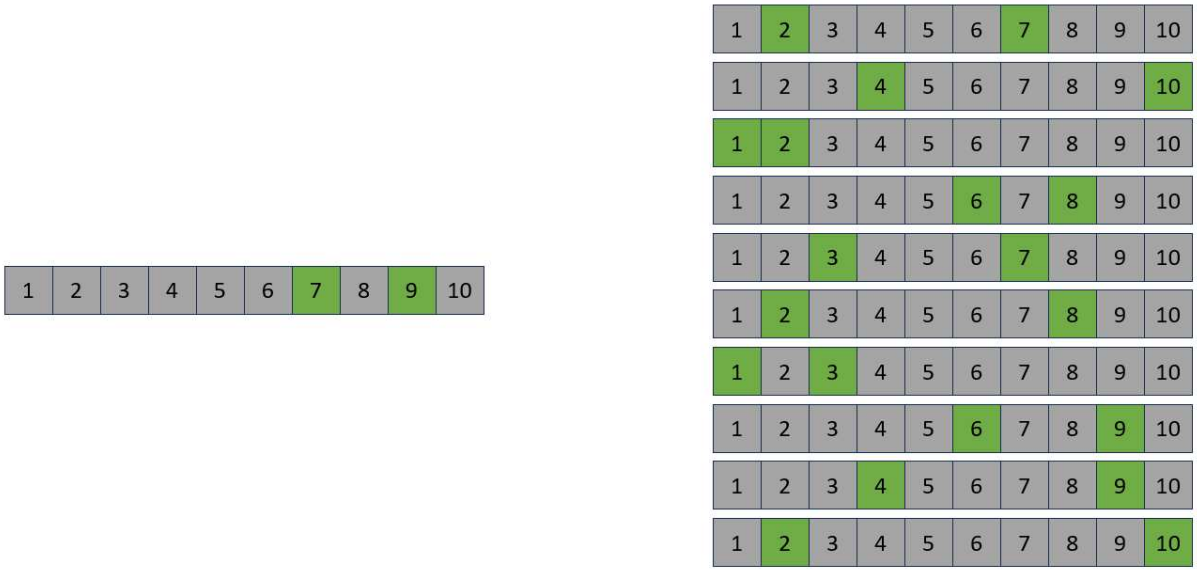
Fonte: Han, Pei e Tong (2012), adaptado pelo autor (2023).

A separação dos dados de teste visa lidar com as imperfeições de um mundo não ideal, como dados e recursos limitados, e a incapacidade de coletar mais dados da distribuição original. Tipicamente, são atribuídos 2/3 dos dados ao conjunto de treinamento e 1/3 ao teste, porém, outras divisões como 60% e 40%, 70% e 30% ou 80% e 20%, respectivamente, também podem ser utilizadas (Raschka, 2018). Porém, a estimativa do método é pessimista, visto que apenas uma parte dos dados originais são utilizados para treinar o modelo (Han; Pei; Tong, 2012).

2.5.2 Submostragem aleatória

Também conhecida como holdout repetido, a subamostragem aleatória é caracterizada pela repetição do método holdout n vezes, cada uma com a própria seleção aleatória de dados de teste. Objetiva-se obter uma estimativa de performance mais robusta e menos variante, com relação ao método holdout convencional (Raschka, 2018). A Figura 5 ilustra as diferenças entre as realizações dos métodos holdout e holdout repetido em dez vezes ($n = 10$), uma divisão dos dados de 20% para teste (quadrados em verde) e 80% para treinamento (quadrados em cinza) foi selecionada.

Figura 5 – Comparação dos métodos holdout convencional (esquerda) e holdout repetido (direita), com divisão de 80% para treino e 20% para teste



Fonte: Elaborado pelo autor (2023).

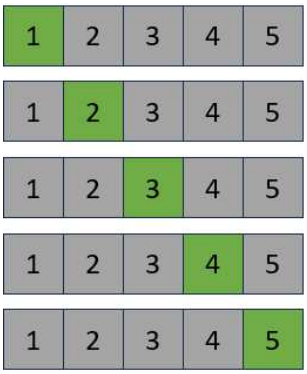
Percebe-se que, mesmo com a realização da subamostragem aleatória dez vezes, os dados presentes no quadrado 5 não foram considerados para teste em qualquer momento. Dessa forma, para garantir que todos os dados serão considerados pelo menos uma vez nos dados de teste, se faz necessário o uso da técnica de validação cruzada.

2.5.3 Validação cruzada

A ideia da validação cruzada é possibilitar que todas as amostras de dados sejam testadas. No caso da validação cruzada k-fold, a validação é realizada k vezes na mesma base de dados. A cada iteração, a base é dividida k vezes: uma parte é utilizada para a validação (teste), enquanto que o restante das bases ($k - 1$ em

quantidade), é utilizado para o treinamento do modelo (Raschka, 2018). A Figura 6 ilustra o processo de validação cruzada 5-fold, no qual o processo descrito é realizado com $k = 5$. Na figura, cada quadrado simboliza 1/5 da base de dados, e as cores verde e cinza significam que a divisão de dados é para teste e treino, respectivamente.

Figura 6 – Validação cruzada k-fold com $k = 5$



Fonte: Raschka (2018), adaptado pelo autor (2023).

Para o exemplo da Figura 6, cinco modelos distintos são gerados a partir de bases de treinamento diferentes, porém, sobrepostas, e suas validações são realizadas em bases de teste completamente distintas, sem sobreposição. Em contrate com os métodos de holdout e amostragem aleatória, a validação cruzada k-fold garante que todas as amostras são utilizadas (Raschka, 2018). De forma geral, é recomendado o uso de k com valor igual a dez, mesmo que o poder computacional permita números maiores para k, devido ao seu baixo viés e variância (Han; Pei; Tong, 2012).

2.5.4 Métricas

Para avaliar a qualidade dos resultados encontrados, se faz necessária uma etapa de avaliação de perfomance dos classificadores, a partir do uso de métricas definidas na literatura. Primeiramente, há necessidade de definir quatro unidades básicas de avaliação de resultados: verdadeiros positivos (VP), que se referem às tuplas valor real positivo corretamente classificadas; verdadeiros negativos (VN), que são as tuplas de valor real negativo corretamente classificados; falsos positivos (FP), tuplas de valor real negativo, porém incorretamente classificados como positivo; e falsos negativos (FN), tuplas incorretamente classificadas como negativas (Han; Pei; Tong, 2012). Por exemplo, em uma loja de televisores, deseja-se saber quais televisores são considerados de tamanho grande. Televisores grandes que são classificados como pequenos são considerados como falsos negativos, televisores pequenos classificados como grandes são falsos positivos, e televisores corretamente classificados como grandes e pequenos são verdadeiros positivos e negativos, respectivamente.

A matriz de confusão é uma ferramenta útil na análise da performance de classificadores e permite a fácil visualização e cálculo de quão bem o classificador acerta (verdadeiros positivos e negativos) ou quão mal está errando (falsos positivos e negativos). O Quadro 1 demonstra a disposição de informações de uma matriz de confusão. A matriz terá, no mínimo, tamanho m (tal que $m \geq 2$), que representa a quantidade de classes a serem classificadas, definiu-se apenas positivo e negativo para o exemplo. Um valor $MC_{i,j}$, indica o número de tuplas de classe i que foram classificadas como classe j (Han; Pei; Tong, 2012).

Quadro 1 – Matriz de confusão genérica

	Positivo real	Negativo real
Positivo previsto	VP	FP
Negativo previsto	FN	VN

Fonte: Elaborado pelo autor (2023).

A partir das quatro unidades básicas de avaliação, pode-se calcular as métricas de acurácia, sensibilidade, precisão e pontuação F1 (Han; Pei; Tong, 2012).

2.5.4.1 Acurácia

A acurácia, representada pela Equação 3, reflete a proporção de tuplas corretamente classificadas pelo classificador e retrata quão bem este é capaz de reconhecer tuplas de diferentes classes (Han; Pei; Tong, 2012).

$$\text{acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}} \tag{3}$$

O uso dessa métrica pode acarretar em um problema de desbalanceamento de classe, no qual a base de dados possui uma distribuição que reflete uma maioria significativa de uma classe e minoria de outra. Em um exemplo que se deseja classificar dados médicos em “cancerígeno” ou “não cancerígeno”, se a segunda classe representar 97% dos dados, um classificador pode atingir 97% de acurácia sem classificar corretamente qualquer tupla que seja de classe “cancerígeno”. Assim, melhor representa a qualidade dos resultados quando as classes de dados são distribuídas uniformemente, quando isso não acontece, se faz necessário o uso métricas que possam avaliar quão bem o classificador reconhece cada classe individualmente, como a sensibilidade e a precisão (Han; Pei; Tong, 2012).

2.5.4.2 Sensibilidade

A métrica de sensibilidade, representada pela Equação 4, também conhecida como a taxa de reconhecimento de verdadeiros positivos, pode ser utilizada para

contornar o problema de desbalanceamento de classes. Seu resultado indica a proporção de tuplas positivas que são corretamente classificadas.

$$\text{sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{4}$$

Uma pontuação de 100% de sensibilidade representa que todos os elementos positivos foram corretamente classificados como tal, porém, não informa quantos elementos negativos também foram considerados como positivos. Para isso, a métrica de precisão deve ser calculada (Han; Pei; Tong, 2012).

2.5.4.3 Precisão

A precisão, representada pela Equação 5, pode ser considerada como uma medida de exatidão do classificador, e representa a proporção de tuplas que foram classificadas como positivas e são de fato positivas (Han; Pei; Tong, 2012).

$$\text{precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{5}$$

Uma pontuação de 100% para a precisão significa que todos os elementos classificados como positivos são realmente positivos, porém, da maneira análoga ao que foi constatado para a sensibilidade, essa métrica não considera a quantidade de positivos que foram incorretamente considerados como negativos. Dessa forma, as métricas de sensibilidade e precisão são comumente utilizadas em conjunto, de maneira que valores de precisão são comparados entre si para um valor fixo de sensibilidade, ou vice-versa (Han; Pei; Tong, 2012).

2.5.4.4 Pontuação F_β

Uma forma alternativa de se utilizar a precisão e sensibilidade, além de fixar um dos valores em uma constante e comparar os valores para o outro, é combiná-los em uma única métrica, chamada de pontuação, ou medição, F_β. Essa métrica, representada pela Equação 6, define um peso β para a sensibilidade e precisão, de forma que a sensibilidade tenha β vezes mais peso que a precisão (Han; Pei; Tong, 2012).

$$F_\beta = \frac{(1 + \beta^2) \times \text{precisão} \times \text{sensibilidade}}{\beta^2 \times \text{precisão} + \text{sensibilidade}} \tag{6}$$

Percebe-se, pela observação da Equação 6, que uma pontuação F_2 faz com que a sensibilidade tenha um peso duas vezes superior à precisão, e $F_{0,5}$ faz com que a precisão tenha peso 2 vezes superior. Caso β seja definido como 1, a pontuação F_1 resultante é uma média harmônica entre a sensibilidade e precisão, e o peso para ambas é o mesmo (Han; Pei; Tong, 2012).

2.6 TRABALHOS SIMILARES

Nesta seção serão apresentadas as sínteses de trabalhos que trazem em si técnicas e raciocínios similares aos utilizados neste. Uma versão completa da análise pode ser visualizada no Apêndice A.

Em Queiroga *et al*. (2018), foi analisada a evasão em um curso de graduação de dois anos utilizando a contagem de interações dos estudantes no ambiente virtual MOODLE como a principal fonte de informação para gerar e comparar modelos analíticos. Os autores utilizaram Bayes Net (BN), Simple Logistic (SL), Multilayer Perceptron (MLP), Random Forest (RF) e J48, sendo que o modelo RF atingiu acurácia de 99% a partir da quinta semana (5% do tempo total do curso).

Burgos *et al*. (2018) exploraram a evasão acadêmica usando uma metodologia incremental com base nas atividades dos estudantes ao longo de um semestre. Foi aplicada a regressão logística (RL) para prever a evasão e compararam-na com outras técnicas de aprendizado de máquina, encontrando uma precisão de 98,95%, sensibilidade de 96,73%, especificidade de 97,14% e acurácia de 97,13% na semana média de evasão. O modelo analisa atividades realizadas pelos alunos e pode identificar a evasão, em média, 1,6 semanas antes de ocorrer.

No estudo de Gottardo, Kaestner e Noronha (2014), o ambiente MOODLE é utilizado com o objetivo de prever desempenho dos estudantes, sem considerar a evasão ou desistência. São aplicados os algoritmos RF e MLP, sendo que o RF alcança acurácia média de 77,4%, enquanto o MLP apresenta acurácia média de 80,1%. Não são mencionadas tentativas de antecipar o desempenho acadêmico usando modelos com uma base de dados reduzida.

No estudo de Viana, Santana e Rabêlo (2022) a avaliação se concentra na evasão do curso de graduação como um todo e a coleta de dados foi realizada por meio de atributos sociais e acadêmicos. RF, MLP, Support Vector Machine (SVM), Árvore de Decisão (AD), Extra Trees (ET), K-Nearest Neighbors (kNN) e Gaussian Naive Bayes (GNB) foram os algoritmos de classificação avaliados, sendo o melhor resultado obtido a partir de RF, com precisão média de 91,55% e sensibilidade de 92%. Não foram realizadas tentativas de antecipar a previsão.

Em Garcia *et al*. (2022), os autores criaram um ambiente em que os professores pudessem prever a probabilidade de um aluno ser aprovado ou reprovado no início do semestre e a coleta de dados se deu a partir de informações acadêmicas passadas e pessoais. Dos algoritmos de aprendizado de máquina avaliados, que incluem Naive Bayes (NB), Instance Based Learner (IBK), JRip, J48, RF e MLP, RF teve melhor resultado, com sensibilidade de 93% e uma acurácia de 81,43% para as turmas em busca da graduação em Matemática, e uma sensibilidade de 80,5% e uma acurácia de 74,01% para o conjunto de alunos do curso de Computação.

O estudo de Manhães *et al*. (2011) visa prever a evasão de alunos em um curso de graduação de 5 anos com dados acadêmicos do primeiro semestre da graduação. Os algoritmos de classificação utilizados foram: One Rule (OneR), JRip, Decision Table (DT), Simple Cart (SC), J48, RF, SL, MLP, NB e BN. Embora alguns algoritmos tenham alcançado altas acurácias, como o SL com 82,29%, eles também apresentaram altas taxas de falsos positivos (36%), levando os pesquisadores a considerá-los inadequados devido ao erro crítico de classificar alunos em risco de evasão como não em risco. Algoritmos como MLP e RF, com acurácias ligeiramente inferiores (74,31% e 80,21%, respectivamente), demonstraram taxas de falsos positivos mais baixas, com 27% e 29%, respectivamente, tornando-os mais apropriados para a tarefa de identificação de alunos em risco.

O Quadro 2 representa a síntese dos trabalhos similares observados, elencando técnicas utilizadas, objetivo da predição, métricas e resultados considerados na avaliação, tipos de dados analisados e tempo de antecedência para obter a predição, caso tenha sido realizado.

A partir dos trabalhos similares é possível perceber que os problemas tratados são, em sua maioria, relacionados à evasão de curso e não à desistência de disciplina. Normalmente, leva-se em consideração o comportamento em vários semestres, bem como dados além daqueles obtidos em disciplinas, como a presença e notas. A maioria dos trabalhos avaliou os resultados não considerando apenas a acurácia, mas também a precisão e sensibilidade. O uso da precisão e sensibilidade é apropriado, uma vez que a evasão/desistência são situações que deveriam ser atípicas, e a acurácia poderia mascarar os resultados obtidos.

Além das métricas de avaliação, é importante considerar a identificação dos alunos desistentes de forma antecipada. Ou seja, a classificação não é útil se considerar todo o desempenho do estudante no semestre, sendo necessário identificar o quanto antes para agir a fim de evitar a desistência. Nesse sentido, apenas os trabalhos Queiroga *et al*. (2018), Burgos *et al*. (2018) e Garcia *et al*. (2022) consideraram suas classificações levando em consideração essa antecipação.

Quadro 2 – Síntese dos trabalhos analisados

Trabalho	Técnica de predição	Objetivo	Resultado	Dados analisados	Tempo de antecipação
Queiroga *et al*. (2018)	BN, SL, MLP, RF e J48	Prever evasão em um curso de 2 anos	99% de acurácia	Contagem de interações no MOODLE	5% do tempo total do curso
Burgos *et al*. (2018)	Regressão logística	Prever evasão ao longo do semestre	Precisão de 98,95%, sensibilidade de 96,73% e acurácia de 97,13%	Notas em atividades	Em média, 1,6 semana antes da semana média de evasão
Gottardo, Kaestner e Noronha (2014)	RF e MLP	Prever nota final dos estudantes	Acurácia média de 80,1%	Interações com ambiente virtual	Não realizado
Viana, Santana e Rabêlo (2022)	RF, MLP, SVM, AD, ET, KNN e GNB	Prever a evasão de alunos em um curso de graduação	Precisão média de 91,55% e sensibilidade de 92%	Atributos de natureza social e acadêmica	Não realizado
Garcia *et al*. (2022)	NB, IBK, JRip, J48, RF, MLP	Prever a aprovação ou reprovação no início do semestre	Recall de 93% e acurácia de 81,43%	Informações acadêmicas passadas e pessoais	Não se aplica
Manhães *et al*. (2011)	OneR, JRip, DT, SC, J48, RF, SL, MLP, NB e BN	Previsão da evasão em um curso de 5 anos	Acurácia de 80,21% e taxa de FP de 29%	Dados acadêmicos do primeiro semestre da graduação do aluno	Não realizado

Fonte: Elaborado pelo autor (2023).

3 MÉTODO

A metodologia empregada nesta pesquisa consiste em um estudo de caso, conforme delineado por Martins (2008), estratégia de se pesquisar no meio das ciências sociais. Esta metodologia envolve a análise de dados referentes a disciplinas específicas, obtidos através do AVA MOODLE. O objetivo principal é avaliar indicadores que possam sinalizar a propensão de desistência de alunos na disciplina em questão e desenvolver um modelo analítico capaz de prever tal comportamento.

3.1 METODOLOGIA

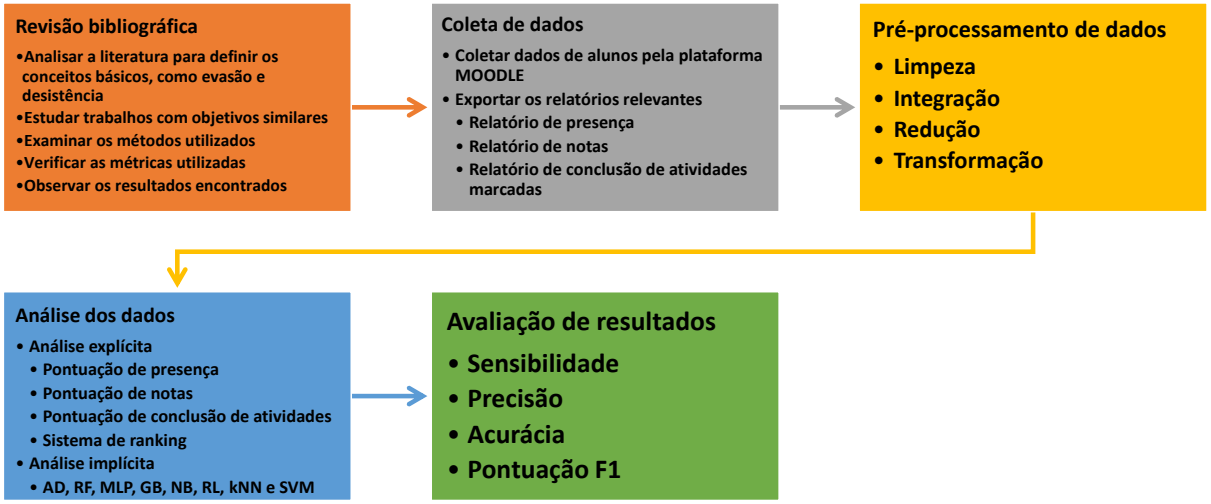
Nesta seção, será descrita a metodologia empregada para o desenvolvimento do trabalho, dividida em cinco estágios: revisão bibliográfica; coleta; pré-processamento; análise; e avaliação de dados. A primeira etapa consiste em definir conceitos básicos, como evasão e desistência, além de buscar e analisar trabalhos com objetivos e desenvolvimento semelhantes aos deste, a fim de comparar os métodos e métricas de avaliação utilizados e os resultados atingidos. A fase de coleta de dados é composta pela análise da plataforma MOODLE e extração dos dados relevantes ao trabalho, os quais foram definidos na primeira etapa.

O pré-processamento dos dados é separado nas etapas de limpeza, integração, redução e transformação. De modo geral, consiste em preparar os dados para a etapa de análise, na qual será necessário providenciar informações sem inconsistências (limpeza), com suas diferentes fontes unificadas (integração), reduzidos de forma que ainda representem o conjunto como um todo (redução) e apresentados de maneira que facilite sua investigação (transformação).

Com os dados pré-processados, sua análise é realizada de forma mais eficiente. Para o presente trabalho, a análise é dividida em explícita e implícita. A primeira objetiva desenvolver equações e métodos de análise a partir de observações realizadas diretamente com os dados, enquanto que a segunda utiliza métodos de aprendizado de máquina para gerar resultados. Por fim, a avaliação dos resultados parte dos frutos da análise de dados, a partir dos quais são calculadas as métricas de sensibilidade, precisão, acurácia e pontuação F_1. A Figura 7 apresenta uma ilustração sequencial da metodologia descrita.

Dessa forma, pôde-se organizar o andamento do trabalho e definir quais etapas são pré-requisito para outras.

Figura 7 – Diagrama que descreve a metodologia empregada



Fonte: Elaborado pelo autor (2023).

3.2 EXTRAÇÃO DE DADOS DO MOODLE

As análises foram baseadas em três relatórios essenciais do MOODLE: presença, notas e conclusão de atividades marcadas como importantes. A escolha desses elementos do ambiente se deu pela facilidade de acesso e porque não exigem plugins adicionais ou acesso ao banco de dados, evitando restrições de instituições usuárias do MOODLE. Assim, os meios propostos permitem uma aplicação mais simples e menos burocrática.

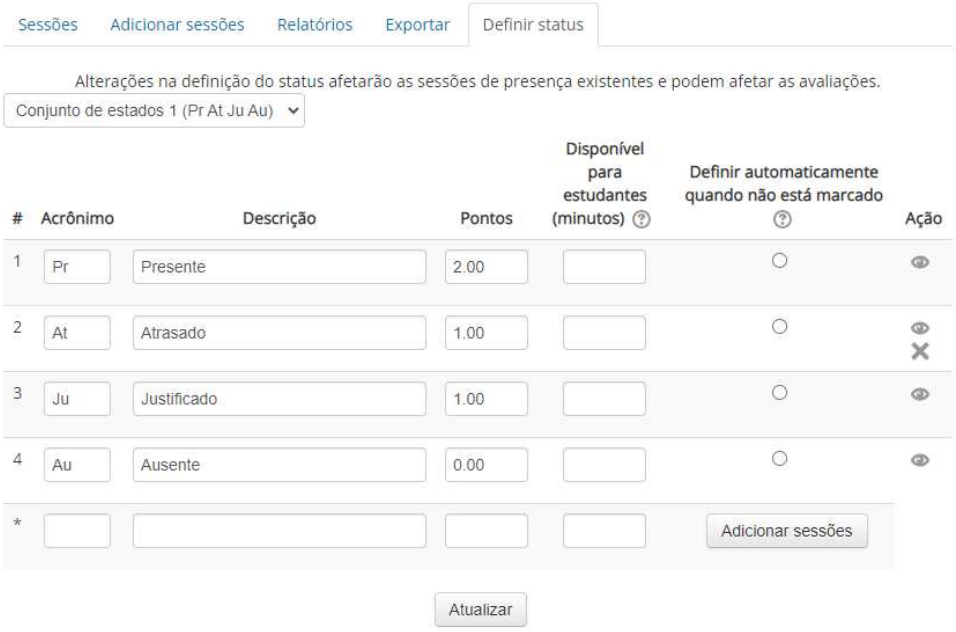
3.2.1 Relatório de presenças

O MOODLE oferece um plugin para gerenciamento de presenças[1], o que facilita o controle de frequência no sistema. A Figura 8 ilustra como esse recurso pode ser configurado, permitindo ajustes em vários elementos, como a pontuação para cada sessão de presença registrada. De maneira geral, a presença é classificada de três formas: alunos que alcançam a pontuação máxima por sessão são considerados presentes; aqueles com pontos intermediários são rotulados como justificados ou atrasados; e aqueles sem pontos são marcados como faltantes ou ausentes. No relatório exportado, cada categoria é representada por uma fração, na qual o denominador é a maior pontuação possível e o numerador indica a quantidade de pontos obtida.

A Figura 9 apresenta um segmento do relatório de presenças do MOODLE, formatado como uma planilha eletrônica. A descrição que acompanha cada tipo de registro de presença pode variar conforme a configuração estabelecida pelo professor.

[1] Link para o plugin de presenças do MOODLE: https://moodle.org/plugins/mod_attendance

Figura 8 – Configuração de presenças no plugin de presenças do MOODLE



Fonte: Elaborado pelo autor (2023).

Contudo, no fim de cada descrição, há sempre uma indicação entre parênteses que mostra o registro de pontos para aquele dia. Na figura exemplificada, as presenças são categorizadas como completa (2/2), justificada (1/2) ou ausência (0/2). Há também situações em que a pontuação é indicada por um ponto de interrogação, que significa que a presença ainda não foi registrada.

Figura 9 – Parte do relatório de presenças em formato de planilha eletrônica

4	ID do Estu	Sobrenom	Nome	7/03/2023	8/03/2023
5	1	11	111	Au (0/2)	Ju (1/2)
6	2	12	112	Au (0/2)	Ju (1/2)
7	3	13	113	Pr (2/2)	Pr (2/2)

Fonte: Elaborado pelo autor (2023).

Destaca-se que o relatório de presenças dispõe a informação de pontuação temporalmente pelas colunas, de forma que a primeira coluna represente o primeiro dia de aulas, a segunda coluna represente o segundo dia, incrementalmente até a última coluna de pontuações, que representa o último dia de aula registrada. Para obter esses relatórios, é necessário acessar o módulo de presenças e, em seguida, selecionar a opção de exportação, que permite salvar os dados de presença no formato de planilha eletrônica padrão, como Microsoft Excel e OpenOffice, e em formato de texto (sendo que esse último não é de interesse deste trabalho).

3.2.2 Relatório de notas

O relatório de notas engloba avaliações individuais, categorias de avaliações e a nota global do curso, proporcionando uma visão abrangente do progresso do aluno. A Figura 10 exibe um exemplo de relatório de notas, que demonstra os variados tipos de atividades que podem ser utilizados para a atribuição de notas. Além disso, podem existir atividades registradas que, embora estejam no sistema, não tenham sido disponibilizadas ou avaliadas.

Figura 10 – Parte de um relatório de notas do MOODLE

Nome	Sobrenome	Matrícula	Tarefa: Prova	Tarefa: 1	Fórum: 1	Fórum: 2	Fórum: 3	P1	P2	Total do curso
1		1	-	-	10	-	-	-	-	1
2		2	-	1,6	10	5	-	0,5	6,5	2,5
3		3	-	4,3	10	8	-	7	10	6
4		4	-	8,8	10	6	-	8	10	9
5		5	-	7,3	10	8	-	8	10	8,5
6		6	-	4	10	-	-	5	1,5	3
7		7	-	4	10	10	-	-	1	2,5
8		8	-	5,7	10	10	-	2	-	4
9		9	-	-	10	-	-	5	-	2

Fonte: Elaborado pelo autor (2023).

O relatório não detalha as ponderações associadas a cada atividade na composição final da nota do curso, tampouco fornece informações sobre as datas de realização das atividades ou se o aluno completou uma atividade que ainda não foi avaliada pelo professor. Ressalta-se que a ordem de colunas é definida a partir da organização realizada pelo professor, portanto, sem padrão estabelecido. Para a obtenção do relatório, é preciso acessar o módulo de notas e, na sequência, selecionar a aba de exportação, também possibilitando salvar os dados no formato Microsoft Excel ou OpenOffice.

3.2.3 Relatório de conclusão de atividades

Dentro do ambiente MOODLE, os responsáveis por uma disciplina têm a opção de destacar certas atividades como significativas. Essa funcionalidade tem como propósito enfatizar atividades que contribuem significativamente para a média do aluno ou que possuem um peso maior em relação às outras. A Figura 11 exibe um trecho do relatório de conclusão de atividades, onde as duas primeiras colunas fornecem dados identificadores dos alunos e as colunas subsequentes representam as atividades enfatizadas pelo docente, incluindo o status de conclusão. As atividades não concluídas são indicadas na respectiva coluna com a expressão “Não concluído”, e a coluna adjacente permanece sem valor. Uma vez completada pelo discente, a coluna adjacente registra o momento da conclusão da atividade.

Figura 11 – Parte de um relatório de conclusão de atividades marcadas do MOODLE

	Endereço de email	Tarefa Importante 1	
Aluno 1	aluno1@ufsc.br	Concluído	Sunday, 12 Mar 2023, 18:58
Aluno 2	aluno2@ufsc.br	Não concluído	
Aluno 3	aluno3@ufsc.br	Não concluído	
Aluno 4	aluno5@ufsc.br	Concluído	Wednesday, 15 Mar 2023, 9:32
Aluno 5	aluno5@ufsc.br	Não concluído	
Aluno 6	aluno6@ufsc.br	Não concluído	
Aluno 7	aluno7@ufsc.br	Concluído	Wednesday, 1 Mar 2023, 11:46
Aluno 8	aluno8@ufsc.br	Concluído	Tuesday, 7 Mar 2023, 12:02
Aluno 9	aluno9@ufsc.br	Não concluído	
Aluno 10	aluno10@ufsc.br	Não concluído	

Fonte: Elaborado pelo autor (2023).

Embora seja uma prática facultativa, a emissão de relatórios de conclusão de atividades pode indicar uma hierarquia de importância entre as atividades. Por exemplo, materiais complementares incluídos na plataforma virtual podem ser úteis para o entendimento da disciplina, mas não compõem o cálculo da nota final. A finalização dessas atividades, demonstrando o engajamento do estudante com o conteúdo, pode ser um fator relevante na análise de desistência, mas não deve ter o mesmo impacto de uma avaliação semestral em termos de peso. Ressalta-se que, assim como no relatório de notas, a disposição das colunas neste relatório não segue um padrão estabelecido. Para acessar o relatório de atividades concluídas, é necessário ir até a seção “Relatórios” na “Administração do curso”, selecionar “Conclusão de atividades” e, em seguida, é possível baixar os dados em formato CSV.

3.2.4 Sobre as turmas analisadas

Para o desenvolvimento do trabalho, um professor realizou a marcação manual de desistentes e não desistentes em duas turmas de semestres diferentes. A primeira turma, que será chamada de Turma 1, possui 41 alunos matriculados e um total de 25 atividades realizadas, dentre essas, nenhuma foi marcada como importante no MOODLE. O tamanho do semestre analisado pelos relatórios extraídos dessa turma equivale a 37 aulas, sendo que 13 dessas não tiveram a presença aferida, resultando em entradas com pontos de interrogação, e três encontros tiveram 100% de presença. Esse conjunto representa uma turma da disciplina de Modelagem de Sistemas do primeiro semestre de 2023.

Na segunda turma, que será referida como Turma 2, foi registrado o ingresso de 86 alunos, com um total de 24 atividades realizadas durante o semestre. Novamente, nenhuma atividade foi marcada como importante pelo docente nesse período. No intervalo de tempo em que a disciplina foi realizada, 26 aulas foram lecionadas, sendo que para seis dessas a presença foi aferida para 100% dos alunos. Os dados dessa

turma são referentes à disciplina de Programação 1, também do primeiro semestre de 2023.

A primeira turma é ofertada para os cursos de Engenharia Mecatrônica, Bacharelado em Ciência e Tecnologia, e Engenharia de Transportes e Logística e a segunda para os cursos de Engenharia Aeroespacial, Engenharia Automotiva, Engenharia Ferroviária e Metroviária, e Engenharia Civil de Infraestrutura, no CTJ da UFSC. Ambas as disciplinas têm uma carga horária de 72 horas-aula[2] no semestre e são ministradas em dois encontros semanais. Para identificar os alunos considerados desistentes ao final do período de aplicação da disciplina, o professor responsável realizou uma marcação manual. Além disso, em ambas as disciplinas, foram aplicadas avaliações contínuas, incluindo questionários, atividades e fóruns, durante o semestre, quase semanalmente.

3.3 PRÉ-PROCESSAMENTO DOS DADOS

Nesta seção, serão descritas as etapas de preprocessamento de dados. A biblioteca pandas, da linguagem de programação Python, providencia uma maneira simplificada de converter planilhas do Microsoft Excel e OpenOffice para objetos de tipo DataFrame, facilitando os processos descritos nessa seção. O procedimento engloba quatro diferentes partes: limpeza, integração, transformação e redução dos dados.

3.3.1 Limpeza dos dados

O primeiro passo na limpeza de dados se deu a partir da anonimização dos dados, de forma que atributos identificadores dos alunos foram alterados para valores genéricos, como *Aluno 1* ao invés do nome real do discente. Em seguida, no caso do relatório de presenças, foram consideradas irrelevantes as informações de identificação numérica do estudante, usuário e curso, sobrenome do discente, e nome do curso, restando apenas as pontuações de presença atribuídas a cada dia marcado, os nomes dos alunos, a quantidade de cada tipo de presença (ausente, presente e atrasado, por exemplo), a quantidade de sessões anotadas, a relação de pontos e a porcentagem de presença.

Para o relatório de notas, removeu-se as informações de endereço de e-mail, identificação numérica do curso, nome do curso, sobrenome do aluno, matrícula do estudante, informações de download do relatório e soma de notas, como médias para categorias de atividades. Assim, restaram os valores das notas definidas para cada atividade e nome do aluno. Finalmente, para o relatório de conclusão de atividades, apenas a coluna com o endereço de e-mail dos alunos foi retirada, restando as

[2] 72 horas-aula são equivalentes a 60 horas-relógio.

informações de nome dos alunos, conclusão ou não de cada atividade marcada como importante e a marcação de tempo de sua conclusão.

3.3.2 Integração dos dados

Após remover informações não relevantes à análise, deu-se início ao processo de combinar os dados de diferentes fontes, neste caso, os três relatórios extraídos do ambiente MOODLE. Conforme descrito na etapa de limpeza dos dados, a informação do nome dos participantes foi mantida em todas as planilhas, devido à necessidade de correlacionar os dados de diferentes relatórios aos mesmos alunos. Dessa forma, a partir do nome dos estudantes, combinaram-se os dados de presença, notas e conclusão de atividades importantes em um objeto da classe *Classroom*[3], que representa a turma. Os atributos da classe que representam os relatórios, são:

- attendance_report: relatório de presenças;
- grade_report: relatório de notas;
- activity_report: relatório de conclusão de atividades marcadas como importantes pelo professor.

3.3.3 Transformação dos dados

Para fornecer os dados às técnicas de aprendizado de máquina, foi necessária uma etapa de transformação dos dados em variáveis e classificações que possam ser interpretadas pelos algoritmos. Como pode-se visualizar na Figura 9, as informações do relatório de presenças estão separadas em células, dentro das quais o valor numérico da pontuação, o valor máximo atingível e a categoria em que se enquadra são apresentados. Para este relatório, as informações de cada célula, referentes a cada aluno, foram organizadas sequencialmente em quatro listas, um tipo de estrutura de dados em Python:

- progression: pontuação atingida a cada dia;
- max_score: máxima pontuação atingível por dia;
- mean_score: média das pontuações da turma em determinado dia;
- missing_rate: taxa de faltantes da turma em determinado dia.

Para o relatório de notas, como ilustrado na Figura 10, e relatório de conclusão de atividades, visualizado na Figura 11, foram criados dicionários, estrutura de dados em Python, onde o nome da atividade serve como chave de acesso aos valores. Para o dicionário do relatório de notas, grade_report, foram criados cinco valores, para cada aluno e atividade:

- grade: nota obtida;

[3] Os códigos desenvolvidos para o presente trabalho podem ser encontrados em https://github.com/bernardodalfovo/tcc

- completed: se a atividade foi realizada pelo aluno, ou não;
- highest_grade: nota mais alta da turma, em determinada atividade;
- mean_grade: média das notas da turma, na mesma atividade;
- completion_rate: taxa de conclusão da atividade, pela turma.

Enquanto que para o relatório de conclusão de atividades, activity_report, apenas dois valores são especificados para cada aluno:

- completed: se a atividade foi realizada, ou não;
- timestamp: marcação de tempo da conclusão da atividade, não definido se a atividade não foi concluída.

Desta forma, os três relatórios, originalmente em formato de linhas e colunas, agora são organizados em estruturas de dados do Python, de forma que podem ser manipulados e acessados por aplicações que utilizem essa linguagem de programação.

3.3.4 Redução dos dados

Com o objetivo de analisar o comportamento de alunos individualmente, torna-se necessário converter os dados já transformados para refletir as atitudes dos discentes. Por conseguinte, as informações foram condensadas em atributos relativos à turma ou ao período analisado como um todo. Tal abordagem permite a generalização dos modelos gerados, assegurando que sejam aplicáveis mesmo em contextos distintos, como turmas com diferentes quantidades de atividades, alunos, ou aulas em um semestre. Para alcançar este fim, foram elaborados 22 atributos para cada aluno, a partir dos relatórios anteriormente mencionados:

- grades_average: média de todas as notas do aluno;
- grades_between_0_2_5: porcentagem de notas entre 0 e 2,5;
- grades_between_2_5_5: porcentagem de notas entre 2,5 e 5;
- grades_between_5_7_5: porcentagem de notas entre 5 e 7,5;
- grades_between_7_5_10: porcentagem de notas entre 7,5 e 10;
- grades_below_5: porcentagem de notas abaixo de 5;
- grades_below_mean: porcentagem de notas abaixo da média da turma;
- important_grades_below_mean: porcentagem de notas em atividades importantes abaixo da média de notas da turma;
- important_activities_complete_majority: porcentagem de atividades importantes completas pelo aluno quando a taxa de conclusão da turma na atividade é superior a 50%;
- important_activities_complete_minority: porcentagem de atividades importantes completas pelo aluno quando a taxa de conclusão da turma na atividade é inferior a 50%;

- important_activities_incomplete: porcentagem de atividades importantes incompletas;
- important_activities_incomplete_majority: porcentagem de atividades importantes incompletas pelo aluno quando a taxa de conclusão da turma na atividade é superior a 50%;
- important_activities_incomplete_minority: porcentagem de atividades importantes incompletas pelo aluno quando a taxa de conclusão da turma na atividade é inferior a 50%;
- activities_complete_majority: porcentagem de atividades completas pelo aluno quando a taxa de conclusão da turma na atividade é superior a 50%;
- activities_complete_minority: porcentagem de atividades completas pelo aluno quando a taxa de conclusão da turma na atividade é inferior a 50%;
- activities_incomplete: porcentagem de atividades incompletas;
- activities_incomplete_majority: porcentagem de atividades incompletas pelo aluno quando a taxa de conclusão da turma na atividade é superior a 50%;
- activities_incomplete_minority: porcentagem de atividades incompletas pelo aluno quando a taxa de conclusão da turma na atividade é inferior a 50%;
- attendance_below_mean: porcentagem de aulas em que a pontuação de presença do aluno foi abaixo da média de pontuações da turma;
- missing: porcentagem de faltas;
- partial_presence: porcentagem de presença parcial;
- sequencial_missing_X: quantidade de faltas sequenciais, compostas por X faltas maiores que um, no período.

Desta forma, o banco de dados que antes era dividido em três relatórios e estruturas de dados diferentes, é reduzido à variáveis numéricas para cada um dos alunos.

3.4 ANÁLISE EXPLÍCITA

Para que a fase de análise de dados possa tomar forma, se faz necessária uma etapa de classificação dos dados, em que se decide quais características definem determinado comportamento. No caso deste trabalho, o principal comportamento a ser identificado é a desistência, porém, durante o desenvolvimento do trabalho e com a extração de diferentes dados a partir do MOODLE, percebeu-se a necessidade de definir outros conceitos, que serão apresentados nesta seção.

3.4.1 Faltas consecutivas

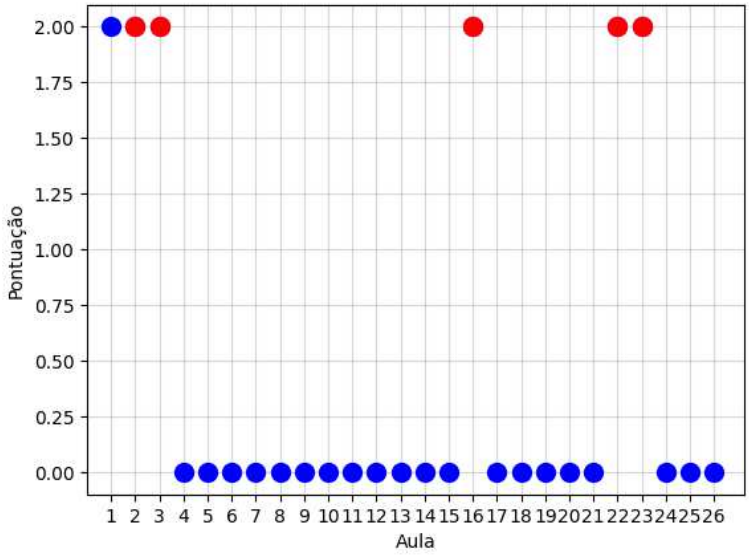
Inicialmente, este trabalho considerou o registro de presenças como principal indicador para identificar a desistência dos alunos, focando mais especificamente na

quantidade de faltas consecutivas do aluno. Durante a análise dos dados de presença, alguns problemas foram identificados em relação à marcação dessas presenças:

- O professor pode não realizar a presença em determinada data ou lançar as presenças obtidas de outra forma;
- O professor pode atribuir presença para a turma toda; e
- Existe a possibilidade dos alunos se auto atribuírem presenças (quando habilitado pelo professor), mas o professor precisa confirmar as ausências e pode não realizar tal ação.

Para o primeiro e terceiro caso, a pontuação de presença é substituída por uma entrada representada por um ponto de interrogação nos relatórios. Assim, mesmo que o aluno não estivesse presente em sala de aula, sua falta não seria computada e o método falharia em verificar a desistência, como pode-se perceber na Figura 12, que mostra em azul a pontuação obtida pelo aluno em dias regulares e em vermelho os dias em que a presença foi verificada para todos.

Figura 12 – Demonstração de dias em que todos os alunos receberam presença



Fonte: Elaborado pelo autor (2023).

Assim, pode-se afirmar que o aluno compareceu às aulas, com certeza, apenas na primeira aferição de frequência. Além disso, percebeu-se a necessidade de considerar fatores externos à sala de aula ao analisar a presença. Há dias em que forças maiores previnem os alunos de chegarem à aula, como desastres naturais e trânsito, ou até mesmo datas próximas à feriados, em que uma parte de alunos pode faltar para a realização de uma viagem. Diante disso, foi necessário desenvolver um método que pudesse abordar e contornar esses problemas.

3.4.2 Faltas consecutivas com média de presença da turma e dados interpolados

Para solucionar o problema de dias em que o professor considera presença para todos os alunos, até para os que não estão presentes, tais dias foram considerados como inválidos e uma interpolação linear foi realizada entre os dados válidos. Para dias em que a presença é indefinida (representada por um ponto de interrogação), definiu-se que caso todas as entradas para o dia sejam indefinidas, tal dia é considerado como presença para todos (ou seja, tratado como um dia inválido comum); caso contrário, a presença indefinida é considerada como falta para alunos que a possuem.

Como descrito na Seção 3.4.1, percebeu-se a necessidade de considerar dias em que uma parcela de alunos da turma não pôde comparecer à aula devido a fatores não relacionados com a aula em si. De maneira que tais fatores externos sejam considerados para a classificação de um aluno desistente, elaborou-se o cálculo da média de presença da turma. Assim, a falta apenas é considerada quando a pontuação está abaixo da média da turma para aquele dia. Adicionalmente, o cálculo do peso para um dia considerado inválido é realizado a partir da relação entre a média global de pontuação dos alunos durante todo o período analisado e a pontuação máxima atingível naquele dia. Para dias válidos, o peso para determinado dia é calculado a partir da média de pontuações da turma para tal dia, dividida pela pontuação máxima naquele dia.

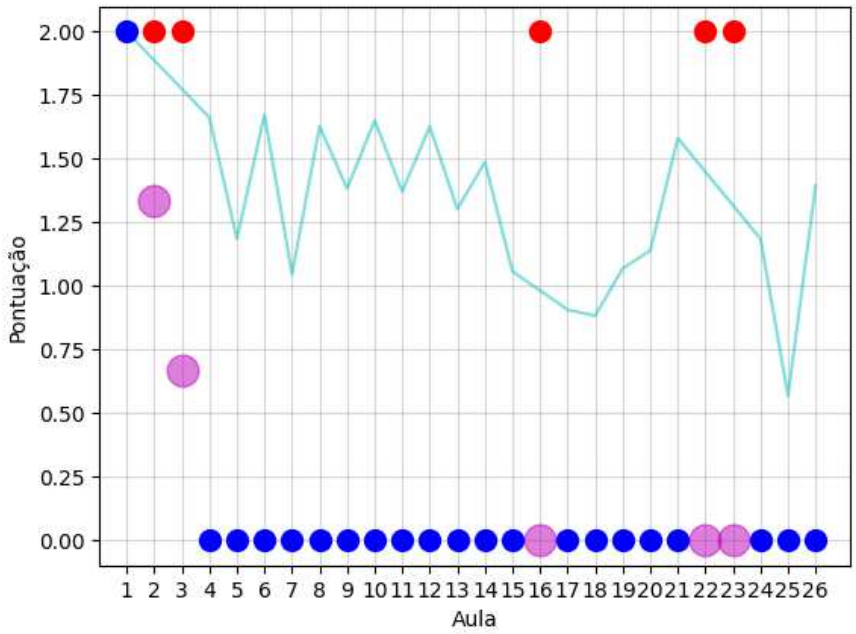
Um exemplo do método pode ser visualizado na Figura 13, na qual a média de pontuações da turma, a pontuação de presença em dias considerados válidos, a pontuação de presença em dias considerados inválidos e a pontuação interpolada para cada dia podem ser visualizados nas cores ciano, azul, vermelho e magenta, respectivamente.

Inicialmente, com base em uma análise empírica dos dados, definiu-se que cinco faltas consecutivas seriam um indicador de desistência de um aluno. Para avaliar a eficácia dessa definição, realizaram-se testes variando o número de faltas sequenciais necessárias para a classificação, considerando três, quatro e cinco faltas.

Porém, a partir de uma análise individual das pontuações interpoladas linearmente de cada aluno, percebeu-se que a utilização exclusiva de dados de presença e sua interpolação pode causar uma tendência a classificar alunos como desistentes. Por exemplo, na Turma 1, onde há três dias inválidos, caso um aluno tenha faltado uma aula antes do primeiro dia inválido, e uma depois, seriam consideradas 5 faltas seguidas, resultando na classificação automática do aluno como desistente, mesmo que essas fossem suas únicas ausências. Tal comportamento pode ser visualizado na Figura 14, que sua codificação por cores na mesma forma que a Figura 13.

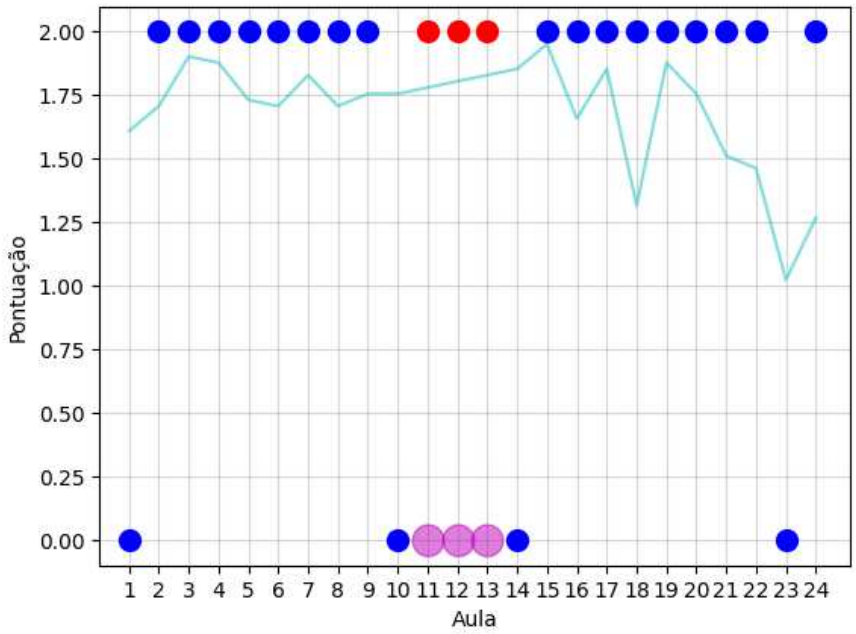
Dessa forma, a interpolação foi descartada, e uma nova técnica que não

Figura 13 – Uso da interpolação e média de pontuações da turma para definir se a pontuação de determinado dia é uma falta ou não



Fonte: Elaborado pelo autor (2023).

Figura 14 – Exemplo de uma falha de classificação no método com a interpolação de dados de presença



Fonte: Elaborado pelo autor (2023).

apenas considerasse a quantidade de faltas consecutivas como fator determinante à desistência foi desenvolvida.

3.4.3 Sistema de pontuação com notas, presença e conclusão de atividades importantes

Buscando aumentar a assertividade do modelo de análise, é adicionado o relatório de notas, descrito na Seção 3.2.2, e o relatório de conclusão de atividades, descrito na Seção 3.2.3, considerando o uso de marcar atividades para conclusão quando esta se destacar das demais, ou seja, utilizar o relatório para determinar que elementos do ambiente virtual são mais importantes que o estudante faça uso. A ideia foi atribuir algum grau de diferenciação entre os conteúdos disponibilizados.

Então, é proposto um sistema de pontuação e ranking para classificar os alunos nas categorias desistente e não desistente. Para essa classificação, utilizou-se uma porcentagem de corte da pontuação dos alunos, que tem um limite inferior de zero, e um limite superior de 100, na qual cada uma das três categorias de dados (presença, notas e conclusão de atividades) representa 1/3 da pontuação total, desta forma os alunos não são julgados por apenas uma competência. Essa distribuição igualitária entre as categorias, e os posteriores indicadores apresentados mais adiante, foram obtidos empírica e experimentalmente.

A motivação principal para a implementação desse método surgiu da observação de alunos que, já aprovados nas notas, mantêm frequência suficiente e optam por faltar quando necessário, sem risco de reprovação. Esse comportamento enfatiza a necessidade de uma avaliação mais abrangente que considere diferentes aspectos do desempenho do aluno.

3.4.3.1 Pontuação de presença

Ao calcular a pontuação de presença, primeiramente é necessário definir os pesos específicos para cada aula, representados pela Equação 7, de forma que se leve em consideração a proporção de alunos que estiveram presentes a cada dia. Desta forma, dias em que uma parcela maior da turma esteve ausente representam menor peso, e instâncias em que houve mais registros de presenças pela turma possuem mais importância ao cálculo.

$$W_i = \frac{Me_i}{Max_i} \tag{7}$$

Para um dia i, Max_i representa a pontuação máxima definida pelo professor no MOODLE, Me_i representa a média das pontuações do MOODLE da classe para aquele dia e W_i é o peso atribuído ao dia. Caso a presença não tenha sido registrada para o dia, Me_i é definido como a média de presenças da turma para o período analisado como

um todo. Assim, é possível calcular, para cada dia em que a presença foi computada para a turma, a pontuação de presença de acordo com a Equação 8.

$$P_f = \sum_{i=1}^{d} \frac{Po_i \cdot W_i}{Max_i} \tag{8}$$

Nessa equação, P_f representa a pontuação final, Po_i é a pontuação atribuída pelo MOODLE, W_i representa o peso calculado, Max_i é a pontuação máxima atribuída pelo MOODLE, i é o dia de presenças analisado e d é o número total de dias no relatório de presença. Caso Po_i seja equivalente a Max_i, o peso não é levado em consideração.

3.4.3.2 Pontuação de conclusão de atividades

Para se levar em consideração a quantidade de atividades concluídas, marcadas como importantes ou não, elaborou-se a Equação 9, que representa a pontuação para a categoria de atividades concluídas. O cálculo segue uma lógica de que cada atividade não realizada pelo aluno resulta em uma redução da sua pontuação, que se inicia em 100. A fórmula também leva em consideração a proporção de conclusão de cada atividade, assim, estudantes que não realizam uma atividade concluída por uma parcela maior da turma, tem um prejuízo maior em sua pontuação, em relação àqueles que não finalizaram uma atividade com baixa interação.

$$P'_f = \frac{Po_{max} - \sum_{j=1}^{n}(W'_j \cdot t_j)}{Po_{max}} \cdot 100 \tag{9}$$

Na qual P'_f representa a pontuação de atividades concluídas final, Po_{max} é soma das pontuações máximas de todas as atividades, W'_j o peso definido para a atividade j, t_j a taxa de conclusão da atividade j, com relação à turma, e n é o número total de atividades realizadas. Se a atividade j foi concluída, W'_j equivale a zero; caso seja concluída e não esteja marcada como importante, é igual a um; caso seja concluída e esteja classificada como importante, possui valor igual a 1,4. Desta forma, atividades classificadas como importantes no MOODLE possuem peso 40% superior em relação às que não são classificadas como tal.

3.4.3.3 Pontuação de notas

Para a categoria de notas, adota-se uma lógica similar à utilizada para a conclusão de atividades. Neste caso, notas relativamente mais baixas em comparação com a média da turma reduzem mais a pontuação de um estudante do que notas de mesmo valor em avaliações onde toda a turma teve desempenho baixo. A fórmula para calcular a pontuação das notas é representada pela Equação 10. Este cálculo considera o peso de cada nota individual de uma atividade, relacionando-a com a maior

nota obtida na turma para essa atividade e multiplicando pelo valor médio das notas da turma na mesma atividade. Após somar os valores para todas as atividades, o total é subtraído da pontuação máxima possível e dividido por este valor máximo.

$$P''_f = \frac{(G_{max} - \sum_{k=1}^{n}(1 - \frac{G_k}{A_k}) \cdot M_k)}{G_{max}} \cdot 100 \quad (10)$$

Nessa equação, P''_f representa a pontuação final para a categoria de notas, G_{max} representa a maior soma de notas possível, G_i equivale à nota individual para a atividade i, A_i simboliza a maior nota obtida entre a classe para a atividade i, M_i é a média de notas da classe para a atividade i, e n é a quantidade total de atividades realizadas.

3.4.3.4 Algoritmo de classificação

A pontuação final de cada aluno é estabelecida mediante o cálculo da média aritmética dos resultados obtidos a partir das equações propostas (Equação 8, Equação 9 e Equação 10) para avaliar a presença em aulas, a conclusão das atividades e as notas obtidas. Uma vez obtida a pontuação final para cada aluno da turma, procede-se à criação de um ranking ordenado, no qual o aluno que alcançar a maior pontuação é posicionado em primeiro lugar, enquanto o aluno com a menor pontuação é colocado no final do ranking.

A fim de determinar a classificação entre desistentes e não-desistentes, é imperativo estabelecer um limite percentual para identificar os desistentes. Isso implica em definir uma porcentagem de corte, de modo que os alunos cujas pontuações se situem abaixo deste percentil sejam classificados como desistentes. Como critério de nota de corte pode ser utilizada a quantidade de alunos que o professor considera que conseguiria dedicar atenção especial para evitar a desistência, ou se basear na desistência histórica da disciplina. Adicionalmente, as análises foram feitas com e sem a interpolação para então verificar sua necessidade.

Além disso, para confirmar a validade da média aritmética como método de cálculo da pontuação final, foi realizada uma análise exaustiva de todas as combinações possíveis de pesos para as pontuações, com resolução de 1% em um intervalo de 0 a 100%, para verificar qual distribuição de peso resultaria em uma classificação com melhores métricas.

3.4.4 Sistema de pontuação e ranking com três classificações

A partir da análise das bases de dados, e conforme as conclusões obtidas em Manhães *et al*. (2011), notou-se a presença de alunos com comportamento considerado atípico. Esse comportamento pode ser separado em alunos com notas geralmente

acima da média da turma, mas que desistiam, e alunos com notas geralmente abaixo da média da turma, mas que eram aprovados na disciplina.

Desta forma, adaptou-se o sistema de pontuação e ranking para classificar os alunos em três categorias, ao invés de duas, sendo assim necessário também adaptar a "verdade"[4] da base de dados. Para classificar os alunos dentre as três categorias, utilizou-se uma separação em percentis da pontuação dos alunos, de forma que 1/3 dos alunos com pontuação mais baixa fosse classificado como em perigo de desistência, alunos entre 33% e 66% fossem classificados com perigo intermediário, e alunos dentre os 33% mais altos do ranking de pontuação fossem considerados não desistentes.

A ideia por trás desse método é verificar o perfil do aluno que não está sendo detectado pelo programa. Constitui uma falha significativa do modelo a categorização incorreta de um estudante que efetivamente abandona o curso como não desistente. Em contrapartida, a classificação equivocada de um aluno intermediário, que exibe padrões comportamentais não convencionais, como não desistente, é considerada um erro de menor magnitude.

3.5 ANÁLISE IMPLÍCITA

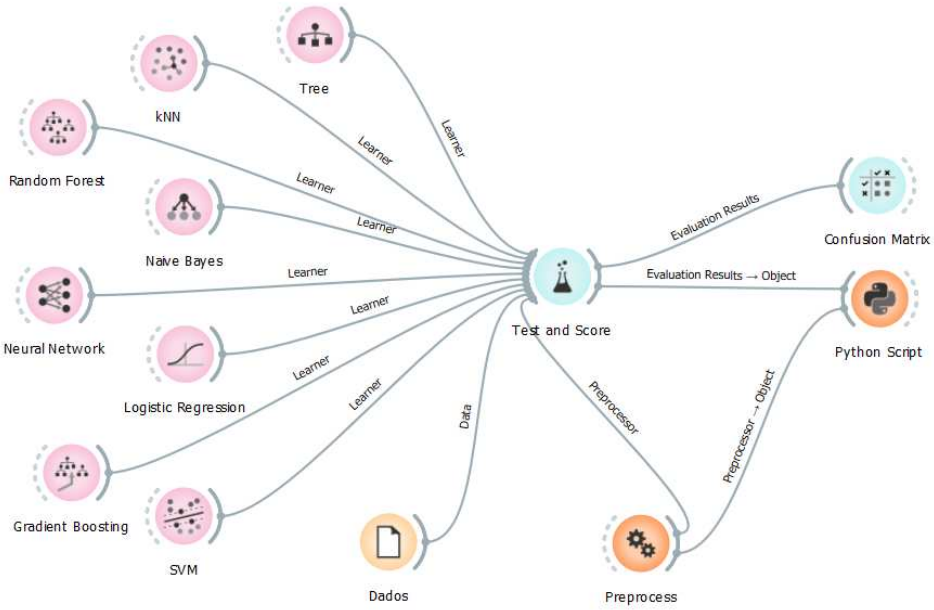
Para a realização da análise implícita, utilizou-se a ferramenta de mineração de dados Orange[5]. Este software oferece ao usuário a possibilidade de criar um fluxo de trabalho personalizado com seus módulos disponíveis. Estes módulos abrangem uma gama de funcionalidades, incluindo modelos de aprendizado de máquina, testes, avaliação e visualização de resultados, bem como preprocessamento de dados. O programa foi escolhido por sua facilidade de utilização, variedade de modelos de aprendizado de máquina nativamente disponíveis, visualização simplificada de árvores de decisão e resultados atingidos, além da compatibilidade com a linguagem de programação Python, possibilitando a importação de código-fonte externo por meio de seu bloco de programação.

3.5.1 Fluxo desenvolvido

A Figura 15 apresenta um fluxo desenvolvido na ferramenta Orange, onde são empregadas oito diferentes técnicas de aprendizado de máquina: AD, KNN, RF, NB, MLP, RL, Gradient Boosting (GB) e SVM. Tais técnicas foram selecionadas a partir de sua disponibilidade na ferramenta escolhida e de uma análise dos trabalhos similares, o que possibilita uma comparação direta entre os resultados alcançados no presente trabalho e os resultados mostrados na Seção 2.6, além de utilizar técnicas não antes verificadas em trabalhos similares.

[4] Entende-se como "verdade", a classificação manual realizada pelo docente durante a elaboração da base de dados fornecida para o desenvolvimento do trabalho.
[5] Link para obtenção do Orange: https://orangedatamining.com/.

Figura 15 – Fluxo desenvolvido na ferramenta Orange



Fonte: Elaborado pelo autor (2023).

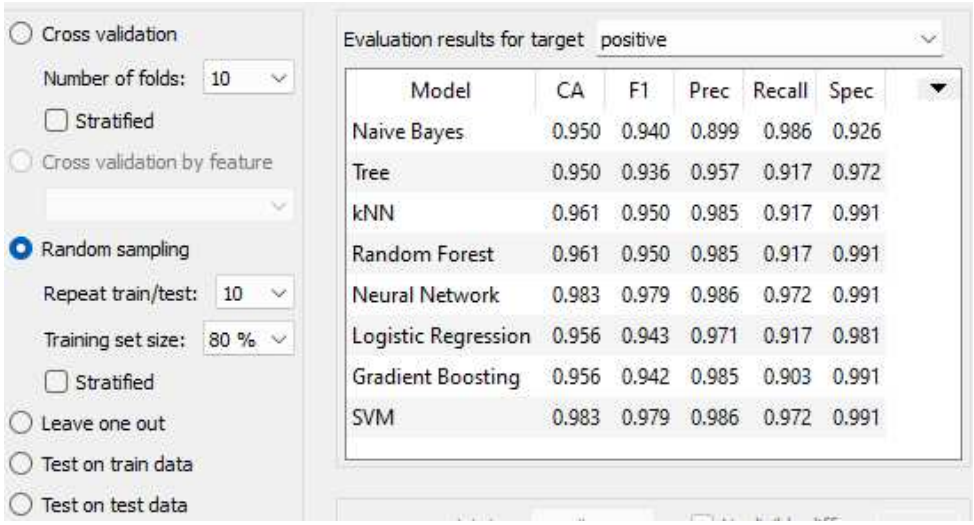
A ferramenta também permite a customização das técnicas de aprendizado de máquina, como limitar a profundidade e número de folhas na Árvore de Decisão. Cada uma das técnicas utilizadas teve suas configurações ajustadas de forma empírica/experimental. As técnicas de aprendizado de máquina escolhidas são ligadas ao bloco de teste e avaliação de resultados que, ao receber dados de entrada e, opcionalmente, métodos de preprocessamento, realiza a aplicação dos modelos aos dados.

A Figura 16 demonstra a visualização de resultados a partir do bloco de teste e avaliação, onde é possível verificar as métricas atingidas por cada técnica de aprendizado de máquina. Também é possível selecionar o método de divisão dos dados, como amostragem aleatória dos dados, divisão entre teste e treinamento (na qual os dados de teste e treinamento devem ser providos e identificados separadamente na ligação ao bloco) e validação cruzada.

Inicialmente, devido à facilidade de utilização e visualização do fluxo, a interface gráfica da ferramenta Orange foi utilizada para gerar resultados a partir das análises dos algoritmos de aprendizado de máquina. Porém, durante o desenvolvimento do projeto, percebeu-se a necessidade de codificar o fluxo gerado no programa para a linguagem de programação Python, visto que simplificou a exportação de resultados, execução de estruturas de repetição (for e while, por exemplo) e edição das bases de dados. Ao invés da utilização da ferramenta, passou-se então a utilizar a biblioteca do Orange3[6] em Python.

[6] Link para a página da biblioteca: https://github.com/biolab/orange3.

Figura 16 – Bloco de teste e avaliação de resultados do Orange



Fonte: Elaborado pelo autor (2023).

3.5.2 Divisão do conjunto de dados

Para realizar testes e avaliar os resultados das técnicas de aprendizado de máquina, a ferramenta disponibiliza diferentes métodos de divisão da base de dados em treinamento (parte dos dados em que os modelos serão treinados) e teste (parte dos dados em que os modelos serão aplicados e avaliados). Foram escolhidos três métodos: amostragem aleatória de dados, com 67% para treinamento e 33% para testes; holdout com 67% dos dados para treinamento e 33% para testes; e validação cruzada 10-fold ($k = 10$).

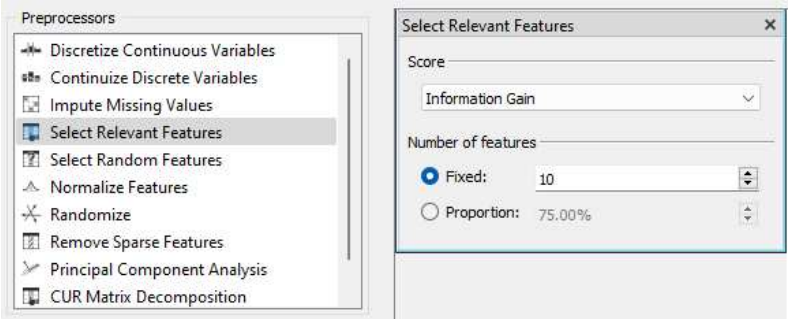
Devido ao caráter aleatório na seleção de estudantes para ambos os conjuntos de dados, implementa-se a técnica de amostragem aleatória dez vezes em cada análise. Este procedimento visa aprimorar a capacidade de representação precisa da base de dados, mitigando as variações inerentes à seleção aleatória e promovendo uma amostra mais representativa da população estudada. Para a validação cruzada, foi definida a quantidade de 10 divisões dos dados. Tanto para a amostragem aleatória quanto para a validação cruzada, não há a necessidade de dividir os dados manualmente, apenas fornece-los aos modelos.

Contudo, ao utilizar o método holdout para a divisão de dados, foi essencial manter a representatividade dos comportamentos de desistência em ambos os conjuntos, treinamento e teste. Desse modo, além de dividir os dados manualmente na proporção definida de 67% para treinamento e 33% para testes, assegurou-se que a proporção de alunos classificados como desistentes fosse equivalente nos dois grupos. Isso significa que tanto o conjunto de treinamento quanto o de teste refletem de forma proporcional a distribuição real de desistentes, garantindo uma análise mais precisa e representativa dos padrões de desistência.

3.5.3 Seleção de atributos

Com a finalidade de aprimorar as análises e identificar os atributos mais relevantes, utilizou-se o módulo de preprocessamento disponível no Orange. A Figura 17 mostra este módulo, que permite aos usuários escolher entre diferentes algoritmos para avaliar a relevância dos atributos e definir o número de atributos a serem selecionados para fornecer aos modelos de aprendizado de máquina.

Figura 17 – Bloco de pré-processamento do Orange



Fonte: Elaborado pelo autor (2023).

Assim como em Viana, Santana e Rabêlo (2022), optou-se pelos algoritmos de seleção Chi2 e ANOVA, e também foram utilizados GI e RG. Para determinar o número ideal de atributos relevantes, considerando os 22 disponíveis, as análises foram conduzidas com quantidades variadas de características, especificamente com conjuntos de cinco e dez atributos.

3.6 REDUÇÃO DA BASE DE DADOS

Como o objetivo principal do trabalho é definir quais alunos apresentam comportamento que aponta à desistência da disciplina *com tempo suficiente* para uma intervenção por parte do professor, da mesma forma que em Kaensar e Wongnin (2023), realizou-se a redução dos conjuntos de dados em porções de 25%, 50% e 75%, além de 100%, que representa toda a informação disponível. Dessa forma, foi possível avaliar as métricas atingidas pelo programa simulando a utilização por parte de um docente no decorrer do semestre, tanto para a análise explícita, quanto para a análise implícita.

Como as informações presentes no relatório de presenças são fornecidas temporalmente, de forma que a primeira coluna represente o primeiro dia, a segunda coluna represente o segundo dia, até a última coluna, que representa o último dia de aulas, a redução dos dados se deu simplesmente a partir da seleção de 25% das colunas, incrementalmente até 100%.

A Figura 18 ilustra um exemplo da ordem de colunas do relatório de notas, a

partir do qual percebeu-se que a sequência era definida a partir da disposição dos elementos na página da disciplina no MOODLE. Tal disposição é organizada pelo professor e pode seguir qualquer formato desejado por este.

Figura 18 – Exemplo da ordem de atividades no relatório de notas exportado

Nome
Laboratório Virtual de Programação: VPL 1 - Fatorial (Real)
Laboratório Virtual de Programação: VPL 2 - Valor de uma série (Real)
Laboratório Virtual de Programação: VPL 3 - Produto interno (Real)
Laboratório Virtual de Programação: VPL 4 - Maior valor do vetor (Real)
Laboratório Virtual de Programação: VPL 5 - Média do vetor (Real)
Laboratório Virtual de Programação: VPL 6 - Valores ímpares de uma matriz (Real)
Laboratório Virtual de Programação: VPL 7 - Palavra ao contrário (Real)
Laboratório Virtual de Programação: VPL 8 - Função para alterar valores do vetor (Real)
Laboratório Virtual de Programação: VPL 9 - Muitos vetores (Real)
Questionário: Avaliação do conteúdo - sistemas de numeração (Real)
Questionário: Avaliação do conteúdo - pseudocódigo (Real)
Questionário: Avaliação do conteúdo - básico C (Real)
Questionário: Avaliação do conteúdo - desvio condicional (Real)
Questionário: Avaliação do conteúdo - estruturas de repetição (Real)
Questionário: Avaliação do conteúdo - vetores (Real)
Questionário: Avaliação do conteúdo - strings (Real)
Questionário: Avaliação do conteúdo - matrizes (Real)
Questionário: Avaliação do conteúdo - ponteiros (Real)
Questionário: Avaliação do conteúdo - funções (Real)
Laboratório Virtual de Programação: VPL final 1 - troca 100 [1,5 ponto] (Real)
Laboratório Virtual de Programação: VPL final 2 - matriz [1,5 ponto] (Real)
Laboratório Virtual de Programação: VPL final 3 - vetores com função [2,0 pontos] (Real)

Fonte: Elaborado pelo autor (2023).

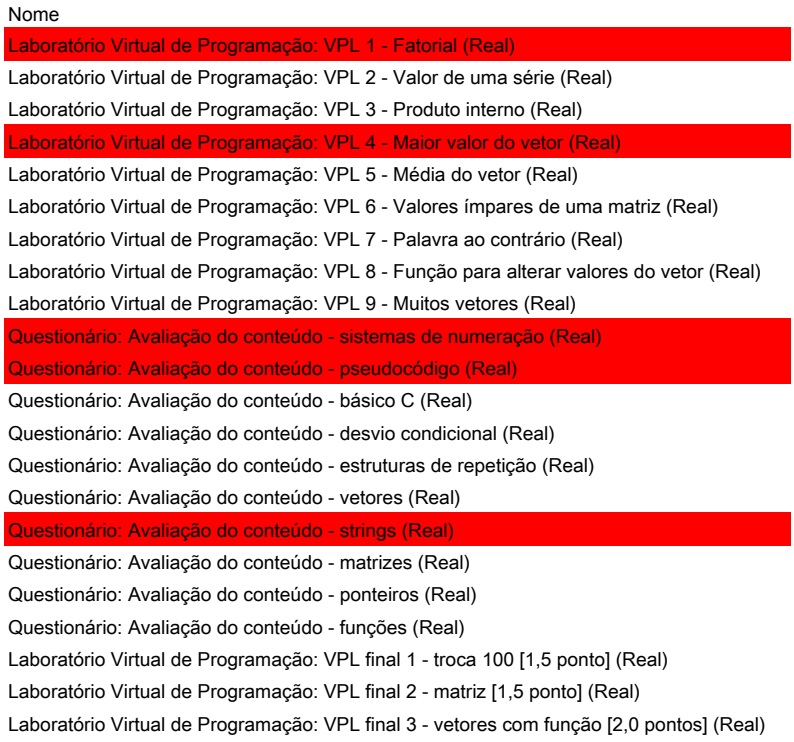
Conforme descrito na Seção 3.2.2, o relatório de notas contém apenas o nome da atividade e as notas obtidas por cada aluno, sem detalhar a contribuição de cada nota para a média final. Essa limitação impede a análise da relevância relativa de diferentes resultados. Portanto, a seleção de atividades para análise foi feita de forma aleatória. Um exemplo disso pode ser visto na Figura 19, onde 5 das 23 atividades, representadas pelas células marcadas em vermelho, foram escolhidas para compor 25% das atividades analisadas.

Quando se busca analisar metade da base de dados do relatório de notas, um exemplo dessa seleção é ilustrado na Figura 20. Neste caso, 11 das 23 atividades foram selecionadas, marcadas em vermelho, para serem utilizadas na análise.

Por fim, para a análise de 75% dos dados disponíveis no relatório de notas, um exemplo de seleção aleatória de atividades é apresentado na Figura 21, onde 17 das 23 atividades foram escolhidas para análise.

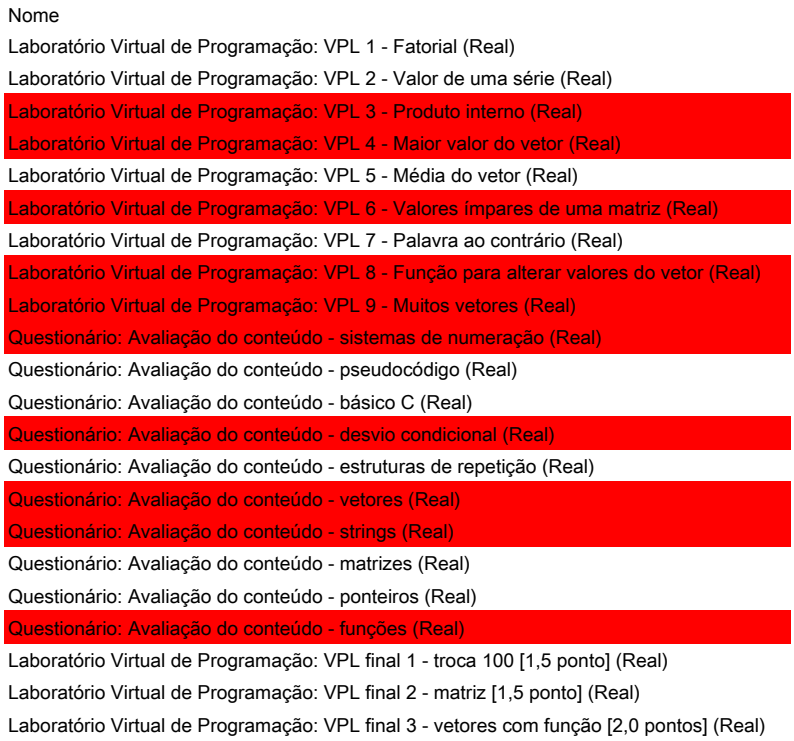
Percebe-se, a partir da análise das figuras, que a seleção de uma atividade com 25% do total da base de dados não garante a seleção da mesma atividade em

Figura 19 – Exemplo de seleção aleatória de 25% das atividades em um relatório de notas



Fonte: Elaborado pelo autor (2023).

Figura 20 – Exemplo de seleção aleatória de 50% das atividades em um relatório de notas



Fonte: Elaborado pelo autor (2023).

Figura 21 – Exemplo de seleção aleatória de 75% das atividades em um relatório de notas

Nome
Laboratório Virtual de Programação: VPL 1 - Fatorial (Real)
Laboratório Virtual de Programação: VPL 2 - Valor de uma série (Real)
Laboratório Virtual de Programação: VPL 3 - Produto interno (Real)
Laboratório Virtual de Programação: VPL 4 - Maior valor do vetor (Real)
Laboratório Virtual de Programação: VPL 5 - Média do vetor (Real)
Laboratório Virtual de Programação: VPL 6 - Valores ímpares de uma matriz (Real)
Laboratório Virtual de Programação: VPL 7 - Palavra ao contrário (Real)
Laboratório Virtual de Programação: VPL 8 - Função para alterar valores do vetor (Real)
Laboratório Virtual de Programação: VPL 9 - Muitos vetores (Real)
Questionário: Avaliação do conteúdo - sistemas de numeração (Real)
Questionário: Avaliação do conteúdo - pseudocódigo (Real)
Questionário: Avaliação do conteúdo - básico C (Real)
Questionário: Avaliação do conteúdo - desvio condicional (Real)
Questionário: Avaliação do conteúdo - estruturas de repetição (Real)
Questionário: Avaliação do conteúdo - vetores (Real)
Questionário: Avaliação do conteúdo - strings (Real)
Questionário: Avaliação do conteúdo - matrizes (Real)
Questionário: Avaliação do conteúdo - ponteiros (Real)
Questionário: Avaliação do conteúdo - funções (Real)
Laboratório Virtual de Programação: VPL final 1 - troca 100 [1,5 ponto] (Real)
Laboratório Virtual de Programação: VPL final 2 - matriz [1,5 ponto] (Real)
Laboratório Virtual de Programação: VPL final 3 - vetores com função [2,0 pontos] (Real)

Fonte: Elaborado pelo autor (2023).

50% ou 75%. Desta forma, colunas que selecionadas em uma divisão, podem não serem selecionadas para outras. Para o relatório de conclusão de atividades, o método de seleção é análogo ao do relatório de notas, uma vez que ambos carecem de uma lógica de organização de colunas pré-definida.

4 RESULTADOS

A partir da disponibilização de dados de duas turmas reais, onde os alunos foram classificados manualmente pelo docente em desistentes ou não desistentes, torna-se possível calcular as métricas de sensibilidade, precisão e acurácia a partir da comparação entre os resultados previstos com a classificação real. Como não houve contato com os dados reais, a análise do classificador automático não possui viés. Além disso, é possível comparar os resultados obtidos pelo classificador explícito (análise explícita) com as técnicas de aprendizado de máquina (análise implícita).

Para avaliar a aplicabilidade prática do programa na prevenção da desistência escolar, as métricas foram analisadas utilizando quatro diferentes proporções do período letivo em duas turmas distintas. O uso de 100% dos dados, embora não seja prático para intervenções imediatas, serve como referência para a melhor métrica possível, pois considera o desempenho completo do aluno durante o semestre. Para simular situações mais realistas, foram adotadas abordagens com 75% e 50% dos dados, que visam medir o desempenho da classificação em momentos críticos do curso. Além disso, a análise com apenas 25% dos dados foi realizada para testar a capacidade do sistema de identificar sinais precoces de desistência, proporcionando uma janela de oportunidade para intervenções eficazes.

Como apresentado na Seção 3.6, quando a base de dados é reduzida, as colunas a serem analisadas são selecionadas aleatoriamente. Uma vez que algumas atividades podem ser mais significativas do que outras para determinar a desistência, para reduzir essa influência, realizou-se a amostragem aleatória de colunas dez vezes para cada porcentagem de dados, na mesma lógica que foi apresentada para a justificativa de realização do método de subamostragem aleatória, na Seção 2.5.2.

4.1 ANÁLISE EXPLÍCITA

Nessa seção, serão apresentados os resultados atingidos pelos métodos descritos na Seção 3.4. Por se tratarem apenas de métodos de análise explícita, os resultados foram organizados na forma de matrizes de confusão e terão suas principais métricas calculadas. Os resultados demonstram a aplicação dos métodos em ambas as turmas.

4.1.1 Faltas consecutivas com média de presença da turma e dados interpolados

Para esse método, deseja-se saber qual escolha de quantidade de faltas consecutivas para se classificar um aluno como desistente apresenta as melhores

métricas. Ressalta-se que as análises para as Turma 1 e Turma 2 foram realizadas de forma independente, para depois terem suas matrizes de confusão somadas e as métricas extraídas. Pode-se visualizar a matriz de confusão dos resultados alcançados com cinco faltas em sequência a partir da Tabela 1.

Tabela 1 – Matriz de confusão para a classificação de desistência com 5 faltas consecutivas e dados interpolados

	Desistente real	Não desistente real
Desistente previsto	19	11
Não desistente previsto	23	74

Fonte: Elaborado pelo autor (2023).

A partir dos valores encontrados, calculou-se as métricas de sensibilidade e precisão, atingindo 45,24% e 63,33%, respectivamente. Em seguida, a Tabela 2 representa a matriz de confusão atingida a partir da classificação de desistentes com quatro faltas seguidas.

Tabela 2 – Matriz de confusão para a classificação de desistência com 4 faltas consecutivas e dados interpolados

	Desistente real	Não desistente real
Desistente previsto	23	24
Não desistente previsto	19	61

Fonte: Elaborado pelo autor (2023).

Para este caso, 54,76% e 48,94% foram os valores atingidos sensibilidade e precisão, respectivamente. Por fim, com a classificação a partir de três faltas consecutivas, a Tabela 3 demonstra os resultados atingidos.

Tabela 3 – Matriz de confusão para a classificação de desistência com 3 faltas consecutivas e dados interpolados

	Desistente real	Não desistente real
Desistente previsto	31	34
Não desistente previsto	11	51

Fonte: Elaborado pelo autor (2023).

Para o último caso, a sensibilidade e precisão encontrados foram de 73,81% e 47,69%, respectivamente. Desta forma, como é preferível para o presente trabalho corretamente classificar a maior quantidade de desistentes, ainda que incorretamente classificando alunos sem tal risco, a quantidade de três faltas sequenciais foi escolhida

para a classificação. Como descrito na Seção 3.4.2, uma tendência à classificação em desistente pode ser verificada na Tabela 3, visto que o método com a maior sensibilidade também é o método com menor precisão. Ademais, para verificar se a hipótese inicial de que a interpolação apresentaria melhorias nas análises, foi realizada uma comparação entre os resultados obtidos a partir do uso ou não da interpolação. Os detalhes de tal comparação podem ser verificados no Apêndice B.

4.1.2 Sistema de pontuação com notas, presença e conclusão de atividades

Nessa seção, serão descritos os resultados encontrados pela análise explícita, utilizando a técnica de se criar um sistema de pontuação a partir das notas, presença e conclusão de atividades dos alunos nas turmas analisadas. Adicionalmente, são realizadas análises complementares para verificarem as hipóteses propostas.

4.1.2.1 Definição dos pesos de cada pontuação

Antes de prosseguir com a análise dos dados, foi conduzida uma análise exploratória, focada em explorar diferentes combinações de pesos para as pontuações detalhadas na Seção 3.4.3. Para isso, utilizou-se uma resolução de 1%, variando os pesos no intervalo de 1% a 100% para cada pontuação, abordagem que permitiu testar todas as combinações possíveis de pesos. O principal objetivo dessa análise era verificar se a hipótese inicial, que consistia em atribuir um peso igual de um terço para cada pontuação, representava de fato a melhor estratégia.

Após a geração de uma planilha com 5.148 combinações para cada proporção de redução da base de dados e para cada turma, foi possível verificar que a hipótese inicial não se encontrava no conjunto de combinações com as melhores métricas de sensibilidade, precisão, acurácia e pontuação F_1. Desta forma, utilizou-se a combinação de 30% para a pontuação de notas, 30% para a pontuação de conclusão de atividades e 40% para a pontuação de presença. Essa combinação foi escolhida por possuir proporções próximas às da hipótese inicial, distribuindo quase igualmente a importância dos resultados.

4.1.2.2 Duas classificações e porcentagem de corte de 50%

Como descrito na Seção 3.6, foram realizadas análises com 25%, 50%, 75% e 100% de proporção em relação a base de dados original, com o intuito de simular uma aplicação real, na qual um professor utilizaria o programa durante o semestre para verificar quais alunos correm o risco de desistir. Ressalta-se também, que as análises foram feitas dez vezes para cada proporção, de forma que a influência da aleatoriedade da escolha de colunas seja reduzida. As métricas atingidas nos resultados da Turma 1 para cada uma das proporções mencionadas podem ser visualizadas na Tabela 4.

Tabela 4 – Métricas de classificação de desistência em duas classes e com porcentagem de corte de 50% para a Turma 1

Proporção do Semestre	Sensibilidade	Precisão	Acurácia	Pontuação F_1
100%	100%	50%	75,61%	66,67%
75%	93%	46,5%	72,2%	62%
50%	90%	45%	70,73%	60%
25%	93%	45%	72,68%	62,42%

Fonte: Elaborado pelo autor (2023).

Percebe-se, a partir da análise das métricas, que, com exceção de 50% dos dados da turma, o incremento da quantidade de dados resulta no aumento das métricas. Para a Turma 2, as métricas resultantes da análise podem ser verificadas na Tabela 5.

Tabela 5 – Métricas de classificação de desistência em duas classes e com porcentagem de corte de 50% para a Turma 2

Proporção do Semestre	Sensibilidade	Precisão	Acurácia	Pontuação F_1
100%	100%	74,42%	87,21%	85,33%
75%	99,69%	74,42%	86,98%	85,07%
50%	99,69%	74,42%	86,98%	85,07%
25%	96,56%	71,86%	84,65%	82,4%

Fonte: Elaborado pelo autor (2023).

A partir da análise dos resultados gerados para a Turma 2, percebe-se que o incremento da quantidade de dados resulta diretamente no aumento de todas as métricas. Os resultados apresentados são diretamente impactados pela escolha da porcentagem de corte, baseada na quantidade de desistentes da turma examinada. Considerando a desistência de cada disciplina no semestre anterior, aproximadamente 25% e 37% para Turma 1 e Turma 2, respectivamente, as análises foram refeitas utilizando tais porcentagens como a porcentagem de corte para desistentes no ranking.

A interpolação, usada na Seção 4.1.1, não se torna mais necessária, visto que as pontuações geradas relativizam a presença para a turma toda e não é realizada uma análise a partir das faltas consecutivas, principal necessidade de se interpolar os dias considerados inválidos. Adicionalmente, para justificar o descarte da interpolação, realizou-se uma comparação entre os resultados encontrados com e sem o uso da técnica para o método descrito na presente seção, os detalhes da comparação podem ser encontrados no Apêndice C.

4.1.2.3 Duas classificações e porcentagem de corte com dados históricos

A partir de informações disponibilizadas pelo professor que lecionou a matéria, definiu-se que a Turma 1 obteve uma porcentagem de desistentes de 25%. Desta

forma, a porcentagem de corte para a análise explícita foi definida nesse número. As métricas de sensibilidade, precisão, acurácia e pontuação F_1 obtidas pelos resultados da classificação de alunos da Turma 1 podem ser verificadas na Tabela 6.

Tabela 6 – Métricas de classificação de desistência em duas classes e com porcentagem de corte de 25% para a Turma 1

Proporção do Semestre	Sensibilidade	Precisão	Acurácia	Pontuação F_1
100%	60,00%	54,55%	78,05%	57,14%
75%	64,00%	58,18%	80,00%	60,95%
50%	63,00%	57,27%	79,51%	60,00%
25%	72,00%	63,16%	82,93%	67,29%

Fonte: Elaborado pelo autor (2023).

Diferentemente de quando utilizado a porcentagem de corte de 50%, a Turma 1 apresentou decrescimento de algumas de suas métricas a medida que mais dados eram disponibilizados. Para 25%, 50%, 75% e 100% dos dados, houve um decréscimo de 21%, 27%, 29%, e 40% para a sensibilidade e acréscimo na métrica de precisão de 18,16%, 12,27%, 11,68% e 4,55%, respectivamente. Para a Turma 2, a mesma observação de dados históricos foi realizada, definindo um valor de 37% de desistentes em turmas anteriores. Os resultados dessa análise podem ser verificados na Tabela 7.

Tabela 7 – Métricas de classificação de desistência em duas classes e com porcentagem de corte de 37% para a Turma 2

Proporção do Semestre	Sensibilidade	Precisão	Acurácia	Pontuação F_1
100%	96,88%	96,88%	97,67%	96,88%
75%	94,38%	94,38%	95,81%	94,38%
50%	89,38%	89,38%	92,09%	89,38%
25%	86,88%	85,80%	89,77%	86,34%

Fonte: Elaborado pelo autor (2023).

A partir da comparação da Tabela 5 e da Tabela 7, percebe-se variações na sensibilidade de -3,12%, -5,31%, -10,31% e -9,68% e na precisão de +22,46%, +19,96%, +14,96% e +13,94%, para 25%, 50%, 75% e 100% dos dados, respectivamente. Em contraste com a Turma 1, a Turma 2 apresentou decréscimos menos acentuados na métrica de sensibilidade e melhorias mais significativas na precisão. Portanto, como a métrica de sensibilidade é a de maior significância ao presente trabalho, e como se faz necessário o uso de dados passados, o método é considerado como não apropriado para os objetivos que se deseja atingir, apesar de apresentar melhora nos índices de precisão para ambas as turmas.

4.1.2.4 Três classificações

A partir dos resultados encontrados para a classificação entre desistente e não desistente, percebeu-se que as métricas com menor valor eram obtidas a partir da Turma 1 e do uso de 50% dos dados. Desta forma, para testar a hipótese de que as métricas são prejudicadas por alunos que apresentem comportamento atípico, como descrito na Seção 3.4.4 e analisado em Manhães *et al*. (2011), a classificação em desistente, intermediário e não desistente para a Turma 1 com os mesmos parâmetros de redução de base de dados foi realizada. Alunos intermediários são aqueles que apresentam comportamento atípico para a desistência, mas ainda a realizam. Um exemplo de comportamento anômalo pode ser um aluno que atinge boas notas nas atividades e participa das aulas, porém desiste da disciplina, tal classificação foi realizada manualmente pelo docente da disciplina.

A matriz de confusão resultante pode ser observada na Tabela 8, nela há três classes: desistentes, não desistentes e intermediários. Alunos intermediários, ou anômalos, serão classificados como tal caso suas pontuações estejam entre 33% e 66% do ranking ordenado, enquanto que desistentes serão classificados caso estejam abaixo do limiar de 33%.

Tabela 8 – Matriz de confusão para a classificação de desistência da Turma 1 com três classes e 50% do total de dados

	Desistente real	Intermediário real	Não desistente real
Desistente previsto	75	5	60
Intermediário previsto	5	14	111
Não desistente previsto	0	1	139

Fonte: Elaborado pelo autor (2023).

Para Manhães *et al*. (2011), é considerado um erro grave do classificador atribuir a classificação de não desistente ao aluno que de fato desiste, sendo tal classificação clara (desistentes) ou com comportamento anômalo (intermediários), pois essa desistência não seria detectada. Porém, classificar alunos não desistentes como desistentes ou intermediários, apesar de constituir um erro, é considerado menos grave, visto que apenas causaria em um atendimento desnecessário por parte do docente, porém aumenta a probabilidade de se detectar a maioria de desistentes. A partir da análise da Tabela 8, percebe-se que a maioria dos erros encontra-se no fato de o classificador considerar alunos não desistentes como desistentes ou intermediários, gerando erros menos graves.

4.2 ANÁLISE IMPLÍCITA

Com o objetivo de avaliar e comparar os resultados obtidos pela análise explícita e outros trabalhos similares, uma análise a partir dos 22 atributos apresentados na Seção 3.3.4 foi conduzida, apesar de que seis desses representam atividades marcadas como importantes, que não foram utilizadas para os conjuntos de dados analisados. As turmas foram analisadas separadamente e foram utilizados três diferentes técnicas de divisão de dados para avaliação dos resultados obtidos com os classificadores: método holdout com 33% dos dados para teste e 67% para treino (HO), mantendo a mesma proporção de desistentes em ambos os conjuntos; subamostragem aleatória com 67% dos dados para treinamento (SA); e validação cruzada com 10 divisões (VC). A seleção de características dos dados foi realizada com os cinco e dez melhores atributos selecionados por quatro técnicas de avaliação: ANOVA, Chi2, GI e RG. Com relação às métricas de avaliação utilizadas, essas foram: acurácia, sensibilidade, precisão e pontuação F_1.

4.2.1 Seleção dos atributos

Com o uso dos métodos de avaliação de atributos, pôde-se verificar quais são as características mais importantes para a análise realizada. A Figura 22 representa os 10 atributos com pontuação mais alta para 100% dos dados em cada um dos métodos de seleção de características para a Turma 1. Adicionalmente, atributos que se repetem entre métodos foram codificados em cores. Percebe-se que *grades_between_0_2_5* é classificado entre os três mais relevantes atributos para todos os métodos e nota-se a relevância de *activities_complete_majority*, *activities_incomplete_majority* e *grades_average*, todos posicionados entre os 5 melhores em todos os casos.

A categoria de atributos que mais aparece entre os 10 atributos mais importantes é a de notas, com 47,5% do total de atributos, seguido pela categoria de atividades concluídas, com 35% dos atributos selecionados, e, por fim, presença representa 17,5% das características escolhidas. Para a Turma 2, diferentemente da Turma 1, apenas *grades_between_0_2_5* é consistentemente considerado como um dos quatro atributos mais relevantes, seguido por *grades_between_7_5_10*, como pode-se perceber na Figura 23, que mostra os 10 atributos com maior pontuação em cada método de seleção de atributos para a Turma 2, novamente codificados em cores, quando se repetem entre colunas.

Nota-se a distribuição de 50% para as categorias de notas e atividades concluídas igualmente para todos os métodos, o que também significa que a categoria de presença não foi selecionada entre as 10 mais relevantes nenhuma vez para os quatro métodos de seleção. Essas diferenças mostram as particularidades de cada turma e aluno matriculado.

Figura 22 – Seleção de atributos para a Turma 1

Rank	Ganho de Informação	Relação de Ganho	Chi2	ANOVA
1	activities_complete_majority	activities_complete_majority	grades_between_0_2_5	activities_incomplete
2	activities_incomplete_majority	activities_incomplete_majority	activities_incomplete_majority	grades_between_0_2_5
3	grades_between_0_2_5	grades_between_0_2_5	activities_complete_majority	activities_complete_majority
4	grades_average	activities_incomplete	grades_below_5	activities_incomplete_majority
5	activities_incomplete	grades_average	grades_average	grades_average
6	grades_below_5	activities_complete_minority	grades_below_mean	grades_below_5
7	grades_between_5_7_5	activities_incomplete_minority	partial_presence	grades_below_mean
8	partial_presence	grades_below_5	grades_between_5_7_5	grades_between_7_5_10
9	grades_below_mean	partial_presence	activities_incomplete	attendance_below_mean
10	attendance_below_mean	grades_between_5_7_5	attendance_below_mean	missing

Fonte: Elaborado pelo autor (2023).

Figura 23 – Seleção de atributos para a Turma 2

Rank	Ganho de Informação	Relação de Ganho	Chi2	ANOVA
1	grades_between_7_5_10	activities_complete_minority	grades_between_0_2_5	activities_complete_majority
2	grades_between_0_2_5	activities_incomplete_minority	grades_below_mean	activities_incomplete_majority
3	grades_below_5	grades_between_7_5_10	grades_between_7_5_10	activities_incomplete
4	grades_below_mean	grades_between_0_2_5	activities_complete_majority	grades_between_0_2_5
5	activities_complete_majority	grades_below_5	grades_average	grades_average
6	activities_incomplete	activities_incomplete	activities_incomplete_majority	grades_below_5
7	activities_incomplete_majority	grades_below_mean	activities_complete_minority	grades_below_mean
8	grades_average	activities_complete_majority	grades_below_5	grades_between_7_5_10
9	activities_complete_minority	activities_incomplete_majority	activities_incomplete	activities_complete_minority
10	activities_incomplete_minority	grades_average	activities_incomplete_minority	activities_incomplete_minority

Fonte: Elaborado pelo autor (2023).

4.2.2 Combinação de algoritmos de aprendizado de máquina, técnicas de divisão do conjunto de dados e seleção de atributos

No decorrer deste estudo, foram empregados oito algoritmos de aprendizado de máquina que são compostos por AD, KNN, RF, NB, MLP, RL, GB e SVM. Estes algoritmos foram combinados com quatro diferentes métodos de avaliação de atributos (GI, RG, Chi2 e ANOVA) e aplicados tanto com a seleção de cinco e dez atributos quanto sem nenhuma seleção de características. Além disso, três técnicas de divisão do conjunto de dados foram utilizadas, que, ao fim da abordagem, gerou um total

de 216 resultados distintos com quatro métricas (sensibilidade, precisão, acurácia e pontuação F_1) calculadas para cada. A Figura 24 ilustra uma parte da planilha gerada a partir das combinações resultantes, para ambas as turmas e para cada proporção do conjunto total de dados (100%, 75%, 50% e 25%).

Figura 24 – Parte da planilha criada a partir das métricas geradas

dataset	ratio_evaluation	evaluation_method	feature_selection	learner	recall	precision	accuracy	f1_score	top 1
turma_1	100,00%	random (n=10, test=33%)	anova (10)	naive_bayes	84,68%	45,92%	71,57%	59,55%	activities_incomplete
turma_1	100,00%	random (n=10, test=33%)	chi2 (10)	naive_bayes	84,48%	50,99%	76,86%	63,60%	grades_between_0_2_5
turma_1	100,00%	random (n=10, test=33%)	info_gain (10)	naive_bayes	83,84%	52,40%	75,93%	64,49%	activities_complete_majority
turma_1	100,00%	random (n=10, test=33%)	gain_ratio (10)	naive_bayes	82,65%	51,09%	76,57%	63,15%	activities_complete_majority
turma_1	100,00%	random (n=10, test=33%)	gain_ratio (5)	naive_bayes	81,79%	55,51%	78,64%	66,14%	activities_complete_majority
turma_1	100,00%	random (n=10, test=33%)	info_gain (5)	naive_bayes	80,73%	58,03%	80,14%	67,52%	activities_complete_majority
turma_1	100,00%	random (n=10, test=33%)	anova (5)	naive_bayes	80,17%	59,03%	80,43%	67,99%	activities_incomplete
turma_1	100,00%	xvalidation (k=10)	chi2 (5)	tree	80,00%	88,89%	92,68%	84,21%	grades_between_0_2_5
turma_1	100,00%	xvalidation (k=10)	info_gain (5)	tree	80,00%	88,89%	92,68%	84,21%	activities_complete_majority
turma_1	100,00%	xvalidation (k=10)	gain_ratio (5)	tree	80,00%	88,89%	92,68%	84,21%	activities_complete_majority
turma_1	100,00%	xvalidation (k=10)	info_gain (10)	gradient_boosting	80,00%	66,67%	85,37%	72,73%	activities_complete_majority
turma_1	100,00%	xvalidation (k=10)	anova (5)	tree	80,00%	72,73%	87,80%	76,19%	activities_incomplete
turma_1	100,00%	xvalidation (k=10)	anova (5)	logistic_regression	80,00%	72,73%	87,80%	76,19%	activities_incomplete

Fonte: Elaborado pelo autor (2023).

A partir de uma análise preliminar dos resultados obtidos a partir de HO, percebeu-se que os subconjuntos de treinamento e teste gerados por essa técnica de divisão de dados não representava precisamente o conjunto total de dados. Desta forma, tal método foi desconsiderado das análises e as comparações dessa seção utilizarão apenas SA e VC.

4.2.3 Métricas e discussões

A seguir, são apresentadas as principais combinações obtidas para cada divisão do banco de dados, das quais apenas as métricas obtidas com divisão de conjunto por método holdout foram suprimidas. Os resultados serão apresentados primeiramente para a Turma 1 e, em seguida, para a Turma 2. Ao fim da seção, uma síntese dos principais resultados será elaborada para facilitar o entendimento das observações.

4.2.3.1 Turma 1

As análises com 100%, 75%, 50% e 25% do conjunto de dados para a Turma 1 serão descritas separadamente nessa seção. Devido à quantidade de possíveis combinações, as métricas para cada proporção de dados foram primeiramente organizadas em função da sensibilidade, métrica de maior interesse do presente trabalho, e em seguida, a partir da precisão. Caso haja um empate entre duas combinações diferentes para a maior sensibilidade, a precisão será usada como critério de desempate. Caso o empate se mantenha, ambas as combinações serão

apresentadas. Haverá também uma solução alternativa, na qual a combinação de maior acurácia será apresentada.

Para o conjunto completo de dados (100%), NB apresentou a maior sensibilidade, com **90,00%**. Suas métricas de precisão, acurácia e pontuação F_1 foram de 52,94%, 78,05% e 66,67%, respectivamente. Para atingir tais métricas, o método de avaliação utilizado foi VC sem qualquer seleção de atributos. Como uma opção alternativa, focada na acurácia, GB, a partir de VC e chi-quadrado com os cinco melhores atributos (Chi2-5) obteve métricas de 76,00%, 97,44%, **93,66%** e 85,39% para sensibilidade, precisão, acurácia e pontuação F_1, respectivamente.

Com 75% do total de dados, a combinação escolhida com ênfase na sensibilidade foi atingida a partir de NB e SA, com chi-quadrado com os dez melhores atributos (Chi2-10) para a seleção de atributos. Suas métricas são **87,43%**, 49,16%, 75,43% e 62,93%, para sensibilidade, precisão, acurácia e pontuação F_1, respectivamente. Como solução alternativa, priorizando a acurácia, GB com método de seleção de atributos por ganho de informação com os cinco melhores atributos (GI-5) e avaliação por VC apresentou sensibilidade de 70,00%, precisão de 74,47%, acurácia de **86,83%** e pontuação F_1 de 72,16%.

Para uma divisão de 50% da base de dados, NB combinado a análise de variância com os dez melhores atributos (ANOVA-10) e avaliação a partir de VC obteve sensibilidade, precisão, acurácia e pontuação F_1 de **86,00%**, 51,50%, 76,83% e 64,42%, respectivamente, que foi definida como a melhor escolha para uma combinação focada em sensibilidade. Alternativamente, com a maior acurácia, LR com características selecionadas por GI-5 ou relação de ganho com os cinco melhores atributos (RG-5) e avaliado com VC obteve métricas de 64,00%, 69,57%, **84,39%** e 66,67%, na mesma ordem da solução principal.

Por fim, com 25% do total dos dados disponíveis e ao utilizar Chi2-10, NB foi escolhida por sua métrica de sensibilidade, avaliado pelo método VC. Suas métricas apresentaram valores de **88,00%**, 56,41%, 80,49% e 68,75%, para sensibilidade, precisão, acurácia e pontuação F_1, respectivamente. Ao invés de escolher a melhor combinação a partir da sensibilidade e focar na acurácia como métrica de importância, o uso de RF com seleção de atributos do método GI-5, avaliado por VC, apresenta sensibilidade de 66,00%, precisão de 75,86%, acurácia de **86,59%** e pontuação F_1 de 70,59%. Com todas as proporções de dados analisadas para a Turma 1, partiu-se para a escolha de combinações para a Turma 2.

4.2.3.2 Turma 2

Da mesma forma que realizado para a Turma 1, as métricas foram analisadas para cada proporção do conjunto de dados total para a Turma 2. Priorizou-se a métrica de sensibilidade, mas os métodos com acurácia mais alta também são apresentados

como uma solução alternativa, na qual se deseja obter resultados mais corretos, do que garantir a detecção da desistência.

Inicialmente todos os dados (100% do conjunto) foram utilizados para gerar as combinações e percebeu-se que o uso de MLP com a escolha de atributos por GI-5 e avaliação por VC apresentava as maiores métricas tanto para sensibilidade quanto para a acurácia, com **100,00%** e **98,14%**, respectivamente. Adicionalmente, obteve precisão de 95,24% e pontuação F_1 de 97,56%.

Em seguida, houve a redução para 75% do total, onde a composição de NB com Chi2-10 e SA apresentaram valores de sensibilidade, precisão, acurácia e pontuação F_1 de **99,54%**, 89,23%, 95,34% e 94,10%, respectivamente. Priorizando a métrica de acurácia, SVM com análise de variância com os cinco melhores atributos (ANOVA-5) e avaliação por VC atingem sensibilidade de 98,44%, precisão de 96,04%, acurácia de **97,91%** e pontuação F_1 de 97,22%.

Ao usar metade da base de dados (50%), NB se destaca ao ser combinado com ANOVA-5 e avaliado por VC, gerando resultados com métricas de sensibilidade, precisão, acurácia e pontuação F_1 de **100,00%**, 86,96%, 94,42% e 93,02%, respectivamente. Por outro lado, a escolha de SVM com ANOVA-5 e VC, apresenta a maior acurácia, com **97,21%**, além de sensibilidade de 97,19%, precisão de 95,40% e pontuação F_1 de 96,28%.

Finalmente, com apenas 25% dos dados disponíveis, NB com seleção de características por RG-5 e avaliação por VC apresentou sensibilidade, precisão, acurácia e pontuação F_1 de **97,19%**, 86,87%, 93,49% e 91,74%, respectivamente, sendo a combinação escolhida para a solução principal. Como alternativa, focada em acurácia, MLP com RG-5 e avaliação por SA apresentou sensibilidade de 95,25%, precisão de 92,01%, acurácia de **95,28%** e pontuação F_1 de 93,60%. Assim, as proporções de dados para ambas as turmas foram analisadas e suas combinações principais, com foco na métrica de sensibilidade, e alternativas, com ênfase em acurácia, foram observadas. Para facilitar a visualização dos dados, será realizada uma síntese das observações de ambas as turmas e, a partir da escolha de uma combinação principal dentre todas as disponíveis, novas métricas serão geradas.

4.2.3.3 Síntese e aplicação dos resultados

A Tabela 9 apresenta a síntese dos resultados discutidos na Seção 4.2.3.1 e Seção 4.2.3.2. Pode-se verificar a principal combinação de método de avaliação de resultados, técnica de escolha de atributos e algoritmo de aprendizado de máquina para cada proporção de dados e a métrica de sensibilidade atingida.

Infere-se, a partir da análise da Tabela 9, que em relação ao total de combinações:

- 87,5% usam o algoritmo de aprendizado de máquina Naive Bayes;

Tabela 9 – Síntese de principais resultados obtidos para a Turma 1 e Turma 2

Turma	Proporção	Avaliação	Seleção de Atributos	Algoritmo	Sensibilidade
1	100%	VC	-	NB	90,00%
1	75%	SA	Chi2-10	NB	87,43%
1	50%	VC	ANOVA-10	NB	86,00%
1	25%	VC	Chi2-10	NB	88,00%
2	100%	VC	GI-5	MLP	100,00%
2	75%	SA	Chi2-10	NB	99,54%
2	50%	VC	ANOVA-5	NB	100,00%
2	25%	VC	RG-5	NB	97,19%

Fonte: Elaborado pelo autor (2023).

- 75% usam o método de avaliação de validação cruzada 10-fold;
- 50% usam os 10 atributos mais bem pontuados pela técnica de seleção de atributos;
- 37,5% usam o cálculo chi-quadrado para pontuar atributos.

A fim de atender às características gerais das turmas analisadas, foi criado um modelo com os aspectos mais comuns de todas as combinações, a partir do qual realizou-se uma nova análise com as quatro proporções de dados usadas ao longo do trabalho. Desta forma, pode-se visualizar na Tabela 10, a síntese dos resultados alcançados com o método de validação cruzada 10-fold, a seleção de atributos Chi2-10 e o algoritmo NB para ambas as turmas.

Tabela 10 – Síntese dos resultados alcançados a partir de um modelo geral para a Turma 1 e Turma 2

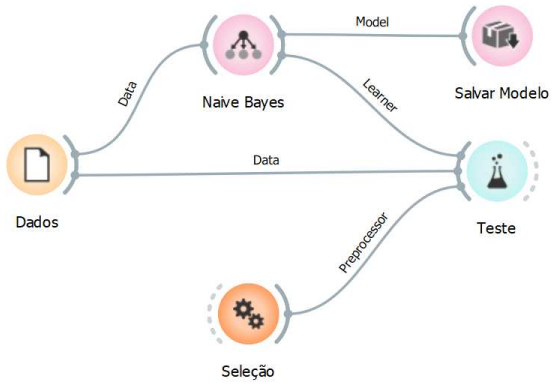
Turma	Proporção de dados	Sensibilidade	Precisão	Acurácia	Pontuação F_1
1	100%	80,00%	53,33%	78,05%	64,00%
1	75%	82,00%	51,90%	77,07%	63,57%
1	50%	82,00%	48,81%	74,63%	61,19%
1	25%	87,00%	55,06%	79,51%	67,44%
2	100%	100,00%	88,89%	95,35%	94,12%
2	75%	100,00%	86,49%	94,19%	92,75%
2	50%	99,38%	85,95%	93,72%	92,17%
2	25%	95,31%	86,40%	92,67%	90,64%

Fonte: Elaborado pelo autor (2023).

Para possibilitar uma visualização de tal modelo, realizou-se no Orange um fluxo, presente na Figura 25, que ilustra o processo desde a obtenção dos dados pré-processados (módulo denominado “Dados”), a conexão da técnica de aprendizado de máquina Naive Bayes, o uso de um módulo de teste, configurado para validação cruzada 10-fold, a utilização da seleção de atributos mais relevantes (correspondente

ao módulo “Seleção”) e, por fim, a conexão resultante ligada ao módulo que permite salvar o modelo gerado (módulo “Salvar Modelo”) [1].

Figura 25 – Fluxo do Orange com o modelo proposto



Fonte: Elaborado pelo autor (2023).

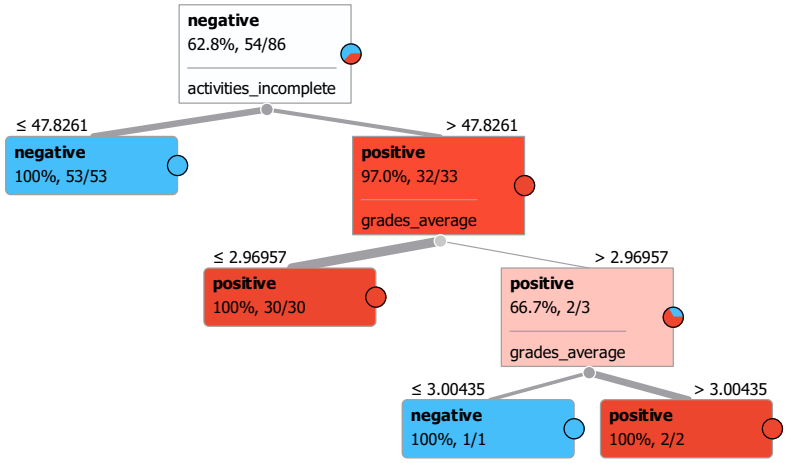
Com tal fluxo, é possível alterar os dados que se deseja fornecer ao modelo para treina-lo novamente, ou exportar o modelo em formato *Pickled model* (.pkcls), que pode ser importado por outros fluxos do Orange ou códigos na linguagem de programação Python. Por fim, para demonstrar um exemplo de lógica de classificação, o algoritmo Árvore de Decisão foi utilizado para gerar uma árvore com o uso do Orange para 100% dos dados da Turma 2, a qual pode ser visualizada na Figura 26. Na árvore gerada, é possível verificar os caminhos (galhos) tomados pelo algoritmo para classificar os alunos em desistentes ou não desistentes (folhas).

O algoritmo utiliza atributos *activities_incomplete* com valores superiores a 47,83% para classificar alunos como desistentes, identificados como positive na árvore. Em seguida, visto que um dentre 33 alunos não pode ser classificado com apenas esse atributo, *grades_average* é inicialmente usado com valores menores ou iguais a 2,97 para classificar 30 alunos em desistentes, sendo necessária mais uma classificação usando valores superiores a três para finalizar a especificação.

Percebe-se que o algoritmo não realiza a última classificação corretamente, visto que utiliza a média de notas com valores menores do que três para designar a classe não desistente aos alunos, o que, logicamente, deveria ser o contrário. Porém, os atributos utilizados seguem uma lógica similar ao que foi encontrado a partir da seleção de atributos, ilustrada na Figura 23, que seleciona apenas atributos das classes de notas e conclusão de atividades.

[1] Ressalta-se que as ligações entre os módulos de “Seleção”, “Dados” e “Teste” seguem as instruções da documentação oficial do Orange (https://orangedatamining.com/widget-catalog/transform/preprocess/).

Figura 26 – Árvore de Decisão ilustrativa do Orange, gerada com 100% dos dados da Turma 2



Fonte: Elaborado pelo autor (2023).

4.3 ANÁLISES E COMPARAÇÕES

Nesta seção, é proposta uma discussão e análise dos resultados alcançados pelas análises e demonstrados na Seção 4.1 e Seção 4.2, para a análise implícita e explícita, respectivamente. Além disso, as métricas atingidas serão comparadas a um trabalho similar ao atual.

Em relação às métricas alcançadas por meio da análise explícita, tanto na Turma 1 quanto na Turma 2, desistentes seriam corretamente classificados como tal em nove de cada dez casos. Entretanto, na primeira turma, a métrica de precisão indica que o classificador erra mais do que acerta, com resultados entre 45% e 46,5%, fazendo com que, em caso de uso em situação real, o docente precise despender mais tempo filtrando ou atendendo alunos que não desistiriam de fato. Por outro lado, na Turma 2, apenas três a cada dez alunos seriam incorretamente classificados como desistentes. A diferença entre as métricas de precisão das turmas pode ser atrelada ao fato de que a Turma 2 possui 45 alunos a mais que a Turma 1, ou seja, mais dados que são levados em consideração nas análises, o que gera resultados mais próximos da realidade do conjunto.

Na análise implícita da Turma 1, similarmente ao que ocorre para sua análise explícita entre 25% e 50% do total de dados, ao aumentar a base de dados, a métrica de sensibilidade apresenta decréscimo, de forma que as métricas com maiores valores são encontradas a partir do uso de apenas 25% dos dados. Fato que não é analisado nos resultados da Turma 2, que apresentam acréscimo a cada aumento da proporção de dados. Comparada à explícita, as métricas da implícita apresentam variações médias

de -11,25% para a sensibilidade e +5,65% para a precisão da Turma 1. Em contraste, para a Turma 2 a sensibilidade varia em média -0,3125% e a precisão +13,15%. Assim, percebe-se um aumento da precisão e diminuição da sensibilidade, na qual esta é mais acentuada para a primeira turma.

Como analisado na Seção 2.6, o estudo realizado por Burgos *et al*. (2018) apresenta objetivo similar ao presente trabalho, visto que pretende prever a evasão ao longo de um semestre de 20 semanas. Os autores atingem, com seu melhor modelo, 96,73% de sensibilidade, 98,35% de precisão e 97,13% de acurácia com 50% dos dados totais. Na mesma proporção de dados, os resultados alcançados no presente estudo com um modelo geral de análise implícita atingem sensibilidade 2,65% superior, porém precisão e acurácia inferiores por 12,4% e 3,41%, respectivamente. O melhor modelo treinado com 50% dos dados do trabalho atual atinge métricas de sensibilidade 0,46% superior, precisão 3,55% inferior e acurácia 0,08% superior. Desta forma, o presente trabalho é capaz de detectar uma maior proporção de desistentes, mas incorretamente classifica alunos não desistentes como desistentes em maior quantidade.

5 CONCLUSÕES

Não há dúvidas quanto ao impacto da desistência escolar na vida pessoal e acadêmica de alunos. Além das possíveis consequências monetárias, esse tipo de fracasso escolar provoca repercussões sociais, psicológicas e interpessoais. Portanto, o presente trabalho foi desenvolvido com o foco em auxiliar professores que lecionam matérias presenciais ou remotas a identificar alunos que demonstram características de abandono das aulas em determinado semestre, para que um atendimento especializado possa ser empregado.

Visando identificar o maior número possível de alunos com comportamentos que indicam uma possível desistência da matéria, a métrica de sensibilidade foi escolhida como o principal critério de avaliação do programa. Esta métrica é crucial pois indica a proporção de alunos desistentes que são erroneamente classificados pelo programa. Em outras palavras, ela reflete a capacidade do programa de identificar corretamente aqueles alunos que, apesar de apresentarem sinais de desistência, poderiam não ser detectados como tais.

Este trabalho conduziu tanto análises explícitas quanto implícitas. Os resultados demonstram que as métricas alcançadas na análise implícita apresentam melhorias significativas em comparação com a análise explícita, especialmente no que se refere à precisão. No entanto, a análise explícita é destacada como uma opção viável, oferecendo métricas satisfatórias e sendo menos complexa do que os métodos tradicionalmente utilizados em estudos semelhantes. Adicionalmente, com base no ranking de pontuações gerado, os professores podem focar seus esforços nos alunos que estão nos limites das classificações. Isso permite identificar aqueles com maior probabilidade de desistência e, consequentemente, maior potencial para intervenções efetivas.

Com relação à configuração de porcentagem de corte da classificação de desistentes da análise explícita, essa pode ser definida de forma arbitrária pelo professor, com base em sua capacidade de atendimento de alunos, e a partir de dados passados de desistência. Porém, uma vez que um objetivo do trabalho era apenas utilizar dados isolados da turma analisada, esse objetivo não pode ser considerado alcançado caso tal método seja utilizado, por depender de informação passada da turma para melhor desempenho do resultado.

A partir da comparação com trabalhos que tenham objetivo similar ao deste, conclui-se que as métricas atingidas são equiparáveis, e considera-se que o objetivo principal do trabalho, que é detectar alunos com tendências à desistência da disciplina a tempo de permitir uma intervenção efetiva do professor, foi atingido. Entretanto, é

importante reconhecer como limitação deste trabalho o fato de que apenas duas turmas foram utilizadas para o treinamento e análise de resultados.

5.1 PUBLICAÇÕES REALIZADAS

No decorrer do presente trabalho, e a partir de análises preliminares de seus resultados, pôde-se elaborar três artigos para publicação, sendo que dois desses foram aprovados e um aguarda o parecer dos revisores. O primeiro artigo, referenciado em Souza, Moreira e Wang (2023), aborda uma comparação com trabalhos semelhantes, detalhados na Seção 2.6, e discute as métricas utilizadas e os resultados preliminares da análise explícita, conforme descrito na Seção 3.4 (no prelo). O segundo artigo, Souza e Moreira (2024a), volta-se novamente para a análise explícita, apresentando aprimoramentos em relação à primeira publicação e alcançando métricas mais significativas para ambas as turmas analisadas, além de levantar a questão da presença de alunos com comportamento atípico nas bases de dados (no prelo). Em contraste com as outras duas publicações, o terceiro artigo, Souza e Moreira (2024b), concentra-se nos resultados, métodos e métricas da análise implícita (submetido à publicação).

REFERÊNCIAS

ADACHI, A. A. C. T. **Evasão e evadidos nos cursos de graduação da Universidade Federal de Minas Gerais**. 2009. Dissertação (Mestrado em Educação) — Faculdade de Educação, Universidade Federal de Minas Gerais, 2009.

AMARAL, F. **Introdução à ciência de dados**: mineração de dados e big data. Rio de Janeiro: Alta Books Editora, 2016.

BRASIL. Ministério da Educação. **Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas**. Brasília, 1996. Disponível em: http://www.dominiopublico.gov.br/download/texto/me001613.pdf. Acesso em: 20 nov. 2022.

BURGOS, C. *et al*. Data mining for modeling students' performance: a tutoring action plan to prevent academic dropout. **Computers & Electrical Engineering**, v. 66, p. 541–556, fev. 2018.

FIALHO, M. G. D.; PRESTES, E. M. T. Evasão escolar no curso de pedagogia da UFPB: na compreensão dos gestores educacionais. **Revista MPGOA**, v. 3, n. 1, p. 42–63, jul. 2014.

FILATRO, A. C. **Data science na educação**: presencial, a distância e corporativa. São Paulo: Saraiva Educação S.A., 2020.

GAIOSO, N. P. L. **O fenômeno da evasão escolar na educação superior no Brasil**. 2005. Dissertação (Mestrado em Educação) — Faculdade de Educação, Universidade Católica de Brasília, 2005.

SILVA GARCIA, L. M. L. da *et al*. Mineração de dados educacionais na predição do desempenho acadêmico: um prognóstico a partir do percurso curricular realizado. *In*: XXXIII SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. **Anais [...]**. Manaus, 2022. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/22483. Acesso em: 31 ago. 2023.

GOTTARDO, E.; KAESTNER, C. A. A.; NORONHA, R. V. Estimativa de desempenho acadêmico de estudantes: análise da aplicação de técnicas de mineração de dados em cursos a distância. **Revista Brasileira de Informática na Educação**, v. 22, n. 1, p. 45–56, 2014.

HALL, M. A. **Correlation-based feature selection for machine learning**. 1999. Tese (Doutorado em Ciência da Computação) — Department of Computer Science, The University of Waikato, 1999.

HAN, J.; PEI, J.; TONG, H. **Data Mining: Concepts and Techniques**. 3. ed. Waltham: Morgan Kaufmann, 2012.

KAENSAR, C.; WONGNIN, W. Analysis and prediction of student performance based on moodle log data using machine learning techniques. **International Journal of Emerging Technologies in Learning (iJET)**, v. 18, n. 10, p. p. 184–203, maio 2023.

LIU, H.; SETIONO, R. Chi2: feature selection and discretization of numeric attributes. *In*: **Proceedings** of the 7TH IEEE INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE. Herndon, p. 388–391, 1995. Disponível em: https://ieeexplore.ieee.org/document/479783. Acesso em: 10 nov. 2023.

MAHBOBI, M.; TIEMANN, T. **Introductory business statistics with interactive spreadsheets**. BCcampus, 2010. (Open textbook library). Disponível em: https://books.google.com.br/books?id=Q0Z_zQEACAAJ.

MALLADA, F. J. R. La gestión del absentismo escolar. **Anuario Jurídico y Económico Escurialense**, n. 44, p. 579–596, 2011.

MANHÃES, L. M. B. *et al*. Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. *In*: XXII SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. **Anais [...]**. Aracaju, 2011. Disponível em: http://ojs.sector3.com.br/index.php/sbie/article/download/1585/1350. Acesso em: 20 nov. 2022.

MARTINS, G. A. Estudo de caso: uma reflexão sobre a aplicabilidade em pesquisas no brasil. **Revista de Contabilidade e Organizações**, v. 2, n. 2, p. 8–18, jan/abr 2008.

MORAIS, A. M. **Abordagem avaliativa multidimensional para previsão da evasão do discente em cursos online**. 2018. Tese (Doutorado em Ciência da Computação) — Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, 2018.

NAGAI, N. P.; CARDOSO, A. L. J. A evasão universitária: uma análise além dos números. **Revista Estudo & Debate**, v. 24, n. 1, p. 193–215, 2017.

PINTO, S. C. **Os custos da evasão de discentes das universidades brasileiras na modalidade de ensino presencial: uma perspectiva de custos contábeis e custos econômicos**. 2021. Dissertação (Mestrado em Ciências Contábeis) — Unidade Acadêmica de Pesquisa e Pós-Graduação, Universidade do Vale do Rio dos Sinos, 2021.

QUEIROGA, E. M. *et al*. Modelo de predição da evasão de estudantes em cursos técnicos a distância a partir da contagem de interações. **Revista Thema**, v. 15, n. 2, p. 425–438, 2018.

RASCHKA, S. Model evaluation, model selection, and algorithm selection in machine learning. **CoRR**, abs/1811.12808, 2018. Disponível em: http://arxiv.org/abs/1811.12808V3.

ROMERO, C.; VENTURA, S. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 3, n. 1, p. 12–27, dez. 2012.

SOUZA, B. D. de; MOREIRA, B. G. Detecção de desistência de estudantes em disciplinas ofertadas com apoio do ambiente moodle: uma abordagem de análise explícita. **Revista Ibérica de Sistemas e Tecnologias de Informação**, Rio Tinto, 2024. No prelo.

SOUZA, B. D. de; MOREIRA, B. G. Detecção de desistência de estudantes em disciplinas ofertadas com apoio do ambiente moodle: uma abordagem de análise implícita. *In*: COMPUTER ON THE BEACH. **Anais [...]**. Balneário Camboriú, 2024. Submetido à publicação.

SOUZA, B. D. de; MOREIRA, B. G.; WANG, C. B. Detecção de desistência de estudantes em disciplinas ofertadas com apoio do ambiente moodle: uma discussão sobre resultados alcançados. *In*: II WORKSHOP DE APLICACÕES PRÁTICAS DE LEARNING ANALYTICS EM INSTITUICÕES DE ENSINO NO BRASIL. **Anais [...]**. SBC, Passo Fundo, 2023. No prelo.

VIANA, F. S.; SANTANA, A. M.; ANDRADE LIRA RABÊLO, R. de. Avaliação de classificadores para predição de evasão no ensino superior utilizando janela semestral. *In*: XXXIII SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. **Anais [...]**. Manaus, 2022. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/22469. Acesso em: 31 ago. 2023.

ZARPELON, E. **Análise do desempenho de alunos calouros de engenharia na disciplina de Cálculo Diferencial e Integral I**: um estudo de caso na UTFPR. 2016. Dissertação (Mestrado em Ensino de Ciência e Tecnologia) — Universidade Tecnológica Federal do Paraná, 2016.

APÊNDICE A - ANÁLISE COMPLETA DE TRABALHOS SIMILARES

Em Queiroga *et al*. (2018), o objetivo é abordar a evasão de disciplinas em contraste com a evasão de cursos de graduação, que é o foco do presente estudo. No trabalho em questão, a principal fonte de informação para criar e comparar modelos analíticos é a contagem de interações dos estudantes no ambiente virtual MOODLE. Os autores utilizaram a biblioteca Waikato Environment for Knowledge Analysis (WEKA) e selecionaram cinco algoritmos de aprendizado de máquina: BN, SL, MLP, RF e J48 (Queiroga *et al*., 2018).

Os autores propuseram dois cenários distintos para gerar os modelos de análise implícita: um cenário usou apenas um curso para treinar e avaliar os modelos, enquanto o outro empregou três cursos diferentes para treinamento e reservou um quarto curso para a avaliação dos dados. As métricas de acurácia utilizadas para avaliar os modelos incluíram a porcentagem de estudantes previstos como evadidos, denominada VP, e a porcentagem de alunos previstos para concluir o curso regularmente, denominada VN (Queiroga *et al*., 2018).

No primeiro cenário, observou-se que apenas o algoritmo SL não atingiu 75% de VP desde a primeira semana, mas gradualmente aumentou para 87% até a semana 25 (final do primeiro semestre), aproximando-se de 94% antes do final do segundo semestre da graduação. No que diz respeito aos valores de VN no primeiro cenário, BN foi o único modelo a superar 75% de acurácia desde a primeira semana, atingindo 95% antes do final do primeiro ano do curso. No segundo cenário, todos os cinco modelos ultrapassaram a marca de 77% de acurácia para VP desde a primeira semana, atingindo 99% na quinta semana com o modelo RF. Esse modelo também obteve resultados semelhantes na avaliação da acurácia de VN, com 97% na primeira semana, 99% na semana cinco e 100% a partir da décima semana (Queiroga *et al*., 2018).

Em Burgos *et al*. (2018), o estudo utiliza as atividades realizadas pelos discentes ao longo de um semestre, no qual foram entregues doze atividades ao longo de 20 semanas. Para construir os modelos, os autores adotam uma metodologia incremental: o primeiro modelo utiliza apenas dados da atividade um; o segundo modelo incorpora dados das atividades um e dois; o terceiro modelo inclui as informações das notas das atividades um, dois e três, e assim por diante até o modelo doze, que utiliza todos os dados disponíveis. Essa abordagem é necessária, pois, segundo os autores, pretende-se aplicar os modelos durante os cursos, fazendo uso do modelo com a maior quantidade de dados disponíveis (Burgos *et al*., 2018).

O trabalho define a evasão de acordo com a legislação espanhola, que exige que o aluno alcance pelo menos 50% de aproveitamento na média das atividades

para participar da avaliação final da disciplina. Para avaliar os resultados dos modelos gerados com o objetivo de antecipar a evasão, os autores estabelecem o momento em que um aluno é considerado um desistente. Além de coletar dados de desempenho acadêmico de turmas de cinco cursos de graduação distintos, os autores classificam os alunos como desistentes e não desistentes, e também registram a semana em que ocorre a evasão, resultando em um banco de dados com 104 estudantes, dos quais 26 evadiram seus cursos (Burgos *et al*., 2018).

Os autores propõem a utilização de regressão logística para criar um algoritmo de classificação de alunos. A cada semana do semestre, uma equação de regressão é gerada com coeficientes calculados pelo método de Newton-Raphson e variáveis que representam as notas das atividades nas semanas anteriores. Dessa forma, o aluno é classificado como desistente ou não desistente com base em todas as suas notas até a semana de análise. Para validar o método proposto, quatro outras técnicas de análise são aplicadas aos dados: uma rede neural feed-forward, SVM, um classificador Probabilistic Ensemble Simplified Fuzzy Adaptive Resonance Theory Mapping (PESFAM) e um sistema de mineração de dados educacionais (Burgos *et al*., 2018).

Os resultados indicam que, em média, o classificador proposto é capaz de detectar a desistência de um aluno até 1,6 semana antes do abandono em quatro dos cinco cursos utilizados como dados de entrada. Isso é alcançado com uma precisão de 98,95%, sensibilidade de 96,73%, especificidade de 97,14% e acurácia de 97,13% na semana 10, que é, em média, quando ocorre a desistência. Comparado às outras quatro técnicas de análise, a classificação por regressão logística apresenta métricas superiores na semana de interesse (Burgos *et al*., 2018).

Gottardo, Kaestner e Noronha (2014) também utilizam o ambiente virtual MOODLE. O trabalho categoriza as interações dos estudantes com o ambiente virtual em três grupos: 1) interações entre estudantes e o conteúdo; 2) interações entre estudantes e professores; e 3) interações entre estudantes, supervisionadas ou não. O objetivo do estudo é prever o desempenho dos estudantes, desconsiderando a evasão. Os autores não abordam tentativas de utilização dos modelos com uma base de dados reduzida, que poderiam permitir a antecipação do desempenho acadêmico ao longo da disciplina (Gottardo; Kaestner; Noronha, 2014).

Após categorizar as variáveis disponíveis, os autores dividem os alunos em três grupos com base nos intervalos de notas: maior que 88, entre 77 e 88, e menor que 77. Em seguida, eles realizam três experimentos com os dados para a criação de modelos analíticos. No primeiro experimento, os dados são usados de forma bruta, ou seja, apenas valores numéricos. No segundo, os atributos são transformados em valores discretos, como rótulos e intervalos numéricos. No terceiro experimento, é utilizado um algoritmo de seleção de atributos supervisionado da ferramenta WEKA para escolher

um subconjunto de atributos que tenham alta correlação com as categorias e baixa correlação entre si (Gottardo; Kaestner; Noronha, 2014).

A análise implícita do trabalho utiliza os algoritmos RF e MLP para gerar os modelos em cada experimento. A avaliação dos resultados é feita por meio de matrizes de confusão, das quais os autores extraem informações sobre a quantidade de alunos corretamente classificados em suas respectivas categorias. O classificador RF apresenta uma acurácia média de 77,4%, 77,2% e 72,7%, com um desvio padrão de 7,78%, 2,99% e 9,92%, para cada experimento, respectivamente. Já o algoritmo MLP resulta em uma acurácia média de 80,1%, 77,2% e 76,9%, com um desvio padrão de 8,88%, 2,99% e 8,07%, para o primeiro, segundo e terceiro experimento, respectivamente (Gottardo; Kaestner; Noronha, 2014).

O estudo de Viana, Santana e Rabêlo (2022), concentra-se na avaliação da evasão no contexto do curso de graduação como um todo. Os autores iniciam sua pesquisa classificando os alunos que frequentaram as turmas de Ciências da Computação de 2012 a 2020 em três categorias: evadidos (aqueles que abandonaram o curso), graduados (aqueles que concluíram o curso) e ainda ativos, resultando em uma base de dados com 287 classificados como evadidos, 92 como graduados e 348 como ainda ativos, totalizando 727 instâncias (Viana; Santana; Rabêlo, 2022).

A coleta de dados abrangeu atributos sociais, como raça, sexo, estado civil, idade e se os alunos foram admitidos por meio de programas de cotas, bem como atributos acadêmicos, como notas, quantidade de reprovações por nota e falta, e quantidade de aprovações. Os atributos acadêmicos coletados foram analisados em seis conjuntos, representando semestres de forma cumulativa para alguns dos atributos (o primeiro conjunto contém informações apenas do primeiro semestre, enquanto o segundo conjunto abrange informações do primeiro e segundo semestre, e assim por diante até o sexto conjunto, que possui informações de todos os seis semestres). A análise foi limitada até o sexto semestre devido à redução no número de alunos a partir do sétimo semestre, o que poderia afetar o desempenho dos algoritmos de aprendizado (Viana; Santana; Rabêlo, 2022).

Ao término da coleta de dados, os autores obtiveram um total de 26 atributos, dos quais 12 eram de natureza social e 11 estavam relacionados ao desempenho acadêmico dos alunos. O algoritmo RF foi empregado para selecionar os atributos mais relevantes à análise. Foram consideradas combinações de 5, 10, 15, 20 e 25 atributos selecionados, e as acurácias foram calculadas para cada um dos seis semestres a serem avaliados, usando três diferentes algoritmos de seleção: Chi2, Mutual Information e ANOVA. Os autores concluíram que a melhor acurácia ao longo dos períodos avaliados foi alcançada com o algoritmo Chi2 e os 15 primeiros atributos, destacando a predominância e maior relevância de atributos acadêmicos (Viana; Santana; Rabêlo, 2022).

Com os atributos definidos para cada um dos semestres a serem avaliados, os autores selecionaram algoritmos de aprendizado de máquina para análise, incluindo RF, AD, ET, MLP, SVM, kNN e GNB, disponíveis na biblioteca Scikit-Learn para a linguagem de programação Python. Cada algoritmo foi validado por meio das métricas de acurácia, precisão, sensibilidade, pontuação F_1, índice Kappa, área da curva ROC e valores da matriz de confusão. Os autores destacaram que, de forma geral, todos os algoritmos apresentaram bom desempenho, com ênfase no RF, que alcançou uma acurácia média de 91,55%, sensibilidade de 0,92, precisão de 0,95, pontuação F_1 de 0,93, índice Kappa de 0,81 e área da curva ROC de 0,91 para os seis semestres avaliados (Viana; Santana; Rabêlo, 2022).

Após validar os resultados obtidos para evadidos e graduados entre 2012 e 2020, os autores propuseram a utilização do algoritmo RF e dos dados previamente avaliados para prever a evasão em novos dados, com base nos alunos classificados como ainda ativos no banco de dados. A escolha do modelo para a análise foi baseada no semestre em que o aluno se encontrava, de modo que os alunos no quarto semestre, por exemplo, utilizaram o modelo treinado para o quarto semestre, e a partir da sétima fase, foi utilizado o modelo treinado com os dados do sexto período. Os resultados da evasão e graduação obtidos com os alunos ativos, que ingressaram entre 2013 e 2020, quando somados aos dados reais de evasão e graduação, demonstraram que as taxas de evasão anuais permanecem consistentes, validando, assim, a proposta inicial dos autores (Viana; Santana; Rabêlo, 2022).

O trabalho conduzido por Garcia *et al*. (2022) tem como objetivo empregar a Mineração de Dados Educacionais para criar um ambiente em que os professores possam calcular a probabilidade de um aluno ser reprovado ou aprovado em suas disciplinas no início do semestre. Isso permite que esses professores intervenham e auxiliem os alunos que apresentam risco de reprovação. Os pesquisadores fazem referência a estudos que indicam que o baixo desempenho acadêmico afeta diretamente a trajetória acadêmica, podendo levar à evasão ou ao atraso na conclusão do curso (Garcia *et al*., 2022).

Para coletar dados, os autores usaram o histórico escolar de estudantes da Universidade do Estado de Mato Grosso, combinando turmas ingressantes entre o segundo semestre de 2013 e o primeiro semestre de 2017 nos cursos de Computação e Matemática. Isso resultou em um banco de dados com 297 alunos e 7.677 matrículas em disciplinas do primeiro curso e 247 estudantes e 5.189 matrículas em disciplinas do segundo curso (Garcia *et al*., 2022).

Durante o processo de tratamento de dados, informações acadêmicas e pessoais dos alunos foram extraídas dos componentes curriculares coletados. Essas informações incluíam a fase letiva em que o aluno estava cursando o componente, o resultado obtido (aprovação ou reprovação, que seria o alvo da previsão), sexo, data

de nascimento e todas as disciplinas cursadas em fases anteriores ao semestre em análise. Após a definição dos atributos, foram criadas quatro configurações de base de treinamento: uma com todos os atributos disponíveis; outra sem a faixa etária; outra sem o sexo; e outra sem a faixa etária e o sexo (Garcia *et al*., 2022).

Os autores utilizaram a biblioteca WEKA da linguagem Java para realizar análises com algoritmos de classificação nas bases de treinamento nas configurações originais e posteriormente equilibradas com os métodos de balanceamento de carga, Class Balancer e Synthetic Minority Oversampling Technique (SMOTE). Os algoritmos usados na análise incluíram NB, IBK, JRip, J48, RF e MLP. Para avaliar os resultados, os autores consideraram a acurácia e a sensibilidade como métricas principais. Assim como no presente trabalho, eles destacaram a importância da sensibilidade para a avaliação, embora a acurácia seja a métrica comumente usada em estudos de mineração de dados. Isso ocorre porque é preferível que a previsão tenha um excelente desempenho na identificação de alunos com risco de reprovação, antecipando essa situação para o maior número possível de alunos e possibilitando uma intervenção adequada (Garcia *et al*., 2022).

Dos resultados obtidos, as melhores configurações foram aquelas que excluíam a faixa etária e o sexo, incluindo apenas atributos acadêmicos. Para ambos os cursos, o algoritmo RF foi selecionado como o melhor, atingindo uma sensibilidade de 93% e uma acurácia de 81,43% na base equilibrada com o método SMOTE para o curso de Matemática, e uma sensibilidade de 80,5% e uma acurácia de 74,01% para o curso de Computação. Vale destacar que, em termos de sensibilidade no curso de Computação e na base não balanceada, os algoritmos NB e MLP apresentaram valores ligeiramente superiores aos do RF, com 82,50% e 81,90%, respectivamente, mas tiveram acurácias significativamente mais baixas, com 66,60% e 69,27%, respectivamente, e esses valores foram prejudicados após o balanceamento com o SMOTE (Garcia *et al*., 2022).

Para aprimorar os resultados, as bases de dados foram divididas em subconjuntos separados por período letivo, considerando apenas matrículas designadas para um determinado semestre de acordo com o currículo do curso. Mais uma vez, o algoritmo RF apresentou os melhores resultados, embora fossem inferiores aos das bases iniciais, com uma sensibilidade máxima de 90,40% na terceira fase letiva para o curso de Matemática e 87,10% na segunda fase letiva do curso de Computação. Assim como em (Viana; Santana; Rabêlo, 2022), após determinado semestre, a classificação é prejudicada pela redução no número de discentes matriculados, que, neste caso, acontece a partir do quinto período (Garcia *et al*., 2022).

No estudo realizado por Manhães *et al*. (2011), cujo objetivo é prever quais alunos correm o risco de não concluir o curso de graduação, a base de dados para o estudo foi desenvolvida a partir de informações coletadas entre os anos de 1994 e 2005. A base de dados resultou em 543 alunos que concluíram o curso e 344 que não

o concluíram, seja por iniciativa própria (abandono ou trancamento) ou por imposição da universidade. Os pesquisadores optaram por usar dados do primeiro semestre, que antecedia o maior número de evasões (Manhães *et al*., 2011).

Os atributos selecionados foram listados em ordem de importância, do mais importante ao menos importante: coeficiente de rendimento do período, nota na disciplina de Cálculo Diferencial e Integral I (CDI-1), situação (aprovado, reprovado por nota ou reprovado por falta) na disciplina de CDI-1, notas de outras disciplinas do primeiro semestre e a situação em disciplinas específicas. A ferramenta de mineração de dados WEKA foi escolhida para a análise de dados. Isso permitiu a seleção de vários algoritmos, incluindo OneR, JRip, DT, SC, J48, RF, SL, MLP, NB e BN. OneR foi escolhido como base de referência para a comparação entre os algoritmos devido à sua simplicidade, custo reduzido e alta acurácia (Manhães *et al*., 2011).

No primeiro experimento, os pesquisadores utilizaram o método de validação cruzada com uma base de dados dividida em dez conjuntos. Nenhum dos algoritmos utilizados apresentou diferença significativa em relação ao OneR, com valores médios de acurácia variando entre 2,03% negativos e 0,53% positivos, em relação a OneR, com 78,39%. No segundo experimento, foi empregado o processo de divisão em teste e treinamento com dados randomizados, que dividiu a base de dados em 66% para treinamento e 34% para teste. Nesse experimento, o algoritmo NB, com uma acurácia de 80,12%, foi o único a apresentar uma melhora em relação ao OneR, com 78,50% (Manhães *et al*., 2011).

No último experimento, o tamanho dos conjuntos de treinamento e teste mantiveram as proporções do segundo. Porém, nesse, ambos os conjuntos possuíam a mesma proporção de alunos que concluíram o curso (61%) e que não o concluíram (39%). Os resultados desse experimento mostraram acurácias variando entre 72,92% e 82,29%. O algoritmo de referência OneR obteve uma acurácia de 81,94% e uma taxa de falsos positivos de 33%, enquanto SL alcançou uma acurácia de 82,29% e uma taxa de falsos positivos de 36% (Manhães *et al*., 2011).

O estudo observou que a alta taxa de falsos positivos torna o classificador inadequado para a solução do problema. Isso se deve ao fato de o algoritmo classificar erroneamente alunos com risco de evasão como não tendo risco, o que é considerado um erro grave. Por outro lado, classificar alunos sem risco de evasão como estando em risco (taxa de falsos negativos) é considerado um erro menos grave. Portanto, apesar de terem acurácia inferior a OneR e SL, MP e RF apresentaram a menor taxa de falsos positivos, com 27% e 29%, respectivamente, enquanto DT teve a maior taxa de erros graves, com 42%. Os outros algoritmos apresentaram taxas de falsos positivos entre 30% e 36%. Por fim, a análise direta da base de dados revelou que alunos com baixo rendimento acadêmico, mas que concluem o curso e alunos que possuem alto rendimento, mas não completam o curso, causam um viés nos resultados de acurácia

e taxa de erro do classificador, pois compõem 23,7% da base de treinamento e 19,44% da base de testes (Manhães *et al*., 2011).

APÊNDICE B - CLASSIFICAÇÃO COM FALTAS CONSECUTIVAS E MÉDIA DE PRESENÇA DA TURMA SEM INTERPOLAÇÃO DOS DADOS DE PRESENÇA

Da mesma forma que foi realizado na Seção 4.1.1, as análises para a Turma 1 e Turma 2 foram realizadas de forma independente, e posteriormente suas matrizes de confusão foram somadas, possibilitando calcular as métricas de forma geral. Entretanto, para as atuais análises, não foi realizada a interpolação dos dados de presença. Para cinco faltas em sequência, a Tabela 11 apresenta a matriz de confusão resultante.

Tabela 11 – Matriz de confusão para a classificação de desistência com 5 faltas consecutivas e dados não interpolados

	Desistente real	Não desistente real
Desistente previsto	14	2
Não desistente previsto	28	83

Fonte: Elaborado pelo autor (2023).

As métricas de sensibilidade e precisão, calculadas a partir da matriz de confusão, que resultam em uma sensibilidade de 33,33% e precisão de 87,50%. A partir da comparação com a Tabela 1, que apresenta a matriz de confusão equivalente, mas com interpolação, percebe-se uma variação de sensibilidade de -11,91% e de precisão de +24,17%. Em seguida, a Tabela 12 representa a matriz de confusão para a classificação de desistentes com quatro faltas seguidas.

Tabela 12 – Matriz de confusão para a classificação de desistência com 4 faltas consecutivas e dados não interpolados

	Desistente real	Não desistente real
Desistente previsto	14	6
Não desistente previsto	28	79

Fonte: Elaborado pelo autor (2023).

A sensibilidade encontrada é, novamente, de 33,33% e a precisão é de 70%. Percebe-se que, com relação aos resultados da presente seção com 5 faltas consecutivas, houve apenas a piora da métrica de precisão em 17,5%. Ao ser comparada com a Tabela 2, nota-se uma variação na sensibilidade de -21,43% e na precisão de +21,06%. Finalmente, a Tabela 13 demonstra os resultados atingidos com a classificação a partir de três faltas consecutivas.

Para o caso que teve as melhores métricas ao se interpolar os dados

Tabela 13 – Matriz de confusão para a classificação de desistência com 3 faltas consecutivas e dados não interpolados

	Desistente real	Não desistente real
Desistente previsto	23	16
Não desistente previsto	19	69

Fonte: Elaborado pelo autor (2023).

de presença, ao optar por não realizar tal técnica de preenchimento de dados, encontra-se uma sensibilidade de 54,76% e uma precisão de 58,97%. Que, ao comparadas à Tabela 3, revelam uma variação de sensibilidade e precisão de -19,05% e +11,28%, respectivamente. Desta forma, além de confirmar a hipótese inicial de que a interpolação resultaria em um aumento na sensibilidade, métrica de avaliação do trabalho, confirma-se também a análise exploratória descrita na Seção 3.4.2, que indicava uma tendência à classificação de desistentes pela interpolação, o que pode ser verificado pelo aumento da precisão nos três casos descritos sem a técnica.

APÊNDICE C - INTERPOLAÇÃO DE DADOS DE PRESENÇA PARA A CLASSIFICAÇÃO COM SISTEMA DE PONTUAÇÃO DE DUAS CLASSIFICAÇÕES E PORCENTAGEM DE CORTE DE 50%

A Tabela 14 apresenta as métricas de sensibilidade, precisão, acurácia e pontuação F_1 resultantes de uma análise explícita com técnica de sistema de pontuação de duas classificações e porcentagem de corte de 50% para a classificação da desistência com dados interpolados para o relatório de presença. Tal análise foi realizada com a mesma metodologia aplicada na Seção 4.1.2.2.

Tabela 14 – Métricas de classificação de desistência em duas classes, com limiar de desistência de 50% e interpolação de dados de presença para a Turma 1 e Turma 2

Turma	Proporção de dados	Sensibilidade	Precisão	Acurácia	Pontuação F_1
1	100%	100,00%	50,00%	75,61%	66,67%
1	75%	93,00%	46,50%	72,20%	62,00%
1	50%	95,00%	47,50%	73,17%	63,33%
1	25%	94,00%	47,00%	72,68%	62,67%
2	100%	96,88%	72,09%	84,88%	82,67%
2	75%	99,06%	73,72%	86,51%	84,53%
2	50%	99,06%	73,72%	86,51%	84,53%
2	25%	87,81%	65,65%	78,37%	75,13%

Fonte: Elaborado pelo autor (2023).

Quando comparada à Tabela 4, para a Turma 1, e Tabela 5, para a Turma 2, percebe-se que os resultados com a interpolação dos dados de presença apresentam incrementos pouco significativos, em um acréscimo médio de 1,5% para a sensibilidade e 1,125% para a precisão da Turma 1, além de ocasionar em um decréscimo médio de 3,28% para a sensibilidade e 2,48% para a precisão da Turma 2.