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

RODRIGO NATHAN FRÉTOLA BAESSA

ESTUDO DE CASO COM ANÁLISE DE DADOS PARA A DETECÇÃO DA DESISTÊNCIA DE ESTUDANTES EM DISCIPLINAS OFERTADAS COM APOIO DO AMBIENTE MOODLE E SENTIMENTOS COLETADOS ATIVAMENTE POR QUESTIONÁRIOS

Joinville
2024

RODRIGO NATHAN FRÉTOLA BAESSA

ESTUDO DE CASO COM ANÁLISE DE DADOS PARA A DETECÇÃO DA DESISTÊNCIA DE ESTUDANTES EM DISCIPLINAS OFERTADAS COM APOIO DO AMBIENTE MOODLE E SENTIMENTOS COLETADOS ATIVAMENTE POR QUESTIONÁRIOS

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

Orientador(a): Dr. Ricardo José Pfitscher

Joinville

2024

RESUMO

O trabalho aborda a análise e predição da desistência de alunos em cursos de graduação, utilizando técnicas de mineração de dados educacionais. Utilizaram-se dados de avaliações de desempenho e sentimentos dos estudantes de duas turmas, que foram analisados com métodos de aprendizado de máquina e técnicas de divisão de dados como *holdout*, subamostragem aleatória e validação cruzada. A seleção de características considerou os melhores atributos segundo ANOVA, Chi2, GI e RG, e as métricas de avaliação incluíram acurácia, sensibilidade, precisão e pontuação F1. A análise evidenciou diferenças significativas entre o desempenho de alunos concluintes e não concluintes, sugerindo a eficácia das variáveis propostas na construção de modelos preditivos principalmente com adição do atributo referente ao sentimentos dos alunos. O estudo contribui para a identificação de fatores críticos que influenciam a desistência dos alunos, permitindo a implementação de intervenções precoces e oferecendo uma abordagem robusta para a predição de desistência em contextos educacionais diversos.

Palavras-chave: Mineração de dados educacionais. Predição de desistência. Análise de desempenho acadêmico. Intervenções precoces.

ABSTRACT

This thesis focuses on analyzing and predicting student dropout rates in undergraduate courses through educational data mining techniques. Data from performance evaluations and student sentiment from two different classes were analyzed using machine learning methods and data division techniques such as holdout, random undersampling, and cross-validation. The feature selection considered the top attributes according to ANOVA, Chi2, GI, and RG, with evaluation metrics including accuracy, sensitivity, precision, and F1 score. The analysis highlighted significant differences between the performance of completers and dropouts, suggesting the effectiveness of the proposed variables especially with the addition of the attribute referring to the students' feelings in constructing predictive models. This study contributes to identifying critical factors influencing student dropout, enabling early interventions, and providing a robust approach to dropout prediction applicable in various educational contexts.

Keywords: Educational data mining. Dropout prediction. Academic performance analysis. Early interventions.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABRUEM	Associação Brasileira dos Reitores das Universidades Estaduais e Municipais
AD	Árvore de Decisão
ANDIFES	Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior
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
AVEA	Ambientes Virtuais de Ensino e 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
CTJ	Centro Tecnológico de Joinville
DEED	Diretoria de Estatísticas Educacionais
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
IGC	Instituto Geográfico e Cartográfico de São Paulo

INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KNN	K-Nearest Neighbors
LSTM	*Long Short Term Memory*
MEC	Ministério da Educação
MLP	Multilayer Perceptron
MOODLE	Modular Object-Oriented Dynamic Learning Environment
NB	Naive Bayes
PLN	Processamento de Linguagem Natural
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
SESU	Secretaria de Educação Superior
SVM	*Support Vector Machine*
UFSC	Universidade Federal de Santa Catarina
VC	Validação cruzada
VP	Verdadeiros positivos
VN	Verdadeiros negativos

LISTA DE SÍMBOLOS

β	Letra grega Beta
W_i	Peso atíbuido a cada dia i
Me_i	Média das pontuações de presença da classe para aquele dia i
Max_i	Pontuação de presença máxima definida para o dia i
P_f	Pontuação da presença
Po_i	Pontuação da presença atribuída no MOODLE
d	Número total de dias do relatório de presença
$P_f^{'}$	Pontuação das notas
n	Quantidade total de avaliações
G_{max}	Maior soma das notas
G_k	Nota da atividade k
A_k	Maior nota da classe na atividade k
M_k	Média das notas da classe na atividade k
$P_f^{''}$	Pontuação das atividade concluídas
Po_{max}	Soma das pontuações máximas das atividades
$W_j^{'}$	Peso da atividade j
t_j	Taxa de conclusão da atividade pela turma
$P_{f_1}^{'''}$	Pontuação do sentimento método binário
$P_{f_2}^{'''}$	Pontuação do sentimento método da porcentagem positiva
$P_{f_3}^{'''}$	Pontuação do sentimento método baseado nas demais pontuações
S_{max}	Somatório máximo atingido pela turma

SUMÁRIO

1 INTRODUÇÃO

A evasão de estudantes é objeto de estudos, análises e debates a algumas décadas, principalmente dentre os países mais desenvolvidos. É um tema comum às instituições universitárias no mundo contemporâneo, principalmente pelo seu teor de complexidade e abrangência. Estudos têm demonstrado que existe tanto uma universalidade do fenômeno quanto uma relativa homogeneidade de seu comportamento em determinadas áreas do saber, apesar das diferenças entre as instituições e das peculiaridades sócio econômico-culturais de cada país (ANDIFES/ABRUEM/SESU/MEC, 1996).

Segundo dados recentes do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), a taxa de evasão nos cursos de graduação tem se mantido preocupantemente alta, os números variam de acordo com a região do país e os tipos de instituições de ensino. Esta realidade não só compromete o acesso à educação superior, mas também desperdiça recursos públicos e privados investidos na formação dos alunos, o que gera uma grande perda para o aluno, para as instituições de ensino e a sociedade como um todo (INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022).

A evasão representa uma série de desafios e prejuízos que ultrapassam o âmbito individual, gerando um impacto negativo no progresso educacional, econômico e social. Para os alunos, a interrupção da caminhada acadêmica pode resultar em dificuldades financeiras, desmotivação, e até mesmo o abandono definitivo da busca pela conclusão do ensino superior, comprometendo suas perspectivas profissionais e socioeconômicas futuras, além disso, a evasão impacta negativamente a autoestima, bem como, o bem-estar emocional dos estudantes, gerando sentimentos de fracasso, inadequação e insuficiência (ANDIFES/ABRUEM/SESU/MEC, 1996).

Para a sociedade, a evasão representa um prejuízo de capital humano e um empecilho para o desenvolvimento econômico e social, os graduados são uma contribuição vital para a mão de obra qualificada de um país e possuem um importante papel na geração de desenvolvimento e inovação. São parte fundamental para a competitividade e o progresso de uma nação, portanto, a evasão compromete não apenas o potencial individual dos alunos, mas também o crescimento e a prosperidade coletiva de uma sociedade inteira (INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022).

Já para as instituições de ensino privadas, a evasão implica em uma série de desafios, incluindo a queda na arrecadação de mensalidades e a perda de reputação e credibilidade das mesmas. As instituições públicas de ensino, financiadas

em grande parte por recursos governamentais, estadual ou federal, também são prejudicadas com a evasão, uma vez que investimentos em infraestrutura e recursos humanos são desperdiçados quando os alunos desistem dos cursos, e estas vagas não necessariamente são reocupadas (INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2022).

No Brasil, o Ministério da Educação (MEC) reconhece que os impressionantes níveis de desistência no sistema de ensino superior são um desafio a ser enfrentado; na última avaliação do Índice Geral de Cursos (IGC), a evasão teve destaque ao ser apontada como um dos indicadores considerados na avaliação da qualidade das instituições (MEC, 2023). Evasão educacional pode ser dividida em três conceitos: evasão de curso, o estudante abandona um curso de graduação; evasão institucional, abandono da instituição de ensino pelo aluno; e evasão do sistema, total desistência do ensino superior por parte do estudante (ANDIFES/ABRUEM/SESU/MEC, 1996).

Ao pesquisar sobre o conceito de evasão dentro da literatura, identifica-se que o termo evasão é atribuído a desistência do curso. No entanto, ao propor um trabalho que realiza a análise de dados para prever a evasão do aluno em disciplinas, o presente trabalho classifica a não-conclusão de uma disciplina qualquer da grade curricular de um curso de ensino superior como desistência ou evasão (DEED - DIRETORIA DE ESTATíSTICAS EDUCACIONAIS. Ministério da Educação, 2017).

Este trabalho tem como foco a análise da não conclusão de disciplinas em um curso de graduação na modalidade presencial, com o propósito de identificar indicadores que possam prever o potencial abandono de uma disciplina acadêmica. Para isso, foi conduzido um estudo de caso, que investiga a integração de dados coletados ativamente ligados aos sentimentos dos estudantes com relação à disciplina e à universidade, com os dados coletados passivamente do desempenho do aluno por meio do ambiente MOODLE.

Neste campo de estudo existem trabalhos que abordam métodos passivos, que são aqueles em que os dados sobre o desempenho do aluno ao longo do semestre ou do curso são coletados sem intervenção do estudante; outra linha de estudos propõem a utilização exclusiva de dados coletados ativamente por meio de questionários, onde os estudantes respondem sobre seu desempenho, sua satisfação e seu sentimento com a disciplina ou universidade. No entanto, o presente trabalho investiga um método combinatório, utilizando tanto a abordagem passiva quanto a ativa, para a classificação.

Trabalhos que utilizam somente um tipo de fonte de dados são limitados a analisar o estudante sobre apenas uma ótica. Por exemplo, um estudante pode estar tendo um desempenho acadêmico considerado positivo, e de mesmo modo estar enfrentando dificuldades dentro da matéria ou universidade, como falta de motivação, ansiedade, medo e sentimentos mais adversos; mesmo tendo um bom desempenho na disciplina ele tem a chance de se evadir da disciplina.

Sendo assim, o presente estudo propõe um modelo de classificação baseado em dois outros trabalhos, o primeiro é um artigo que elabora uma classificação de sentimentos dos alunos por meio de Processamento de Linguagem Natural (PLN) PFITSCHER et al. (2023). O segundo trata-se de um trabalho de conclusão de curso que investiga a criação de modelos de classificação implícitos e explícitos utilizando dados coletados de maneira passiva do estudante ao decorrer do semestre letivo SOUZA (2023).

A proposta deste trabalho é não é simples, uma vez que os dados ativos são coletados por meio de respostas dos questionários, e essas respostas são esperadas na forma de texto livre, que são de âmbito facultativo e anonimizadas. O trabalho analisa métodos de transformação e adequação destas respostas obtidas para um conjunto de dados compatíveis, que posteriormente serão utilizadas para alimentar o modelo de classificação desenvolvido.

A não obrigatoriedade e anonimato das respostas dificultam a atribuição de uma resposta específica a um indivíduo da turma, sendo assim, é necessário generalizar os sentimentos extraídos de maneira que representem a turma como um conjunto. Neste trabalho serão abordadas algumas estratégias e métodos de como realizar esta generalização e atribuição.

1.1 OBJETIVO

Para resolver a problemática da falta de ferramentas disponíveis aos docentes para a predição da desistência dos estudantes nas disciplinas de cursos superiores, propõe-se neste trabalho os seguintes objetivos.

1.1.1 Objetivo Geral

Identificar, por meio da análise de dados passivos e sentimentos ativos, alunos que estejam propensos a desistir da disciplina, permitindo assim que o docente intervenha da maneira mais rápida possível para auxiliar esses alunos a garantir sua permanência e sucesso dentro do curso.

1.1.2 Objetivos Específicos

- Elaborar métodos para converter sentimentos em atributos úteis para o modelo de classificação;
- Analisar a conjunção de dados provenientes de mais de uma fonte;
- Elaborar modelos para classificar alunos desistentes utilizando dados combinados;
- Identificar, por meio da análise de dados passivos extraídos do Ambiente Virtual de Aprendizagem MOODLE do estudante e sentimentos ativos adquiridos por

meio de questionários combinados, alunos que estejam propensos a desistir da disciplina.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo visa apresentar os conceitos fundamentais utilizados na construção do estudo, por meio de uma revisão bibliográfica da literatura. Sendo assim, serão abordados os principais conceitos desde a definição de evasão, etapas de pré-processamento de dados, quais são as ferramentas comumente utilizadas na classificação de dados, até como avaliar os resultados obtidos por meio de métricas. Por último, este capítulo apresenta os trabalhos que serão utilizados como base para a condução do presente estudo.

2.1 EVASÃO E DESISTÊNCIA

A publicação de Vincent Tinto em 1975, o livro "*Dropout in Higher Education: A Theoretical Synthesis of Recent Research*" é base para muitos estudos sobre evasão de estudantes nas universidades. No texto, o autor procura entender o motivo que leva um estudante a tomar a decisão de evadir a universidade e entender um pouco sobre este processo, já que na época pouca atenção era destinada à conceituação do fenômeno da evasão do ensino superior (TINTO, 1975).

Tinto (1975) propõem um modelo teórico de integração institucional para entender a evasão universitária, e afirma que a evasão decorre das influências que as comunidades sociais e intelectuais exercem sobre a vontade dos estudantes em permanecer na faculdade. Na pesquisa, o autor define o termo evasão como um termo relacionado às pessoas que abandonam permanentemente a instituição na qual estão matriculados (TINTO, 1975).

A Comissão Especial de Estudos sobre a evasão nas Universidades Públicas Brasileiras publicou um documento que apresenta um estudo que reúne um conjunto significativo de dados sobre o desempenho das universidades públicas brasileiras relativo aos índices de diplomação, retenção e evasão dos estudantes de seus cursos de graduação (ANDIFES/ABRUEM/SESU/MEC, 1996).

Neste documento, é abordada a distinção entre evasão e exclusão, afirmando que a evasão corresponde a uma postura ativa do aluno que decide desligar-se por sua própria responsabilidade. Já a exclusão, implica a admissão de uma responsabilidade da Universidade e de tudo que a cerca, por não ter mecanismos de aproveitamento e direcionamento do estudante para a formação (ANDIFES/ABRUEM/SESU/MEC, 1996).

A Comissão, mesmo reconhecendo as limitações possíveis, decidiu por caracterizar evasão distinguindo:

1. **Evasão de curso:** Quando o estudante desliga-se do curso superior em situações diversas tais como: abandono (deixa de matricular-se), desistência (oficial),

transferência ou reopção (mudança de curso), exclusão por norma institucional;
2. **Evasão da instituição:** Quando o estudante desliga-se da instituição na qual está matriculado;
3. **Evasão do sistema:** Quanto o estudante abandona de forma definitiva ou temporária o ensino superior.

De acordo com a diretoria de estatísticas educacionais (DEED) do INEP, a evasão pode ser definida como a saída antecipada, antes da conclusão do ano, série ou ciclo, por desistência, representando, portanto, condição terminativa de insucesso em relação ao objetivo de promover o aluno a uma condição superior à de ingresso, no que diz respeito à ampliação do conhecimento, ao desenvolvimento cognitivo, de habilidades e de competências almejadas para o respectivo nível de ensino (DEED - DIRETORIA DE ESTATíSTICAS EDUCACIONAIS. Ministério da Educação, 2017).

No entanto, segundo MOURA e SILVA (2007) o termo evasão é carregado de um sentido que culpabiliza o indivíduo que, por várias razões, interrompeu definitivamente sua trajetória em uma determinada oferta educacional. Assim, o termo contribui para a isenção da instituição e o respectivo sistema educacional de qualquer responsabilidade sobre esse fenômeno. É preciso ter claro que o afastamento definitivo de um estudante é fruto de múltiplos fatores sociais, econômicos, familiares, institucionais e pessoais, os quais se reforçam mutuamente e resultam na chamada evasão (MOURA; SILVA, 2007).

A partir disso, julga-se correta a utilização do termo desistência, que é considerado um conceito que carrega menor culpa para o indivíduo, neste trabalho utiliza-se o termo desistência para abordar a interrupção da trajetória escolar do aluno. Sendo assim, o termo evasão ou desistência se referem a não conclusão de uma disciplina por parte do estudante, por quaisquer motivos que o leve a tomar esta atitude sem investigar a atribuição de culpa de tal decisão.

2.2 COLETA DE DADOS

A coleta de dados é um processo utilizado para captar informações geradas por pessoas ou processos, e que posteriormente servirão de insumos para um modelo de análise que busca extrair informações ocultas nestes dados. Os dados podem ser coletados de diversas maneiras, sendo comumente empregadas ferramentas de mineração de dados por meio de plataformas específicas para coletas, formulários, sites, históricos e outras metodologias.

Na coleta ativa de dados, o público alvo é estimulado pelo proprietário da pesquisa a responder questionários com perguntas específicas ou abertas sobre o tópico em questão. Desta maneira, o participante necessariamente precisa responder às questões solicitadas de maneira direta (ESOMAR, 2020).

São todos os dados coletados sem utilizar o sistema tradicional de questionários (de perguntar e responder às perguntas), estes dados são coletados geralmente por relatórios de ações, comportamentos ou resultados. Assim, são coletadas informações indiretas sobre o indivíduo, por exemplo, quantos questionários ativos ele respondeu, as datas destes questionários, entre outros (ESOMAR, 2020).

2.3 PRÉ-PROCESSAMENTO DE DADOS

A fase de pré-processamento inicia-se depois que os dados são coletados e organizados na forma de um conjunto de dados. Dentro desta fase tem-se diversos objetivos a serem cumpridos, que podem variar de acordo com os dados obtidos e resultados desejados. As ações desempenhadas no pré-processamento tem como objetivo preparar os dados para as próximas etapas, onde serão realmente processados e interpretados para assim extrair conclusões e/ou resultados.

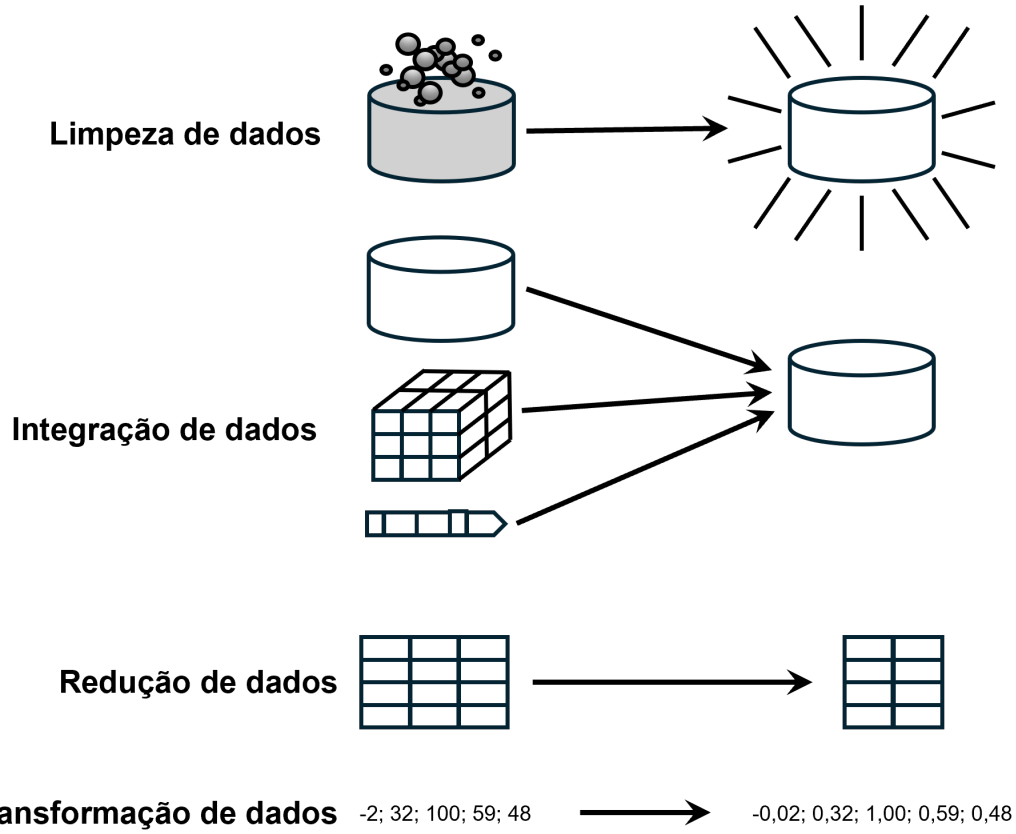
Muitas vezes, é necessário solucionar problemas no banco de dados, tais como identificar e tratar dados corrompidos, imprecisos ou inconsistentes, valores desconhecidos e atributos desnecessários. Ou então, pode ser interessante alterar a estrutura dos dados coletados de tal maneira que facilite a sua manipulação, e contribua para a qualidade do conjunto de dados e, consequentemente, com a eficiência do processamento dos dados (HAN; KAMBER, 2006).

Pré-processar dados é uma fase crucial e complexa, que frequentemente demanda um grande período de tempo e esforço, e no ambiente educacional não é diferente, essa complexidade é oriunda da natureza inicialmente inadequada dos dados educacionais para questões específicas de análise. A seleção e transformação dos dados em um formato apropriado dependem do objetivo da análise, exigindo coleta meticulosa e alinhamento com os objetivos propostos (HAN; KAMBER, 2006).

Ambientes educacionais geram grandes volumes de dados provindos de múltiplas fontes, o que necessita um processo de 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 como nomes e *e-mails* (HAN; KAMBER, 2006).

O pré-processamento é constituído de quatro principais etapas no seu processo: limpeza, integração, redução e transformação. A limpeza de dados é o primeiro passo deste processo, e consiste tanto em preencher valores ausentes e suavizar dados ruidosos quanto identificar e remover pontos inconsistentes e discrepantes, de maneira que aprimore a qualidade e confiança nos resultados obtidos a partir da mineração desses dados (HAN; KAMBER, 2006).

Figura 1 – Etapas do pré-processamento de dados



Fonte: Han e Kamber (2006), adaptado pelo autor (2024).

Pode ser conveniente a integração de dados provindos de múltiplas fontes para a análise, no entanto, os atributos que representam um mesmo conceito podem estar nomeados de diferentes maneiras dentro destes conjuntos de dados, causando assim inconsistências e redundâncias. O processo que visa melhorar a acurácia e velocidade da mineração de dados por meio da união consciente de diferentes bases de dados é conhecido como integração de dados (HAN; KAMBER, 2006).

A estratégia da redução de dados se baseia em obter uma representação reduzida da base de dados em volume de ocupação, e mesmo assim, produzir o mesmo (em alguns casos, quase o mesmo) resultado analítico. Existem alguns métodos para a redução de dados, a redução por dimensão, redução por número e compressão de dados, todos visam reduzir a complexidade e os tempos de execução atrelados a análise e mineração de dados (HAN; KAMBER, 2006).

A transformação de dados 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.

2.4 ANÁLISE DE SENTIMENTOS

O grande propósito da análise de sentimentos é definir técnicas automáticas que sejam capazes de extrair informações subjetivas de textos escritos utilizando linguagem natural, de modo a criar conhecimento estruturado que possa ser convertido em uma tomada de decisão. A identificação de sentimentos em textos é uma das áreas de pesquisa mais destacadas em Processamento de Linguagem Natural desde o início do século (BENEVENUTO; RIBEIRO; ARAÚJO, 2015)(BRITTO; PACÍFICO, 2019).

Dentre os objetivos da análise de sentimentos estão determinar a polaridade (classificar em positivo, negativo ou neutro), identificar emoções, quantificar sentimentos (pontuação sobre a intensidade do sentimento), monitorar tendências e extrair *insights*. Para atingir tais objetivos podem ser empregadas diversas técnicas que variam em complexidade, sendo elas métodos utilizando dicionários, aprendizado de máquina supervisionado e não-supervisionado e modelos de *Deep Learning* (BENEVENUTO; RIBEIRO; ARAÚJO, 2015).

Em resumo, a análise de sentimentos é uma ferramenta poderosa para entender a percepção e emoções das pessoas em relação a diversos tópicos, permitindo que organizações tomem decisões mais informadas e ajustem suas estratégias com base no *feedback* real dos usuários (BRITTO; PACÍFICO, 2019).

2.5 FERRAMENTAS DE CLASSIFICAÇÃO

Jiawei Han (2006) define a classificação e a predição como duas formas de análise de dados que podem ser utilizadas tanto para extrair modelos que podem ser utilizados para descrever classes de dados importantes quanto para prever tendências futuras dos dados. Tal análise pode ajudar em uma melhor compreensão dos conjuntos de dados, enquanto a classificação prevê rótulos categóricos (discretos, não ordenados), a previsão modela funções de valor contínuo (HAN; KAMBER, 2006).

A classificação é uma das ferramentas mais comuns dentro da mineração de dados, visa identificar a classe ao qual determinado dado pertence. O modelo analisa o conjunto de dados disponível, com cada um dos dados previamente categorizados em suas respectivas classes, a fim de aprender a categorizar novos dados baseados no conjunto fornecido, esse método é conhecido como aprendizado supervisionado.

A tarefa de predição é similar às tarefas de classificação, no entanto, ela visa antecipar o resultado futuro de um determinado atributo. Métodos de classificação podem ser utilizados para predição de rótulos categóricos, este processo de predição por categorização pode ser dividido em etapas de aprendizagem (também conhecida como treinamento), e de classificação (onde o modelo gerado é utilizado na predição de rótulos dos dados de teste) (HAN; KAMBER, 2006).

A primeira etapa do processo de classificação compreende a construção

do modelo classificador, este modelo é obtido a partir de uma base de dados de treinamento, composta por tuplas e suas respectivas classificações. As tuplas, são conjuntos de valores de atributos que representam um dado específico, as tuplas podem ser consideradas como elementos de dados estruturados ou registros em uma tabela de banco de dados relacional.

Na etapa seguinte, o modelo gerado é usado para classificar uma base de dados de teste, também composta de tuplas e seus rótulos de classe associados. Essas tuplas são selecionadas aleatoriamente do conjunto de dados gerais e são independentes das tuplas de treinamento, o que significa que não são usadas para construir o classificador, tendo em vista que o classificador tende a realizar um sobre ajuste aos dados durante o aprendizado.

2.5.1 Classificadores Baseados em PLN

O Processamento de Linguagem Natural (PLN) compreende a área da ciência da Computação dedicada ao desenvolvimento de métodos que tem por finalidade analisar, classificar, reconhecer e/ou gerar textos baseados na linguagem humana, ou linguagem natural. No entanto, o PLN pode ser uma tarefa árdua, devido ao nível de complexidade e ambiguidade atribuída à linguagem natural, dessa maneira criaram-se diversas técnicas computacionais para tentar viabilizar sua interpretação(VIEIRA; LOPES, 2010).

Tradicionalmente, o trabalho no Processamento de Linguagem Natural é decomponível em etapas, representando as distinções linguísticas teóricas divididas entre sintaxe, semântica e pragmática. No entanto, se faz necessário uma decomposição mais detalhada, adicionando mais estágios, ao considerar o atual estado da arte da combinação de dados de linguagem reais, como ilustrado na Figura 2.

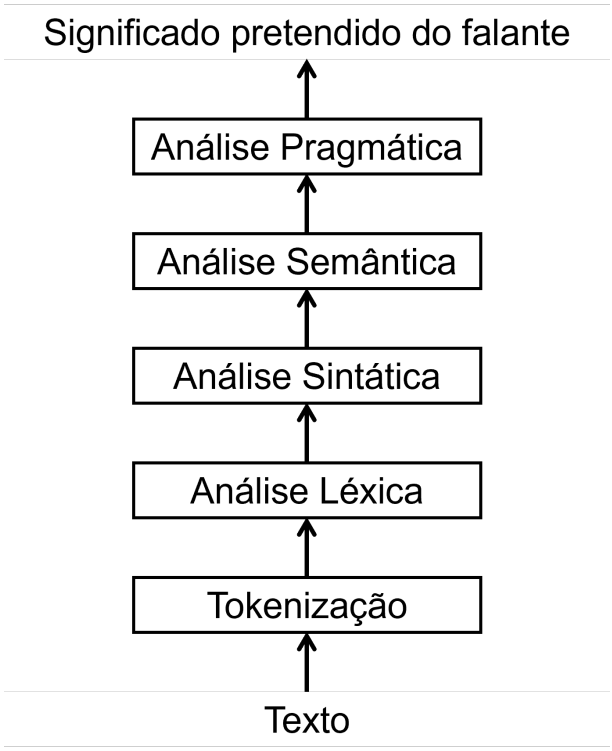
Observa-se que o PLN pode ser dividido em cinco fases de análise, a primeira destas fases é a *tokenização*, também conhecida como segmentação de palavras, quebra a sequência de caracteres em um texto localizando o limite de cada palavra, ou seja, os pontos onde uma palavra termina e outra começa. Para fins de linguística computacional, as palavras assim identificadas são frequentemente chamadas de *tokens* (BARBOSA et al., 2017).

O processo subsequente em questão é a análise a nível de palavras, a análise léxica, a tarefa básica da análise léxica é relacionar variantes morfológicas aos seus *lemmas*, que nada mais são do que as formas canônicas das palavras, ou a forma em que as palavras se encontram no dicionário. A lematização é utilizada de diferentes formas, as quais dependem da tarefa a ser realizada pelo sistema de processamento de linguagem natural (BARBOSA et al., 2017).

Uma frase expressa uma proposição, uma ideia ou um pensamento, além de dizer algo sobre o mundo real ou imaginário, sendo assim, extrair o significado de

uma frase é, portanto, uma questão crucial. Para realizar essa tarefa é necessário uma análise aprofundada de cada frase, as quais determinam suas estruturas de uma maneira ou de outra, desta forma caracteriza-se a análise sintática (BARBOSA et al., 2017).

Figura 2 – Fases da análise do Processamento de Linguagem Natural



Fonte: Barbosa et al. (2017), adaptado pelo autor (2024).

Na linguística, a análise semântica refere-se à análise do significado das palavras, expressões fixadas, sentenças inteiras e enunciados no contexto, já na prática, isso significa traduzir as expressões originais em um tipo de metalinguagem. Em termos gerais, a evidência primária para a linguística semântica vem das interpretações do orador nativo no uso das expressões no contexto padrões de uso, colocação e frequência, que são detectáveis usando técnicas linguísticas (BARBOSA et al., 2017).

A análise pragmática foge à estrutura de apenas uma frase, uma vez que ela busca nas demais frases a compreensão do texto que falta àquela frase em análise. Sendo assim, a análise pragmática se concentra na compreensão do significado contextual e na interpretação das intenções por trás das expressões linguísticas em um dado contexto comunicativo (BARBOSA et al., 2017).

2.6 AVALIAÇÃO DE RESULTADOS

Esta seção tem por finalidade descrever diferentes métodos de separação de dados entre treinamento e validação, bem como definir métricas que serão utilizadas

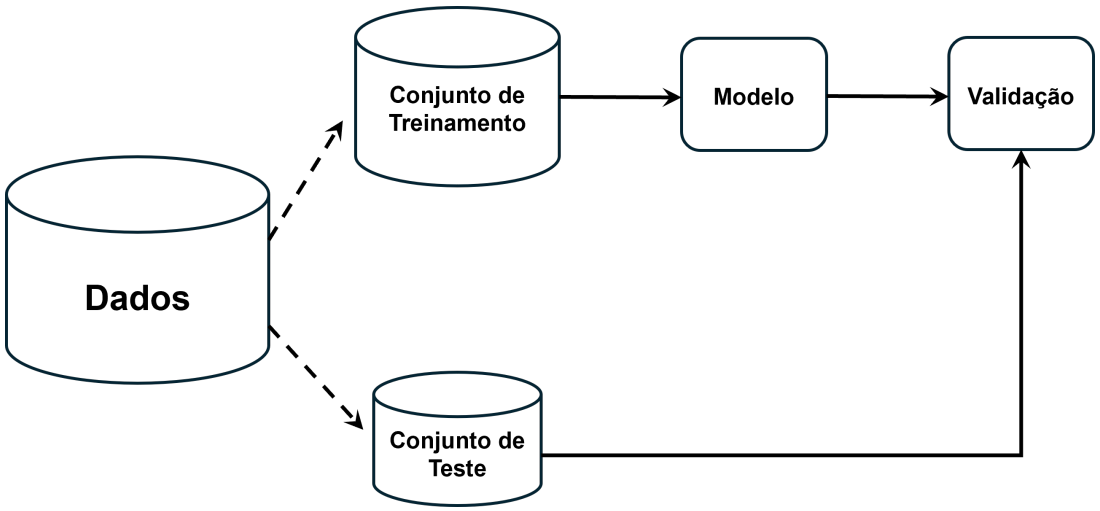
para avaliar a performance dos resultados obtidos por meio desses métodos e discutir qual o melhor modelo estudado para realizar a classificação dos dados.

Método *holdout*, subamostragem aleatória e validação cruzadas são algumas das técnicas mais comumente utilizadas para avaliar a precisão com base em partições amostradas aleatoriamente dos dados fornecidos, e serão abordadas nesta seção. O uso de tais técnicas para estimar a precisão aumenta o tempo total de computação, mas é extremamente útil para a seleção de melhores modelos (HAN; KAMBER, 2006).

2.6.1 Método *Holdout*

O método *holdout*, também conhecido como validação simples, particiona aleatoriamente a base de dados em dois conjuntos independentes de dados, um conjunto destinado para treinamento e outro reservado para a avaliação/teste. A literatura menciona que é comumente utilizado dois terços dos dados para realizar o treinamento do modelo e o restante, um terço, para estimar a precisão do modelo gerado, no entanto, outras divisões são possíveis (HAN; KAMBER, 2006).

Figura 3 – Validação de modelos a partir do método *holdout*



Fonte: Han e Kamber (2006), adaptado pelo autor (2024).

A separação dos dados de teste visa lidar com as imperfeições de um mundo não ideal, e a incapacidade de coletar mais dados da distribuição original. Com o método *holdout* não é possível assegurar que o subconjunto de dados de treinamento é representativo para o conjunto total da base de dados, sendo assim, a estimativa do método é pessimista, uma vez que apenas parte dos dados totais são utilizados para treinar o modelo (HAN; KAMBER, 2006).

2.6.2 Subamostragem Aleatória

A subamostragem aleatória pode ser encontrada na literatura pelo nome de *holdout* repetido, e é caracterizada pela iteração do método *holdout* η vezes, sendo cada uma dessas iterações uma nova separação aleatória dos conjuntos de dados de treinamento e teste. Com isso, a estimativa da precisão geral é considerada como a média das precisões obtidas em cada iteração realizada pelo método, o que visa obter uma estimar uma performance mais robusta e de com menor variância (HAN; KAMBER, 2006).

2.6.3 Validação Cruzada

A proposta do método da validação cruzada é que todas as amostras de dados devem ser testadas pelo modelo. Na avaliação cruzada k-*fold*, o conjunto de dados é particionado de forma aleatória em k subconjuntos exclusivos de dados de maneira que possuam tamanhos aproximadamente iguais, o treinamento e o teste são realizados k vezes, onde em cada iteração apenas um dos subconjuntos criados, sem que se repita, é utilizado para a avaliação do modelo (HAN; KAMBER, 2006).

2.6.4 Métricas

A etapa de avaliação dos resultados obtidos é imprescindível, é o momento de avaliar a performance dos modelos de classificação gerados por meio de métricas bem definidas pela literatura. Ao analisar-se os resultados de um algoritmo de aprendizado de máquina, especialmente quando relacionados a previsão de desistência de estudantes, as métricas estão relacionadas com o número de previsões:

1. **Verdadeiras Positivas (VP):** Referentes as tuplas valor real positivo que foram corretamente classificadas, ou seja, quando um estudante que desistiu da disciplina é corretamente classificado como desistente;
2. **Verdadeiras Negativas (VN):** Referentes as tuplas valor real negativo que foram corretamente classificadas, ou seja, quando um estudante que não desisteiu da disciplina é corretamente classificado como não desistente;
3. **Falsas Positivas (FP):** Referentes as tuplas valor real negativo que foram erroneamente classificadas, ou seja, quando um estudante que não desistiu da disciplina é classificado erroneamente como desistente gerando um resultado falso positivo;
4. **Falsas Negativas (FN):** Referentes as tuplas valor real positivo que foram erroneamente classificadas, ou seja, quando um estudante que desistiu da disciplina é classificado erroneamente como não desistente gerando um resultado falso negativo.

A ferramenta normalmente utilizada para representar os resultados das previsões é a matriz de confusão, ela auxilia na compreensão e na análise da performance de classificadores, pode-se observar um exemplo na Figura 4. Esta matriz, permite mensurar e visualizar de maneira fácil o quão assertivo é o modelo classificador, considera-se como acertos previsões dadas como verdadeiras positivas e negativas (diagonal primária), e erros são previsões falsas positivas e negativas (diagonal secundária) (OLIVEIRA, 2023).

Figura 4 – Matriz de confusão das previsões

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Oliveira (2023), adaptado pelo autor (2024).

Sendo assim, com as quatro unidades básicas de avaliação, é possível estabelecer métricas de acurácia, sensibilidade, precisão e pontuação F1, que serão abordadas a seguir.

2.6.4.1 Acurácia

A acurácia representa a proporção de tuplas que o modelo classificador foi capaz de classificar corretamente, desta forma sinaliza o quão bem o modelo reconhece tuplas de diferentes classes. Calcula-se a partir da divisão entre o número de previsões verdadeiras (VP e VN) e o número total de previsões, como descrito na Equação 1 (OLIVEIRA, 2023).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

A acurácia é mais aconselhada para conjuntos de dados que estejam balanceados, esta métrica melhor representa a qualidade dos resultados obtidos quando as classes de dados são distribuídas uniformemente. Caso isso não aconteça, se faz necessário o uso métricas que possam avaliar quão bem o classificador reconhece cada classe individualmente, como a sensibilidade e a precisão (HAN; KAMBER, 2006).

2.6.4.2 Sensibilidade

A sensibilidade é a taxa de reconhecimento de verdadeiros positivos, ou seja, a proporção de tuplas positivas corretamente identificadas pelo modelo. Esta métrica é

uma das alternativas para contornar o problema de desbalanceamento de classes, é representada na Equação 2 (OLIVEIRA, 2023).

$$Sensibilidade = \frac{VP}{VP + FN} \tag{2}$$

Um resultado com sensibilidade de 1 representa a correta classificação de todos os elementos positivos, no entanto, para saber a quantidade de elementos negativos que foram classificados como positivos erroneamente, a métrica de precisão deve ser calculada (HAN; KAMBER, 2006).

2.6.4.3 Precisão

Precisão é uma métrica para aferir a exatidão do modelo classificador, refere-se a proporção de tuplas classificadas como positivas que de fato são positivas. O valor é calculado a partir da divisão entre as previsões positivas verdadeiras (VP) e o total de previsões positivas (VP + FP), como pode ser observado na Equação 3 (OLIVEIRA, 2023).

$$Precisão = \frac{VP}{VP + FP} \tag{3}$$

Um resultado de precisão elevado, indica a presença de um valor baixo de previsões positivas falsas (FP), porém, analogamente ao que foi apresentado para a sensibilidade, essa métrica não considera a quantidade de positivos que foram considerados erroneamente como negativos. Assim, as métricas de sensibilidade e precisão são utilizadas em conjunto, comparando valores de precisão para um valor fixo de sensibilidade, ou vice-versa (HAN; KAMBER, 2006).

2.6.4.4 Pontuação F_β

A pontuação F_β é uma forma de combinar as métricas de precisão e sensibilidade em uma única métrica. Essa métrica define um peso para a sensibilidade e a precisão, de maneira que a sensibilidade possua um peso β vezes maior que a precisão, como é representado na Equação 4 (HAN; KAMBER, 2006).

$$F_\beta = \frac{(1 + \beta^2) \times Precisão \times Sensibilidade}{\beta^2 \times Precisão \times Sensibilidade} \tag{4}$$

Caso β seja definido como 1, a pontuação F1 resultante é uma média harmônica entre a sensibilidade e precisão, e o peso para ambas é o mesmo (HAN; KAMBER, 2006).

2.7 TRABALHOS SIMILARES

Esta seção apresenta um resumo de trabalhos utilizados como base de técnicas e raciocínios para a elaboração deste estudo. O estudo do presente trabalho é derivado do artigo “Análise de sentimentos em turmas de programação com vistas ao apoio à permanência estudantil” (PFITSCHER et al., 2023) e o trabalho de conclusão de curso “Estudo de caso com análise de dados para detecção da desistência de estudantes em disciplinas ofertadas com apoio do ambiente MOODLE” (SOUZA, 2023).

No entanto, existem outros trabalhos que possuem abordagens semelhantes à proposta deste estudo e também serão citados a seguir.

2.7.1 Análise de sentimentos em turmas de programação com vistas ao apoio à permanência estudantil

O artigo "Análise de sentimentos em turmas de programação com vistas ao apoio à permanência estudantil” tinha o objetivo de utilizar de ferramentas baseadas em PLN para gerar modelos que classificam o sentimento do estudante referente à disciplina e à universidade. Esta classificação tinha o intuito de futuramente ser utilizada para investigar as possíveis correlações entre os sentimentos coletados dos alunos e as desistências da disciplina, bem como as suas razões (PFITSCHER et al., 2023).

O estudo de caso relatado pelo artigo foi direcionado às turmas de programação de computadores de duas instituições de ensino superior, Universidade Federal de Santa Catarina e Centro Universitário Católica de Santa Catarina, em que dados foram coletados de forma ativa por meio de questionários não obrigatórios em momentos específicos do semestre. Posteriormente, estes dados foram classificados utilizando uma metodologia que combinou análise psicopedagógica e automatizada baseada em Processamento de Linguagem Natural.

A classificação psicopedagógica foi realizada manualmente por duas psicopedagogas, e para esta classificação, primeiramente as avaliadoras definiram as categorias e estabeleceram as listas de itens sobre os aspectos positivos e negativos na experiência com o curso/universidade. Os sentimentos e emoções extraídos dos alunos posteriormente foram categorizados em cinco categorias: muito satisfeito, satisfeito, indiferente, insatisfeito e muito insatisfeito.

Para a criação dos modelos de classificação foram utilizadas duas bases de dados distintas, os autores selecionaram as duas maiores bases de dados em português e que apresentavam uma classificação binária. A primeira delas foi baseada em resenhas de filmes do site **IMDb** e a segunda base escolhida foi da companhia **SenticNet**.

Os principal resultado deste estudo relevanta para o presente trabalho é uma acurácia de 68% com PLN, indicando possível utilização para ações de mitigação da

desistência.

2.7.2 Estudo de caso com análise de dados para detecção da desistência de estudantes em disciplinas ofertadas com apoio do ambiente MOODLE

Em relação ao trabalho “Estudo de caso com análise de dados para detecção da desistência de estudantes em disciplinas ofertadas com apoio do ambiente MOODLE”, apresenta-se o desenvolvimento de um software que analisa dados obtidos a partir do MOODLE, dados passivos dos estudantes, identificando fatores que apontem a possibilidade de desistência de alunos em determinada disciplina da Universidade Federal de Santa Catarina (SOUZA, 2023).

O trabalho sugere duas análises, a primeira a partir de uma análise direta dos dados coletados, e a segunda utilizando algoritmos de aprendizado de máquina. O autor construiu alguns conceitos para auxiliar na classificação dos estudantes desistentes considerando os dados passivos obtidos com relação a presença de cada estudante, notas obtidas e informações sobre conclusão de atividades importantes no ambiente do MOODLE.

A primeira análise sugerida pelo autor objetiva desenvolver equações e métodos de análise diretamente a partir de observações realizadas dos dados, enquanto que a segunda análise utiliza métodos de aprendizado de máquina para gerar resultados. Por fim, a avaliação dos resultados da análise de dados, a partir dos quais são calculadas as métricas de sensibilidade, precisão, acurácia e pontuação F_β.

O autor comenta que os estudantes foram classificados manualmente pelo docente em desistentes ou não desistentes, o que tornou possível o cálculo das métricas de sensibilidade, precisão e acurácia a partir da comparação entre os resultados previstos pelo modelo criado e a classificação real. Souza (2023) comenta que não houve contato com os dados reais, assim 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).

Os resultados mostraram que a análise implícita proporcionou melhorias significativas nas métricas de precisão, embora a análise explícita também tenha oferecido métricas satisfatórias e seja menos complexa. Com base nas pontuações geradas, professores podem focar seus esforços nos alunos com maior probabilidade de desistência, otimizando as intervenções.

Além disso, o estudo comparou os resultados obtidos com outros trabalhos semelhantes, concluindo que as métricas alcançadas são equiparáveis. O trabalho foi capaz de identificar uma maior proporção de desistentes em comparação a estudos anteriores, embora tenha classificado incorretamente um número maior de alunos não desistentes como desistentes. Apesar das limitações, como a dependência de dados

históricos e a aplicação em apenas duas turmas, o objetivo principal de detectar alunos com tendências à desistência e permitir uma intervenção efetiva foi atingido.

2.7.3 Análise e predição de evasão dos alunos de um curso de Graduação em Sistemas de Informação por meio da mineração de dados educacionais

Noetzold e Pertile (2021) em seu artigo propõe o desenvolvimento de um estudo sobre os padrões da evasão escolar no ensino superior, que se baseia na análise de dados fornecidos pelo curso de Sistemas de Informação da Universidade Federal de Santa Maria - Campus Frederico Westphalen. Os dados do estudo passaram por um rigoroso pré-processamento de dados, a fim de apontar indicadores relacionados a fatores que classificam possíveis evasões durante o curso (NOETZOLD; PERTILE, 2021).

A etapa de pré-processamento de dados resultou em 409 registros de alunos, onde os estudantes foram categorizados em três tipos: "regular", "formado"ou "evadido". Além disso, os dados foram discretizados em conceitos, classificados como "MUITO BAIXO", "BAIXO", "MÉDIO", "ALTO"e "MUITO ALTO", o estudo analisou a relação entre a evasão escolar no curso de Sistemas de Informação e a distância entre a instituição e a moradia do estudante, utilizando a API do Google Maps para calcular a distância aproximada em quilômetros.

Os resultados foram gerados por meio de árvores de decisão que apontaram dados referentes ao aluno e seu desempenho acadêmico como fatores importantes para identificação da evasão escolar no ensino superior. A pesquisa aborda a relevância da investigação da evasão escolar no ensino superior, considerando seus impactos econômicos e na educação em geral.

Na conclusão geral dos casos, o algoritmo acabou por classificar atributos relacionados a média de notas, aprovação, idade e presença do aluno, ou seja, atributos ligados diretamente ao estudante e seu desempenho acadêmico, que se mostraram fortes indicadores de evasão. O estudo também apontou que a distância entre a moradia do estudante e a Instituição não atuou como um fator relevante para a evasão neste estudo.

2.7.4 Um estudo sobre as variáveis para predição de alunos não concluintes em cursos suportados por Ambientes Virtuais de Ensino e Aprendizagem

O estudo conduzido por Silva e Imran (2015) investiga a utilização de um conjunto de sete variáveis para distinguir estudantes entre concluintes e não concluintes em seus respectivos cursos que fazem uso de Ambientes Virtuais de Ensino e Aprendizagem (AVEA). De modo a verificar a eficácia dessas variáveis, os autores realizaram uma análise estatística comparativa dos dados de 1168 alunos, abrangendo

89 disciplinas em 5 cursos de uma instituição de ensino (SILVA; IMRAN, 2015).

Os resultados obtidos com esta análise evidenciaram diferenças significativas no desempenho de alunos concluintes e não concluintes, sugerindo que as variáveis propostas podem ser empregadas na construção de modelos de previsão de alunos não concluintes. As variáveis de interesse foram selecionadas baseadas na literatura existente e nas características mais comuns dos AVEAs, assim os dados foram coletados a partir das interações dos alunos nas plataformas e dos registros acadêmicos da instituição. Para facilitar a análise, os valores das variáveis foram normalizados em uma escala de 0 a 1, garantindo uma comparação coerente do desempenho dos estudantes.

As conclusões desta pesquisa têm o potencial de identificar fatores importantes que auxiliam na previsão do desempenho dos alunos em cursos que fazem uso de AVEAs, sem depender unicamente de variáveis demográficas. Esse estudo contribui com informações preciosas para educadores e instituições de ensino, enriquecendo a compreensão dos fatores que impactam na desistência de alunos.

Uma limitação do estudo reside no fato de que apenas um número restrito de registros continha dados completos, destacando a necessidade de pesquisas futuras que explorem técnicas de mineração de dados para lidar com a ausência de dados em contextos semelhantes. A falta de dados sobre os demais aspectos do curso pode mostrar uma limitação, onde apenas dados com relação ao desempenho do aluno foram levados em conta.

3 MÉTODO

Neste capítulo são apresentados os métodos de pesquisa utilizados durante o presente estudo de caso. A metodologia utilizada é baseada na integração de dados provenientes de duas fontes distintas, e posteriormente uma análise do conjunto de dados resultante com o objetivo de avaliar indicadores que possam sinalizar a possibilidade de desistência de estudantes dentro da disciplina e obter um modelo analítico responsável por prever estes comportamentos.

Figura 5 – Diagrama do desenvolvimento da metodologia

Fonte: Elaborado pelo autor (2024).

3.1 EXTRAÇÃO DE DADOS

Os dados coletados para a realização das análises deste trabalho são obtidos por meio de dois processos distintos. O primeiro, refere-se a coleta de sentimentos dos estudantes com relação a disciplina, já o segundo método de aquisição de dados é responsável pelas informações referentes ao desempenho do estudante ao longo do semestre, como por exemplo presença, notas e atividades finalizadas.

3.1.1 Turmas analisadas

Na elaboração do estudo deste trabalho, será utilizado dados provenientes de duas turmas de programação de níveis diferentes e de dois semestres distintos. Em ambas as turmas o professor responsável fez a marcação manual dos alunos desistentes e não desistentes.

A turma 1, uma turma de programação 2 do segundo semestre de 2022 que possui 83 alunos matriculados e um total de 26 atividades avaliativas ao decorrer do semestre em questão, sendo que nenhuma atividade foi marcada como importante para ser acompanhar o engajamento dos estudantes. Esta disciplina tem uma carga

horária de 54 horas-aula e é realizada em um encontro semanal. Além disso, foram obtidas 28 respostas nos questionários.

Conforme o relatório de presenças foram anotadas presenças em 26 seções ao longo do semestre, ou seja, 26 dias de aulas nos quais a presença foi aferida pelo professor, no entanto, são duas turmas de programação 2 de dias diferentes que tiveram presenças anotadas em conjunto. Como resultado disso, os alunos obtiveram diversas ausências anotadas mesmo não tendo aula no dia em questão.

A segunda turma em questão, que será titulada de Turma 2, é da disciplina de programação 1 do segundo semestre de 2023, que possui inicialmente 93 alunos matriculados e foram realizadas 25 atividades na totalidade do semestre, assim como na turma 1 nenhuma atividade foi marcada como importante pelo docente responsável para que se possa acompanhar o comprometimento dos estudantes ao longo do semestre. A disciplina tem 72 horas-aula e tem dois encontros semanais. E foram registradas 43 respostas aos questionários neste semestre. Conforme o relatório de presenças foram anotadas presenças em 29 seções ao longo do semestre, ou seja, 29 dias de aulas nos quais a presença foi aferida pelo professor

Estas turmas são oferecidas na Universidade Federal de Santa Catarina (UFSC) no campus Joinville (CTJ). Sendo a Turma 1 composta pelos cursos de Engenharia Aeroespacial, Engenharia Mecatrônica, Ciência e Tecnologia, Engenharia de Transportes e Logística, Engenharia Automotiva e a segunda turma (Turma 2) por todos os anteriores e mais os cursos de Engenharia Ferroviária e Metroviária, Engenharia Civil de Infraestrutura e Engenharia Naval.

3.1.2 Dados de sentimentos

Como descrito por PFITSCHER et al. (2023) no artigo “Análise de sentimentos em turmas de programação com vistas ao apoio à permanência estudantil” foi realizada uma coleta ativa de sentimentos dos estudantes. A aquisição dos dados referentes aos sentimentos dos alunos com relação à disciplina foi realizada por meio de questionários com perguntas abertas, os mesmos elaborados em conjunto com psicopedagogas.

Uma das perguntas é a seguinte: “Como você está se sentindo em relação à disciplina?”. Dessa forma, as respostas desejadas são pequenos textos elaborados livremente pelos estudantes que buscam descrever o seu sentimento perante a disciplina em questão. No entanto, algumas frases podem possuir termos relacionados a sentimentos e outras não, bem como, conter emoções que podem causar confusão e incertezas no momento de identificar o real sentimento do estudante.

Segundo o trabalho de PFITSCHER et al. (2023) após a coleta das respostas, foi realizada manualmente uma classificação psicopedagógica, que empiricamente busca distinguir termos atribuídos aos sentimentos dos demais em cada frase obtida. O primeiro passo desta classificação é a definição de categorias, posteriormente lista-se

os itens sobre aspectos positivos e negativos na experiência com o curso/disciplina utilizando a análise de conteúdo, e em seguida é realizada uma pré-análise flutuante construindo categorias de classificação.

Para automatizar a classificação se faz necessário empregar ferramentas de Processamento de Linguagem Natural (PLN) para análise dos dados, uma vez que a coleta de sentimentos foi realizada por meio de questionários abertos com respostas livres. Neste caso, optou-se por uma arquitetura de rede neural recorrente *Long Short Term Memory* (LSTM) que é capaz de capturar dependências de longo prazo permitindo melhor entendimento do contexto das palavras.

A primeira etapa é o pré-processamento dos dados coletados, que envolve a limpeza de caracteres considerados indesejáveis ou sem informação útil para análise. Em seguida, aplica-se a ferramenta *Tokenizer* para dividir os textos em unidades menores chamadas *tokens*, que são aplicados a uma camada de *Embedding* [1], responsável por mapeá-las em um espaço vetorial numérico de acordo com suas relações semânticas e contexto.

Por último é aplicada uma função de ativação *Softmax* [2] na camada de saída do modelo, essa função é encarregada de converter as saídas do modelo em uma distribuição de probabilidade sobre as classes de classificação. Sendo assim, ao utilizar os modelos treinados, a saída gerada apresenta qual a probabilidade da entrada está associada a uma determinada classificação.

Para o treinamento dos modelos, buscou-se na literatura bases de dados existentes que rotulam palavras, termos, expressões ou frases em negativo e positivo. Foram selecionadas as duas maiores bases em português e que apresentam classificação binária (positiva ou negativa), sendo a primeira baseada em resenhas de filmes do site **IMDb** e a segunda base escolhida foi da companhia **SenticNet**.

Então, os dados coletados foram submetidos aos dois modelos desenvolvidos (um treinado para cada base de dados citada). Com o resultados da respostas de ambos os modelos gerou-se uma terceira classificação com os seguintes critérios:

- Caso os dois modelos acusassem positivo: positivo;
- Caso os dois modelos acusassem negativo: negativo;
- Caso um modelo acusasse positivo e outro negativo: escolheu-se aquele que apresentasse maior porcentagem de probabilidade;
- Caso um modelo acusasse 100% positivo e outro 100% negativo, optou-se por classificar como negativo levando em consideração o pior cenário no contexto de desistência estudantil.

[1] Os *embeddings* representam objetos do mundo real, como palavras, imagens ou vídeos, em uma forma que os computadores podem processar.

[2] É uma função que converte um vetor de K números reais em uma distribuição de probabilidade de K resultados possíveis.

3.1.3 Dados de desempenho

Baseado no estudo de caso realizado por SOUZA (2023), os dados de desempenho do estudante são coletados por meio de três relatórios essenciais dentro do Ambiente Virtual de Aprendizagem do MOODLE. Sendo estes relatórios de: presença, notas e conclusão de atividades estabelecidas como importantes pelo docente, optou-se por estes elementos dentro do ambiente pela facilidade de acesso aos dados e de maneira a evitar restrições de instituições usuárias do MOODLE, tornando o estudo simples e menos burocrático.

3.1.3.1 Relatório de presenças

O AVA MOODLE fornece um *plugin* que permite o gerenciamento das presenças da disciplina, facilitando o controle da assiduidade dos estudantes no sistema pelo professor, e ainda o recurso é customizável pelo docente de acordo com as necessidades. Porém, normalmente a presença é registrada de três formas: pontuação máxima para discentes presentes; pontos intermediários para atrasados ou justificados; pontuação zerada para ausentes e faltantes.

No relatório de presença cada uma das categorias é representada por uma fração, onde no denominador se encontra o valor máximo de pontuação para aquela sessão e no numerador o valor atribuído ao estudante. O relatório é constituído de informações sobre a pontuação classificada temporalmente pelas colunas, de maneira que a primeira coluna indique o primeiro dia de aula e subsequentemente até o último dia de aula. É possível exportar estes dados no formato de planilha eletrônica padrão, como *Microsoft Excel* e *OpenOffice*, ou em formato de texto.

3.1.3.2 Relatório de notas

O relatório de notas é constituído de notas atingidas por meio de avaliações individuais, categorias de avaliação e a nota global do curso. O mesmo não fornece informações sobre datas de realização das atividades, sobre a conclusão de uma atividade não avaliada pelo professor e nem detalha as ponderações associadas a cada atividade definida pelo docente na composição final da nota do curso.

Neste relatório, a ordem das colunas é definida por meio da organização do docente, sendo assim, não possui quaisquer padrões estabelecidos. É possível exportar o relatório de notas no formato de planilha eletrônica padrão *Microsoft Excel* ou *OpenOffice*, ou em formato de texto.

3.1.3.3 Relatório de conclusão de atividades

O MOODLE permite que os administradores do curso marquem atividades como importantes, o que permite enfatizar atividades que devem ser realizadas durante

o semestre pelos estudantes. Dentro de cada atividade marcada como importante, o responsável pode acompanhar o estado da atividade de cada um de seus alunos, onde pode-se estar marcado com a expressão “Concluído” ou “Não concluído"e também a informação da data de conclusão da tarefa.

Por se tratar de um recurso optativo, muitos professores não possuem o hábito de utilizar a ferramenta em suas turmas. No entanto, vale ressaltar a importância destes dados para a avaliação do comportamento do estudante durante o semestre, por meio desta ferramenta pode-se acompanhar o engajamento e comprometimento do aluno para com a matéria em diversos momentos ao longo do semestre, o que torna este um indicador relevante para a previsão de desistência. Da mesma maneira que os relatórios anteriores, este pode ser exportado nos mesmos formatos.

3.2 PRÉ-PROCESSAMENTO DE DADOS

Similar ao procedimento descrito por SOUZA (2023), serão abordados nesta seção os passos utilizados para o pré-processamento de dados deste estudo. Dentro da linguagem de programação *Python*, uma das bibliotecas que são amplamente utilizadas para manipulação de planilhas é a pandas, esta biblioteca converte os arquivos em estruturas denominadas de *DataFrame*, o que auxilia nas etapas de pré-processamento de dados que serão implementadas.

A técnica envolve quatro importantes passos: limpeza, integração, transformação e redução de dados que serão descritos a seguir.

3.2.1 Limpeza de dados

Já que os dados utilizados neste estudo de caso são de cunho pessoal, o primeiro passo desta etapa de limpeza dos dados é a anonimização, sendo assim, atributos utilizados para identificar o aluno são alterados para valores genéricos. Subsequentemente, foram removidos quaisquer dados irrelevantes para o estudo como por exemplo, endereço de *e-mail*, sobrenome, número de matrícula, entre outros atributos similares.

No relatório de presença foram levados em conta apenas os nomes genéricos dos alunos, a quantidade de cada tipo de presença atribuída (ausente, atrasado e ausente), a quantidade de sessões anotadas, a relação de pontos e a porcentagem de presença. Já no relatório de notas são preservados os atributos referentes ao nome genérico do aluno e os valores das notas atribuídas para cada uma das atividades realizadas pelo estudante.

O relatório de conclusão de atividades de ambas as turmas estão anulados, pois o professor não marcou atividades como importantes e os dados não puderam ser levantados. Similar aos dados de desempenho, os dados de sentimentos coletados

foram limpos, no entanto, os dados já são anonimizados e possuem apenas o sentimento expresso pelo estudante.

3.2.2 Integração de dados

Subsequente a limpeza de dados, começou a etapa de integração de dados provenientes de diferentes fontes, no caso deste estudo de caso, os três relatórios extraídos no ambiente MOODLE e as respostas obtidas pelos questionários. Com relação aos relatórios, a informação sobre o nome dos estudantes foi mantida generalizada em todos os arquivos, sendo assim, podemos correlacionar todos os dados de presença, notas e atividades concluídas por indivíduo.

As respostas com relação aos sentimentos dos alunos obtidos por meio dos questionários abertos não totalizam o número de participantes da turma e nem podem ser atrelados diretamente a um indivíduo único. Desta maneira, será adicionado a nossa classe como valores sem atribuição específica e será tratado posteriormente. Com isso, criou-se uma classe *Classroom*[3], que representa uma turma com os seguintes atributos:

- **attendance_report:** Referentes ao relatório de presenças;
- **grade_report:** Referente ao relatório de notas;
- **activity_report:** Referente ao relatório de conclusão de atividades marcadas como importantes pelo professor;
- **feeling_report:**Referente as respostas obtidas dos questionários abertos e classificadas como positivas ou negativas.

3.2.3 Transformação de dados

Da mesma maneira que SOUZA (2023) realizou em seu estudo, para utilizar as técnicas de aprendizado de máquina com os dados obtidos, se fez necessária uma etapa de transformação de dados em variáveis e classificações que possam ser interpretadas pelos algoritmos. As informações referentes a presença de cada estudante foram organizadas em quatro listas, que é um tipo de estrutura de dados em *Python*:

- **progression:** Pontuação máxima 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.

Com relação ao relatório de notas e de conclusão de atividades, SOUZA (2023) criou dicionários, que é uma estrutura de dados utilizada em *Python*, onde o nome da

[3] Os dados e códigos utilizados no desenvolvimento do presente estudo podem ser encontrados em https://github.com/rodrigo-nfb/TCC.git

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;
- **completed:** Se a atividade foi realizada pelo aluno, ou não;
- **highest_grade:** Nota mais alta da turma, em determinada atividade;
- **completion_rate:** Taxa de conclusão da atividade, pela turma.

Para o relatório de conclusão de atividades, somente dois atributos são especificados por 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.

E por fim, um valor é especificado com relação ao sentimento da turma. Este valor terá seu método de cálculo estudado mais adiante.

- **class_feeling:** É o sentimento da turma com relação a matéria obtido por meio da classificação dos sentimentos coletados pelos questionários.

Assim, os relatórios extraídos de diferentes formas foram transformados em estruturas de dados do *Python*, o que viabiliza o acesso e interpretação dos dados pela linguagem de programação.

3.2.4 Redução de dados

Visando uma representação mais sucinta dos dados que representam cada um dos estudantes ou o sentimento da turma, converteu-se os dados transformados em atributos individuais do aluno ou coletivo da turma com valores numéricos, na expectativa de auxiliar na análise de desistentes. Assim como descrito por SOUZA (2023) em seu trabalho, esta abordagem permite a generalização dos modelos desenvolvidos com o intuito de serem aplicados em contextos distintos.

Para alcançar este fim, foram elaborados 23 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;
- **class_feeling:** é o sentimento da turma com relação a matéria obtido por meio da classificação dos sentimentos coletados pelos questionários. Podendo ser atribuído segundo um cálculo binário, porcentagem positiva ou ponderado com outras pontuações.

Com isso, o conjunto de dados que antes era dividido em quatro relatórios e estruturas de dados diferentes, é reduzido à variáveis numéricas para cada um dos estudantes.

3.3 ESTRUTURAÇÃO DA ANÁLISE

Esta seção formula a construção dos passos necessários para a elaboração dos modelos de análise de dados de modo explícito e implícito, utilizados na identificação de estudantes desistentes da turma. Para a elaboração deste modelos foram criados sistemas de pontuações para cada um dos relatórios levados em consideração nestes estudo de caso.

3.3.1 Sistema de pontuação de notas, presença, conclusão de atividades importantes e sentimento da turma com relação a matéria

Cada um dos relatórios de desempenho e o relatório de sentimento precisam ser transformados em uma nota que pode ser manipulada e comparada. Para isso, criou-se sistemas de pontuação baseados nas informações presentes nestes relatórios.

3.3.1.1 Pontuação de presença

Para o cálculo da pontuação referente a presença do discente, é levado em conta primeiramente o peso para cada uma das aulas anotadas, de maneira que em dias com um maior número de alunos presentes, represente um peso maior para a pontuação final da presença. Este método tem o propósito de mitigar efeitos externos, como por exemplo efeitos climáticos adversos ou então eventos concomitantes às aulas que possam dificultar a presença.

$$W_i = \frac{Me_i}{Max_i} \tag{5}$$

Conforme mostrado na na Equação 5, o peso W atribuído a cada dia i é obtido pela divisão da média das pontuações do MOODLE da classe para aquele dia pela pontuação máxima definida pelo responsável da matéria dentro do ambiente do MOODLE. Sendo assim, pode-se calcular a presença computada para a turma para cada um dos dias anotados de acordo com a equação a seguir.

$$P_f = \sum_{i=1}^{d} \frac{Po_i \times W_i}{Max_i} \tag{6}$$

Onde na Equação 6, o termo P_f corresponde ao valor da pontuação final, Po_i é a pontuação atribuída pelo MOODLE e d é o número total de dias anotados no relatório de presença.

3.3.1.2 Pontuação de notas

O cálculo da pontuação das notas leva em consideração a média da turma na avaliação em questão, caso o estudante obteve uma nota relativamente mais baixa

com relação a média da turma, isso reduz a pontuação final mais do que em avaliações em que a média da turma foi mais baixa. Esse recurso é utilizado para ponderar e acompanhar o desempenho do aluno conforme o resultado geral da turma.

O cálculo baseia-se em considerar a ponderação de cada nota individualmente de uma atividade, relacionando-a com a maior nota obtida na turma para esta avaliação e multiplicando pelo valor médio das notas geral da turma para a mesma avaliação. Soma-se os valores de todas as avaliações, e em seguida o total é subtraído da pontuação máxima possível é dividido pelo valor máximo possível também, como mostrado na Equação 7 a seguir.

$$P_f^{'} = \frac{G_{max} - \sum_{k=1}^{n}(1 - \frac{G_k}{A_k}) \times M_k}{G_{max}} \times 100 \tag{7}$$

Onde, $P_f^{'}$ é atribuído a pontuação final das notas, G_{max} é a maior soma de notas possível, G_k representa à nota individual para a atividade k, a maior nota obtida entre a classe na atividade é representada por A_k, a média das notas da classe na atividade é simbolizada por M_k e n é a quantidade total de atividades avaliativas realizadas ao decorrer do semestre.

3.3.1.3 Pontuação de conclusão de atividades

A construção da pontuação de conclusão de atividades segue um raciocínio similar às demais, onde cada atividade não realizada pelo estudante resulta na redução de sua pontuação. A formulação da pontuação considera a proporção de alunos que concluíram a atividade na turma, sendo assim, a não conclusão de uma atividade na qual a maioria da turma concluiu tem um impacto maior na pontuação final do que uma atividade que teve uma proporção menor de conclusão.

$$P_f^{''} = \frac{Po_{max} - \sum_{j=1}^{n}(W_j^{'} \times t_j)}{Po_{max}} \times 100, \;\; com \;\; W_j^{'} = \begin{cases} 0 & \text{atividade j não concluída} \\ 1 & \text{concluída e não importante} \\ 1,4 & \text{concluída e importante} \end{cases} \tag{8}$$

A Equação 8 expõe o fórmula do cálculo da pontuação, assim $P_f^{i'}$ representa a pontuação final de atividades concluídas, Po_{max} indica a soma das pontuações máximas de todas as atividades, $W_j^{'}$ é o peso dependente da conclusão da atividade j, t_j representa a taxa de conclusão da atividade pela turma e n é o número de atividades totais realizadas, essa fórmula é inversamente proporcional, ou seja, o aluno que concluiu todas as atividades pontua 0.

3.3.1.4 Pontuação com base no sentimento da turma

O quarto aspecto a ser estudado e o principal enfoque deste trabalho é a pontuação construída a partir dos dados coletados sobre os sentimentos dos estudantes com relação a disciplina. Diferentemente das pontuações anteriores, a pontuação de sentimentos exige a conversão de dados qualitativos para uma forma de escala similar ao adotado nas demais pontuações. Os sentimentos anteriormente foram classificados entre positivos e negativos.

Na Tabela 1 pode-se observar um exemplo de classificação das respostas obtidas, onde a classificação *"Positivo"* significa que o sentimento descrito pelo estudante com relação a disciplina é tido como otimista e negativo como pessimista.

Tabela 1 – Exemplo de resultados da classificação de sentimentos

Respostas	Classificação
Resposta 1	Positivo
Resposta 2	Negativo
Resposta 3	Positivo
Resposta 4	Positivo
...	...
Resposta n	Negativo

A partir disso, serão estudadas três abordagens distintas de como transformar estas classificações dos sentimentos em um atributo que possa ser utilizado no aprendizado de máquina para a previsão de desistentes.

3.3.1.4.1 Atribuição binária

O primeiro método abordado é a distribuição de um valor binário de acordo com o sentimento geral das respostas obtidas na turma. São contabilizadas as respostas obtidas e suas classificações, o sentimento que predomina, ou seja, o sentimento mais recorrente dentre todos na turma é utilizado no atributo de sentimento, como mostrado na equação abaixo.

$$P'''_{f_1} = \beta \times 100, \;\; com \;\; \beta = \begin{cases} 0 & \text{Negativo predominante} \\ 1 & \text{Positivo predominante} \end{cases} \tag{9}$$

Conforme a Equação 9 mostra, a pontuação atribuída será de 0 em caso do sentimento negativo predominar nas respostas (mais classificações negativas do que positivas nas respostas) e o valor de 1 em caso de sentimento predominantemente otimista dos estudantes da turma.

3.3.1.4.2 Porcentagem otimista

Outra abordagem estudada é a atribuição de um valor numérico que representa a porcentagem de alunos otimistas com a disciplina dentro da turma. Similarmente ao método binário é contabilizado a quantidade de respostas com registros de sentimentos positivos dos estudantes, e então é convertido em uma porcentagem com relação ao número de respostas obtidas.

$$P'''_{f_2} = \frac{n_{positivos}}{n_{respostas}} \times 100 \tag{10}$$

Na Equação 10 temos o valor da pontuação P'''_{f_2} que é resultado da divisão da quantidade de respostas positivas ($n_{positivos}$) pelo número total de respostas obtidas por meio dos questionários ($n_{respostas}$).

3.3.1.4.3 Ponderação baseado nas demais pontuações

O último método empregado é um passo subsequente ao método denominado de porcentagem de otimismo. Assim, neste método a porcentagem obtida pelo método anterior terá seu peso modificado de acordo com as demais pontuações adquiridas pelo estudante (presença, notas e conclusão de atividades), este recurso visa diferenciar os valores atribuídos ao sentimento dentre os participantes da mesma turma.

Com isso, quanto maior for a relação entre a soma das pontuações anteriores e a soma máxima obtida na turma (o que indica que o aluno está indo melhor no semestre), maior será o peso atribuído ao otimismo do estudante. Seguindo esta mesma lógica, quanto menor a relação entre a pontuação somada do estudante e a pontuação máxima da turma, menor será o sentimento de otimismo atribuído ao estudante em questão.

$$P'''_{f_3 i} = \frac{(P_{fi} + P'_{fi} + P''_{f\,i})}{S_{max}} \times P'''_{f_2} \tag{11}$$

Na Equação 11 pode-se observar o cálculo empregado para a criação da pontuação de sentimento segundo o método de ponderação, assim, o resultado da pontuação para o aluno i é dado por $P'''_{f_3 i}$, que calcula-se pela soma das pontuações de presença (P_{fi}), notas (P'_{fi}) e de conclusão de atividades ($P''_f i$), dividido pelo somatório máximo atingido pela turma $Smax$ e multiplicado pela porcentagem de otimismo da turma com relação a matéria P'''_{f_2}.

3.3.2 Análise Explícita

A construção dos passos necessários para a elaboração do modelo de análise de dados de modo explícito para identificação de alunos desistentes da disciplina

é descrito abaixo. O classificador gerado foi implementado em *Python* e levou em consideração um sistema de ranqueamento dos alunos baseado em seu desempenho ao longo do semestre.

3.3.2.1 Sistema de pontuação com notas, presença, conclusão de atividades importantes e sentimento da turma com relação a matéria

Para distinguir alunos desistentes dos não desistentes é proposto um sistema de pontuação e ranking, o qual realiza a classificação dos estudantes em categorias. A pontuação atribuída aos estudantes é baseada nas quatro categorias de dados que foram descritas por este trabalho (presença, notas, conclusão de atividades e sentimento da turma), sendo que cada uma delas compõem uma porcentagem dessa pontuação.

A ponderação dos quatro aspectos busca uma avaliação abrangente do aluno, que por quaisquer motivos pode vir a receber uma pontuação baixa em um dos quesitos, no entanto, mesmo assim não desistiu da disciplina. No primeiro momento, as ponderações dos aspectos vai ser igualitária, sendo 1/4 do peso total para cada, e a pontuação final deve ser um valor dentro da escala de 0 a 100, estes valores foram definidos empiricamente e baseados no trabalho desenvolvido por SOUZA (2023).

3.3.2.2 Algoritmo de classificação

Com os critérios de pontuação de cada um dos relatórios bem definidos, pode-se estabelecer a pontuação final de cada um dos estudantes, que é calculada com a média aritmética simples dos resultados obtidos na avaliação de presença, notas, conclusão de atividades e sentimento da turma. Com a obtenção das pontuações individuais finais dos estudantes cria-se um ranking ordenado, assim, o aluno com maior pontuação é posicionado em primeiro lugar e subsequentemente.

Com o propósito de classificar estudantes desistentes e não-desistentes da disciplina na turma é definido um limite percentual para distinguir as categorias. Se faz necessário a definição de uma porcentagem de corte, de modo que alunos piores classificados e que estejam abaixo do limite de corte sejam considerados desistentes pelo modelo. Alguns dos critérios possíveis de serem utilizados são a quantidade de alunos que o professor espera conseguir dedicar atenção especial para evitar a desistência, ou o embasamento na desistência histórica da disciplina.

3.3.3 Análise Implícita

A elaboração do método de análise de modo implícito de dados para identificação de estudantes desistentes da turma é descrita a seguir. O classificador gerado foi implementado em *Python* utilizando uma biblioteca de ferramentas de

visualização de dados, aprendizado de máquina e mineração de dados de código aberto chamado de *Orange*, a biblioteca possui uma ampla gama de ferramentas que realizam preprocessamento de dados, testes, avaliações e visualização de resultados.

Para a utilização do modelo implícito foram usados os atributos discutidos e listados na Seção 3.2.4, sendo que o atributo que descreve o sentimento da turma é atribuído por meio de um dos 3 métodos de pontuação descritos anteriormente (binário, porcentagem positiva e ponderação com demais notas).

3.3.3.1 Algoritmo de classificação

Na elaboração do modelo de classificação, foram empregados diferentes técnicas de aprendizado de máquina: Árvore de decisão (AD), *K-Nearest Neighbors* (KNN), *Random Forest* (RF), *Naive Bayes* (NB), *Multilayer Perceptron* (MLP), Regressão Logística (RL), *Gradient Boosting* (GB) e *Support Vector Machine* (SVM). As escolhas de tais técnicas se deu pela disponibilidade e a possibilidade de comparação dos resultados alcançados no presente trabalho e os resultados apresentados no trabalho desenvolvido por SOUZA (2023), presentes na Seção 2.7.

Com o intuito de testar as diferentes técnicas de aprendizado de máquina e seus resultados, utilizou-se três métodos de divisão de dados em treinamento (parcela dos dados que os modelos utilizarão para treinar) e teste (parcela dos dados em que o modelo será aplicado e avaliado). O método de amostragem aleatória de dados (sendo 67% dos dados para treinamento e 33%para testes); o método *holdout* (67% dos dados para treinar e 33% para testar); e o método de validação cruzada 10-*fold* ($k = 10$).

Para a análise implícita optou-se por unir as duas turmas em um grande grupo de dados, desta maneira, o modelo tem uma variedade maior de dados para o treinamento. Por causa da seleção aleatória dos estudantes para os conjuntos de dados, a técnica de amostragem aleatória foi implementada dez vezes em cada análise, tal procedimento objetiva melhorar a capacidade de representação da base de dados, buscando minimizar as variações naturais à seleção aleatória e promovendo uma amostra mais fidedigna da turma em questão. No método de validação cruzada 10-*fold*, foram definidas 10 divisões dos dados.

Tanto na amostragem aleatória quanto na validação cruzada, não é necessário dividir os dados manualmente, basta fornecê-los aos modelos. No entanto, ao utilizar o método *holdout* para dividir os dados, foi necessário assegurar a representatividade dos comportamentos de desistência em ambos os conjuntos. Assim, além de dividir os dados manualmente na proporção de 67% para treinamento e 33% para teste, garantiu-se que a proporção de alunos classificados como desistentes fosse equivalente nos dois grupos.

3.4 CONSIDERAÇÕES DO CAPÍTULO

Neste capítulo, foram descritos os métodos de pesquisa utilizados, com base na integração de dados provenientes de duas fontes distintas. Inicialmente, os dados são coletados a partir dos sentimentos dos estudantes em relação à disciplina e do desempenho dos alunos ao longo do semestre, incluindo presença, notas e atividades concluídas. A análise é realizada com o objetivo de avaliar indicadores que possam sinalizar a possibilidade de desistência dos estudantes e desenvolver um modelo analítico para prever esses comportamentos.

A extração de dados é detalhada, incluindo a coleta de informações de duas turmas de programação de níveis e semestres diferentes. Em ambas as turmas, o professor responsável registrou manualmente os alunos desistentes e não desistentes. Os dados de presença, notas, atividades concluídas e sentimentos dos estudantes foram anonimizados e organizados para posterior análise. O pré-processamento dos dados envolveu limpeza, integração, transformação e redução, garantindo a preparação adequada dos dados para a análise.

A estrutura da análise envolveu a criação de pontuações que representam cada um dos aspectos descritos pelos relatórios do estudantes. Elaborou-se um modelo de análise explícita e outro de análise implícita, sendo que o primeiro é um sistema de pontuação que considerou a presença, notas, conclusão de atividades e sentimentos dos alunos. A pontuação foi calculada com base na média da turma e na conclusão das atividades, ponderando a importância de cada aspecto.

Já para o segundo modelo, análise implícita, foi criado um modelo de aprendizado de máquina utilizando os atributos criados para descrever o estudante e seu desempenho. Em ambos os casos as respostas aos questionários sobre sentimentos foram tratadas de forma a não identificar os estudantes individualmente, mas contribuindo para a análise geral.

4 ANÁLISE DE DADOS

Para avaliar os modelos desenvolvidos neste trabalho, calculou-se as métricas de sensibilidade, precisão e acurácia a partir da comparação entre os resultados das previsões dos modelos e a classificação real dos estudantes, além disso, pode-se comparar os resultados obtidos pelo classificador explícito com as técnicas de aprendizado de máquina. Os dados de turmas reais são disponibilizados pelo docente encarregado da disciplina, que classifica manualmente os alunos em desistentes ou não desistentes.

Com a finalidade de avaliar a aplicabilidade prática do modelo na previsão da desistência escolar, investigou-se quatro diferentes proporções do período letivo em duas turmas distintas. Embora o uso de 100% dos dados não seja factível para intervenções precoces, ele serve como referência para a melhor métrica possível. Para simular situações letivas reais, foram adotadas porcentagem de 75% e 50% dos dados, com o objetivo de medir o desempenho da classificação em momentos críticos do curso. Além disso, uma análise com apenas 25% dos dados foi realizada para testar a capacidade do sistema de identificar sinais precoces de desistência.

4.1 ANÁLISE EXPLÍCITA

Nesta seção, serão apresentados os resultados obtidos por meio da análise explícita do conjunto de dados de cada turma separadamente, utilizando o sistema de pontuação com notas, presença, conclusão de atividades e sentimentos dos estudantes nas turmas analisadas, com uma comparação dentre diferentes modelos de cálculo do atributo referente ao sentimento da turma.

Com os dados disponibilizados pelo docente responsável pela disciplina, foi possível verificar que a matéria tem em média 33% de estudantes desistentes em ambas as turmas analisadas. Sendo assim, a porcentagem de corte para a análise explícita foi definida com este valor. As métricas de sensibilidade, precisão, acurácia e pontuação F1 obtidas pelos resultados da classificação de alunos da Turma 1 podem ser verificadas na Tabela 2.

A Tabela 2 expõe a comparação entre os métodos de cálculos descritos na Seção 3.3.1.4, onde o modelo denominado de ’Sem’ refere-se ao método considerando apenas notas, atividades concluídas e presença (sem considerar sentimentos da turma); o modelo chamado de ’Binário’ considera os sentimentos da turma pelo cálculo binário; o modelo ’%’ leva em consideração o método de porcentagem positiva para atribuir valor ao atributo do sentimento; e o modelo denominado ’Ponderar’ considera o sentimento da turma sobre as métricas passivas de cada indivíduo. Pode-se observar

que os valores das métricas não tiverem mudanças expressivas entre os modelos.

Tabela 2 – Comparação dos resultados das métricas da Turma 1

Modelo	Proporção do Semestre	Sensibilidade	Precisão	Pontuação F1	Acurácia
Sem	100%	89,29%	89,29%	89,29%	92,77%
	75%	88,57%	88,57%	88,57%	92,29%
	50%	**89,64**%	89,64%	89,64%	93,01%
	25%	**82,50**%	82,50%	82,50%	88,19%
Binário	100%	89,29%	89,29%	89,29%	92,77%
	75%	**89,64**%	89,64%	89,64%	93,01%
	50%	86,43%	86,43%	86,43%	90,84%
	25%	80,00%	79,72%	79,86%	86,39%
%	100%	89,29%	89,29%	89,29%	92,77%
	75%	88,57%	88,57%	88,57%	92,29%
	50%	85,00%	85,00%	85,00%	89,88%
	25%	80,36%	80,07%	80,21%	86,63%
Ponderar	100%	89,29%	89,29%	89,29%	92,77%
	75%	87,86%	87,86%	87,86%	91,81%
	50%	85,71%	85,71%	85,71%	90,36%
	25%	79,29%	77,89%	78,58%	85,42%

Ao observar a Tabela 2 pode-se concluir que para uma análise de 100% dos dados todos os modelos obtiveram os mesmos valores para as métricas. Somente a partir de uma análise igual ou inferior a 75% dos dados temos uma diferença perceptível entre os modelos, sendo que o modelo considerando os sentimentos pelo método binário obteve as melhores métricas para 75% dos dados e nas demais proporções o método sem considerar sentimentos foi o melhor.

O melhor resultado encontrado na análise explícita da Turma 1 foi de 89,64% de sensibilidade, o que indica que 25 dos 28 estudantes desistentes foram detectados pelo modelo desenvolvido. Um bom resultado, no entanto, sem tanta eficácia do novo atributo relacionado ao sentimento da turma. Essa métrica foi alcançada por dois modelos, o primeiro foi o método ’Sem’ com 50% dos dados do semestre, e o segundo o método ’Binário’ utilizando 75% dos dados disponíveis.

As métricas de sensibilidade, precisão, acurácia e pontuação F1 obtidas pelos resultados da classificação de alunos da Turma 2 podem ser verificadas na Tabela 3. Observa-se que os valores das métricas não tiverem mudanças expressivas entre os modelos, assim como na Turma 1.

Com a Tabela 3 pode-se aferir que para uma análise de 100% dos dados todos os modelos obtiveram os mesmos valores para as métricas assim como para a turma 1. Somente a partir de uma análise igual ou inferior a 75% dos dados temos uma diferença perceptível entre os modelos, sendo que os modelos considerando os sentimentos pelo método de porcentagem positiva e a ponderação obtiveram as

melhores métricas para 75% dos dados e nas demais proporções o método sem considerar sentimentos foi o melhor.

Tabela 3 – Comparação dos resultados das métricas da Turma 2

Modelo	Proporção do Semestre	Sensibilidade	Precisão	Pontuação F1	Acurácia
Sem	100%	32,14%	29,03%	30,51%	55,91%
	75%	28,57%	25,81%	27,12%	53,76%
	50%	**28,93**%	26,05%	27,41%	53,87%
	25%	**29,64**%	26,77%	28,14%	54,41%
Binário	100%	32,14%	29,03%	30,51%	55,91%
	75%	28,57%	25,81%	27,12%	53,76%
	50%	28,93%	26,12%	27,46%	53,98%
	25%	28,57%	25,72%	27,07%	53,66%
%	100%	32,14%	29,03%	30,51%	55,91%
	75%	**28,93**%	26,12%	27,46%	53,98%
	50%	28,21%	25,48%	26,78%	53,55%
	25%	28,21%	25,24%	26,64%	53,23%
Ponderar	100%	32,14%	29,03%	30,51%	55,91%
	75%	**28,93**%	26,13%	27,46%	53,98%
	50%	28,57%	25,81%	27,12%	53,76%
	25%	28,57%	25,81%	27,12%	53,76%

O melhor resultado encontrado na análise explícita da Turma 2 foi de 32,14% de sensibilidade, o que indica que cerca de 5 dos 17 estudantes desistentes foram detectados pelo modelo desenvolvido. Um resultado não satisfatório, uma vez que muitos alunos desistentes foram considerados como não desistentes (falso negativo). Essa métrica foi alcançada por todos os modelos utilizando 100% dos dados disponíveis do semestre.

Avaliando o valor da sensibilidade, para considerar que mais um estudante desistente foi detectado na Turma 1 pelo modelo é necessário um aumento de 3,56% na sensibilidade registrada. Já para a turma 2 este valor deve ser de 5,88%, desta forma, para considerar que o modelo foi capaz de prever um estudante a mais ele precisa registrar um aumento de 5,88% de sensibilidade.

Em ambas as turmas não foi possível notar ganho significativo com a adição do novo atributo relacionado ao sentimento da turma, em alguns casos até piorou a performance (principalmente para 25% dos dados). Uma vez que o atributo é igual em dois dos três modelos para todos os alunos (modelo binário e porcentagem positiva), é notável que a adição aritmética simples não resultaria em um ganho de informação, mas devido a arredondamentos pode prejudicar o algoritmo.

Nota-se que quanto maior é a proporção do semestre analisado, maiores são os valores de métricas obtidos. Em contra partida, o tempo disponível para o docente intervir e tentar ajudar o estudantes a reverter este quadro é menor.

4.2 ANÁLISE IMPLÍCITA

Para avaliar e comparar os resultados obtidos pela análise explícita e outros trabalhos similares, foi realizada uma análise usando os 23 atributos apresentados na Seção 3.2.4. As turmas citadas na Seção 3.1.1, foram agrupadas em uma única turma e analisadas em conjunto, e então três técnicas diferentes de divisão de dados foram empregadas para avaliar os resultados dos classificadores tendo como métricas de avaliação a acurácia, a sensibilidade, a precisão e a pontuação F1.

As técnicas utilizadas para divisão de dados foram o 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 treino (SA); e validação cruzada com 10 divisões (VC). A seleção de características dos dados foi feita com os cinco e dez melhores atributos escolhidos por quatro técnicas de avaliação: ANOVA, Chi2, GI e RG.

4.2.1 Seleção de atributos

Utilizando os métodos de avaliação de atributos, foi possível identificar as características mais importantes para a análise realizada. A Figura 6 mostra os 10 atributos com a pontuação mais alta, considerando 100% dos dados em cada método de seleção de características para a turma sem considerar os dados de sentimentos dos estudantes, assim como SOUZA (2023) realizou em seu estudo.

Figura 6 – Seleção de atributos para a turma sem considerar os sentimentos

Rank	Ganho de Informação	Relação de Ganho	Chi2	ANOVA
1	grades_between_7_5_10	sequencial_missing_22	grades_below_5	grades_below_5
2	grades_below_5	sequencial_missing_11	grades_average	grades_between_7_5_10
3	grades_average	activities_complete_majority	grades_between_0_2_5	grades_average
4	grades_between_0_2_5	activities_incomplete_majority	grades_between_7_5_10	grades_between_0_2_5
5	activities_complete_minority	grades_between_7_5_10	grades_below_mean	grades_below_mean
6	activities_incomplete	grades_below_5	important_grades_below_mean	important_grades_below_mean
7	grades_below_mean	grades_average	important_activities_complete_majority	important_activities_complete_majority
8	activities_incomplete_minority	grades_between_0_2_5	important_activities_complete_minority	important_activities_complete_minority
9	activities_complete_majority	activities_complete_minority	important_activities_incomplete	important_activities_incomplete
10	activities_incomplete_majority	activities_incomplete	important_activities_incomplete_majority	important_activities_incomplete_majority

Fonte: Elaborado pelo autor (2024).

Os atributos foram identificados com diferentes cores para facilitar a visualização e comparação entre diferentes métodos. Nota-se que *grades_between_7_5_10* é o único atributo entre os cinco principais em todos os casos. A categoria de atributos mais frequente entre os 10 mais importantes é a de notas, representando 52,5% do total, seguida por atividades concluídas com 42,5%, e presença com 5%.

Após a avaliação utilizando apenas dados sobre notas, atividades concluídas

e presença dos estudantes, incluiu-se o atributo calculado referente aos dados dos sentimentos dos estudantes, e então realizou-se uma nova análise dos atributos, similar a anterior, para investigar os diferentes métodos de seleção de atributos. Variou-se a maneira na qual o atributo é calculado segundo a Seção 3.3.1.4. O intuito é investigar o impacto do atributo para a previsão dos estudantes desistentes.

Nas Figuras 7, 8 e 9 pode-se observar os resultados do ranking dos 10 melhores atributos obtidos por meio de diferentes formas de cálculo do atributo *class_feeling*. Nota-se que os resultados obtidos por meio do método da atribuição binária e da porcentagem positiva são idênticos, ambos os resultados combinaram no mesmo ranking de atributos. O novo atributo criado para descrever os sentimentos dos estudantes ficou em 3º lugar no método de seleção de relação de ganho e em 9° lugar no ganho de informação.

Figura 7 – Seleção de atributos para a turma utilizando atribuição binária

Rank	Ganho de Informação	Relação de Ganho	Chi2	ANOVA
1	grades_between_7_5_10	sequencial_missing_22	grades_below_5	grades_below_5
2	grades_below_5	sequencial_missing_11	grades_average	grades_between_7_5_10
3	grades_average	class_feeling	grades_between_0_2_5	grades_average
4	grades_between_0_2_5	activities_complete_majority	grades_between_7_5_10	grades_between_0_2_5
5	activities_complete_minority	activities_incomplete_majority	grades_below_mean	grades_below_mean
6	activities_incomplete	grades_between_7_5_10	important_grades_below_mean	important_grades_below_mean
7	grades_below_mean	grades_below_5	important_activities_complete_majority	important_activities_complete_majority
8	activities_incomplete_minority	grades_average	important_activities_complete_minority	important_activities_complete_minority
9	class_feeling	grades_between_0_2_5	important_activities_incomplete	important_activities_incomplete
10	activities_complete_majority	activities_complete_minority	important_activities_incomplete_majority	important_activities_incomplete_majority

Fonte: Elaborado pelo autor (2024).

Figura 8 – Seleção de atributos para a turma utilizando porcentagem positiva

Rank	Ganho de Informação	Relação de Ganho	Chi2	ANOVA
1	grades_between_7_5_10	sequencial_missing_22	grades_below_5	grades_below_5
2	grades_below_5	sequencial_missing_11	grades_average	grades_between_7_5_10
3	grades_average	class_feeling	grades_between_0_2_5	grades_average
4	grades_between_0_2_5	activities_complete_majority	grades_between_7_5_10	grades_between_0_2_5
5	activities_complete_minority	activities_incomplete_majority	grades_below_mean	grades_below_mean
6	activities_incomplete	grades_between_7_5_10	important_grades_below_mean	important_grades_below_mean
7	grades_below_mean	grades_below_5	important_activities_complete_majority	important_activities_complete_majority
8	activities_incomplete_minority	grades_average	important_activities_complete_minority	important_activities_complete_minority
9	class_feeling	grades_between_0_2_5	important_activities_incomplete	important_activities_incomplete
10	activities_complete_majority	activities_complete_minority	important_activities_incomplete_majority	important_activities_incomplete_majority

Fonte: Elaborado pelo autor (2024).

Já para o método de cálculo baseado em ponderação com as demais pontuações resultou em um ranking diferente, e uma melhor classificação do atributo *class_feeling*. O novo atributo criado para descrever os sentimentos dos estudantes ficou em 3º lugar no método de seleção de relação de ganho, similar aos cálculos anteriores, no entanto, ficou em 1° lugar no ganho de informação.

Figura 9 – Seleção de atributos para a turma utilizando ponderação baseada nas demais pontuações

Rank	Ganho de Informação	Relação de Ganho	Chi2	ANOVA
1	class_feeling	sequencial_missing_22	grades_below_5	grades_below_5
2	grades_between_7_5_10	sequencial_missing_11	grades_average	grades_between_7_5_10
3	grades_below_5	class_feeling	grades_between_0_2_5	grades_average
4	grades_average	activities_complete_majority	grades_between_7_5_10	grades_between_0_2_5
5	grades_between_0_2_5	activities_incomplete_majority	grades_below_mean	grades_below_mean
6	activities_complete_minority	grades_between_7_5_10	important_grades_below_mean	important_grades_below_mean
7	activities_incomplete	grades_below_5	important_activities_complete_majority	important_activities_complete_majority
8	grades_below_mean	grades_average	important_activities_complete_minority	important_activities_complete_minority
9	activities_incomplete_minority	grades_between_0_2_5	important_activities_incomplete	important_activities_incomplete
10	activities_complete_majority	activities_complete_minority	important_activities_incomplete_majority	important_activities_incomplete_majority

Fonte: Elaborado pelo autor (2024).

Com estes resultados pode-se constatar que os métodos de classificação de atributos: ganho de informação e relação de ganho, consideram o atributo *class_feeling* relevante para a determinação dos estudantes desistentes, especialmente no caso da ponderação baseada nas demais pontuações.

4.2.2 Integrando algoritmos de aprendizado de máquina, técnicas de divisão de dados e seleção de atributos

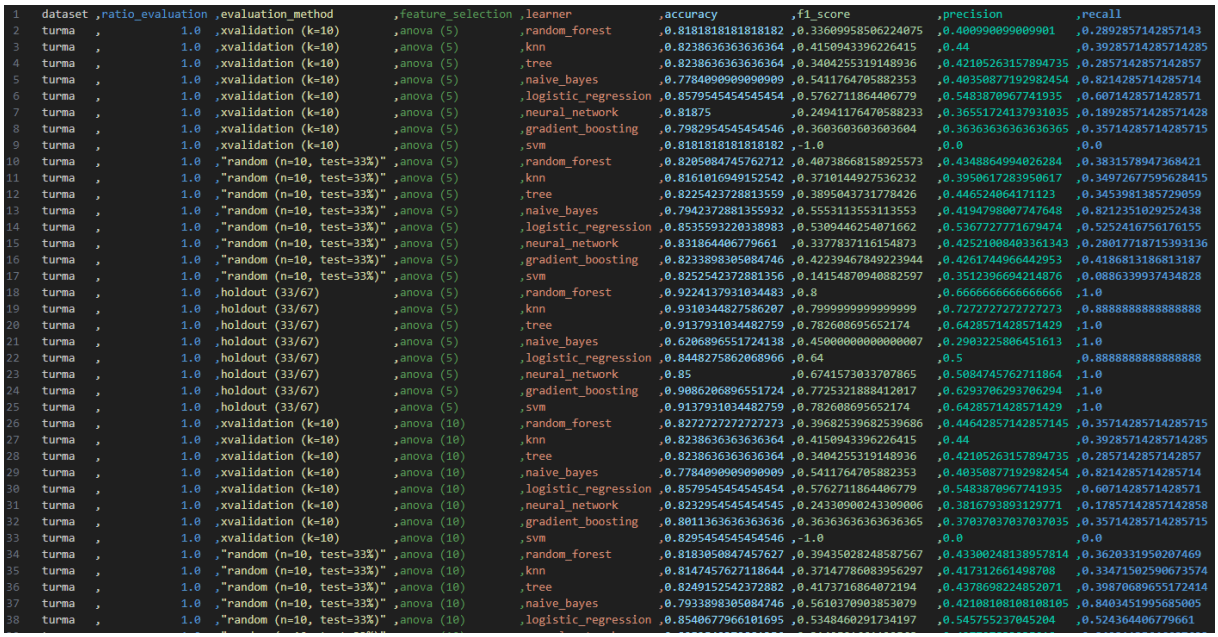
Ao longo deste trabalho, empregou-se oito algoritmos diferentes de aprendizado de máquina que são compostos por AD, KNN, RF, NB, MLP, RL, GB e SVM. E foram combinados com quatro diferentes métodos de avaliação de atributos (GI, RG, Chi2 e ANOVA) e aplicados tanto com a seleção de cinco e dez atributos quanto sem nenhuma seleção de características.

Além disso, utilizou-se três técnicas de divisão do conjunto de dados que geraram um total de 216 resultados diferentes com quatro métricas de avaliação (sensibilidade, precisão, acurácia e pontuação F1) calculadas. A Figura 10 ilustra uma parte da planilha gerada a partir das combinações resultantes, para ambas as turmas analisadas em conjunto e para cada proporção do conjunto total de dados (25%, 50%, 75% e 100%).

4.2.3 Métricas e discussões

As principais combinações obtidas para cada divisão do banco de dados foram apresentadas e discutidas, optou-se por realizar a análises com 100%, 75%, 50% e 25% do conjunto de dados da turma (combinação da Turma 1 e Turma 2) de maneira separada. As métricas para cada proporção de dados foram primeiramente organizadas em função da sensibilidade, métrica de maior interesse do presente trabalho, e subsequentemente, em função da precisão.

Figura 10 – Planilha parcial criada a partir das métricas geradas



dataset	ratio_evaluation	evaluation_method	feature_selection	learner	accuracy	f1_score	precision	recall
turma	1.0	xvalidation (k=10)	anova (5)	random_forest	0.8181818181818182	0.33609958506224075	0.400990099009901	0.2892857142857143
turma	1.0	xvalidation (k=10)	anova (5)	knn	0.8238636363636364	0.4150943396226415	0.44	0.3928571428571428
turma	1.0	xvalidation (k=10)	anova (5)	tree	0.8238636363636364	0.3404255319148936	0.42105263157894735	0.2857142857142857
turma	1.0	xvalidation (k=10)	anova (5)	naive_bayes	0.7784090909090909	0.5411764705882353	0.40350877192982454	0.8214285714285714
turma	1.0	xvalidation (k=10)	anova (5)	logistic_regression	0.8579545454545454	0.5762711864406779	0.5483870967741935	0.6071428571428571
turma	1.0	xvalidation (k=10)	anova (5)	neural_network	0.81875	0.24941176470588233	0.36551724137931035	0.18928571428571428
turma	1.0	xvalidation (k=10)	anova (5)	gradient_boosting	0.7982954545454546	0.3603603603603604	0.36363636363636365	0.3571428571428571
turma	1.0	xvalidation (k=10)	anova (5)	svm	0.8181818181818182	-1.0	0.0	0.0
turma	1.0	"random (n=10, test=33%)"	anova (5)	random_forest	0.8205084745762712	0.40738668158925573	0.4348864994026284	0.3831578947368421
turma	1.0	"random (n=10, test=33%)"	anova (5)	knn	0.8161016949152542	0.3710144927536232	0.3950617283950617	0.34972677595628415
turma	1.0	"random (n=10, test=33%)"	anova (5)	tree	0.8225423728813559	0.3895043731778426	0.446524064171123	0.3453981385729059
turma	1.0	"random (n=10, test=33%)"	anova (5)	naive_bayes	0.7942372881355932	0.5553113553113553	0.4194798007747648	0.8212351029252438
turma	1.0	"random (n=10, test=33%)"	anova (5)	logistic_regression	0.8535593220338983	0.5309446254071662	0.5367727771679474	0.5252416756176155
turma	1.0	"random (n=10, test=33%)"	anova (5)	neural_network	0.831864406779661	0.3377837116154873	0.42521008403361343	0.28017718715393136
turma	1.0	"random (n=10, test=33%)"	anova (5)	gradient_boosting	0.8233898305084746	0.42239467849223944	0.4261744966442953	0.4186813186813187
turma	1.0	"random (n=10, test=33%)"	anova (5)	svm	0.8252542372881356	0.14154870940882597	0.3512396694214876	0.0886339937434828
turma	1.0	holdout (33/67)	anova (5)	random_forest	0.9224137931034483	0.8	0.6666666666666666	1.0
turma	1.0	holdout (33/67)	anova (5)	knn	0.9310344827586207	0.7999999999999999	0.7272727272727273	0.8888888888888888
turma	1.0	holdout (33/67)	anova (5)	tree	0.9137931034482759	0.782608695652174	0.6428571428571429	1.0
turma	1.0	holdout (33/67)	anova (5)	naive_bayes	0.6206896551724138	0.4500000000000007	0.2903225806451613	1.0
turma	1.0	holdout (33/67)	anova (5)	logistic_regression	0.8448275862068966	0.64	0.5	0.8888888888888888
turma	1.0	holdout (33/67)	anova (5)	neural_network	0.85	0.6741573033707865	0.5084745762711864	1.0
turma	1.0	holdout (33/67)	anova (5)	gradient_boosting	0.9086206896551724	0.7725321888412017	0.6293706293706294	1.0
turma	1.0	holdout (33/67)	anova (5)	svm	0.9137931034482759	0.782608695652174	0.6428571428571429	1.0
turma	1.0	xvalidation (k=10)	anova (10)	random_forest	0.8272727272727273	0.396825396825396	0.44642857142857145	0.3571428571428571
turma	1.0	xvalidation (k=10)	anova (10)	knn	0.8238636363636364	0.4150943396226415	0.44	0.3928571428571428
turma	1.0	xvalidation (k=10)	anova (10)	tree	0.8238636363636364	0.3404255319148936	0.42105263157894735	0.2857142857142857
turma	1.0	xvalidation (k=10)	anova (10)	naive_bayes	0.7784090909090909	0.5411764705882353	0.40350877192982454	0.8214285714285714
turma	1.0	xvalidation (k=10)	anova (10)	logistic_regression	0.8579545454545454	0.5762711864406779	0.5483870967741935	0.6071428571428571
turma	1.0	xvalidation (k=10)	anova (10)	neural_network	0.8232954545454545	0.24330900243309006	0.3816793893129771	0.17857142857142858
turma	1.0	xvalidation (k=10)	anova (10)	gradient_boosting	0.8011363636363636	0.36363636363636365	0.37037037037037035	0.3571428571428571
turma	1.0	xvalidation (k=10)	anova (10)	svm	0.8295454545454546	-1.0	0.0	0.0
turma	1.0	"random (n=10, test=33%)"	anova (10)	random_forest	0.8183050847457627	0.3943502824858757	0.43300248138957814	0.3620331950207469
turma	1.0	"random (n=10, test=33%)"	anova (10)	knn	0.8147457627118644	0.37147786083956297	0.417312661498708	0.33471502590673574
turma	1.0	"random (n=10, test=33%)"	anova (10)	tree	0.8249152542372882	0.4173716864072194	0.4378698224852071	0.39870689655172414
turma	1.0	"random (n=10, test=33%)"	anova (10)	naive_bayes	0.7933898305084746	0.5610370903853079	0.42108108108108105	0.8403451995685005
turma	1.0	"random (n=10, test=33%)"	anova (10)	logistic_regression	0.8540677966101695	0.5348460291734197	0.545755237045204	0.5243644067796610

Fonte: Elaborado pelo autor (2024).

4.2.3.1 Modelo utilizando somente notas, conclusão de atividades e presença

Analisando a totalidade dos dados (100%), o método de aprendizagem de máquina NB apresentou a maior sensibilidade, com um valor de 98,36%. Suas demais métricas (precisão, acurácia e pontuação F1) foram de 43,90%, 80,19% e 60,71%, e para atingir as métricas citadas foi utilizado o método de avaliação SA sem quaisquer métodos de seleção de atributos.

Uma opção alternativa, com enfoque na maior acurácia gerada pelo modelo, é o método de aprendizagem GB, com o modelo de classificação VC e sem quaisquer métodos de seleção de atributos, que obteve uma acurácia de 97,56%. Os valores de precisão, sensibilidade e pontuação F1 foram respectivamente 92,78%, 91,79% e 92,28%.

Com 75% dos dados analisados, a combinação escolhida com ênfase na sensibilidade foi atingida a partir do método de aprendizagem de máquina NB e o método de avaliação SA sem quaisquer métodos de seleção de atributos, que apresentou a maior sensibilidade, com um valor de 98,10%. Suas demais métricas de precisão, acurácia e pontuação F1 foram de 44,22%, 79,83% e 60,96% respectivamente.

Priorizando a acurácia, a melhor opção alternativa é o método de aprendizagem GB, com o modelo de classificação VC, sem quaisquer métodos de seleção de atributos. Este modelo obteve uma acurácia de 95,68% e as demais métricas com valores 85,92%, 87,14% e 86,52% sendo respectivamente a precisão, sensibilidade e pontuação F1.

Os resultados utilizando somente a metade dos dados disponíveis (50%),

resultaram na maior sensibilidade utilizando o método de aprendizagem de máquina NB e o método de avaliação SA sem quaisquer métodos de seleção de atributos, o valor foi de 96,69%. Suas demais métricas de precisão, acurácia e pontuação F1 nesta configuração foram de 45,92%, 80,80% e 62,27% respectivamente.

Ao investigar a maior acurácia obtida para 50% dos dados, encontra-se o modelo que utiliza o método de aprendizado de máquina MLP com o método de avaliação SA, sem um modelo de seleção de atributos. Os valores de acurácia, precisão, sensibilidade e pontuação F1 são respectivamente 92,51%, 75,98%, 77,91% e 76,93%.

Por último, a análise utilizando apenas 25% dos dados, a melhor configuração encontrada para obter-se a maior sensibilidade foi utilizando o método de aprendizagem de máquina NB, com avaliação SA sem quaisquer seleção de atributos. Suas métricas foram 93,96%, 43,58%, 79,59% e 59,54%, sendo estes valores referentes a sensibilidade, precisão, acurácia e pontuação F1 respectivamente.

A opção alternativa, focando na maior acurácia é o método de aprendizagem MLP, com o modelo de classificação VC e sem quaisquer métodos de seleção de atributos, que obteve uma acurácia de 91,31%. Os valores de precisão, sensibilidade e pontuação F1 foram respectivamente 73,70%, 67,44% e 70,43%.

4.2.3.2 Modelo utilizando notas, conclusão de atividade, presença e sentimentos calculados pelo método binário

Analisando o conjunto inteiro de dados disponíveis, o método de aprendizagem de máquina NB apresentou a maior sensibilidade, com 100%. Suas demais métricas (precisão, acurácia e pontuação F1) foram de 33,73%, 68,75% e 50,45%, e para atingir as métricas citadas foi utilizado o método de avaliação VC com o método de seleção de atributos *gain_ratio(5)*, que classificou o atributo *class_feeling* como 3º atributo mais relevante.

Uma opção alternativa, com enfoque na maior acurácia gerada pelo modelo, é o método de aprendizagem GB, com o modelo de classificação VC e sem quaisquer métodos de seleção de atributos, que obteve uma acurácia de 97,73%. Os valores de precisão, sensibilidade e pontuação F1 foram todos de 92,86%.

Com 75% dos dados analisados, a combinação escolhida com ênfase na sensibilidade foi atingida a partir do método de aprendizagem de máquina NB e o método de avaliação VC sem quaisquer métodos de seleção de atributos, que apresentou a maior sensibilidade, com 100%. Suas demais métricas de precisão, acurácia e pontuação F1 foram de 48,11%, 82,84% e 64,97% respectivamente.

Priorizando a acurácia, a melhor opção alternativa é o método de aprendizagem SVM, com o modelo de classificação SA, com o método de seleção de atributos *info_gain(10)* que classificou o sentimento da classe como o 10° atributo mais relevante. Este modelo obteve uma acurácia de 96,39% e as demais métricas com valores 87,32%,

91,36% e 89,29% sendo respectivamente a precisão, sensibilidade e pontuação F1.

Os resultados obtidos utilizando somente 50% dos dados disponíveis, resultaram na maior sensibilidade utilizando o método de aprendizagem de máquina NB e o método de avaliação SA sem quaisquer métodos de seleção de atributos, o valor foi de 98,80%. Suas demais métricas de precisão, acurácia e pontuação F1 nesta configuração foram de 43,96%, 80,17% e 60,84% respectivamente.

Ao investigar a maior acurácia obtida para 50% dos dados, encontra-se o modelo que utiliza o método de aprendizado de máquina SVM com o método de avaliação SA, com o modelo de seleção de atributos *gain_ratio(10)* que classificou o sentimento da turma como o atributo mais relevante para a análise. Os valores de acurácia, precisão, sensibilidade e pontuação F1 são respectivamente 95,73%, 84,29%, 90,17% e 87,13%.

Por fim, a análise utilizando apenas 25% dos dados disponíveis, a melhor configuração encontrada para obter-se a maior sensibilidade foi utilizando o método de aprendizagem de máquina NB, com avaliação SA sem quaisquer seleção de atributos. Suas métricas foram 94,60%, 44,49%, 79,47% e 60,52%, sendo estes valores referentes a sensibilidade, precisão, acurácia e pontuação F1 respectivamente.

Alternativamente, focando na maior acurácia possível, encontramos a configuração utilizando o método de aprendizagem RL, com o modelo de classificação VC e o método de seleção de atributos *gain_ratio(5)* que classificou o atributo *class_feeling* como o mais relevante. Obteve-se uma acurácia de 93,92%. e os valores de precisão, sensibilidade e pontuação F1 foram respectivamente 82,64%, 78,21% e 80,37%.

4.2.3.3 Modelo utilizando notas, conclusão de atividade, presença e sentimentos calculados pelo método porcentagem positiva

A partir da análise total dos dados (100%), o método de aprendizagem de máquina NB apresentou a maior sensibilidade, 100%. Suas demais métricas (precisão, acurácia e pontuação F1) foram de 34,28%, 68,39% e 51,06%, e para atingir as métricas citadas foi utilizado o método de avaliação VC com o método de seleção de atributos *gain_ratio(5)*, que classificou o atributo *class_feeling* como 3º atributo mais relevante.

Uma opção alternativa, com enfoque na maior acurácia gerada pelo modelo, é o método de aprendizagem GB, com o modelo de classificação VC e sem quaisquer métodos de seleção de atributos, que obteve uma acurácia de 97,73%. Os valores de precisão, sensibilidade e pontuação F1 foram todos de 92,86%.

Analisando 75% dos dados disponíveis, a combinação escolhida com ênfase na sensibilidade foi atingida a partir do método de aprendizagem de máquina NB e o método de avaliação VC sem quaisquer métodos de seleção de atributos, que apresentou a maior sensibilidade, com 98,57%. Suas demais métricas de precisão,

acurácia e pontuação F1 foram de 48,51%, 83,13% e 65,02% respectivamente.

Priorizando a acurácia, a melhor opção alternativa é o método de aprendizagem SVM, com o modelo de classificação SA, com o método de seleção de atributos *info_gain(10)* que classificou o sentimento da classe como o atributo mais relevante. Este modelo obteve uma acurácia de 96,46% e as demais métricas com valores 86,51%, 91,97% e 89,15% sendo respectivamente a precisão, sensibilidade e pontuação F1.

Os resultados obtidos utilizando somente 50% dos dados disponíveis, resultaram na maior sensibilidade utilizando o método de aprendizagem de máquina NB e o método de avaliação SA sem quaisquer métodos de seleção de atributos, o valor foi de 98,15%. Suas demais métricas de precisão, acurácia e pontuação F1 nesta configuração foram de 45,19%, 81,19% e 61,88% respectivamente.

Ao investigar a maior acurácia obtida para 50% dos dados, encontra-se o modelo que utiliza o método de aprendizado de máquina MLP com o método de avaliação SA, com o modelo de seleção de atributos *gain_ratio(5)* que classificou o sentimento da turma como o atributo mais relevante para a análise. Os valores de acurácia, precisão, sensibilidade e pontuação F1 são respectivamente 95,36%, 82,92%, 88,96% e 85,83%.

Por fim, a análise utilizando apenas 25% dos dados disponíveis, a melhor configuração encontrada para obter-se a maior sensibilidade foi utilizando o método de aprendizagem de máquina NB, com avaliação SA sem quaisquer seleção de atributos. Suas métricas foram 94,12%, 48,31%, 82,47% e 63,85%, sendo estes valores referentes a sensibilidade, precisão, acurácia e pontuação F1 respectivamente.

Alternativamente, focando na maior acurácia possível, encontramos a configuração utilizando o método de aprendizagem SVM, com o modelo de classificação VC e o método de seleção de atributos *gain_ratio(10)* que classificou o atributo *class_feeling* como o 9º mais relevante. Obteve-se uma acurácia de 94,89%. e os valores de precisão, sensibilidade e pontuação F1 foram respectivamente 83,45%, 84,64% e 84,04%.

4.2.3.4 Modelo utilizando notas, conclusão de atividade, presença e sentimentos calculados pelo método de ponderação baseada nas demais pontuações

Com 100% dos dados analisados, o método de aprendizagem de máquina NB apresentou a maior sensibilidade, 100%. Suas demais métricas (precisão, acurácia e pontuação F1) foram de 71,79%, 93,75% e 83,58%, e para atingir as métricas citadas foi utilizado o método de avaliação VC juntamente com o método de seleção de atributos *gain_ratio(10)* que classificou o sentimento da turma como o 3º atributo mais relevante para a análise.

Uma opção alternativa, com enfoque na maior acurácia gerada pelo modelo, é o método de aprendizagem RF, com o modelo de classificação VC e o método de

seleção de atributos *gain_ratio(10)*, que obteve uma acurácia de 97,78%. Os valores de precisão, sensibilidade e pontuação F1 foram respectivamente 93,82%, 92,14% e 92,97%.

Com 75% dos dados analisados, a combinação escolhida com ênfase na sensibilidade foi atingida a partir do método de aprendizagem de máquina NB e o método de avaliação VC sem quaisquer métodos de seleção de atributos, que apresentou a maior sensibilidade, com um valor de 99,64%. Suas demais métricas de precisão, acurácia e pontuação F1 foram de 46,81%, 81,93% e 63,70% respectivamente.

Priorizando a acurácia, a melhor opção alternativa é o método de aprendizagem RF, com o modelo de classificação VC, utilizando o método de seleção de atributos *info_gain(10)* que classificou o atributo *class_feeling* como o mais relevante. Este modelo obteve uma acurácia de 96,31% e as demais métricas com valores 89,96%, 86,43% e 88,16% sendo respectivamente a precisão, sensibilidade e pontuação F1.

Os resultados utilizando somente a metade dos dados disponíveis, resultaram na maior sensibilidade utilizando o método de aprendizagem de máquina NB e o método de avaliação VC sem quaisquer métodos de seleção de atributos, o valor foi de 98,21%. Suas demais métricas de precisão, acurácia e pontuação F1 nesta configuração foram de 45,91%, 81,31% e 62,57% respectivamente.

Ao investigar a maior acurácia obtida para 50% dos dados, encontra-se o modelo que utiliza o método de aprendizado de máquina MLP com o método de avaliação VC e o modelo de seleção de atributos *info_gain(5)*, que classificou o sentimento da turma como o atributo mais relevante. Os valores de acurácia, precisão, sensibilidade e pontuação F1 são respectivamente 95,28%, 82,51%, 89,29% e 85,76%.

Para finalizar, a análise utilizou apenas 25% dos dados disponíveis, e a melhor configuração encontrada para obter-se a maior sensibilidade foi utilizando o método de aprendizagem de máquina NB, com avaliação SA com a seleção de atributos usando *info_gain(10)* que classificou o atributo *class_feeling* como o mais relevante. Suas métricas foram 93,78%, 45,73%, 80,80% e 61,48%, sendo estes valores referentes a sensibilidade, precisão, acurácia e pontuação F1 respectivamente.

A opção alternativa, focando na maior acurácia é o método de aprendizagem RF, com o modelo de classificação SA utilizando o método de seleção de atributos *info_gain(10)* que classificou o atributo referente ao sentimento da turma como o mais relevante, e obteve uma acurácia de 93,88%. Os valores de precisão, sensibilidade e pontuação F1 foram respectivamente 81,32%, 79,48% e 80,39%.

4.2.3.5 Síntese dos resultados

Os resultados obtidos com a análise implícita podem ser observados de maneira sintetizada nas Tabelas 4 e 5. Assim, é possível apurar as principais

combinações de métodos de avaliação de resultados, técnicas de escolha de atributos e algoritmos de aprendizado de máquina para cada uma das proporções de dados e as métricas de sensibilidade e acurácia atingidas.

Tabela 4 – Comparação entre os diferentes métodos com relação a sensibilidade

Modelo	**%**	**Algori.**	**Avali.**	**Seleção Atributos**	**Rank**	**Sensibilidade**	**Precisão**
Sem	100%	NB	SA	-	-	98,36%	43,90%
	75%	NB	SA	-	-	98,10%	44,22%
	50%	NB	SA	-	-	96,69%	45,92%
	25%	NB	SA	-	-	93,96%	43,58%
Binário	100%	NB	VC	Gain Ratio-5	3°	**100%**	33,73%
	75%	NB	VC	-	-	**100%**	48,11%
	50%	NB	SA	-	-	**98,80%**	43,96%
	25%	NB	SA	-	-	**94,60%**	44,49%
%	100%	NB	VC	Gain Ratio-5	3°	**100%**	34,28%
	75%	NB	VC	-	-	98,57%	48,51%
	50%	NB	SA	-	-	98,15%	45,19%
	25%	NB	SA	-	-	94,12%	48,31%
Ponderar	100%	NB	VC	Gain Ratio-10	3°	**100%**	71,79%
	75%	NB	VC	-	-	99,64%	46,81%
	50%	NB	VC	-	-	98,21%	45,91%
	25%	NB	SA	Gain Info-10	1°	93,98%	45,73%

Pode-se extrair a partir da Tabela 4, que em relação ao total de combinações:

- 100% utilizam o algoritmo de aprendizado de máquina *Naive Bayes* (NB);
- 56,25% utilizam o método de avaliação de validação cruzada 10-*fold* (VC);
- 25% utilizam um método de seleção de atributos.

Ao analisar os resultados de sensibilidade obtidos, podemos constatar que os modelos que utilizaram o atributo referente ao sentimentos dos estudantes com relação a matéria tiverem uma sensibilidade maior em todos os casos. Ressaltando que a sensibilidade representa a porcentagem de desistentes classificados corretamente como desistentes, e a precisão indica a porcentagem de previsões positivas consideradas falsas. Quando existe um método de seleção de atributos selecionado é mostrado a posição do *ranking* do atributo *class_feeling* em "Rank".

O método que resultou nos melhores resultados, considerando a precisão, foi o modelo utilizando notas, conclusão de atividade, presença e sentimentos calculados pelo método de ponderação baseado nas demais pontuações, onde 100% dos estudantes desistentes foram identificados pelo modelo e um aumento de 27,89% na

precisão, o que resulta em uma diminuição dos falsos positivos, e ajuda o docente a focar apenas nos alunos com real desejo de desistência. Todos os valores obtidos para quaisquer porcentagem de dados foram iguais ou superiores ao modelo que não levou em conta a análise de sentimentos, mostrando assim que o novo atributo trouxe um ganho de informação para a melhor classificação dos estudantes.

Tabela 5 – Comparação entre os diferentes métodos com relação a acurácia

Modelo	**%**	**Algori.**	**Avali.**	**Seleção Atributos**	**Rank**	**Acurácia**	**Precisão**
Sem	100%	GB	VC	-	-	97,56%	92,78%
	75%	GB	VC	-	-	95,68%	85,92%
	50%	MLP	SA	-	-	92,51%	75,98%
	25%	MLP	VC	-	-	91,31%	73,70%
Binário	100%	GB	VC	-	-	97,73%	92,86%
	75%	SVM	SA	Gain Info-10	10°	96,39%	87,32%
	50%	SVM	SA	Gain Ratio-10	1°	**95,73%**	84,29%
	25%	RL	VC	Gain Ratio-5	1º	93,92%	82,64%
%	100%	GB	VC	-	-	97,73%	92,86%
	75%	SVM	SA	Gain Info-10	1°	**96,46%**	86,51%
	50%	MLP	SA	Gain Ratio-5	1°	95,36%	88,96%
	25%	SVM	VC	Gain Ratio-10	9°	**94,89%**	83,45%
Ponderar	100%	RF	VC	Gain Ratio-10	3°	**97,78%**	93,82%
	75%	RF	VC	Gain Info-1	1º	96,31%	89,96%
	50%	MLP	VC	Gain Info-5	1°	95,28%	82,51%
	25%	RF	SA	Gain Info-10	1°	93,88%	81,32%

O pior valor de sensibilidade obtido foi de 93,98% para uma análise de apenas 25% dos dados do semestre, o que mostra a eficácia do do algoritmo, que mesmo em um estágio inicial do semestre é capaz de apontar a grande maioria dos alunos desistentes. Com 50% dos dados do semestre, o método binário demonstrou o maior ganho de sensibilidade do presente estudo de caso, obtendo um ganho de 2,11% na sensibilidade com relação ao método que não utiliza os sentimentos (considerando o conjunto total de dados da turma, para identificar um estudante desistente a mais é necessário um aumento de 1,61% no valor da sensibilidade). Sendo assim, tem-se um

bom resultado, já que foi possível a identificação de mais um aluno.

Os demais modelos que consideram o sentimento em sua análise, apresentaram o mesmo comportamento. Tanto na análise com 100% dos dados quanto na análise utilizando 50% dos dados disponíveis, obtiveram uma precisão inferior ao modelo sem considerar sentimentos. No entanto, não foram valores discrepantes, são valores bem próximos entre os modelos analisados.

Com a Tabela 5, pode-se inferir que em relação ao total de combinações temos:

- A utilização de 5 diferentes métodos de aprendizado de máquina;
- 62,5% utilizando o método de avaliação de validação cruzada 10-fold (VC);
- 62,5% utilizando um método de seleção de atributos.

Ao analisar os resultados obtidos considerando os melhores cenários para o valor da acurácia, é perceptível que temos uma maior variação nos métodos de aprendizado de máquina, métodos de avaliação e selecionadores de atributos empregados. Recordando que a acurácia representa a proporção de desistentes e não desistentes que o modelo classificador foi capaz de classificar corretamente (verdadeiros positivos e negativos), desta forma sinaliza o quão bem o modelo previu ambas as classificações.

Similarmente ao ocorrido na análise dos resultados da sensibilidade, podemos constatar que os modelos que utilizaram o atributo referente ao sentimento dos estudantes com relação a matéria tiveram uma acurácia maior em todos os casos. E o modelo que obteve os melhores resultados não foi unânime desta vez, os três modelos que utilizam os sentimentos foram melhores em pelo menos uma das porcentagens de dados analisados, mas não destoando dos demais.

O novo atributo agregou principalmente para as proporções iniciais do semestre, sendo que aumentou a acurácia na análise de 50% do semestre com o método Binário em 3,22% e a precisão em 8,31%. Com 25% do semestre analisado, o método de porcentagem positiva teve um ganho de 3,58% na acurácia e 9,75% na precisão das classificações verdadeiras com relação ao modelo que não considera atributo relacionado ao sentimento. Assim, o atributo demonstrou que principalmente em estágios iniciais do semestre melhora a classificação dos estudantes.

Referente a acurácia, para considerar que quaisquer modelo identificou pelo menos uma classificação verdadeira a mais, sendo ela positiva ou negativa, considerando a totalidade de alunos da turma é necessário que a métrica da acurácia tenha um ganho de aproximadamente 0,57%. A partir disso é possível afirmar que o novo atributo referente aos sentimentos da turma ajudou a classificar novos estudantes corretamente, ou seja, teve um impacto positivo e agregou informação ao modelo.

O atributo *class_feeling* mostrou-se um atributo muito relevante em todas as análises utilizando algum método de seleção de atributos, especialmente do modelo que pondera com as demais notas do estudante. Com isso, pode-se notar que é um

atributo importante na predição de estudantes desistentes, uma vez que melhorou todas as métricas calculadas neste trabalho.

4.3 CONSIDERAÇÕES DO CAPÍTULO

O capítulo descrito acima tem como objetivo avaliar a performance dos modelos desenvolvidos no trabalho e avaliar o impacto do novo atributo atrelado ao sentimento da turma. Os dados utilizados na análise provêm de duas turmas distintas, coletados pelo docente responsável, que classifica manualmente os alunos em desistentes ou não desistentes.

As análises foram feitas em quatro diferentes proporções do período letivo (100%, 75%, 50% e 25%) para avaliar a aplicabilidade prática do modelo na previsão da desistência escolar. O uso de 100% dos dados serve como uma referência para a melhor métrica alcançável, enquanto proporções menores simulam intervenções em momentos críticos do curso e com tempo hábil para intervenções.

Na análise explícita, foram avaliados os dados de cada turma separadamente, utilizando os sistemas de pontuações desenvolvidos. As métricas de sensibilidade, precisão, acurácia e pontuação F1 foram calculadas para diferentes modelos (incluindo e excluindo o atributo referente ao sentimento da turma), ao analisar os resultados obtidos com o foco em sensibilidade, acurácia e precisão não notou-se ganho significativo nos resultados.

Os modelos que consideraram de alguma forma o sentimento da turma obtiveram resultados iguais ou até mesmo piores do que o modelo que desconsideram, exceto quando o modelo foi aplicado com 75% dos dados do semestre, onde o modelo binário teve uma pequena melhora com relação ao modelo sem considerar sentimentos. Sendo assim, o novo atributo não tem impacto positivo na previsão de alunos desistentes na análise explícita dos dados.

Com a análise implícita foi constatado que, o novo atributo referente ao sentimento da turma mostrou-se relevante na previsão de estudantes desistentes, ao contribuir para a melhora de todas as métricas calculadas para quaisquer porcentagens do semestre. Foi impactante principalmente para a previsão de desistentes com 50% dos dados do semestre, utilizando o método de cálculo binário (aumento de 2,11% na sensibilidade e 3,22% na acurácia), momento onde o docente tem bastante tempo disponível para intervir e reverter o quadro do estudante.

5 CONCLUSÕES

A desistência de estudantes é objeto de estudos, análises e debates a algum tempo, principalmente dentre os países mais desenvolvidos. É um tema comum às instituições universitárias no mundo contemporâneo, principalmente pelo seu teor de complexidade e abrangência ANDIFES/ABRUEM/SESu/MEC (1996). Sendo assim, o presente trabalho objetivou analisar o impacto de uma nova métrica relacionada ao sentimento da turma sobre a disciplina e a assertividade dos modelos preditivos de desistência estudantil, foram desenvolvidas tanto uma análise explícita quanto implícita dos dados, cumprindo com o objetivo de *elaborar modelos para classificar alunos desistentes utilizando dados combinados*.

Os modelos desenvolvidos procuram identificar, por meio da análise de dados passivos extraídos diretamente do Ambiente Virtual de Aprendizagem MOODLE do estudante e sentimentos ativos coletados por meio de questionários combinados alunos que estejam propensos a desistir da disciplina, alcançando desta maneira o objetivo de *analisar a conjunção de dados provenientes de mais de uma fonte*. No primeiro momento do estudo de caso, o objetivo é *elaborar métodos para converter sentimentos em atributos úteis para o modelo de classificação*, assim foram elaborados métodos de transformação dos dados referentes ao sentimentos em atributos que pudessem ser utilizados pelos modelos de classificação e aprendizado de máquina gerados.

Com o intuito de identificar o maior número possível de alunos desistentes, bem como seus comportamentos que representam uma possível desistência da matéria, a métrica de sensibilidade foi escolhida como o principal critério de avaliação. A sensibilidade é uma métrica importante pois indica a porcentagem de alunos desistentes que são erroneamente classificados pelo algoritmo, ou seja, ela reflete a capacidade do classificador de identificar corretamente aqueles alunos que apresentam sinais de desistência e podem não ser detectados como tais.

Referente a análise explícita, os resultados obtidos por meio deste estudo de caso mostraram que, apesar das diferentes abordagens para o cálculo do atributo referente ao sentimento da turma, a métrica de sensibilidade não apresentou variação significativa entre os modelos desenvolvidos. Desta maneira, a inclusão do atributo de sentimento, não altera a eficácia da classificação com relação ao modelo gerado somente com os atributos de desempenho do aluno.

Com a seleção de atributos, dentro da análise implícita, pode-se observar que dois dos quatro métodos de seleção de atributos consideraram o novo atributo criado a partir dos sentimentos dos estudantes entre os 10 atributos mais relevantes para a análise. Sendo que o método de ganho de informação classificou o atributo

class_feeling como o mais relevante no modelo utilizando notas, conclusão de atividade, presença e sentimentos calculados pelo método de ponderação baseado nas demais pontuações.

Na análise implícita foi constatado que, o novo atributo referente ao sentimento da turma mostrou-se relevante na previsão de estudantes desistentes, ao contribuir para a melhora de todas as métricas calculadas, inclusive a da sensibilidade, para quaisquer porcentagens do semestre avaliadas. Destacando o método de cálculo binário que obteve as maiores métricas de sensibilidade dentre os modelos desenvolvidos, tendo uma sensibilidade de 94,60% com apenas 25% dos dados do semestre utilizados.

Tabela 6 – Síntese dos resultados obtidos

Análise	**Melhor Modelo (50%)**	**Sensibilidade**	**Precisão**	**Impacto do Sentimento**
Explícita	Sem considerar sentimentos	1: 89,64% 2: 28,93%	1: 89,64% 2: 26,05%	Neutro
Implícita	Atribuição binária	98,80%	43,96%	Melhora

Com isso, foi possível analisar o impacto do novo atributo relacionado ao sentimento da turma sobre a disciplina e a assertividade dos modelos preditivos de desistência estudantil em ambas as análises, como mostrado na Tabela 6. Com o novo atributo se mostrando dispensável na classificação dos alunos desistentes na análise explícita e muito importante para a análise implícita, onde obteve ganhos com relação a sensibilidade.

Para trabalhos futuros, sugere-se a elaboração de um modelo que procure individualizar o sentimento e atribuí-lo a estudantes específicos, podendo ser implementado por métodos de similaridade de linguagem natural da escrita dos estudantes. Desta forma, pode-se mesmo que de forma anonimizada, saber o sentimento especifico de cada um dos estudantes perante a disciplina.

REFERÊNCIAS

ANDIFES/ABRUEM/SESU/MEC. **Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas**. 1996. Disponível em: https://www.andifes.org.br/wp-content/files_flutter/Diplomacao_Retencao_Evasao_Graduacao_em_IES_Publicas-1996.pdf. Acesso em: 10 abr. 2024.

BARBOSA, J. L. N. et al. Introdução ao processamento de linguagem natural usando python. **Escola Regional de Informática do Piauí**, v. 1, n. 1, p. 336–360, jan. 2017.

BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. Métodos para análise de sentimentos em mídias sociais. **Sociedade Brasileira de Computação**, 2015.

BRITTO, L.; PACÍFICO, L. Análise de sentimentos para revisoes de aplicativos mobile em português brasileiro. In: SBC. **Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2019. p. 1080–1090.

DEED - DIRETORIA DE ESTATíSTICAS EDUCACIONAIS. Ministério da Educação. **Metodologia de Cálculo dos Indicadores de Fluxo da Educação Superior**. Brasília, 2017. Disponível em: https://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2017/metodologia_indicadores_trajetoria_curso.pdf. Acesso em: 10 abr. 2024.

ESOMAR. **Guia de proteção de dados - LGPD para profissionais de pesquisa de mercado, opinião e mídia**. [S.l.], 2020.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. 2. ed. São Francisco, Estados Unidos: Elsevier, 2006.

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Censo da Educação Superior 2021**. 2022. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior. Acesso em: 10 abr. 2024.

MOURA, D. H.; SILVA, M. D. S. A evasão no curso de licenciatura em geografia oferecido pelo cefet-rn. 2007.

NOETZOLD, E.; L. PERTILE, S. de. Análise e predição de evasão dos alunos de um curso de graduação em sistemas de informação por meio da mineração de dados educacionais. **Revista Novas Tecnologias na Educação**, v. 19, n. 1, p. 351–360, jul. 2021. Disponível em: https://seer.ufrgs.br/index.php/renote/article/view/118525.

OLIVEIRA, R. d. S. **Modelo de Predição de Evasão Escolar com Base em Dados de Autoavaliação de Cursos de Graduação**. Dissertação de Mestrado (Mestrado em Tecnologia da Informação) — Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, 2023.

PFITSCHER, R. J. et al. Análise de sentimentos em turmas de programação com vistas ao apoio à permanência estudantil. **Congresso Brasileiro de Informática na Eduacação**, 2023.

SILVA, J. Marques Carvalho da; IMRAN, H. Um estudo sobre as variáveis para predição de alunos não concluintes em cursos suportados por ambientes virtuais de ensino e aprendizagem. **Revista Novas Tecnologias na Educação**, v. 13, n. 2, dez. 2015. Disponível em: https://seer.ufrgs.br/index.php/renote/article/view/61427.

SOUZA, B. D. d. **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 (Graduação em Engenharia Mecatrônica) — Centro Tecnológico de Joinville, Universidade Federal de Santa Catarina, Joinville, 2023.

TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. **Review of Educational Research**, v. 45, n. 1, p. 89–125, 1975. Disponível em: https://doi.org/10.3102/00346543045001089.

VIEIRA, R.; LOPES, L. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. In: **Linguagens Especializadas em Corpora Modos de Dizer Interface de Pesquisa**. Porto Alegre: ediPUCRS, 2010. p. 183–202. Disponível em: https://editora.pucrs.br/edipucrs/acessolivre//livros/linguagensespecializadasemcorpora.pdf. Acesso em: 10 abr. 2024.