UNIFACCAMP
Centro Universitário Campo Limpo Paulista

Análise de Desempenho de Alunos de um Curso Técnico de Nível Médio Utilizando Algoritmos de Mineração de Dados

Renan Aleixo Paganatto

Maio / 2023

Dissertação de Mestrado em Ciência da Computação

Análise de Desempenho de Alunos de um Curso Técnico de Nível Médio Utilizando Algoritmos de Mineração de Dados

Esse documento corresponde à Dissertação apresentada à Banca Examinadora para Defesa de Dissertação no curso de Mestrado em Ciência da Computação da UNIFACCAMP – Centro Universitário Campo Limpo Paulista.

Campo Limpo Paulista, 29 de maio de 2023.

Renan Aleixo Paganatto

Profa. Dra. Ana Maria Monteiro (Orientadora)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Ficha catalográfica elaborada pela

Biblioteca Central da Unifaccamp

P147a

Paganatto, Renan Aleixo

Análise de desempenho de alunos de um curso técnico de nível médio utilizando algoritmos de mineração de dados / Renan Aleixo Paganatto. Campo Limpo Paulista, SP: Unifaccamp, 2023.

Orientadora: Profª. Drª. Ana Maria Monteiro

Dissertação (Programa de Mestrado Profissional em Ciência da Computação) – Centro Universitário Campo Limpo Paulista – Unifaccamp.

1. Previsão do desempenho de estudantes. 2. Mineração de dados educacionais. 3. Árvores de decisão. 4. Algoritmo *Naive Bayes*. 5. KNN. I. Monteiro, Ana Maria. II. Centro Universitário Campo Limpo Paulista. III. Título.

CDD – 006.312

Resumo: A educação tem um papel fundamental no desenvolvimento de uma sociedade, em especial em um país como o Brasil. Para os encarregados de gerenciar atividades educacionais e para os professores é muito importante poder prever possíveis problemas no desempenho dos estudantes o que, eventualmente, pode fazer com que alguns deles abandonem o ambiente educacional. Em geral, professores experientes conseguem predizer o desempenho acadêmico futuro de seus estudantes. Além da contribuição desses professores, para entender melhor como antever o desempenho dos alunos, pode ser utilizado o volume crescente de dados eletrônicos disponíveis nas instituições de ensino. A partir da aplicação de técnicas de mineração nos dados disponíveis, pode ser extraída informação significativa a ser utilizada por gestores educacionais e professores, para tomar decisões e aplicar medidas preventivas, quando necessário, para melhorar o desempenho dos alunos e a qualidade do processo educacional. Nesta pesquisa foram utilizados dados provenientes de um curso de Informática para Internet Integrado ao Ensino Médio de uma Escola Técnica Estadual (ETEC). Os dados são originários do primeiro, segundo e terceiro ano do curso, coletados entre 2014 e 2020. Para a previsão de desempenho dos alunos foram realizados experimentos utilizando os algoritmos Naive Bayes, J48 (árvore de decisão) e IBk (KNN), todos disponíveis no ambiente WEKA. Os alunos são classificados em quatro categorias podendo ser aprovados, aprovados com pendências para o próximo ano, reprovados ou evadidos. O objetivo dos experimentos foi determinar o desempenho dos algoritmos e técnicas utilizadas para classificar o desempenho dos alunos que estão cursando na instituição e, a partir das análises realizadas e descritas neste trabalho foi desenvolvido um sistema para auxiliar os gestores da instituição na previsão do desempenho dos alunos com auxílio dos algoritmos de classificação abordados na pesquisa.

Palavras-chave: Previsão do desempenho de estudantes, mineração de dados educacionais, árvores de decisão, algoritmo Naive Bayes, KNN.

Abstract: *Education plays a fundamental role in the development of a society, especially in a country like Brazil. For those in charge of managing educational activities and for teachers, it is very important to be able to predict possible problems in students' performance, which eventually may cause some of them to leave the educational environment. In general, experienced teachers are able to predict the future academic performance of their students. In addition to the contribution of these teachers, to better*

understand how to predict student performance, the growing volume of electronic data available in educational institutions can be used. From the application of mining techniques on the available data, meaningful information can be extracted to be used by educational managers and teachers to make decisions and apply preventive measures, when necessary, to improve student performance and the quality of the educational process. In this research, data from an Integrated High School Internet Computing course at a State Technical School (ETEC) were used. The data originated from the first, second, and third year of the course, collected between 2014 and 2020. To predict student performance, experiments were conducted using the Naive Bayes, J48 (decision tree) and IBk (KNN) algorithms, all available in the WEKA environment. The students are classified into four categories, being approved, approved with pendencies for the next year, failed, or dropped out. The objective of the experiments was to determine the performance of the algorithms and techniques used to classify the performance of the students who are attending the institution and, from the analyses performed and described in this work, a system was developed to help the institution's managers in predicting the students' performance with the help of the classification algorithms addressed in the research.

Keywords: *Student performance prediction, educational data mining, decision trees,* Naive Bayes *algorithm*, KNN.

Sumário

Glossário

ACM	*Association for Computing Machinery*
ADASYN	Abordagem de Amostragem Sintética Adaptativa
AM	Aprendizado de Máquina
AUC	*Area Under the Curve*
ARFF	*Attribute-Relation File Format*
CART	*Classification and Regression Trees*
PC	Computador Pessoal
ETEC	Escola Técnica
ESP	Especificidade
FN	Falso Negativo
FP	Falso Positivo
GNU	*General Public License*
INEP	Instituo Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IEEE	*Institute of Electrical and Electronics Engineers*
IA	Inteligência Artificial
ID3	*Iterative Dichotomiser 3*
KNN	*K-nearest neighbor*
KDD	*Knowledge Discovery in Databases*
MP	Média Ponderada
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
NNSEARCH	*Nearest Neighbor Search*
PREC	Precisão
ROC	*Receiver Operating Characteristic*
RBIE	Revista Brasileira de Informática na Educação

Recall	Revocação
SBIE	Simpósio Brasileiro de Informática na Educação
SVM	*Support Vector Machine*
SMOTE	*Synthetic Minority Oversampling Technique*
$T_{ACURACIA}$	Taxa de Acurácia
ERR	Taxa de Erro Total
TFN	Taxa de Falso Negativo
TFP	Taxa de Falso Positivo
TVP	Taxa de Verdadeiros Positivos
UNB	Universidade de Brasília
UFRJ	Universidade Federal do Rio de Janeiro
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
WEKA	*Waikato Environment for Knowledge Analysis*
WEE	WEKA *Experiment Environment*
WE	WEKA *Explorer*

Lista de Tabelas

Lista de Figuras

1. Introdução

Como a desistência escolar está se tornando algo mais comum entre os estudantes, identificar alunos com problemas no seu desempenho escolar é um processo fundamental que pode auxiliar educadores e administradores educacionais a tomarem medidas apropriadas, com o objetivo de orientar os alunos para melhorar seu rendimento ou superar seus problemas e, assim, evitar o abandono do curso.

Todos os anos a educação sofre constantes modificações, em parte ocasionadas pelo avanço tecnológico que revoluciona cada dia mais o setor educacional, tanto pelo gerenciamento digital dos registros acadêmicos dos estudantes quanto pelo uso da Internet como um veículo para ajudar a melhorar o processo de ensino-aprendizado. Esses fatores impulsionaram um crescimento exponencial no volume de dados educacionais digitais. Para que esse grande volume de dados seja analisado é imprescindível contar com recursos computacionais, caso contrário a tarefa torna-se muito difícil e, dependendo da situação, impraticável (Baker *et al*., 2015). No contexto de tratamento de grandes volumes de dados surgiu a área de pesquisa de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases - KDD), como a confluência de diversas outras áreas, em especial da área de Aprendizado de Máquina (AM). KDD tem como objetivo tentar interpretar e analisar grandes volumes de dados e extrair o conhecimento contido neles (Han *et al.*, 2012).

A descoberta de conhecimento envolve várias etapas, desde a definição e compreensão do domínio do qual surgiram os dados, até a análise e interpretação dos resultados obtidos. Vale salientar que em alguns contextos o termo Mineração de Dados (MD) e descoberta de conhecimento em bases de dados são utilizados como termos equivalentes. Analisando esse cenário, destaca-se a Mineração de Dados Educacionais (MDE) que utiliza as técnicas de MD para extrair informações relevantes de conjuntos de dados educacionais com o objetivo de melhorar a educação como um todo. Em outras palavras, a MD refere-se a um conjunto de técnicas computacionais para extrair informações de dados, e quando os dados analisados são provenientes de contextos educacionais, chama-se MDE (Romero e Ventura, 2013).

Para os encarregados de gerenciar atividades educacionais e para os professores é muito importante poder prever possíveis problemas no desempenho dos estudantes, o que, eventualmente, pode fazer com que alguns deles abandonem o ambiente educacional. Em geral, professores experientes conseguem predizer o desempenho futuro de seus estudantes. Além da contribuição desses professores, para entender melhor como antever o desempenho dos alunos, pode ser utilizado o crescente volume de dados eletrônicos disponíveis nas instituições de ensino. A partir da aplicação de técnicas de MD pode ser extraída informação significativa a ser utilizada por gestores educacionais e professores para tomar decisões e aplicar medidas preventivas, quando necessário, para melhorar o desempenho dos alunos e a qualidade do processo educacional.

Os dados utilizados na pesquisa realizada e descrita nesta dissertação são originários do curso de Informática para Internet Integrado ao Ensino Médio da Escola Técnica Estadual (ETEC) Bartolomeu Bueno da Silva – Anhanguera. Nesta modalidade de ensino, o aluno cursa o Ensino Médio em conjunto com a formação de Técnico em Informática para Internet, numa jornada de até 40 aulas semanais (até 8 aulas diárias), em cada uma das 3 séries no formato de ensino anual (primeiro, segundo e terceiro ano). Ao final do curso, com duração total de 3 anos, o aluno além de receber o diploma de conclusão do Ensino Médio obterá o diploma de Técnico em Informática para Internet, com validade nacional, de acordo com o perfil profissional a seguir: O Técnico em Informática para Internet é o profissional que desenvolve e realiza manutenção em websites e portais na Internet e Intranet. Utiliza ferramentas de desenvolvimento de projetos para construir soluções que auxiliam o processo de criação de interfaces e aplicativos empregados no comércio e marketing eletrônicos (ETEC 2022).

Os dados originais utilizados na pesquisa são provenientes do primeiro, segundo e terceiro ano do curso entre 2014 e 2020 (a cada ano, após o ano letivo, os dados são coletados para as três turmas do curso) e foram obtidos a partir de planilhas da Microsoft Excel fornecidas pela instituição. Os dados estavam organizados em 4 arquivos, um contendo informações pessoais e socioeconômicas dos alunos regularmente matriculados e os outros três (um para cada ano do curso) com informação, em sua maioria, referente às notas e frequência do aluno no curso.

Para a previsão de desempenho dos alunos foram realizados experimentos utilizando os algoritmos Naive Bayes e J48 (árvore de decisão) (Han *et al.*, 2011) e IBk (KNN) (Witten *et al.*, 2011), todos disponíveis no ambiente WEKA (WEKA 2022).

1.1. Objetivos e Método

O objetivo dos experimentos foi utilizar os algoritmos de classificação para prever o desempenho de alunos a partir de dados pessoais e de desempenho acadêmico e também comparar o desempenho dos algoritmos utilizados para classificar os alunos da instituição. Esse entendimento pode ser utilizado por gestores educacionais e professores para tomar decisões e aplicar medidas preventivas quando necessário para melhorar o rendimento dos alunos. Os alunos são classificados em quatro categorias, dependendo de eles terrem sido aprovados, aprovados com pendências para o próximo ano, reprovados ou evadidos. Os alunos não aprovados precisam ser identificados e oferecidos a eles os cuidados necessários para melhorar seu aproveitamento escolar.

Para entender melhor como antever o desempenho dos alunos para assim tomar medidas quando necessário foi realizada uma revisão sistemática da literatura para determinar o estado da arte nesse tema e apresentar uma visão muito mais fidedigna da literatura e menos suscetível à influência de possíveis vieses por parte dos pesquisadores (Kitchenham, 2007; Neiva e Silva, 2016).

Para atingir o objetivo proposto neste trabalho, foi adotada uma metodologia constituída das seguintes etapas:

- Levantamento da bibliografia em relação ao desempenho de estudantes utilizando mineração de dados.
- Análise da problemática na instituição escolhida.
- Obtenção e análise dos dados a serem utilizados nos experimentos.
- Tratamento dos dados obtidos, referenciada como pré-processamento, consistindo de aplicação da limpeza, integração, transformação e redução nos dados obtidos junto à instituição.
- Aplicação dos algoritmos.

- Análise dos resultados obtidos na etapa anterior quanto à efetividade do estudo conduzido.

Deste modo, o estudo fornece à instituição informações compreensíveis e relevantes por meio de um conjunto de experimentos computacionais. Essas informações poderão auxiliar as autoridades e professores da instituição na tomada de decisão.

1.2. Organização e Estrutura do Trabalho

Este trabalho está estruturado da seguinte maneira: o Capítulo 2 apresenta a revisão da literatura resumindo os principais pontos relacionados aos trabalhos selecionados na execução da revisão sistemática; no Capítulo 3 é apresentado o referencial teórico que aborda a área de pesquisa conhecida como *Knowledge Discovery in Databases* (KDD) e as diferentes etapas envolvidas no processo; no Capítulo 4 são abordados os algoritmos de classificação Naive Bayes, J48 e KNN que foram utilizados para predizer o desempenho dos alunos do ensino médio-técnico da instituição; no Capítulo 5 é apresentado o processo de pré-processamento dos dados utilizados na pesquisa assim como a ferramenta WEKA e as configurações aplicadas nos algoritmos utilizados; no Capítulo 6 são apresentados os experimentos e análises dos resultados para os algoritmos Naive Bayes, J48 e IBk, e respectivas análises para cada cenário considerado; no Capítulo 7 é apresentado um protótipo de um sistema para prever o desempenho dos alunos; por fim, no Capítulo 8 são apresentadas as conclusões da pesquisa e os trabalhos futuros.

2. Revisão da Literatura

Neste capítulo são apresentados os trabalhos que embasaram a pesquisa, tanto em temas relacionados ao desempenho dos alunos no ensino presencial quanto ao uso mineração de dados como auxílio na detecção desempenho. A Seção 2.1 descreve o planejamento da revisão sistemática realizada e as principais questões associadas. A Seção 2.2 descreve os passos para execução da revisão sistemática. A Seção 2.3 apresenta a análise dos trabalhos selecionados, as respostas às questões da Seção 2.1, assim como uma síntese dos artigos mais relevantes.

2.1. Planejamento

A etapa de planejamento é importante para definir a forma como a revisão sistemática é executada, os critérios levados em consideração para a inclusão e exclusão de trabalhos (Biolchini *et al*. 2005), assim como as questões que guiam o trabalho de revisão. As questões definidas para a revisão aqui apresentada estão listadas a seguir:

- Questão Principal: Quais são as técnicas e algoritmos utilizados na predição do desempenho escolar?
- Questões Secundárias:
    1. Quais dados são importantes para essa predição?
    2. Quais os mecanismos para avaliar os resultados da predição?
    3. Quais as vantagens de utilizar uma técnica ou algoritmo para predição?

Na etapa de planejamento foi decidido que só seriam considerados os trabalhos publicados entre 2008 e 2022 nos repositórios que contém o maior número de trabalhos na área de tecnologia sendo: ACM Digital Library[1] (ACM), IEEE Xplore Digital Library[2] (IEEE), Revista Brasileira de Informática na Educação (RBIE)[3], Simpósio Brasileiro de Informática na Educação (SBIE)[4] e Springer Link[5] (SPRINGER).

[1] https://dl.acm.org/
[2] https://ieeexplore.ieee.org/Xplore/home.jsp
[3] https://www.br-ie.org/pub/index.php/rbie
[4] https://www.br-ie.org/pub/index.php/sbie/index
[5] https://link.springer.com/

Incialmente foi realizada uma pesquisa preliminar utilizando as palavras chaves abaixo:

(education AND (data OR task) AND (mining OR processing OR analytics)) OR (performance AND prediction)

Porém com essas palavras o volume de artigos recuperados a partir das bases da ACM, IEEE e SPRINGER foi bem alto, a saber, 110.474, 31.814 e 62.848, respectivamente.

A partir do volume obtido foi então necessária uma filtragem mais elaborada das palavras-chaves com o objetivo de reduzir o volume de artigos e, também para que trouxessem um conteúdo mais relevante. A nova *string* de busca definida foi:

(students AND performance) OR (students AND prediction) OR (students AND dropout)

Também foi estipulado que se as buscas devolvessem uma grande quantidade de estudos, após a elaboração das novas palavras-chaves, como ocorrido novamente para as bases da ACM (7.108), IEEE (754) e SPRINGER (2.433), seriam considerados para leitura e análise apenas os primeiros 40 trabalhos, ordenados por data de publicação em ordem decrescente e relevância em cada uma das bases.

Por outro lado, para realizar a busca nos repositórios da RBIE e SBIE foi necessária uma tradução das palavras da língua inglesa para a portuguesa, conforme apresentado a seguir:

(estudantes *AND* desempenho) *OR* (estudantes *AND* predição) *OR* (estudantes *AND* evasão)

Os critérios de inclusão e exclusão pré-definidos para obter um conjunto de trabalhos mais relevantes dentre aqueles retornados pela busca são apresentados na Tabela 2.1.

Tabela 2.1: Critérios de seleção de trabalhos relevantes.

Critério	Identificador	Descrição
Inclusão	I1	Artigos de revistas e conferências que trabalhem com predição de desempenho em cursos presenciais.
	I2	Artigos na língua nativa (português) ou em língua estrangeira (inglês).

	I3	Artigos publicados em Anais de Conferências ou Revistas.
Exclusão	E1	Artigos voltados para cursos exclusivamente online.
	E2	Artigos que não estejam na língua nativa ou na língua inglesa.
	E3	Artigos e resumos com menos de 4 páginas.

2.2. Execução da Revisão Sistemática

Após o planejamento, os passos seguidos nesta revisão foram:

1. Em cada um dos 5 repositórios foi realizada a busca utilizando as palavras-chave, contidas no Título (*Title*), Resumo (*Abstract*) ou Corpo (*Full Text*) de cada trabalho.
2. Após a realização da busca foram aplicados os critérios de inclusão e exclusão apresentados na Tabela 1, a partir da leitura do Título, Resumo, Palavras-Chave e a Conclusão. Conforme mencionado anteriormente, foram considerados somente os 40 primeiros estudos obtidos como resultado da busca.
3. Os artigos que resultaram da aplicação dos critérios de inclusão e exclusão foram objeto de uma análise apurada para poder responder à questão primária e as questões secundárias apresentadas no início da Seção 2.1.

A Tabela 2.2 apresenta a quantidade de trabalhos recuperados, excluídos e escolhidos utilizando os critérios mencionados anteriormente e a Figura 2.1 apresenta essa informação graficamente.

Tabela 2.2: Trabalhos escolhidos e suas bases de pesquisa.

Base de Busca	Trabalhos Recuperados	Trabalhos Excluídos	Trabalhos Escolhidos
ACM	40	28	12
IEEE	40	16	24
RBIE	23	16	7
SBIE	29	19	10
SPRINGER	40	24	16
TOTAL	**172**	**103**	**69**

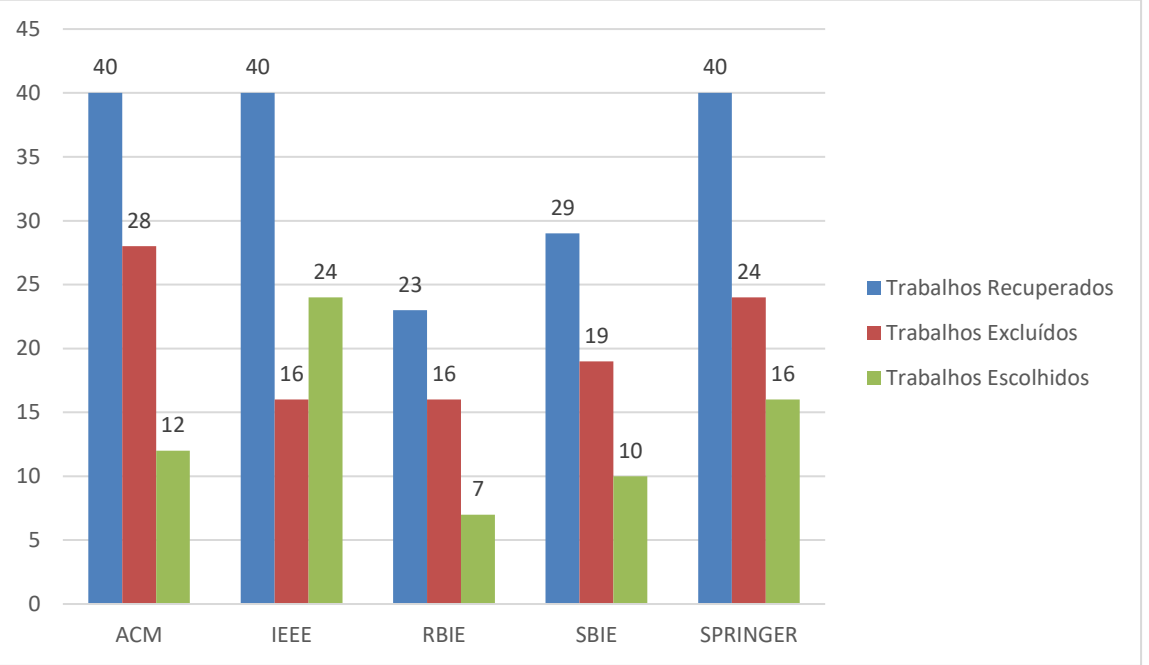


Figura 2.1: Totais de trabalhos recuperados, excluídos e escolhidos.

2.3. Respostas para a Questão Principal e as Questões Secundárias

Todos os artigos escolhidos foram analisados a partir das questões apresentadas na Seção 2.1. Com relação à questão principal, na leitura dos artigos foram identificadas diversas técnicas aplicadas para a detecção de desempenho escolar sendo elas: árvores de decisão (23%), técnicas baseadas na teoria de Bayes (18%), máquinas de vetor de suporte (14%), *k*-vizinhos mais próximos (9%), floresta aleatória (9%), redes neurais artificiais (9%), modelos de regras (5%), regressão logística (5%), ensacamento (*bagging*) (5%) e impulsionador (*boosting*) (5%). Os algoritmos abordados e aplicados são: *naive bayes* (9%), *support vector machine* (8%), J48 (6%), *random forest* (6%), *logistic regression* (5%), *multilayer perceptron* (5%), *decision tree* (3%), *JRip* (3%), *k-Nearest Neighbors* (3%), *linear regression* (3%), *random tree* (3%), *ZeroR* (3%), *ADABoost* (2%), *ADTree* (2%), *averaged one-dependence estimators* (*AODE*) (2%), *bagging* (2%), *BayesNet* (2%), *boosted decision trees* (2%), *boosting* (2%), *ClassificationViaRegression* (2%), *decision forest* (2%), *decision table* (2%), *gaussian processes* (2%), *hoeffiding tree* (2%), *IBk* (2%), *iterative classifier optimizer* (2%), *KStar* (2%), *locally weighted learning* (*LWL*) (2%), *logistic model tree* (2%), *naive bayes updateable* (2%), *OneR* (2%), *PART* (2%), *REPtree*

(2%), *ridor* (2%), *simple linear regression* (2%), *simple logistic* (2%), *Simplecart* (2%), *generalized linear model* (2%) e *deep learning* (2%).

No que concerne à questão secundária referente aos dados utilizados para a previsão de desempenho dos alunos, nos artigos analisados são utilizados dados socioeconômicos, demográficos, fatores psicológicos, culturais, regionais, vocacionais, institucionais e comportamentais para descrever o perfil do aluno. O gênero, idade/data de nascimento, estado civil, quantidade de filhos, origem, informações institucionais, participações curriculares e extracurriculares, curso, as notas/média das notas, disciplinas, frequência, conhecimentos específicos e créditos concluídos são alguns dos atributos relevantes encontrados nos artigos selecionados.

Por sua vez, para avaliar os resultados obtidos, os autores utilizam, na maioria dos trabalhos, precisão (33%), acurácia (22%), revocação (11%), medida-F (11%), curva ROC (6%), curva AUC (6%), taxa de acerto (6%), falso negativo (6%).

Por fim, o trabalho de Bujang *et al.* (2021) aborda as vantagens de utilizar uma técnica ou algoritmo que são:

- O algoritmo *J48* é amplamente utilizado em várias classificações multiclasse que podem lidar com valores ausentes e dados com muitos atributos (dimensionalidade). O algoritmo tem sido utilizado eficazmente para dar um resultado de precisão ideal com um número mínimo de características.
- A técnica de sobreamostragem (SMOTE) e a seleção de características (*Features Selection*) integradas com o algoritmo KNN permitiram obter melhora significativas na precisão dos resultados obtidos.
- O algoritmo *Logistic Regression* utiliza uma função logística para representar a modelagem matemática para resolver problemas de classificação. O modelo realiza uma grande análise contextual de dados categóricos para entender a relação entre as variáveis.
- O algoritmo *Naive Bayes*, baseado no teorema de Bayes, é amplamente utilizado, pois é simples e capaz de fazer previsões rápidas. É adequado para pequenos conjuntos de dados que combinam complexidade com um modelo probabilístico flexível.

Já o trabalho de Arun *et al.* (2021) descreve o algoritmo *Support Vector Machines* como um algoritmo supervisionado simples de aprendizado de máquina que pode ser usado tanto para classificação quanto para regressão.

2.4. Análise dos Trabalhos Escolhidos

A seguir, na Tabela 2.3, são apresentados os trabalhos mais relevantes para a pesquisa e uma descrição dos mesmos.

Tabela 2.3: Trabalhos mais relevantes

Referência	Título	Descrição
[Arun *et al.* 2021]	*Student Academic Performance Prediction using Educational Data Mining*	O trabalho tem como objetivo identificar os estudantes com riscos de obter uma baixa média de pontos. Foram utilizados vinte e três algoritmos disponíveis na ferramenta WEKA, com dados acadêmicos dos alunos dos primeiros, segundos e terceiros anos da Faculdade de Engenharia BMS, na Índia, no ramo de Ciência da Computação e Engenharia. Os desempenhos obtidos com os algoritmos *naive bayes updateable*, *hoeffding tree*, *random forest*, *logistic regression* e *classification via regression* foram os melhores, com acurácia de 86,09%.
[Brito *et al.* 2014]	Predição de Desempenho de Alunos do Primeiro Período Baseado nas Notas de Ingresso Utilizando Métodos de Aprendizagem de Máquina	A pesquisa tem como objetivo obter estimativas sobre o desempenho dos alunos nas disciplinas do primeiro período do curso de Ciência da Computação da Universidade Federal da Paraíba (UFPB), com auxílio da ferramenta WEKA, através dos algoritmos *naive bayes*, *IBk*, *support vector machine*, *random forest* e *multilayer perceptron*, sendo este último algoritmo o que obteve uma maior precisão de 75%.
[Bujang *et al.* 2021]	*Multi-class Prediction Model for Student Grade Prediction using Machine Learning*	O objetivo desse trabalho é melhorar o desempenho, de uma das escolas Politécnicas da Malásia, para prever as notas dos alunos do primeiro semestre do curso de Arquitetura de Sistemas de Computadores e Introdução a Sistemas de Computadores do Departamento de Tecnologia e Comunicação. As técnicas SMOTE (*Synthetic Minority Oversampling Technique* e seleção de características são integradas ao algoritmo KNN e apresentaram a maior precisão de 99,6%.
[Gil *et al.* 2021]	*A Data-Driven Approach to Predict First-Year Students' Academic Success in Higher Education Institutions*	Para determinar o sucesso acadêmico dos alunos dos primeiros anos foi analisado um conjunto de dados de 10 anos letivos de licenciaturas de uma Instituição de Ensino Superior. Árvore de decisão, floresta aleatória, rede neural artificial e máquina de vetor de suporte (SVM) foram as técnicas aplicadas à pesquisa. Os autores mencionam que a técnica SVM obteve melhor desempenho em comparação às outras técnicas.

[Hasib *et al.* 2022]	*A Machine Learning and Explainable AI Approach for Predicting Secondary School Student Performance*	O trabalho tem como objetivo prever o sucesso dos estudantes do ensino médio de duas escolas portuguesas, utilizando os algoritmos *Logistic Regression*, KNN, *XGBoost*, *Naive Bayes* e SVM, sendo que o último obteve a melhor acurácia (96,89%). A idade, educação da mãe, educação do pai, nota do primeiro período, nota do segundo período e nota final são alguns dos dados utilizados na pesquisa. Segundo os autores, para balancear os dados utilizaram a técnica SMOTE.
[Meedech *et al.* 2016]	*Prediction of Student Dropout Using Personal Profile and Data Mining Approach*	O trabalho se concentra no uso de algoritmos de classificação clássicos na ferramenta WEKA com a finalidade de prever o interesse e desempenho acadêmico dos alunos da Universidade Mae Fah Luang na Tailândia. As técnicas aplicadas na pesquisa foram árvores de decisão e modelos de regras. Com auxílio dá técnica SMOTE, para resolver o problema de desequilíbrio de dados, a precisão obtida, segundo autores, é em torno de 80%.
[Nahar *et al.* 2021]	*Mining Educational Data to Predict Students Performance*	Foram aplicados seis algoritmos de classificação populares, sendo *J48*, *NaiveBayes*, *PART*, *bagging*, *boosting* e *random forest*, para prever o desempenho dos estudantes de Engenharia da Universidade Notre Dame de Bangladesh, na Índia. Os algoritmos J48 e *NaiveBayes* obtiveram, respectivamente 64,3% e 75% de precisão.
[Ramaphosa *et al.* 2018]	*Educational Data Mining to Improve Learner Performance in Gauteng Primary Schools*	Neste trabalho foram utilizados dados obtidos de quatro escolas primárias na África do Sul. Os autores criaram um modelo experimental focado principalmente em analisar a precisão da previsão do desempenho acadêmico dos alunos, usando algoritmos de classificação sendo *BayesNet*, *NaiveBayes*, *JRip* e *J48*. O algoritmo J48 obteve a maior taxa de acerto de 99,13%.
[Razak *et al.* 2021]	*Prediction of Secondary Students Performance: A Case Study*	A pesquisa tem como objetivo prever a performance acadêmica dos estudantes em um curso de Matemática nas escolas secundárias Portuguesas. Para essa análise são utilizados os algoritmos *logistic regression*, *boosted decision tree*, *decision forest* e *support vector machine*. O melhor algoritmo na previsão do desempenho dos alunos é o *boosted decision tree*, com precisão de 50,6%.
[Saa *et al.* 2020]	*Mining Student Information System Records to Predict Students' Academic Performance*	O trabalho tem como objetivo prever o desempenho acadêmico dos alunos de uma universidade privada nos Emirados Árabes Unidos. *Decision tree*, *random forest*, *gradient boosted trees*, *deep learning*, *naive bayes*, *logistic regression* e *generalized linear model* são os algoritmos aplicados ao trabalho. Os resultados indicam que o algoritmo *random forest*, com 75.52% de acurácia, foi a técnica de mineração de dados mais adequada utilizada para prever o desempenho acadêmico dos alunos.
[Yağcı *et al.* 2022]	*Educational data mining: prediction of students'*	Para prever as notas dos exames finais dos alunos, neste trabalho, são utilizados e comparados os algoritmos RF,

	academic performance using machine learning algorithms	SVM, *Naïve Bayes* e *k-nearest neighbour* (KNN). O conjunto de dados utilizado nas previsões consiste das notas de desempenho acadêmico dos alunos que fizeram o curso de Língua Turca-I em uma universidade estadual na Turquia durante o semestre de outono de 2019 a 2020. Para avaliar os resultados é utilizada a acurácia, precisão, revocação, medida-F e a área sob a curva ROC. Os resultados mostram que os modelos obtidos tiveram uma acurácia de classificação de 70% a 75%.

A revisão sistemática realizada nesta pesquisa apresenta uma visão mais detalhada da aplicabilidade de técnicas, fundamentalmente da área de mineração de dados, para poder prever, o mais cedo possível, problemas que podem levar ao abandono escolar, garantindo que medidas necessárias sejam tomadas em uma tentativa de melhorar o desempenho dos alunos e eventualmente evitar desistências. Grande parte dos trabalhos abordados nesta pesquisa também identificam diversos atributos relevantes para o desenvolvimento e aproveitamento dos alunos, permitindo identificar os fatores de sucesso e insucesso específicos para cada curso e relacionar estes fatores ao currículo do curso.

3. Referencial Teórico

Nas últimas décadas houve um grande crescimento na capacidade de gerar e coletar grandes volumes de dados em virtude do avanço tecnológico advindo do aumento do poder de processamento dos computadores e da capacidade crescente de armazenamento de dados, tudo isso associado ao surgimento de novas tecnologias para a transmissão e o processamento de dados, a automação e ao uso da Internet. Os dados estão invariavelmente presentes em quantidades substanciais e à medida que o volume de dados aumenta diminui de forma alarmante o entendimento que as pessoas têm desses dados que usualmente contém informações potencialmente úteis que raramente são explicitadas ou aproveitadas (Witten *et al.*, 2011).

Nesse contexto surgiu a área de pesquisa de Descoberta de Conhecimento em Bases de Dados conhecido como *Knowledge Discovery in Databases* (KDD), cujo objetivo é desenvolver técnicas para tentar interpretar e analisar grandes volumes de dados e extrair conhecimento embutido neles de forma automática ou semiautomática. Esta ideia não é nova visto que economistas, estatísticos, analistas e engenheiros de comunicação há muito tempo trabalham com a ideia de que padrões em dados armazenados eletronicamente podem ser identificados, validados e usados para previsões. A novidade é o aumento impressionante de oportunidades para encontrar padrões nos dados e, a partir desses dados, surge o trabalho do cientista que dá sentido aos dados, descobrindo padrões que mostram como o mundo funciona e que podem ser usados para prever o que acontecerá em novas situações.

Está área surgiu como a confluência de diversas outras áreas como Bases de Dados, Estatística, Visualização de Dados, Inteligência Artificial (IA), em especial a subárea de Aprendizado de Máquina (AM).

3.1. Aprendizado de Máquina

A área de AM estuda e desenvolve métodos computacionais capazes de adquirir conhecimento, habilidades e meios para organizar o conhecimento pré-existente. Um

sistema de AM aprende a partir de fatos e/ou problemas que foram resolvidos com sucesso anteriormente (Monard e Baranauskas, 2003).

Os seres humanos possuem a habilidade de criar generalizações a partir de fatos, buscando padrões para uma coleção de observações aparentemente caóticas; essa habilidade é conseguida por meio de um processo indutivo, que é um dos tópicos centrais do aprendizado automático (Witten *et al.*, 2011).

No trabalho de Han *et al.* (2011), os autores descrevem a inferência indutiva como um dos principais métodos utilizados para derivar novo conhecimento. A indução é a forma de inferência que, a partir da capacidade de observação de um conjunto de exemplos (fatos, dados) fornecidos por um processo externo ao sistema de aprendizado, permite obter conclusões que partem do específico para o geral, da parte para o todo, das consequências ao princípio. O aprendizado indutivo pode ser dividido em supervisionado, não-supervisionado e semi-supervisionado.

O aprendizado supervisionado requer, como o nome indica, algum tipo de supervisão. Neste tipo de aprendizado é encontrado um modelo ou hipótese (função) que descreve e distingue classes de dados ou conceitos para futuras predições. O modelo ou hipótese surge a partir da análise de um conjunto de dados ou exemplos (conjunto de treinamento), cuja classificação é previamente conhecida. Em geral, o exemplo é apresentado como um conjunto de atributos (valores de características do exemplo) e o rótulo da classe associada ao exemplo. O modelo obtido deve ser capaz de predizer a classificação de novos dados. A capacidade do modelo para predizer a classificação deve ser avaliada utilizando um conjunto de dados de teste cuja classificação é conhecida. Se os rótulos de classe são discretos, o problema é conhecido como classificação e se os rótulos preditos são contínuos fala-se de regressão (Witten *et al.*, 2011).

No aprendizado não supervisionado, também conhecido como aprendizado por observação e descoberta, exemplos não classificados são analisados e o processo de indução tenta agrupá-los com base em critérios baseados nos atributos desses dados (similaridade). Os algoritmos utilizados neste tipo de aprendizado tentam descobrir novos padrões nos dados a partir de alguma caracterização de regularidade (Witten *et al.*, 2011).

O aprendizado supervisionado, em geral, precisa de uma quantidade significativa de exemplos rotulados para a indução de um bom classificador e em muitos casos essa quantidade de exemplos não está disponível. Para lidar com estes casos foi proposto um terceiro tipo de AM denominado aprendizado semi-supervisionado que tem o potencial de reduzir a necessidade de uma quantidade considerável de exemplos rotulados com as classes (Blum e Mitchell, 1998; Matsubara, 2004).

3.2. Descoberta de Conhecimento

Uma definição comumente aceita de KDD é a de Fayyad (Fayyad *et al*. 1996): "Descoberta de conhecimento em bases de dados é o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados". Nesse cenário os dados representam um conjunto de fatos ou casos num repositório, a partir dos quais é possível achar padrões ou modelos que permitem abstrair esses dados ou um subconjunto deles. Esses padrões ou modelos poderão então ser utilizados no domínio de aplicação com certo grau de certeza, fornecendo nova informação ou conhecimento de valia.

A descoberta de conhecimento envolve várias etapas, desde a definição e compreensão do domínio do qual os dados foram coletados, a seleção e amostragem de dados, a limpeza, preparação e mineração dos dados (*Data Mining*), até finalmente a análise e interpretação dos resultados obtidos. Vale destacar que em alguns contextos o termo mineração de dados (MD) e extração ou descoberta de conhecimento em bases de dados são utilizados como termos equivalentes. Neste trabalho MD só é considerada como uma das etapas do processo de KDD, que é descrito na Figura 3.1.

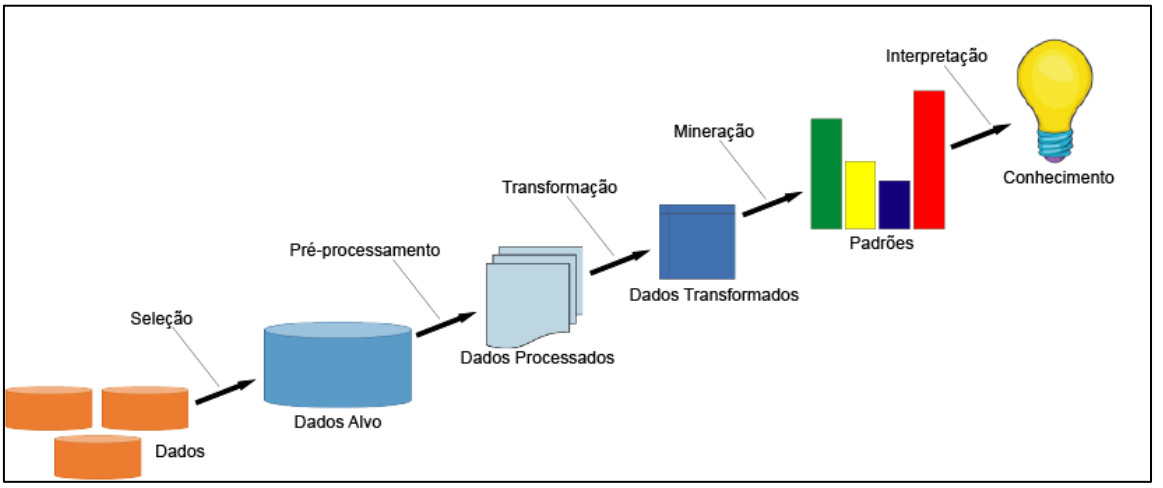


Figura 3.1: Etapas do processo de descoberta de conhecimento (Fayyad *et al.*, 1996).

Na etapa da seleção, os dados relevantes para o processo de KDD são recuperados de acordo com critérios definidos para o domínio de aplicação, ou seja, do problema. Nessa etapa é importante contar com pessoas com conhecimento do domínio pois é necessário saber o que se deseja encontrar.

O pré-processamento é a etapa na qual os dados são submetidos a um processo de limpeza por meio do qual as informações desnecessárias são removidas e há a reescrita dos dados assegurando formatos consistentes. Essa limpeza está diretamente relacionada a detectar, remover ou tratar os dados inválidos, eliminar dados inconsistentes ou duplicados que não serão utilizados durante o processo de descoberta.

No processo de transformação dos dados são usados métodos de redução de dimensionalidade ou transformação que os coloca em um formato adequado para que sejam usados pelos algoritmos da etapa da mineração. O formato depende diretamente da técnica de mineração utilizada.

Na etapa de MD é feita a extração dos padrões de comportamento dos dados utilizando algoritmos provenientes da área de AM.

Na etapa final, os resultados obtidos na mineração são interpretados e analisados os padrões identificados são interpretados com o objetivo de gerar conhecimento que dará suporte à tomada de decisão humana para a resolução de problemas.

Os algoritmos de AM utilizados na etapa de MD, na maioria das vezes, não podem utilizar os dados tais como estão armazenados nas bases de dados. Esses dados devem ser preparados previamente para deixá-los prontos para os algoritmos. As etapas que antecedem o uso dos algoritmos utilizam diferentes técnicas para pré-processar os dados. Além das técnicas, a qualidade dos dados também deve ser avaliada já que os resultados obtidos dependem, em parte, da qualidade dos dados utilizados.

Os conceitos fundamentais do pré-processamento de dados são apresentados na seção a seguir.

3.3. Pré-processamento de Dados e Qualidade dos Dados

O pré-processamento manipula e transforma os dados brutos de maneira que o conhecimento neles contido possa ser obtido de forma mais fácil e correta (Pyle, D., 1999). Os dados do mundo real são, em sua maioria, considerados “sujos”, ou seja, de baixa qualidade. A baixa qualidade dos dados ocorre quando estão incompletos, tem ruído e/ou são inconsistentes.

Os dados são incompletos quando faltam alguns dos valores para os atributos que constituem o dado, mas os atributos também podem ter valores incorretos e a origem desses problemas pode ser falha humana, de hardware, software, indisponibilidade durante a coleta dos dados.

Por sua vez, dados inconsistentes são aqueles que têm valores conflitantes entre seus atributos, por exemplo, o ano atual é o ano 2023, o atributo idade tem o valor 20, sendo que o ano de nascimento é o 2012. As inconsistências, em geral, surgem quando ocorrem violações de dependências funcionais ou surgem do processo de integração de diferentes fontes de dados.

Os dados têm ruídos quando contém erros ou apresentam instâncias que não parecem ter a mesma origem que a maioria dos dados. Os ruídos podem originar-se durante a coleta, entrada ou transmissão dos dados. Também pode haver dados redundantes ou atributos redundantes, valores de um atributo que podem ser deduzidos de um outro, como por exemplo idade e data de nascimento.

Para tratar destes problemas existem várias técnicas de pré-processamento dos dados e a partir dessas técnicas torna-se possível trabalhar com esses dados de forma menos complexa (Han *et al.*, 2011). As técnicas de pré-processamento são a limpeza de dados, integração de dados, redução de dados e transformação de dados. O que torna essas técnicas eficazes é que podem trabalhar juntas, por exemplo, a limpeza de dados pode envolver a redução dos dados para corrigir atributos ausentes. Essas técnicas quando aplicadas antes da mineração podem melhorar substancialmente a qualidade geral dos padrões extraídos e/ou o tempo necessário para a mineração (Han *et al.*, 2011).

As técnicas de pré-processamento e transformação são comumente utilizadas para aumentar o poder e a qualidade dos dados a serem minerados. Para medir a qualidade dos dados, como apontado em Witten *et al.* (2011), existem muitos fatores importantes, incluindo:

- **Precisão ou acurácia (*accuracy*):** o dado é correto, preciso.
- **Completeza (*completeness*):** o dado está disponível, registrado.
- **Consistência (*consistency*):** alguns dados são alterados, outros não.
- **Atualidade (*timeliness*):** o dado está temporalmente atualizado.
- **Credibilidade ou confiabilidade (*believability*):** o dado é verdadeiro, possível, provável.
- **Interpretabilidade (*interpretability*):** o dado pode ser interpretado de forma fácil.
- **Acessibilidade (*acessibility*):** o dado está acessível, disponível.

3.4. Principais Tarefas de Pré-processamento

As principais tarefas de pré-processamento podem ser agrupadas em tarefas de integração, limpeza, redução, amostragem, balanceamento e transformação de dados.

3.4.1. Integração de Dados

Os dados originalmente selecionados para uma aplicação de KDD podem ter sua origem em fontes diferentes (arquivos, base de dados, *data warehouse*, etc), com

organizações e formatos diferentes. Se assim for, é necessário realizar a extração e integração dos dados dessas fontes diferentes para uma única fonte. Cada objeto ou entidade de interesse para a aplicação pode ter atributos que estão presentes em fontes diferentes, eventualmente o mesmo atributo pode estar presente em fontes diferentes com nomes diferentes, um mesmo atributo pode estar representado de forma diferente (tipo de atributo) nas diferentes fontes, ou ainda um atributo pode ter sido atualizado em momentos diferentes. Tudo pode dificultar a integração e o processo pode requerer conhecimento específico do domínio de aplicação.

Como resultado da integração tem-se, em geral, um conjunto de dados (*dataset*) no qual cada dado é representado como um vetor de atributos ou tupla. A escolha de quais atributos são relevantes ou irrelevantes para a aplicação pode ser feita com o auxílio de especialistas do domínio ou utilizando técnicas de seleção de atributos (Han *et al*, 2011).

3.4.2. Limpeza de Dados

A limpeza de dados realiza o tratamento dos dados com o objetivo de melhorar sua qualidade, porém, a participação de um especialista do domínio é essencial nesta fase. Neste processo as principais tarefas são preencher os valores faltantes, identificar e suavizar os ruídos, corrigir informações errôneas ou inconsistentes e por fim resolver a redundância causada pela integração dos dados. A seguir serão abordados alguns métodos básicos para limpeza de dados.

- **Valores Ausentes:** Os valores ausentes ou dados incompletos ocorrem porque os dados nem sempre estão disponíveis, faltam valores nos atributos ou estão incompletos (Witten *et al.*, 2011). As ações comumente usadas para tratar valores ausentes devem levar em consideração a quantidade de registros afetados e a natureza do atributo envolvido, bem como os registros nos quais tal atributo comparece com valor presente. Um valor ausente caracteriza um valor ignorado ou que não foi observado, e, neste sentido, a substituição desses valores, também conhecida como imputação, tem como objetivo estimar esses valores com base nas informações disponíveis no conjunto de dados (Little *et al.*, 2010). Os métodos tradicionais de imputação são: eliminar a tupla que representa o dado, preencher

valores manualmente, usar uma constante global para preencher o valor ausente, usar uma medida de tendência central para o atributo, usar uma medida de tendência central do atributo para todas as amostras pertencentes à mesma classe, usar modelos preditivos (regressão, redes bayesianas e árvores de decisão) para preencher o valor ausente.

- **Dados com Ruídos:** Um indicador de ruído é a presença de valores que estão além dos limites ou são muito diferentes dos valores observados para um atributo. Para a detecção e remoção de ruídos podem ser usadas técnicas que utilizam algoritmos de agrupamento de dados, técnicas baseadas em distância ou particionamento, técnicas baseadas em regressão ou classificação e até técnicas provenientes da área de estatística que requerem supor que os dados correspondem a uma certa distribuição (Han *et al.*, 2011).

3.4.3. Redução de Dados

É intuitivo pensar que, quanto maior a quantidade de dados e atributos maior será a informação disponível para o algoritmo de MD, porém o aumento do volume dos atributos pode fazer com que os dados disponíveis fiquem dispersos e a análise das medidas matemáticas inconstantes, tornando o processamento dos algoritmos mais complexos assim como os modelos gerados. Para algumas aplicações torna-se necessário diminuir o volume dos dados.

Dentre os métodos de redução de dados, segundo Han *et al.* (2011), destacam-se a redução da dimensionalidade, redução de objetos e a compressão de dados.

- **Redução da dimensionalidade**: é o processo no qual ocorre a detecção de atributos irrelevantes, pouco relevantes ou redundantes, que podem ser eliminados.
- **Redução de dados**: nesta técnica o volume de dados é removido, substituído ou estimado por representações mais simples, armazenando apenas parâmetros do modelo em vez dos dados reais (métodos paramétricos) ou como agrupamento, amostragem (escolha de um subconjunto representativo de dados) e histogramas (métodos não paramétricos).

- **Compressão de dados**: são técnicas úteis para tratar a redundância de informações e ruídos, pois neste processo efetua-se a redução da dimensionalidade empregando algoritmos de transformação de dados.

3.4.4. Transformação de Dados

As bases de dados que são integradas a partir de outras bases de dados e, também, as bases de dados brutas sofrem, além dos problemas já mencionados, com a não padronização por causa de dados com diferentes escalas, unidades e com a falta de uniformidade. As técnicas de transformação de dados consolidam e transformam os dados para que o processo da mineração seja mais eficiente, mapeando todo o conjunto de valores de um dado atributo e transformando-o em um novo conjunto de valores. Os principais métodos aplicados em transformação de dados são a discretização e a normalização (Witten *et al.*, 2011).

A técnica de discretização é utilizada quando alguns algoritmos de MD não trabalham com dados numéricos contínuos, apenas com dados discretos. Nesses casos os atributos numéricos devem ser discretizados. Este método é comumente aplicado no momento da coleta de dados dividindo o domínio de um certo atributo em intervalos iguais. Os métodos mais utilizados em discretização são o particionamento (*binning*), análise de histograma, agrupamento e árvore de decisão (Han *et al.*, 2011).

Por outro lado, a normalização é uma técnica que objetiva tornar os dados mais apropriados para a aplicação de alguns algoritmos de classificação. Com esta técnica os dados são transformados para uma escala menor, não afetando a análise de dados. A normalização dos dados busca atribuir a todos os atributos um peso igual, portanto normalizar os valores de entrada dos atributos nos dados de treinamento ajudará a acelerar a fase de aprendizado (Han *et al.*, 2011).

Existem muitos métodos para normalização de dados, os métodos mais utilizados são *min-max*, *z-score* e escalonamento decimal apresentados a seguir.

- **Min-max:** Dado um atributo numérico $\boldsymbol{A}$ com $\boldsymbol{n}$ valores observados, $\boldsymbol{v_1}$, $\boldsymbol{v_2}$, ..., $\boldsymbol{v_n}$, a normalização *min-max* mapeia os valores $\boldsymbol{v_i}$ do atributo $\boldsymbol{A}$ nos valores $\boldsymbol{v_i'}$ em

uma nova escala estabelecida como [$\boldsymbol{novo\_min_A}$, $\boldsymbol{novo\_max_A}$], utilizando a Equação 3.1:

$$v_i' = \frac{v_i - min_A}{max_A - min_A} \times (novo\_max_A - novo\_min_A) + novo\_min_A \quad (3.1)$$

Sendo que min_A e max_A são, respectivamente os valores mínimo e máximo de um atributo A e a nova escala é definida pelo usuário.

Nessa normalização, o mais frequente é colocar todos os atributos de uma base de dados sob o mesmo intervalo de valores, por exemplo [0,1].

- **Z-score:** A normalização conhecida como z-score, também conhecida como normalização de média zero, é útil quando os valores mínimo e máximo de um atributo **A** são desconhecidos ou quando há *outliers* predominantes na normalização min-max. Os valores $\mathbf{v_i}$ de um atributo **A** são normalizados para $\mathbf{v_i'}$ tendo como base a média ($\boldsymbol{\mu_A}$) e desvio padrão, $\boldsymbol{\sigma_A}$conforme Equação 3.2:

$$v_i' = \frac{v_i - \mu_A}{\sigma_A} \quad (3.2)$$

- **Escalonamento Decimal:** A normalização por escalonamento decimal move as casas decimais dos valores do atributo $\boldsymbol{A}$. O número de casas decimais movidas é dependente do valor absoluto máximo de $\boldsymbol{A}$. O novo valor $\boldsymbol{v_i}$ do atributo $\boldsymbol{A}$ é normalizado para $\boldsymbol{v_i'}$, em que $\boldsymbol{j}$ é o menor inteiro tal que $\boldsymbol{max(|v_i'|) < 1}$, como mostra a Equação 3.3.

$$v_i' = \frac{v_i}{10^j} \quad (3.3)$$

4. Algoritmos de Classificação

Neste capítulo são apresentados os algoritmos utilizados na previsão do desempenho de alunos assim como alguns conceitos relacionados. A Seção 4.1 descreve o conhecimento teórico referente aos algoritmos de classificação. A Seção 4.2 descreve o algoritmo Naive Bayes. A Seção 4.3 apresenta o conceito de árvore de decisão e o algoritmo C4.5 (Quinlan 1993), cuja implementação em Java está disponível no ambiente WEKA com o nome de J48. Por último, a Seção 4.4 apresenta os conceitos de distâncias e o uso dessas distâncias no algoritmo IBk, também disponível no ambiente WEKA. A Seção 4.5 apresenta medidas para a avaliação dos resultados obtidos pelos algoritmos de classificação e, por último, a Seção 4.6 discorre sobre desbalanceamento de dados.

4.1. Classificação

A classificação é uma forma de aprendizado supervisionado que extrai modelos que atribuem classes (rótulos) a um dado de entrada. Esses modelos, chamados de classificadores, preveem rótulos de classes categóricas (discretas, não ordenadas). Muitos algoritmos de classificação foram propostos por pesquisadores em AM, reconhecimento de padrões e estatística. Um algoritmo é treinado utilizando um conjunto de dados com classes conhecidas (conjunto de treinamento), podendo estes dados estar divididos só em duas classes (classificação binária) ou em várias classes (classificação multiclasse). Para estimar o desempenho do classificador é usado um conjunto de teste e existem diversas métricas que podem ser utilizadas (Han *et al*., 2011).

A classificação tem inúmeras aplicações, incluindo detecção de fraude, marketing direcionado ao tipo de cliente, previsão de desempenho e diagnóstico médico (Han *et al*., 2011). As três seções a seguir descrevem os algoritmos utilizados neste trabalho.

4.2. Naive Bayes

O aprendizado bayesiano é um tipo de aprendizado supervisionado que faz uso de cálculo de probabilidades e fórmulas estatísticas para realizar classificação. Um dos classificadores mais simples proposto é o classificador bayesiano ingênuo (Naive Bayes),

um classificador baseado na aplicação do teorema de Bayes para o cálculo das probabilidades necessárias para determinar a classe mais provável para uma nova instância, fazendo a suposição (ingênua) de independência de valores dos atributos (Han *et al.*, 2011).

Seja X um dado representado como uma tupla com medições feitas em um conjunto de n atributos. Em termos bayesianos, X é considerado "evidência". Seja H alguma hipótese tal que a tupla de dados X pertence a uma classe específica C. Para problemas de classificação, o objetivo é determinar a probabilidade $P(H|X)$, ou seja, a probabilidade de que a hipótese H seja satisfeita dada a "evidência" ou tupla de dados observados X (Han *et al.*, 2011). As probabilidades $P(H)$, $P(X|H)$ e $P(X)$ são estimadas a partir dos dados fornecidos. O teorema de Bayes é útil porque fornece uma maneira de calcular a probabilidade $P(H|X)$, a partir de $P(H)$, $P(X|H)$ e $P(X)$ conforme a Equação 4.1:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{4.1}$$

O classificador Naive Bayes calcula a probabilidade de todas as possíveis classe e escolhe a classe com maior probabilidade como rótulo da nova instância, trabalhando da seguinte forma:

- Seja D um conjunto de tuplas de treinamento e seus rótulos de classe associados. Cada tupla é representada por um vetor $X = [x_1, x_2, ..., x_n]$, que representam os valores de cada um dos n atributos $A_1, A_2, ..., A_n$. Sejam $C_1, C_2, \dots, C_m$ as m possíveis classes associadas às tuplas.
- Dada uma tupla X com classe desconhecida, o classificador bayesiano prevê que a tupla X pertence à classe C_i se e somente se a condição estabelecida na Equação 4.2 for verdadeira:

$$P(C_i|X) > P(C_j|\mathrm{X}), \forall\, j \neq i \tag{4.2}$$

- Aplicando o Teorema de Bayes, a expressão de $P(C_i|X)$ é dada pela Equação 4.3:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (4.3)$$

Como $P(X)$ é constante para todas as classes, apenas $P(X|C_i)P(C_i)$ precisa ser maximizado assumindo, em geral, que as classes são igualmente prováveis, ou seja, $P(C_1) = P(C_2) = ... = P(C_m)$, e, portanto, o objetivo torna-se maximizar $P(X|C_i)$, sendo que as probabilidades a priori da classe podem ser estimadas por $P(C_i) = |C_{i,D}|/|D|$, onde $|C_{i,D}|$ é o número de tuplas de treinamento da classe C_i em D e $|D|$ é a quantidade de elementos do conjunto de treinamento.

- Para conjunto de dados com muitos atributos o cálculo de $P(X|C_i)$ seria extremamente caro do ponto de vista computacional e, por isso, é assumida a premissa da independência condicional de classe para os atributos conforme a Equação 4.4:

$$P(C_i) = \prod_{k=1}^{m} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times ... \times P(x_n|C_i) \quad (4.4)$$

Sendo que as probabilidades $P(x_1|C_i)$, $P(x_2|C_i)$, ..., $P(x_2|C_i)$ podem ser estimadas a partir das tuplas de treinamento. Para cada atributo, verifica-se se o atributo é categórico ou de valor contínuo.

Se A_k for categórico, então $P(x_k|C_i)$ é o número de tuplas da classe C_i em D tendo o valor x_k para A_k, dividido por $|C_{i,D}|$, o número de tuplas da classe C_i em D.

Se A_k tiver um valor contínuo, então o atributo assume tipicamente uma distribuição de probabilidade gaussiana dada pela Equação 4.5:

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma_{C_i}} e^{-\frac{(x_k-\mu_{C_i})^2}{2\sigma_{C_i}^2}} \quad (4.5)$$

Onde $g(x_k, \mu_{C_i}, \sigma_{C_i})$ é a função densidade gaussiana ou normal para o atributo A_k, enquanto μ_{C_i} e σ_{C_i} são a média e o desvio padrão, respectivamente, do atributo A_k para as amostras da classe C_i.

- Para prever o rótulo da classe de X, $P(X|C_i)P(C_i)$ é avaliado para cada classe C_i. O rótulo de classe previsto é a classe C_i para a qual $P(X|C_i)P(C_i)$ é máximo.

4.3. Árvores de Decisão

Os conceitos relacionados a árvores de decisão são baseados no trabalho de Han *et al.* (2011), que classifica a árvore de decisão como um método de aprendizado supervisionado que induz como modelo uma estrutura de árvore semelhante a um fluxograma na qual cada nó interno corresponde a uma escolha (teste de um atributo) entre várias alternativas e cada nó folha uma decisão (um rótulo de classe). Cada arco que desce de um nó interno corresponde a um dos possíveis valores desse atributo. O resultado final é uma árvore de decisão em que cada ramo representa um cenário possível de decisão e seu resultado.

Ainda, segundo Han *et al.* (2011), uma árvore de decisão pode ser usada para classificar um novo dado de classe desconhecida, basta que os valores nos atributos da árvore sejam testados e a árvore seja percorrida até encontrar um nó folha, que corresponde à classe predita para aquele dado. A construção de classificadores de árvore de decisão não requer nenhum conhecimento de domínio ou configuração de parâmetros e, portanto, é apropriada para a descoberta de conhecimento exploratório. As árvores de decisão podem lidar com dados multidimensionais. Sua representação do conhecimento adquirido na forma de árvore é intuitiva e geralmente fácil de interpretar pelas pessoas.

4.3.1. Algoritmos de Árvores de Decisão

A maioria dos algoritmos para indução de árvore de decisão adotam uma abordagem gulosa em que as árvores de decisão são construídas de uma maneira recursiva de divisão e conquista de cima para baixo. John Ross Quinlan, um pesquisador em AM, desenvolveu uma primeira versão de algoritmo de árvore de decisão conhecido como ID3 e posteriormente o aperfeiçoou dando origem ao algoritmo C4.5. Por sua vez, um grupo de estatísticos publicaram o livro *Classification and Regression Trees* (CART), que descreve a geração de árvores de decisão binárias. Os algoritmos ID3 e CART foram idealizados independentemente um do outro na mesma época, e seguem uma abordagem

semelhante para aprender árvores de decisão a partir de dados de treinamento (Han *et al.*, 2011).

Na Figura 4.1 é apresentado um algoritmo básico de árvore de decisão adaptado de Han *et al.* (2011).

Árvore_Decisão (D, LA)

Entrada: o conjunto de dados de treinamento $D = \{d_1, d_2, \dots, d_m\}$, com os rótulos de classe utilizados para gerar a árvore de decisão e LA a lista de atributos $a_1, a_2, \dots, a_k$ que compõem os dados de D. Cada $d_i \in D$ (i = 1, ..., m) é uma tupla composta pelos valores dos k atributos para esse dado e o valor da classe associada, $d = (v_1, v_2, \dots, v_k, C)$.

Saída: a árvore de decisão.

Início

Criar um nó N;

Se os dados em D são todos da mesma classe, C, **então** retornar N como um nó folha rotulado com classe C;

Se LA estiver vazia, **então** retornar N como um nó folha rotulado com a classe majoritária em D;

Aplicar o **método de seleção de atributos** (D, LA) para encontrar o "melhor" critério de divisão;

Rotular o nó N com o critério de divisão;

Se o atributo de divisão é de valor discreto **e** múltiplas divisões são permitidas **então**

Lista de atributos recebe a lista de atributos menos o atributo de divisão;

Para cada valor possível v do atributo A **fazer**

seja D_v o conjunto de dados em D que tem o valor v para o atributo A;

se D_v estiver vazio **então** associar ao nó N o rótulo da classe majoritária em D;

senão anexe o nó retornado pela geração da **árvore de decisão** (D_v, LA);

Fim para

retornar N;

Fim

Figura 4.1: Algoritmo básico de árvore de decisão.

A seguir é apresentado o processo de cálculo de ganho de informação, que é utilizado pelo algoritmo de árvore de decisão para a seleção de atributos.

4.3.2. Ganho de Informação

O ganho de informação, $Ganho(A)$ de um atributo A, em relação a um conjunto de dados D tem como objetivo determinar o melhor atributo para atribuir a um nó específico na árvore, e é definido pela Equação 4.6:

$$Ganho(A) = Info(D) - Info_A(D) \tag{4.6}$$

Neste contexto, cada dado $d \in D$ é uma tupla composta pelos valores dos k atributos para esse dado e o valor da classe dentre *n* classes possíveis.

$Info(D)$ representa a informação esperada para classificar uma tupla em D. No caso do algoritmo ID3, essa informação é calculada utilizando a entropia que caracteriza a não homogeneidade de um conjunto de tuplas com relação ao valor das classes.

A entropia para um determinado conjunto de dados D é calculada com base na distribuição de *n* classes em D, conforme Equação 4.7:

$$Info(D) = Entropia(D) = \sum_{i=1}^{n} -p_i log_2(p_i) \tag{4.7}$$

Onde p_i é a probabilidade da classe i em D, determinada pela divisão do número de dados da classe i em D pelo número total de dados do conjunto D.

Por sua vez $Info_A(D)$, conforme Equação 4.8, é a informação necessária esperada para classificar uma tupla de D com base no particionamento que os *v* valores do atributo A produzem em D. Quanto menor a informação necessária esperada, mais homogênea será a partição, gerada pela escolha do atributo A, com relação as classes (Han *et al.*, 2011).

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \tag{4.8}$$

Uma vez que o termo $\frac{|D_j|}{|D|}$ atua como o peso da j-ésima componente da partição.

O atributo com maior ganho de informação é o escolhido para rotular o nó na árvore.

4.4. O algoritmo KNN

O algoritmo KNN (*K-nearest neighbors*) pertence à família de algoritmos baseados em instâncias *(Instance-based Learning)* (Cover e Hart, 1967). Estes algoritmos se caracterizam por armazenar todos os dados de treinamento e, para classificar um novo dado, esses algoritmos recuperam um conjunto de dados similares ao novo dado e, em função desses dados recuperados, o novo dado é classificado.

Para calcular a similaridade ou proximidade entre dois dados são utilizadas métricas de distância. As distâncias mais utilizadas para o cálculo são a Euclidiana, a Manhattan e a Chebyshev.

Cada dado pode ser considerado como uma tupla que representa um ponto num espaço *m*-dimensional, sendo *m* a quantidade de atributos do dado. Assim, dados dois pontos representados pelas tuplas $P_i = (p_{i_1}, p_{i_2}, \dots p_{i_m})$ e $P_j = (p_{j_1}, p_{j_2}, \dots p_{j_m})$ pertencentes a um espaço m-dimensional, a distância Euclidiana (dist), é calculada pela Equação 4.9.

$$dist(P_i, P_j) = \sqrt{\sum_{n=1}^{m} (p_{i_n} - p_{j_n})^2} \tag{4.9}$$

Uma alternativa a esta distância é a distância Euclidiana ao quadrado, definida pela Equação 4.10. Segundo Han *et al.* (2011), a diminuição do tempo computacional para efetuar o cálculo é a grande vantagem dessa medida.

$$dist(P_i, P_j) = \sum_{n=1}^{m} (p_{in} - p_{jn})^2 \tag{4.10}$$

Uma outra distância usada é a distância Manhattan, definida pela Equação 4.11:

$$dist(P_i, P_j) = \sum_{n=1}^{m} |p_{in} - p_{jn}| \tag{4.11}$$

Por fim, a distância Chebyshev, apresentada na Equação 4.12, representa a máxima diferença absoluta entre os valores de cada atributo (Han *et al.*, 2011).

$$dist(P_i, P_j) = max_{n=1}^{m} |p_{in} - p_{jn}| \quad (4.12)$$

Essas medidas são as mais utilizadas, porém existem outras formas de calcular distância como, por exemplo, a distância Minkowski e a Mahalonobis, entre outras.

O KNN (*K-nearest neighbors*) é um algoritmo de aprendizado supervisionado amplamente usado em MD. Quando um novo dado deve ser classificado, o algoritmo KNN calcula a distância entre o novo dado e cada um dos dados do conjunto de treinamento, recuperando os *k* dados (vizinhos) mais próximos e atribui ao novo dado, a classe mais frequente entre esses *k* vizinhos (Faria, 2016).

Para determinar os vizinhos mais próximos ou similares ao novo dado a ser classificado é utilizando o conceito de distância. A medida mais usada para determinar a proximidade ou similaridade é a distância Euclidiana, mas essa escolha deve considerar o contexto da aplicação e as características dos dados. Na Figura 4.2 é apresentado o algoritmo KNN, denominado IBk no ambiente WEKA.

Classificador_KNN (D, k, t)

Entrada: D, o conjunto de tuplas que representam os dados do conjunto de treinamento, o valor de *k* e a tupla *t* que representa o novo dado a ser classificado.

Saída: a classe atribuída a *t.*

Início

Calcule $d(x, t)$, a distância entre t e cada uma das tuplas x ∈ D.

Selecione $D_t \subseteq D$, o conjunto de k tuplas de treinamento mais próximas de t.

Retorne *c'* a classe majoritária das tuplas no conjunto D_t.

Fim

Figura 4.2: Algoritmo classificador KNN (*K-nearest neighbors*) (Wu *et al.*, 2007).

4.5. Avaliação do Desempenho dos Algoritmos de Classificação

Uma vez aplicados os algoritmos, o desempenho de cada um deles para predizer a classe de um novo dado deve ser avaliado. Uma medida de desempenho usualmente utilizada é a taxa de erro calculada, em função das classificações incorretas, como a proporção de dados classificados incorretamente do conjunto de dados submetidos ao classificador. A taxa de acerto é o complemento da taxa de erro.

Uma alternativa para analisar o desempenho de um classificador é usar uma tabela chamada matriz de confusão:

- Para k classes a matriz é uma matriz de $k \times k$ que apresenta o número de predições corretas e incorretas de cada classe.
- As linhas da matriz representam as classes verdadeiras e as colunas as classes preditas pelo classificador
- O elemento m_{ij} de uma matriz de confusão M apresenta o número de dados da classe i classificados como pertencentes à classe j.
- Os elementos da diagonal principal representam os acertos e os demais campos da matriz representam os erros.

Sem perder generalidade, e por simplicidade, pode-se pensar em termos de dados positivos (os dados da classe de interesse) e os negativos (todos os outros). Nesse caso podemos falar de uma matriz de confusão de 2×2 como a apresentada na Figura 4.3.

		Classe Prevista		
		+	**-**	**Total**
Classe Verdadeira	+	VP	FN	P
	-	FP	VN	N
	Total	P’	N’	

Figura 4.3: Matriz de confusão de 2 *x* 2.

Nessa matriz, P representa a quantidade de dados positivos, N a quantidade de dados negativos, P' é a quantidade de dados que foram classificadas como positivos e N' a quantidade de dados que foram classificados como negativos. Já VP a quantidade de dados positivos que foram corretamente classificados (Verdadeiros Positivos), VN a quantidade de dados negativos que foram corretamente classificados (Verdadeiros Negativos), FP a quantidade de dados negativos que foram incorretamente classificados como positivos (Falsos Positivos) e FN a quantidade de dados positivos que foram incorretamente classificados como negativos (Falsos Negativos).

A partir da matriz de confusão podem ser calculadas várias medidas de desempenho:

- **Taxa de erro na classe negativa:** proporção de dados da classe negativa incorretamente classificados (taxa de falsos positivos – TFP).

$$TFP = \frac{FP}{FP + VN} \tag{4.13}$$

- **Taxa de erro na classe positiva:** proporção de dados da classe positiva incorretamente classificados (taxa de falsos negativos – TFN).

$$TFN = \frac{FN}{VP + FN} \tag{4.14}$$

- **Taxa de erro total (ERR):**

$$ERR = \frac{FP + FN}{P + N} \tag{4.15}$$

- **Taxa de acerto ou acurácia total (AC):**

$$AC = \frac{VP + VN}{P + N} \tag{4.16}$$

- **Precisão ($PREC$):** proporção de dados positivos classificados corretamente entre todos aqueles preditos como positivos.

$$PREC = \frac{VP}{VP + FP} \tag{4.17}$$

- **Sensibilidade ou revocação (recall):** taxa de acerto da classe positiva (taxa de verdadeiros positivos – TVP, REV).

$$REV = TVP = \frac{VP}{VP + FN} = \frac{VP}{P} \tag{4.18}$$

- **Especificidade (ESP):** taxa de acerto da classe negativa cujo complemento corresponde a TFP.

$$ESP = \frac{VN}{VN + FP} = \frac{VN}{N} = 1 - TFP \tag{4.19}$$

- **Medida-F:** é uma média harmónica ponderada da revocação e a precisão.

$$F\alpha = \frac{(1 + \alpha) \times REV \times PREC}{REV + \alpha \times PREC} \tag{4.20}$$

Se α=1 fica:

$$F1 = \frac{2 \times REV \times PREC}{REV + PREC} = \frac{2}{\frac{1}{PREC} + \frac{1}{REV}} \tag{4.21}$$

Uma outra forma alternativa de comparar classificadores é usando análise ROC **(***Receiver Operating Characteristics*) que fornece uma medida de desempenho originária da área de processamento de sinais muito utilizada nas áreas médica e biológica.

O gráfico ROC é um gráfico bidimensional em um espaço denominado espaço ROC. Os eixos X e Y representam as medidas de taxa de falsos positivos (TFP) e a taxa de verdadeiros positivos (TVP). O desempenho de um classificador pode ser plotado nessa curva equivalendo a um ponto no espaço bidimensional.

Na Figura 4.4, a diagonal representa um classificiador aleatório. O que está por baixo da diagonal é pior que o aleatório. O ponto (0,1) representa a classificação perfeita em que todos os exemplos positvos e negativos são classificados corretamente. O ponto (1,1) representa classificações sempre positivas e o (0, 0) classificações sempre negativas. Os classificadores que estão na região proxima do (1,1) quase sempre rotulam os exemplos como positivos.

Um classificador é considerado melhor que outro se seu ponto no espaço ROC encontra-se acima e a esquerda do ponto correspondente ao segundo classificador.

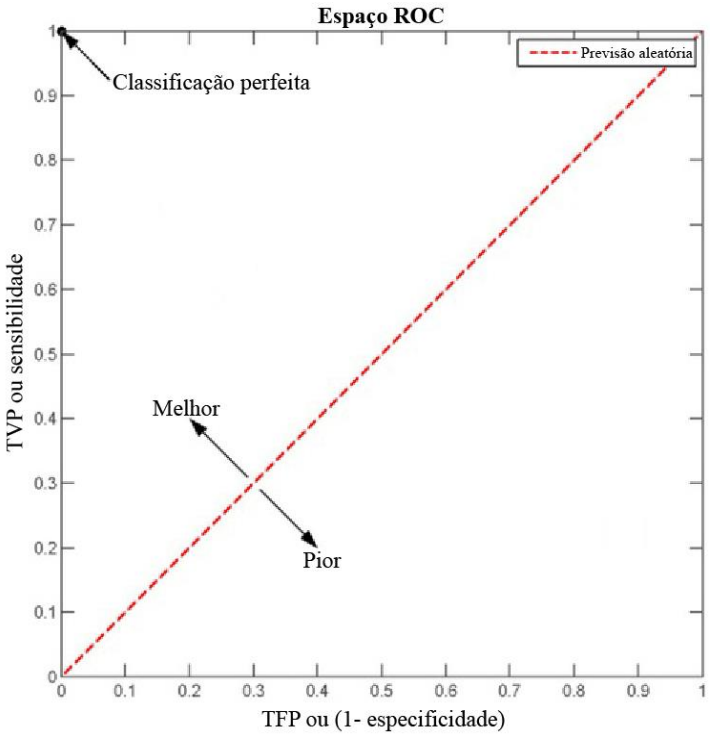


Figura 4.4: Gráfico ROC bidimensional.

4.6. Desbalanceamento de Dados

Quando o conjunto de dados utilizados por um algoritmo de classificação tem uma distribuição desbalanceada entre as classes, a capacidadede de predição desse algoritmo pode ser prejudicada com uma tendência a classificar exemplos da classes minoritárias como sendo das maioritárias. Normalmente, os resultados obtidos com esse tipo de conjunto de dados apresentam uma boa acurácia para as classes majoritárias, e uma acurácia baixa para as classes minoritárias, o que resulta um problema quando as classes de interesse são as minoritárias.

A comunidade de aprendizado de máquina abordou a questão do desbalanceamento de classes de duas maneiras. Uma consiste em reamostrar o conjunto de dados original, seja por sobreamostragem da classe minoritária e/ou subamostragem da classe majoritária (Kubat & Matwin, 1997; Japkowicz, 2000; Lewis & Catlett, 1994; Ling & Li, 1998).

Uma outra consiste em atribuir custos distintos aos dados de treinamento (Pazzani, *et al*., 1994; Domingos, 1999). Uma abordagem é a técnica de aprendizado sensível ao custo (*Cost-sensitive Learning*) e é conhecida por usar modelos que penalizam os erros gerados pelos algoritmos na classificação. Essa técnica aplica diferentes funções de custo, podendo penalizar as classes de forma igual ou diferente para cada classe. Ela pode ser utilizada em conjunto com sobreamostragem e/ou subamostragem.

4.6.1. Técnicas de Sobreamostragem para Problemas de Classificação

As técnicas de sobreamostragem, também conhecidas como *oversampling techniques*, podem modificar um conjunto de dados desbalanceado e gerar uma nova distribuição das classes minoritárias existentes tornando-as mais balanceadas (Han *et al*., 2011).

Algumas das técnicas de sobreamostragem mais conhecidas são Sobreamostragem Aleatória, Sobreamostragem Minoritária Sintética e ADASYN.

- **Sobreamostragem Aleatória:** A sobreamostragem aleatória envolve complementar os dados de treinamento com várias cópias de dados de algumas das classes minoritárias. A sobreamostragem pode ser feita mais de uma vez. Este é um dos primeiros métodos propostos e, em geral aumenta a probabilidade de ocorrer sobreajuste (*overftting*) (Ling & Li, 1998).

- **Sobreamostragem Minoritária Sintética:** A técnica SMOTE (*Synthetic Minority Oversampling Technique*), conhecida como técnica de sobreamostragem minoritária sintética é a técnica mais usada para melhorar o problema de sobreajuste (Chawla *et al.*, 2002). A técnica funciona criando dados sintéticos baseados em dados minoritários e seus *k* vizinhos mais próximos.

- **ADASYN:** A abordagem de amostragem sintética adaptativa, ou algoritmo ADASYN, baseia-se na metodologia do algoritmo SMOTE. ADASYN tem por objetivo reduzir o viés introduzido pela distribuição desbalanceada deslocando a importância do limite de classificação para aquelas classes minoritárias que são mais difíceis de serem aprendidas. O ADASYN, segundo He *et al.* (2008), usa uma

distribuição de densidade $\hat{r}_i$ como critério para decidir automaticamente o número de amostras sintéticas que precisam ser geradas para cada exemplo de dados minoritários. Fisicamente, $\hat{r}_i$ é uma medida da distribuição de pesos para diferentes exemplos de classes minoritárias de acordo com seu nível de dificuldade de aprendizado. Essa é uma grande diferença em relação ao algoritmo SMOTE, no qual são gerados números iguais de amostras sintéticas para cada exemplo dos dados minoritários.

4.6.2. Técnica de Modelos Penalizados

Geralmente os modelos de aprendizado de máquina penalizam erros de classificação da mesma forma para todas as classes, isto é, classificar *X* como *Y* ou *Y* como *X* será penalizado da mesma maneira pela função de custo.

Para tornar um modelo mais resistente ao desbalanceamento de dados é necessário adicionar uma penalidade maior para erros de classificação dos algoritmos de aprendizado de máquina, o que faz com que o algoritmo "tome mais cuidado" em errar a classe predita, evitando que o modelo entregue resultados que ignoram as classes com menos observações.

O *cost-sensitive learning* é um subcampo do aprendizado de máquina que opera com o custo de predições incorretas no treinamento de um modelo. O principal objetivo do *cost-sensitive learning* é atribuir diferentes custos para os diferentes tipos de classificações incorretas que podem acontecer em dados desbalanceados.

Em classificação, o *cost-sensitive learning* usa a matriz de custos, em cada posição na matriz contém os valores dos pesos atribuídos a cada classificação. Esses pesos tem como objetivo penalizar o algoritmo por fazer predições incorretas, e o valor das predições incorretas devem ser sempre maiores que as predições corretas.

5. Considerações sobre o Conjunto de Dados Utilizados e Configurações do Ambiente WEKA

Neste capítulo na Seção 5.1, é apresentado o conjunto de dados com os quais foi realizado o trabalho de pesquisa. Inicialmente é feita uma descrição dos dados disponibilizados pela instituição de ensino e é feita uma descrição de cada uma das partes que constituem os dados (atributos) utilizados no trabalho. Logo após é apresentada a análise e preparação do conjunto de dados, na Seção 5.2 e nas Subseções 5.2.1, 5.2.2, 5.2.3 e 5.2.4, são apresentadas as técnicas de limpeza, integração, transformação e redução de dados, respectivamente. Por último na Seção 5.3 é apresentada uma introdução ao ambiente WEKA[6] utilizado para executar os algoritmos de classificação.

5.1. Conjunto de Dados

Os dados utilizados na pesquisa são dados provenientes de um curso de Informática para Internet Integrado ao Ensino Médio da Escola Técnica Estadual (ETEC) Bartolomeu Bueno da Silva – Anhanguera. Os dados são originários do primeiro, segundo e terceiro ano do curso, e foram obtidos a partir de planilhas do Microsoft Excel fornecidas pela instituição e correspondem a dados coletados, após o ano letivo, entre 2014 e 2020.

Os conjuntos de dados originais, ou seja, que foram disponibilizados pela instituição e não sofreram nenhuma alteração estavam disponíveis em quatro conjuntos: um conjunto de dados com informações pessoais dos alunos, conhecido como dados socioeconômicos e os outros três conjuntos de dados, conhecidos como dados acadêmicos, que apresentam dados separados pelos anos em específico (primeiro, segundo e terceiro ano), em sua maioria, referentes as notas dos alunos das disciplinas do eixo comum e as disciplinas técnicas e podem ter um dos seguintes valores: “I” quando o aluno apresenta conceito “Insuficiente” na disciplina, ou seja, quando o aluno não atinge o mínimo necessário para aprovação, “R” quando o aluno tem o conceito “Regular” na disciplina, “B” quando o aluno tem o conceito “Bom”, “MB” quando o aluno tem o conceito “Muito Bom”, “NA” quando o aluno não realizou as atividades/provas da disciplina ou seja

[6] https://www.cs.waikato.ac.nz/ml/weka/

quando o conceito é “Não Avaliado” e por fim “AE” (“Aproveitamento de Estudos”) quando o aluno é transferido para a instituição ou quando o aluno reprova e retorna à instituição para cursar novamente no próximo ano, então possui notas reaproveitadas na disciplina já cursada e aprovada anteriormente.

O primeiro conjunto de dados, conhecido como dados socioeconômicos, contém os dados que contém informações pessoais dos alunos matriculados no curso de Informática para Internet Integrado ao Ensino Médio da Escola Técnica Estadual (ETEC) Bartolomeu Bueno da Silva – Anhanguera. Os dados correspondem aos 80 alunos que a instituição comumente acolhe no primeiro ano e aos alunos regularmente matriculados no segundo e terceiro ano. Estes dados estão armazenados em planilhas do Microsoft Excel e correspondem a um total de 571 alunos, cada um deles representado por valores associados a 43 atributos. Os atributos e suas descrições são listados na Tabela 5.1.

Tabela 5.1: Atributos dos dados socioeconômicos.

Atributo	Descrição do Atributo	Formato
RM	Registro de matrícula do aluno.	Numérico
NOME	Nome do aluno.	Nominal
RG	Registro geral do aluno com a unidade federativa.	Nominal
RNE	Carteira de identidade do aluno para estrangeiros.	Nominal
RA	Registro acadêmico do aluno.	Numérico
CPF	Cadastro de pessoa física do aluno.	Numérico
DN	Data de nascimento do aluno.	Numérico
IDADE	Idade do aluno.	Numérico
FONE	Telefone do aluno.	Numérico
CEL	Telefone celular do aluno.	Numérico
MAIL	E-mail pessoal do aluno.	Nominal
IMAIL	E-mail institucional do aluno.	Nominal
NR	Nome do responsável pelo aluno.	Nominal
ER	E-mail do responsável pelo aluno.	Nominal
TR1	Primeiro telefone do responsável pelo aluno.	Numérico
TR2	Segundo telefone do responsável pelo aluno.	Numérico
CR	Telefone celular do responsável pelo aluno.	Numérico
RGR	Registro geral do responsável pelo aluno.	Nominal
CPFR	Cadastro de pessoa física do responsável pelo aluno.	Numérico

END	Endereço do aluno.	Nominal
APTO	Variável utilizada quando o aluno reside em apartamento.	Nominal
BL	Variável para aluno que reside em apartamento, armazena o bloco.	Nominal
BAIRRO	Bairro ao qual reside o aluno.	Nominal
CIDADE	Cidade com a unidade federativa.	Nominal
CEP	Código de endereçamento postal do aluno.	Numérico
DIST	Distância da residência do aluno até a unidade escolar em KMs.	Numérico
SEXO	Gênero do aluno.	Nominal
NP	Nome do pai do aluno.	Nominal
NM	Nome da mãe do aluno.	Nominal
AFRO	Se o aluno é considerado afrodescendente. (Sim ou não ou descrição textual)	Nominal
EP	Se o aluno estudou em escola pública. (Sim ou não ou descrição textual)	Nominal
NAT	Naturalidade do aluno.	Nominal
RC	Raça ou cor do aluno.	Nominal
QPCF	Quantidade de pessoas que compõe a família do aluno.	Nominal
PFEAR	Quantidade de pessoas que exercem atividade remunerada na família do aluno.	Nominal
RFSM	Renda familiar do aluno em salários-mínimos.	Nominal
GR	Direito a guarda religiosa.	Nominal
PS	Se o aluno apresenta ou não algum problema de saúde.	Nominal
MED	Se necessário qual medicação o aluno utiliza.	Nominal
ECE	Informações de contato para emergências referentes ao aluno.	Nominal
DEF	Se o aluno apresenta ou não alguma deficiência.	Nominal
QDEF	Qual deficiência é apresentada pelo aluno.	Nominal
NOMEREG	Fator de identificação pessoal do aluno com relação ao gênero.	Nominal

Os dados socioeconômicos contêm alguns valores sensíveis (dados pessoais dos alunos e dos pais/parentes), porém não foram incorporados ao conjunto de dados final utilizado para treinamento e testes dos algoritmos de mineração de dados, sendo esses valores irrelevantes para a pesquisa que visa identificar padrões com a finalidade de prever o desempenho de alunos. Estes atributos serão eliminados no processo de preparação dos dados.

Por sua vez, os conjuntos de dados dos anos em específico (primeiro, segundo e terceiro ano), denominados dados acadêmicos, também estão armazenados em planilhas

do Microsoft Excel. No total são 477 de alunos do primeiro ano cada um deles com 28 atributos**,** 396 alunos no segundo ano com 27 atributos cada e 385 alunos com 27 atributos no terceiro ano, sendo que o atributo da classe “RESULTADO” também está presente em cada conjunto de dados de cada ano em específico da instituição.

Os atributos dos dados acadêmicos referentes ao primeiro ano são apresentados na Tabela 5.2.

Tabela 5.2: Atributos dos dados acadêmicos dos alunos do primeiro ano.

Atributo	Descrição do Atributo	Formato
RM	Registro de matrícula do aluno.	Numérico
NOME	Nome do aluno.	Nominal
TURMA	Turma a qual o aluno está matriculado.	Nominal
AULDAD	Quantidade de aulas ministradas.	Numérico
FALTAS	Quantidades de falta do aluno referentes a todas as disciplinas.	Numérico
FREQ	Frequência do aluno baseado nas aulas ministradas e faltas.	Numérico
PORT	Nota na disciplina de Português.	Nominal
ING	Nota na disciplina de Inglês.	Nominal
ART	Nota na disciplina de Artes.	Nominal
EF	Nota na disciplina de Educação Física.	Nominal
HIS	Nota na disciplina de História.	Nominal
GEO	Nota na disciplina de Geografia.	Nominal
FILO	Nota na disciplina de Filosofia.	Nominal
SOCI	Nota na disciplina de Sociologia.	Nominal
FIS	Nota na disciplina de Física.	Nominal
QUI	Nota na disciplina de Química.	Nominal
BIO	Nota na disciplina de Biologia.	Nominal
MAT	Nota na disciplina de Matemática.	Nominal
LP	Nota na disciplina de Lógica de Programação.	Nominal
IMC	Nota na disciplina de Instalação e Manutenção de Computadores.	Nominal
OSA	Nota na disciplina de Operações de Software Aplicativos.	Nominal
ECO	Nota na disciplina de Ética e Cidadania Organizacional.	Nominal
AD	Nota na disciplina de Aplicativos de Design.	Nominal
GSO	Nota na disciplina de Gestão de Sistemas Operacionais.	Nominal
PP	Identificação das disciplinas parcialmente pendentes, ou seja, que o aluno obteve menção insatisfatória nas disciplinas do ano anterior	Nominal

	(neste atributo os dados são sempre zero, pois os alunos ingressantes eram de outra instituição).	
RDSSEA	Resumo referente as orientações pedagógicas e comportamentais do aluno na instituição, tanto nas atividades de classe como extraclasse.	Nominal
RESULTADO	Resultado do aluno quanto à sua matrícula na instituição no ano.	Nominal
SITUACAO	Situação do aluno quanto à sua matrícula na instituição durante todo o curso.	Nominal

No conjunto de dados acadêmicos do primeiro ano, a disciplina do eixo comum de artes, representada pelo atributo "ART", assim como as disciplinas técnicas representadas pelos atributos "LP", "IMC", "OSA", "ECO", "AD" e "GSO", são especificas para o primeiro ano da instituição.

Os atributos dos dados acadêmicos referentes aos alunos do segundo ano são apresentados na Tabela 5.3.

Tabela 5.3: Atributos dos dados acadêmicos dos alunos do segundo ano.

Atributo	Descrição do Atributo	Formato
RM	Registro de matrícula do aluno.	Numérico
NOME	Nome do aluno.	Nominal
TURMA	Turma a qual o aluno está matriculado.	Nominal
AULDAD	Quantidade de aulas ministradas.	Numérico
FALTAS	Quantidades de falta do aluno referentes a todas as disciplinas.	Numérico
FREQ	Frequência do aluno baseado nas aulas ministradas e faltas.	Numérico
PORT	Nota na disciplina de Português.	Nominal
ING	Nota na disciplina de Inglês.	Nominal
ESP	Nota na disciplina de Espanhol.	Nominal
EF	Nota na disciplina de Educação Física.	Nominal
HIS	Nota na disciplina de História.	Nominal
GEO	Nota na disciplina de Geografia.	Nominal
FILO	Nota na disciplina de Filosofia.	Nominal
SOCI	Nota na disciplina de Sociologia.	Nominal
FIS	Nota na disciplina de Física.	Nominal
QUI	Nota na disciplina de Química.	Nominal
BIO	Nota na disciplina de Biologia.	Nominal
MAT	Nota na disciplina de Matemática.	Nominal
DDW	Nota na disciplina de Desenvolvimento e Design de Websites.	Nominal
CPA	Nota na disciplina de Composição, Projeto e Animação.	Nominal

MDBD	Nota na disciplina de Modelagem e Desenvolvimento de Banco de Dados.	Nominal
FRLR	Nota na disciplina de Fundamentos de Redes Locais e Remotas.	Nominal
PWI	Nota na disciplina de Programação para Web I.	Nominal
PP	Identificação das disciplinas parcialmente pendentes, ou seja, que o aluno obteve menção insatisfatória nas disciplinas do primeiro ano, após o ano corrente levito.	Nominal
RDSSEA	Resumo referente as orientações pedagógicas e comportamentais do aluno na instituição, tanto nas atividades de classe como extraclasse.	Nominal
RESULTADO	Resultado do aluno quanto à sua matrícula na instituição no ano.	Nominal
SITUACAO	Situação do aluno quanto à sua matrícula na instituição durante todo o curso.	Nominal

Nos dados acadêmicos do segundo ano, a disciplina do eixo comum de espanhol, representada pelo atributo “ESP”, assim como as disciplinas técnicas representadas pelos atributos “DDW”, “CPA”, “MDBD”, “FRLR” e “PWI”, são específicas do segundo ano da instituição.

Por fim, os atributos dos dados acadêmicos do terceiro ano são apresentados na Tabela 5.4.

Tabela 5.4: Atributos dos dados acadêmicos dos alunos do terceiro ano.

Atributo	Descrição do Atributo	Formato
RM	Registro de matrícula do aluno.	Numérico
NOME	Nome do aluno.	Nominal
TURMA	Turma a qual o aluno está matriculado.	Nominal
AULDAD	Quantidade de aulas ministradas.	Numérico
FALTAS	Quantidades de falta do aluno referentes a todas as disciplinas.	Numérico
FREQ	Frequência do aluno baseado nas aulas ministradas e faltas.	Numérico
PORT	Nota na disciplina de Português.	Nominal
ING	Nota na disciplina de Inglês.	Nominal
EF	Nota na disciplina de Educação Física.	Nominal
HIS	Nota na disciplina de História.	Nominal
GEO	Nota na disciplina de Geografia.	Nominal
FILO	Nota na disciplina de Filosofia.	Nominal
SOCI	Nota na disciplina de Sociologia.	Nominal
FIS	Nota na disciplina de Física.	Nominal
QUI	Nota na disciplina de Química.	Nominal

BIO	Nota na disciplina de Biologia.	Nominal
MAT	Nota na disciplina de Matemática.	Nominal
PWII	Nota na disciplina de Programação para Web II.	Nominal
AW	Nota na disciplina de Aplicativos para Web.	Nominal
EI	Nota na disciplina de Empreendedorismo e Inovação.	Nominal
MW	Nota na disciplina de Marketing para Web.	Nominal
PAW	Nota na disciplina de Projeto de Aplicações para Web	Nominal
PDTCC	Nota na disciplina de Planejamento e Desenvolvimento do Trabalho de Conclusão de Curso	Nominal
PP	Identificação das disciplinas parcialmente pendentes, ou seja, que o aluno obteve menção insatisfatória nas disciplinas do primeiro e segundo ano após o ano corrente levito.	Nominal
RDSSEA	Resumo referente as orientações pedagógicas e comportamentais do aluno na instituição, tanto nas atividades de classe como extraclasse.	Nominal
RESULTADO	Resultado do aluno quanto a sua matrícula na instituição no ano.	Nominal
SITUACAO	Situação do aluno quanto à sua matrícula na instituição durante todo o curso.	Nominal

As disciplinas técnicas representadas pelos atributos “AW”, “EI”, “MW”, “PAW” e “PDTCC” na Tabela 5.4 são especificas do terceiro ano da instituição.

Na próxima seção será abordado o processo de preparação realizado para obter os dados utilizados pelos algoritmos de MD.

5.2. Processo de Análise e Preparação dos Dados

Nesta seção serão detalhados os processos de análise e preparação dos dados dos alunos fornecidos pela instituição. Serão aplicadas as técnicas de limpeza de dados, integração de dados, transformação dos dados e redução de dados. Comumente a etapa de limpeza de dados é a primeira aplicada com a finalidade de tratar valores ausentes e ruídos, assim evitando que esses problemas venham a interferir nas próximas etapas conduzidas neste trabalho.

5.2.1. Limpeza de Dados

A primeira técnica aplicada foi a imputação de dados para suprir a ausência de valores de atributos. O método de imputação foi aplicado a alguns atributos conforme apresentados na Tabela 5.5.

Tabela 5.5: Atributos com seus respectivos valores e imputação de dados para valores ausentes ou dados com ruídos.

Atributo	Descrição do Atributo	Valores	Imputação
DIST	Distância da residência do aluno até a unidade escolar em KMs.	Todos os 571 alunos presentes no banco de dados apresentaram valores nulos para este atributo.	Aplicada uma técnica no Microsoft Excel com a API do GOOGLE MAPS, assim imputando os valores ausentes a partir do atributo "CEP" do aluno com a finalidade de calcular a distância até a instituição.
RC	Raça ou cor do aluno.	Do total de alunos, 568 apresentavam valores válidos, os 3 alunos restantes apresentaram valores ausentes ou dados ruidosos.	O valor de imputação usado para a raça nos dados ausentes e dados "não declara" foi "branca", este valor foi obtido pelo cálculo da moda para os valores distintos encontrados nos dados.
QPCF	Quantidade de pessoas que compõe a família do aluno.	567 alunos apresentavam valores válidos e 4 alunos restantes apresentaram valores ausentes ou dados ruidosos.	O valor de imputação aplicado foi "4a6" e foi obtido pelo cálculo da moda do número de membros da família.
PFEAR	Quantidade de pessoas que exercem atividade remunerada na família do aluno.	472 alunos apresentavam valores válidos e os 99 alunos restantes apresentaram valores ausentes ou dados ruidosos.	O valor de imputação aplicado foi 2 e este valor também foi obtido pelo cálculo da moda da quantidade de pessoas que exercem atividade remunerada na família do aluno.
RFSM	Renda familiar do aluno em salários-mínimos.	562 alunos com valores válidos e 9 alunos apresentaram valores ausentes ou dados ruidosos.	O valor de imputação aplicado "[1a2[" foi relacionado ao intervalo onde o colchete terminal pertence ao intervalo 1 e o colchete aberto significa que o limite 2 não pertence ao intervalo, esse valor de intervalo é obtido pelo cálculo da moda da quantidade em salários-mínimos da renda familiar dos alunos.
PS	Se o aluno apresenta ou não algum problema de saúde.	41 alunos valores válidos e os 530 alunos restantes apresentaram valores ausentes.	O valor de imputação "nao" foi utilizado para substituir os valores ausentes, pois quando o aluno apresenta problema de saúde, o tipo de problema era apresentado em detalhe nos dados.

MED	Se necessário qual medicação o aluno utiliza.	40 alunos com valores válidos e os 531 alunos restantes apresentaram valores ausentes.	O valor de imputação "nao" foi utilizado para substituir os valores ausentes, pois quando o aluno utiliza alguma medicação este dado é apresentado nos dados.

Já nas atas, o método de imputação foi aplicado apenas ao atributo que representa pendências dos alunos de disciplinas dos anos anteriores ("PP"). As Tabelas 5.6, 5.7 e 5.8 contém, respectivamente, os dados dos alunos e a imputação para o primeiro, segundo e terceiro ano.

Tabela 5.6: Atributos com seus respectivos valores e imputação de dados para valores ausentes ou dados com ruídos do primeiro ano.

Atributo	Descrição do Atributo	Valores	Imputação
PP	Identificação das disciplinas parcialmente pendentes, ou seja, que o aluno obteve menção insatisfatória nas disciplinas do ano anterior (neste atributo os dados são sempre zero, pois os alunos ingressantes eram de outra instituição).	Não foram obtidos valores para este atributo pois o atributo PP faz referência a progressões parciais dos alunos com relação ao ano anterior cursado na instituição	O valor de imputação 0 foi utilizado para substituir todos os valores ausentes já que se os dados correspondem a alunos do primeiro ano na instituição.

Tabela 5.7: Atributos com seus respectivos valores e imputação de dados para valores ausentes ou dados com ruídos do segundo ano.

Atributo	Descrição do Atributo	Valores	Imputação
PP	Identificação das disciplinas parcialmente pendentes, ou seja, que o aluno obteve menção insatisfatória nas disciplinas do primeiro ano, após o ano corrente letivo.	Foram obtidos 8 alunos com valores válidos e os 391 alunos restantes apresentaram valores ausentes, pois quando o aluno apresenta alguma pendência, o campo é marcado com valor deste atributo.	O valor de imputação 0 foi utilizado para substituir os valores ausentes.

Tabela 5.8: Atributos com seus respectivos valores e imputação de dados para valores ausentes ou dados com ruídos do terceiro ano.

Atributo	Descrição do Atributo	Valores	Imputação
PP	Identificação das disciplinas parcialmente pendentes, ou seja, que o aluno obteve menção insatisfatória nas disciplinas do primeiro e segundo ano, após o ano corrente letivo.	Foram obtidos 9 alunos com valores nominais, os 376 alunos restantes apresentaram valores ausentes, pois quando o aluno apresenta alguma progressão parcial o campo é marcado com valor deste atributo.	O valor de imputação 0 foi utilizado para substituir os valores ausentes.

Após esta etapa, os dados estão prontos para a integração de dados que tem como objetivo agrupar os dados pessoais dos alunos com os dados acadêmicos disponibilizados nas atas.

5.2.2. Integração de Dados

A integração de dados consiste em realizar ações que permitam integrar, de forma adequada, os dados oriundos de diversas fontes. Neste trabalho a integração é aplicada após o processo de limpeza de dados; a integração ocorre entre os dados que contém a informação pessoal dos alunos e as atas do primeiro ano, do segundo ano e do terceiro ano.

O atributo “RM”, presente nos dados socioeconômicos e acadêmicos, foi utilizado como atributo chave para a integração entre os dados socioeconômicos com os dados acadêmicos de cada ano em específico. A partir deste atributo foi possível realizar a junção entre os conjuntos de dados com a finalidade de aplicar os algoritmos de mineração e assim poder auxiliar na previsão do desempenho dos alunos.

A seguir são apresentados os atributos referente aos dados socioeconômicos com os dados acadêmicos do primeiro ano, que apresentaram e, após a etapa de integração, foram obtidos 552 registros de alunos com os seguintes atributos: “RM”, “NOME”, “RG”, “RNE”, “RA”, “CPF”, “DN”, “IDADE”, “FONE”, “CEL”, “MAIL”, “IMAIL”, “NR”, “ER”, “TR1”, “TR2”, “CR”, “RGR”, “CPFR”, “END”, “APTO”, “BL”, “BAIRRO”, “CIDADE”, “CEP”, “DIST”, “SEXO”, “NP”, “NM”, “AFRO”, “EP”, “NAT”, “RC”, “QPCF”, “PFEAR”, “RFSM”, “GR”, “PS”, “MED”, “ECE”, “DEF”, “QDEF”, “NOMEREG”, “RM”, “NOME”, “TURMA”, “AULDAD”, “FALTAS”, “FREQ”, “PORT”, “ING”, “ART”, “EF”, “HIS”, “GEO”, “FILO”, “SOCI”, “FIS”, “QUI”, “BIO”, “MAT”, “LP”, “IMC”, “OSA”, “ECO”, “AD”, “GSO”, “PP”, “RDSSEA”, “RESULTADO” e “SITUACAO”.

A segunda integração, referente aos dados socioeconômicos com os dados acadêmicos do segundo ano, foram obtidos 486 registros de alunos com os seguintes atributos: “RM”, “NOME”, “RG”, “RNE”, “RA”, “CPF”, “DN”, “IDADE”, “FONE”, “CEL”, “MAIL”, “IMAIL”, “NR”, “ER”, “TR1”, “TR2”, “CR”, “RGR”, “CPFR”,

“END”, “APTO”, “BL”, “BAIRRO”, “CIDADE”, “CEP”, “DIST”, “SEXO”, “NP”, “NM”, “AFRO”, “EP”, “NAT”, “RC”, “QPCF”, “PFEAR”, “RFSM”, “GR”, “PS”, “MED”, “ECE”, “DEF”, “QDEF”, “NOMEREG”, “RM”, “NOME”, “TURMA”, “AULDAD”, “FALTAS”, “FREQ”, “PORT”, “ING”, “ESP”, “EF”, “HIS”, “GEO”, “FILO”, “SOCI”, “FIS”, “QUI”, “BIO”, “MAT”, “DDW”, “CPA”, “MDBD”, “FRLR”, “PWI”, “PP”, “RDSSEA”, “RESULTADO” e “SITUACAO”.

Por fim, os atributos referentes aos dados socioeconômicos com os dados acadêmicos do terceiro ano, após a integração dos dados, foram obtidos 387 registros de alunos com os seguintes atributos: “RM”, “NOME”, “RG”, “RNE”, “RA”, “CPF”, “DN”, “IDADE”, “FONE”, “CEL”, “MAIL”, “IMAIL”, “NR”, “ER”, “TR1”, “TR2”, “CR”, “RGR”, “CPFR”, “END”, “APTO”, “BL”, “BAIRRO”, “CIDADE”, “CEP”, “DIST”, “SEXO”, “NP”, “NM”, “AFRO”, “EP”, “NAT”, “RC”, “QPCF”, “PFEAR”, “RFSM”, “GR”, “PS”, “MED”, “ECE”, “DEF”, “QDEF”, “NOMEREG”, “RM”, “NOME”, “TURMA”, “AULDAD”, “FALTAS”, “FREQ”, “PORT”, “ING”, “ART”, “EF”, “HIS”, “GEO”, “FILO”, “SOCI”, “FIS”, “QUI”, “BIO”, “MAT”, “PWII”, “AW”, “EI”, “MW”, “PAW”, “PDTCC”, “PP”, “RDSSEA”, “RESULTADO” e “SITUACAO”.

Após a integração de dados, eles estão prontos para a etapa de transformação de dados, que tem como finalidade a padronização dos dados que é apresentada na Seção 5.2.3.

5.2.3. Transformação de Dados

O objetivo da transformação dos dados é mapear todo o conjunto de valores de um determinado atributo, substituindo-o por um novo conjunto de valores.

Nos dados fornecidos para a pesquisa, os atributos não estavam padronizados e a técnica de transformação dos dados é um dos processos fundamentais nessa etapa. Conforme apresentado na Tabela 5.9, a padronização foi aplicada aos seguintes atributos do conjunto de dados socioeconômicos dos alunos:

Tabela 5.9: Atributos socioeconômicos no processo de transformação dos dados.

Atributo	Descrição do Atributo	Tipos de dados	Valores aceitáveis
IDADE	Idade do aluno.	Numérico	{15, 16, 17, 18, 19, 20, 21, 22}
DIST	Distância da residência do aluno até a unidade escolar em KMs.	Numérico	Valores >= 0
SEXO	Gênero do aluno.	Nominal	{masculino, feminino}
AFRO	Se o aluno é considerado afrodescendente. (Sim ou não ou descrição textual)	Nominal	{sim, nao}
EP	Se o aluno estudou em escola pública. (Sim ou não ou descrição textual)	Nominal	{sim, nao}
RC	Raça ou cor do aluno.	Nominal	{“branca”, “parda”, “preta”, “amarela”, “indigena”}
QPCF	Quantidade de pessoas que compõe a família do aluno.	Nominal	{“1a3”, “4a6”, “7a9”}
PFEAR	Quantidade de pessoas que exercem atividade remunerada na família do aluno.	Numérico	Valores >= 0
RFSM	Renda familiar do aluno em salários-mínimos.	Nominal	{“[0,1[”, “[1,2[”, “[2,3[”, “[3,5[”, “[5,7[”, “[7,10[”, “[10,15[”, “[15,30[”, “[30,45] ”}
MED	Se necessário qual medicação o aluno utiliza.	Nominal	{sim, nao}
GR	Direito a guarda religiosa.	Nominal	{sim, nao}
PS	Se o aluno apresenta ou não algum problema de saúde.	Nominal	{sim, nao}
DEF	Se o aluno apresenta ou não alguma deficiência.	Nominal	{sim, nao}
FREQ	Frequência do aluno baseado nas aulas ministradas e faltas.	Numérico	Valores >= 0 e Valores <=100
RESULTADO	Resultado do aluno quanto a sua matrícula na instituição no ano.	Nominal	{“aprovado”, “reprovado”, “pendente”, “evadido”}

O atributo “RESULTADO” representa a classe associada com cada um dos registros de dados. As definições para cada um dos valores deste atributo são descritas na Tabela 5.10.

Tabela 5.10: Siglas e definições para os atributos de "RESULTADO".

Valores do atributo classe	Definição
aprovado	Aprovação no ano sem qualquer pendência.
reprovado	Reprovação no ano.
pendente	Aprovação no ano, com pendências para o próximo.
evadido	Evasão.

Por sua vez, com o intuito de refinar os resultados, levando em consideração que os alunos de segundo e terceiro ano podem apresentar pendências nos anos anteriores, analisando essa problemática foram realizadas transformações nos dados com a finalidade de padronizá-los. Os alunos do segundo ano só podem ter pendências do primeiro ano, então para eles o atributo "PP" (relativo às disciplinas cursadas com resultado de reprovação) foi dividido em 4 (quatro) novos atributos que representam a quantidade de disciplinas pendentes por área de conhecimento (humanas, exatas, biológicas e técnicas) (Tabela 5.11). Já para os dados de alunos dos terceiros anos, o atributo "PP" foi dividido em 8 (oito) novos atributos que representam a quantidade de disciplinas pendentes por área referentes às pendências do primeiro e segundo ano (Tabela 5.12).

Tabela 5.11: Subatributos do atributo PP para o segundo ano.

Atributo	Tipos de dados	Valores aceitáveis
1EIIHUMANAS	Numérico	[0;10]
1EIIEXATAS	Numérico	[0;10]
1EIIBIOLOGICAS	Numérico	[0;10]
1EIITECNICAS	Numérico	[0;10]

Tabela 5.12: Subatributos do atributo PP para o terceiro ano.

Atributo	Tipos de dados	Valores aceitáveis
1EIIHUMANAS	Numérico	[0;10]
1EIIEXATAS	Numérico	[0;10]
1EIIBIOLOGICAS	Numérico	[0;10]
1EIITECNICAS	Numérico	[0;10]
2EIIHUMANAS	Numérico	[0;10]
2EIIEXATAS	Numérico	[0;10]
2EIIBIOLOGICAS	Numérico	[0;10]
2EIITECNICAS	Numérico	[0;10]

Os atributos com o prefixo “1EII” (Primeiro Ano do ETIM de Informática para Internet) e “2EII” (Segundo Ano do ETIM de Informática para Internet) são utilizados nos nomes dos atributos que representam quantidade de pendências que o aluno apresenta para disciplinas do eixo de humanas, exatas, biológicas e disciplinas técnicas. Assim, quando o aluno frequentar o segundo ano poderá apresentar pendências nas disciplinas do primeiro ano e quando o aluno frequentar o terceiro ano poderá apresentar pendências nas disciplinas do primeiro e do segundo ano.

Após as etapas mencionadas é normalmente realizado outro processo, conhecido como normalização, já que quando são executados os algoritmos, um dos parâmetros que podem ser modificados é o de aplicar ou não normalização nos dados. Neste caso este processo é feito pela ferramenta WEKA, no filtro conhecido como “*Standardize*” (normalização *z-score*), essa opção está disponível na ferramenta WEKA em *weka.filters.unsupervised.attribute.Standardize*.

Após a etapa de transformação de dados, os dados estão prontos para o processo de redução de dados, que é um processo fundamental para evitar que os dados se tornem esparsos.

5.2.4. Redução de Dados

Com o objetivo de manter as características dos dados originais e aumentar o desempenho computacional na aplicação dos algoritmos de mineração de dados, a redução vertical de dados é aplicada, com a finalidade de obter a maior quantidade de atributos relevantes possível e suprimir atributos menos relevantes. Neste estudo, alguns atributos foram suprimidos por caracterizar dados sensíveis, sendo eles: “NOME”, “RG”, “RNE”, “RA”, “CPF”, “TEL”, “MAIL”, “IMAIL”, “NR”, “ER”, “TR1”, “TR2”, “CR”, “RGR”, “CPFR”, “END”, “APTO”, “BL”, “BAIRRO”, “CIDADE”, “NP”, “NM”, “NAT”, “ECE”. O atributo “DN” (data de nascimento) também é suprimido, pois o atributo “IDADE” contém os valores das idades dos alunos do ensino médio-técnico no ano em que o aluno estiver cursando na instituição e, neste estudo, não causa mudanças relevantes diretamente nos resultados dos algoritmos utilizados. Outros atributos suprimidos são

“QDFE” e “NOMEREG”, pois apresentam valores nulos ou alguns dados inconsistentes e não foi achado um método adequado para fazer a imputação.

Após a etapa de integração e transformação dos dados, o atributo “TURMA” foi suprimido e os atributos “AULDAD” e “FALTAS” foram suprimidos pois o atributo “FREQ” resume ambos os atributos. Por sua vez, o atributo “RDSSEA” (Resumo ao comportamento em geral do aluno) é suprimido, pois as informações apresentadas nos dados deste atributo sobre a situação escolar do aluno estão presentes nos atributos “RESULTADO” e “SITUACAO”.

Por fim, o atributo “SITUACAO” é suprimido, pois na etapa de transformação os valores são agregados aos valores do atributo “RESULTADO”, assim, resultando em novos valores para este atributo.

Como foi escolhida a ferramenta WEKA para realizar o trabalho, os dados resultantes do processo de preparação foram transformados em três arquivos no formato ARFF suportados pela ferramenta, onde cada arquivo contém os dados socioeconômicos associados aos dados acadêmicos (um para cada ano do curso) dos alunos da instituição.

A Figura 5.1 apresenta o processo que sofreram os dados disponibilizados pela instituição em planilhas EXCEL (Dados Socioeconômicos + Dados Acadêmicos), que depois de um processo de integração e pré-processamento foram transformados em três arquivos em formato ARFF para cada um dos três anos do curso.

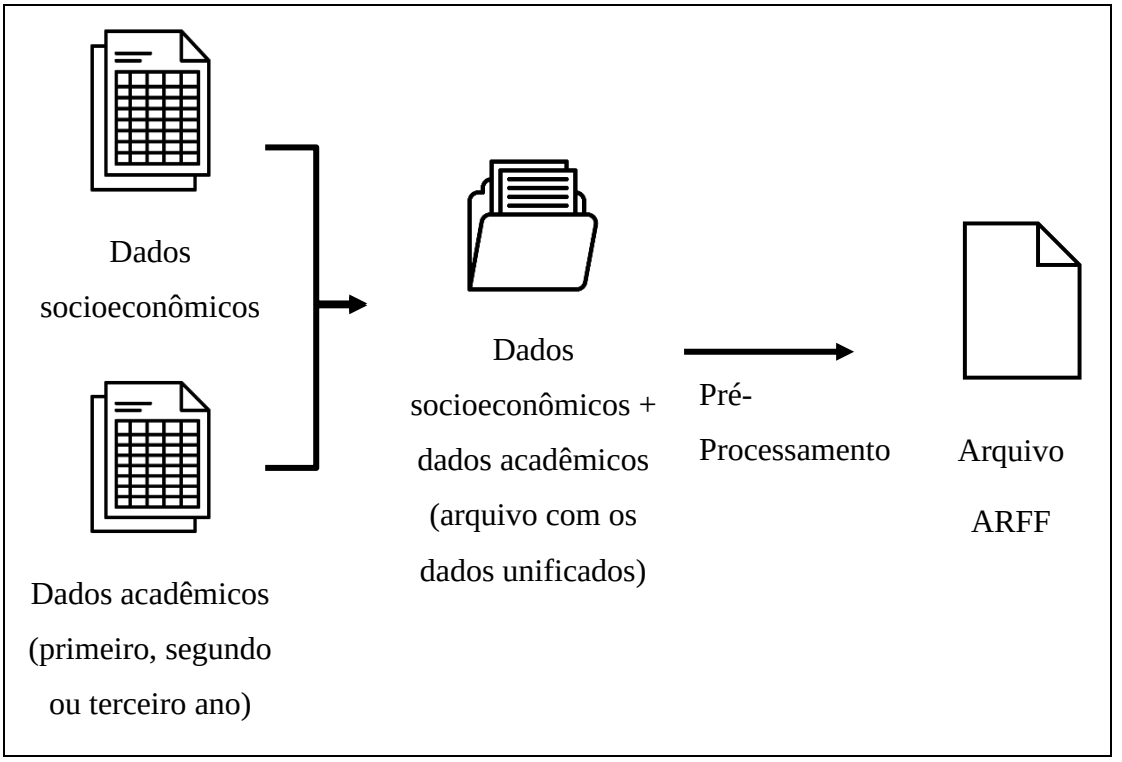


Figura 5.1: Processo de transformação dos dados dos alunos.

A Tabela 5.13 apresenta os atributos que estão presentes no arquivo de dados socioeconômicos, referentes a todos os alunos matriculados no curso.

Tabela 5.13: Atributos e descrições do conjunto de dados socioeconômicos.

Atributo	Descrição do Atributo
IDADE	Idade do aluno.
DIST	Distância da residência do aluno até a unidade escolar em KMs.
SEXO	Gênero do aluno.
AFRO	Se o aluno é considerado afrodescendente.
EP	Se o aluno estudou em escola pública.
RC	Raça ou cor do aluno.
QPCF	Quantidade de pessoas que compõe a família do aluno.
PFEAR	Quantidade de pessoas que exercem atividade remunerada na família do aluno.
RFSM	Renda familiar do aluno em salários-mínimos.
GR	Direito a guarda religiosa
PS	Se o aluno apresenta ou não algum problema de saúde
MED	Se necessário qual medicação o aluno utiliza.
DEF	Se o aluno apresenta ou não alguma deficiência.

Os atributos apresentados na Tabela 5.14 junto com os atributos apresentados na Tabela 5.13 constituem os atributos do conjunto de dados unificados (dados socioeconômicos + dados acadêmicos) dos alunos do primeiro ano.

Tabela 5.14: Atributos e descrições do conjunto de dados acadêmicos dos alunos do primeiro ano.

Atributo	Descrição do Atributo
PORT	Nota na disciplina de Português.
ING	Nota na disciplina de Inglês.
EF	Nota na disciplina de Educação Física.
HIS	Nota na disciplina de História.
GEO	Nota na disciplina de Geografia.
FILO	Nota na disciplina de Filosofia.
SOCI	Nota na disciplina de Sociologia.
FIS	Nota na disciplina de Física.
QUI	Nota na disciplina de Química.
BIO	Nota na disciplina de Biologia.
MAT	Nota na disciplina de Matemática.
ART	Nota na disciplina de Artes.
LP	Nota na disciplina de Lógica de Programação.
IMC	Nota na disciplina de Instalação e Manutenção de Computadores.
OSA	Nota na disciplina de Operações de Software Aplicativos.
ECO	Nota na disciplina de Ética e Cidadania Organizacional.
AD	Nota na disciplina de Aplicativos de Design.
GSO	Nota na disciplina de Gestão de Sistemas Operacionais.
FREQ	Frequência do aluno baseado nas aulas ministradas e faltas.
RESULTADO	Resultado do aluno quanto a sua matrícula na instituição no ano.

Já a Tabela 5.15 junto com a Tabela 5.13 compõem os atributos do conjunto de dados unificados (dados socioeconômicos + dados acadêmicos) dos alunos do segundo ano.

Tabela 5.15: Atributos e descrições do conjunto de dados acadêmicos dos alunos do segundo ano.

Atributo	Descrição do Atributo
PORT	Nota na disciplina de Português.
ING	Nota na disciplina de Inglês.
EF	Nota na disciplina de Educação Física.
HIS	Nota na disciplina de História.

GEO	Nota na disciplina de Geografia.
FILO	Nota na disciplina de Filosofia.
SOCI	Nota na disciplina de Sociologia.
FIS	Nota na disciplina de Física.
QUI	Nota na disciplina de Química.
BIO	Nota na disciplina de Biologia.
MAT	Nota na disciplina de Matemática.
ESP	Nota na disciplina de Espanhol.
DDW	Nota na disciplina de Desenvolvimento e Design de Websites.
CPA	Nota na disciplina de Composição, Projeto e Animação.
FRLR	Nota na disciplina de Modelagem e Desenvolvimento de Banco de Dados.
MDBD	Nota na disciplina de Fundamentos de Redes Locais e Remotas.
PWI	Nota na disciplina de Programação para Web I.
1EIIHUMANAS	Quantidade de disciplinas pendentes no primeiro ano para a área de humanas.
1EIIBIOLOGICAS	Quantidade de disciplinas pendentes no primeiro ano para a área de biológicas.
1EIIEXATAS	Quantidade de disciplinas pendentes no primeiro ano para a área de exatas.
1EIITECNICAS	Quantidade de disciplinas pendentes no primeiro ano para a área técnica.
FREQ	Frequência do aluno baseado nas aulas ministradas e faltas.
RESULTADO	Resultado do aluno quanto a sua matrícula na instituição no ano.

Por fim, a Tabela 5.16 junto com a Tabela 5.13 apresentam os atributos do conjunto de dados unificados (dados socioeconômicos + dados acadêmicos) dos alunos do terceiro ano.

Tabela 5.16: Atributos e descrições do conjunto de dados acadêmicos para os alunos do terceiro ano.

Atributo	Descrição do Atributo
PORT	Nota na disciplina de Português.
ING	Nota na disciplina de Inglês.
EF	Nota na disciplina de Educação Física.
HIS	Nota na disciplina de História.
GEO	Nota na disciplina de Geografia.
FILO	Nota na disciplina de Filosofia.
SOCI	Nota na disciplina de Sociologia.
FIS	Nota na disciplina de Física.
QUI	Nota na disciplina de Química.
BIO	Nota na disciplina de Biologia.

MAT	Nota na disciplina de Matemática.
PWII	Nota na disciplina de Programação para Web II.
AW	Nota na disciplina de Aplicativos para Web.
EI	Nota na disciplina de Empreendedorismo e Inovação.
MW	Nota na disciplina de Marketing para Web.
PAW	Nota na disciplina de Projeto de Aplicações para Web
PDTCC	Nota na disciplina de Planejamento e Desenvolvimento do Trabalho de Conclusão de Curso
1EIIHUMANAS	Quantidade de disciplinas pendentes no primeiro ano para a área de humanas.
1EIIBIOLOGICAS	Quantidade de disciplinas pendentes no primeiro ano para a área de biológicas.
1EIIEXATAS	Quantidade de disciplinas pendentes no primeiro ano para a área de exatas.
1EIITECNICAS	Quantidade de disciplinas pendentes no primeiro ano para a área técnica.
2EIIHUMANAS	Quantidade de disciplinas pendentes no segundo ano para a área de humanas.
2EIIBIOLOGICAS	Quantidade de disciplinas pendentes no segundo ano para a área de biológicas.
2EIIEXATAS	Quantidade de disciplinas pendentes no segundo ano para a área de exatas.
2EIITECNICAS	Quantidade de disciplinas pendentes no segundo ano para a área técnica.
FREQ	Frequência do aluno baseado nas aulas ministradas e faltas.
RESULTADO	Resultado do aluno quanto a sua matrícula na instituição no ano.

5.3. WEKA

A ferramenta WEKA (Waikato Environment for Knowledge Analysis) começou a ser desenvolvida na Universidade de Waikato na Nova Zelândia utilizando a linguagem Java em 1993. Adquirida posteriormente por uma empresa no final de 2006, WEKA encontra-se licenciada pela GNU (*General Public License*) sendo, portanto, público o acesso ao código fonte.

Uma das grandes vantagens do ambiente WEKA está relacionada à facilidade de uso em função de suas interfaces gráficas desenvolvidas em Java e às técnicas de modelagem, facilidade na portabilidade, além de ser um software livre.

Esta ferramenta tem como objetivo agrupar algoritmos provenientes de diferentes abordagens da área da inteligência artificial dedicada ao estudo de AM. A Figura 5.2 representa a tela inicial da ferramenta e disponibiliza dois ambientes, o primeiro é o WEKA *Experiment Environment* (WEE) (Figura 5.3), que é apropriado para realizar comparações entre o desempenho de vários algoritmos de mineração de dados, o outro

ambiente é o WEKA *Explorer* (WE) (Figura 5.4) que permite a seleção e execução dos algoritmos disponibilizados (Bouckaert *et al.*, 2010). Também é importante destacar que existe a opção de selecionar os tipos de filtros (*Filter*) que podem ser usados no conjunto de dados. Alguns dos principais filtros de processamento são normalização, discretização, entre outros.

Figura 5.2: Tela inicial do WEKA e opções iniciais para atividades de mineração.

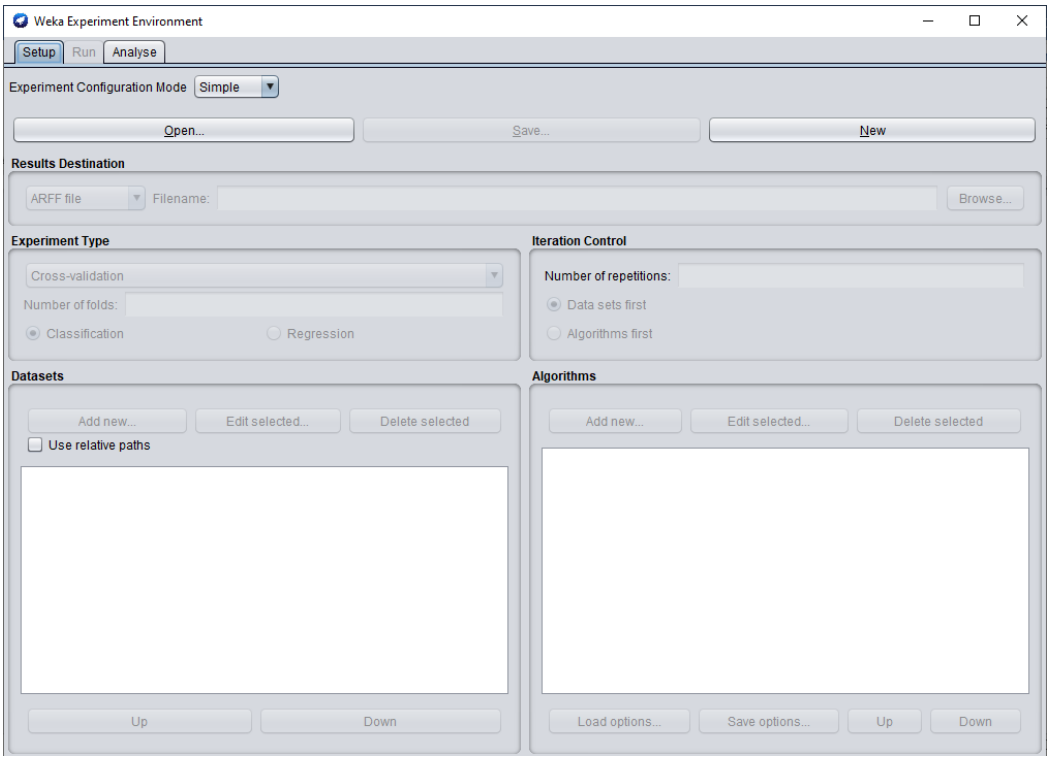


Figura 5.3: Tela do ambiente de experimento da ferramenta WEKA que realiza comparações entre algoritmos de mineração de dados.

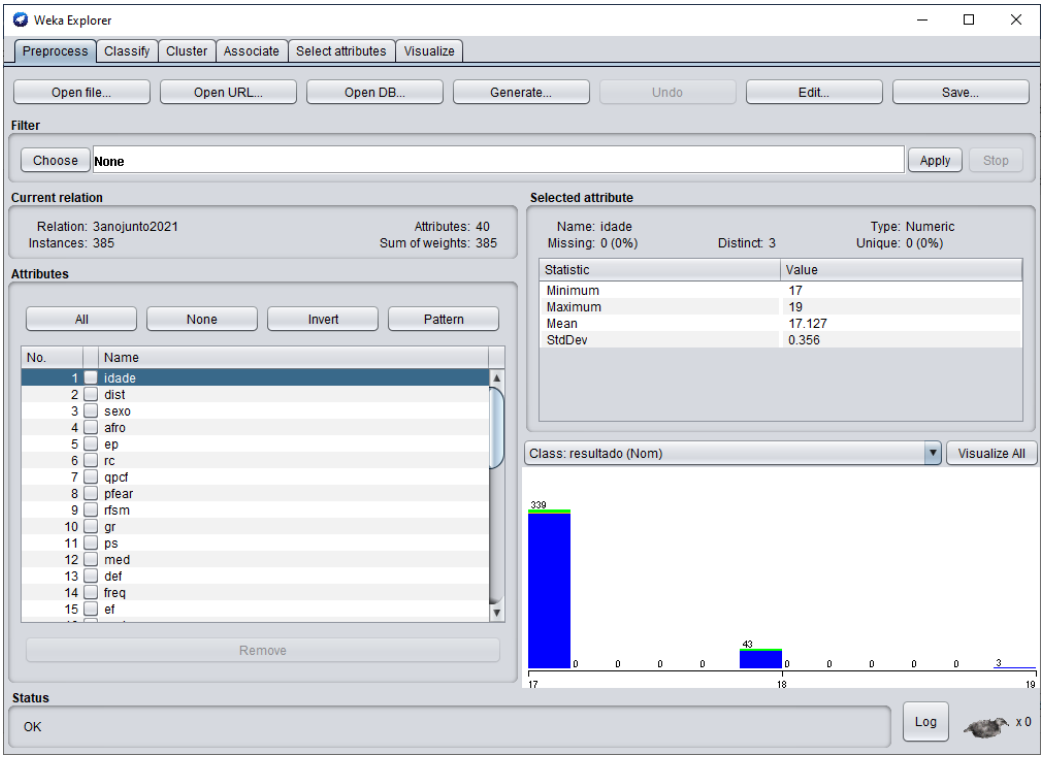


Figura 5.4: Tela do explorador do WEKA que permite a aplicação dos filtros e algoritmos no processamento.

O WEE permite selecionar um ou mais algoritmos disponíveis na ferramenta e analisar os resultados de modo a identificar se um classificador é, estatisticamente, melhor do que os demais, oferecendo três opções de estratificação da base de dados:

1. *Cross-validation* (validação cruzada, que é a opção padrão).
2. *Train/Test Percentage Split (data randomized)* (dados processados em ordem aleatória).
3. *Train/Test Percentage Split (order preserved)* (dados processados na ordem em que aparecem no arquivo de dados).

Por sua vez, no ambiente WE a comparação dos resultados dos classificadores não é efetuada de forma automática como no ambiente WEE, o resultado obtido representa a acurácia de apenas uma execução do algoritmo, oferecendo quatro opções de estratificação dos dados, conforme apresentado na Tabela 5.17.

Tabela 5.17: Opções de teste no ambiente WEKA para algoritmos de classificação.

Use training set	Utiliza a própria base de dados como treino e teste.
Supplied test set	Permite carregar outra base de teste diferente do treino, facilitando a seleção dos dados que vão compor o conjunto de treinamento e testes.
Cross-validation	Por exemplo, os dados podem ser divididos em 10 conjuntos de tamanho n/10, treinando com 9 conjuntos de dados e 1 de teste e seu desempenho final é obtido a partir da média das execuções.
Percentage split	Particiona o conjunto de dados em dois conjuntos, um conjunto para treinamento segundo uma percentagem pré-estabelecida e outro para teste sobre a percentagem restante.

Uma desvantagem do WEE está no grande consumo de memória que limita o tamanho do conjunto de dados que será processado pelos algoritmos na utilização da interface gráfica, ficando em evidência quando se tem algoritmos supervisionados que, no processamento, necessitam de uma etapa de treinamento e outra de teste. Para minimizar esse problema deve-se utilizar o modo simplificado de execução permitindo que mais memória seja direcionada para o processamento dos conjuntos de dados, ou seja, o processamento do conjunto de dados utilizados pelos algoritmos deve ser realizado sem o ambiente gráfico.

Finalizados os processos de preparação dos dados e da escolha da ferramenta a ser utilizada, é necessário selecionar os classificadores e as métricas utilizadas para avaliar o desempenho dos classificadores, tema que é abordado no Capítulo 6.

6. Experimentos e Resultados

Neste capítulo são apresentados os experimentos que foram realizados para detectar o desempenho dos alunos da ETEC Bartolomeu Bueno da Silva – Anhanguera, utilizando os algoritmos Naive Bayes, J48 e IBk em conjunto com as técnicas SMOTE (Técnica de Sobreamostragem Minoritária Sintética) e Modelo Penalizado (*Cost-sensitive Learning*) apresentados no Capítulo 4. Também são apresentados a matriz de confusão, as fórmulas utilizadas na avaliação dos classificadores, os resultados obtidos, suas respectivas interpretações e uma discussão comparativa dos resultados.

6.1. Descrição Geral dos Experimentos

Esta seção tem como objetivo apresentar a descrição geral dos experimentos. Todos os experimentos foram realizados utilizando um computador (PC) com processador AMD FX 8320e de 3,2 Giga-hertz, 16 Gigabytes de memória e HD de 200 Gigabytes SSD.

Os experimentos utilizaram os algoritmos Naive Bayes, J48 e IBk (para k = 1, 3 e 5) junto com as técnicas SMOTE e a técnica de modelo penalizado (custo), todos disponíveis no ambiente WEKA. Para a realização dos experimentos com os algoritmos de natureza supervisionada citados anteriormente foram utilizados os dados dos alunos de primeiro, segundo e terceiro ano, disponíveis em três arquivos no formato requerido pela ferramenta ("1anosrelacaoeatas.arff", "2anosrelacaoeatas.arff" e "3anorelacaoeatas.arff"). Um exemplo dos atributos e conteúdo das amostras com seus conjuntos de treinamento e teste do primeiro ano pode ser observado na Figura 6.1.

IDADE,DIST,SEXO,AFRO,EP,RC,QPCF,PFEAR,RFSM,GR,PS,MED,DEF,FREQ,PORT,ING,EF,HIS,GEO,FILO,SOCI,FIS,QUI,BIO,MAT,ART,LP,IMC,OSA,ECO,AD,GSO,RESULTADO

16,14,feminino,nao,nao,parda,[4a6],2,[1a2[,nao,nao,nao,nao,96,mb,b,b,na,mb,r,mb,mb,mb,b,b,r,mb,mb,b,b,b,0,0,0,0,**aprovado**

17,1,masculino,nao,nao,branca,[1a3],1,[2a3[,nao,nao,nao,nao,78,i,i,b,na,b,r,mb,b,r,r,b,i,r,r,b,i,r,0,0,0,0,**reprovado**

16,12,masculino,nao,sim,parda,[1a3],2,[3a5[,nao,nao,nao,nao,79,i,b,r,b,b,r,b,b,r,r,r,i,b,b,r,b,r,0,0,0,0,**pendente**

16,5,masculino,sim,sim,parda,[4a6],1,[2a3[,nao,nao,nao,nao,33,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,0,0,1,0,**evadido**

Figura 6.1: Amostras do conjunto de treinamento e teste do primeiro ano.

Foi estabelecido um esquema de trabalho que incluiu a realização de 3 experimentos:

- O primeiro experimento realizado utilizou os dados disponíveis em cada um dos três arquivos resultado do processo de preparação de dados no formato requerido pela ferramenta com extensão ".arff" e foram calculadas métricas para avaliar os classificadores. Os algoritmos utilizaram cada um dos três arquivos correspondentes aos dados dos alunos de cada uma das séries do curso.
- O segundo experimento levou em consideração que os dados disponíveis são desbalanceados com relação à distribuição das quatro classes de dados de alunos classificados como pendentes, reprovados e evadidos têm uma menor incidência com relação aos aprovados, o que pode ocasionar um viés nas predições, favorecendo as classes majoritárias. Para evitar possíveis problemas ocasionados com o desbalanceamento o segundo experimento foi realizado utilizando a Técnica de Sobreamostragem Minoritária Sintética (SMOTE) que aumenta a porcentagem apenas das classes minoritárias (He *et al.*, 2009 e Fernandez *et al.*, 2018).
- O terceiro experimento utilizou também a técnica SMOTE em conjunto com a técnica de modelos penalizados (*Cost-sensitive Learning*) (Ling *et al.*, 2010) como apresentado na Seção 6.5.

O objetivo dos experimentos é determinar qual dos algoritmos utilizados consegue classificar melhor as instâncias para os alunos que estão cursando regularmente na instituição que forneceu os dados. Esses alunos são classificados como "aprovado", "reprovado", "pendente" e "evadido" para o primeiro e segundo ano e "aprovado", "reprovado" e "evadido" para o terceiro ano. As siglas apresentadas anteriormente estão descritas na Tabela 5.10 da Seção 5.2.3.

6.2. Configuração dos Algoritmos no Ambiente WEKA

A seguir são apresentados os parâmetros básicos, a combinação dos parâmetros e detalhes de pré-processamento utilizados nos experimentos realizados neste trabalho para o funcionamento dos algoritmos Naive Bayes, J48 e IBk.

6.2.1. Configuração do Algoritmo Naive Bayes

Na parametrização do algoritmo, nos experimentos no ambiente WEKA, foi utilizada a configuração mostrada na Figura 6.2, que apresenta a tela de configuração do algoritmo.

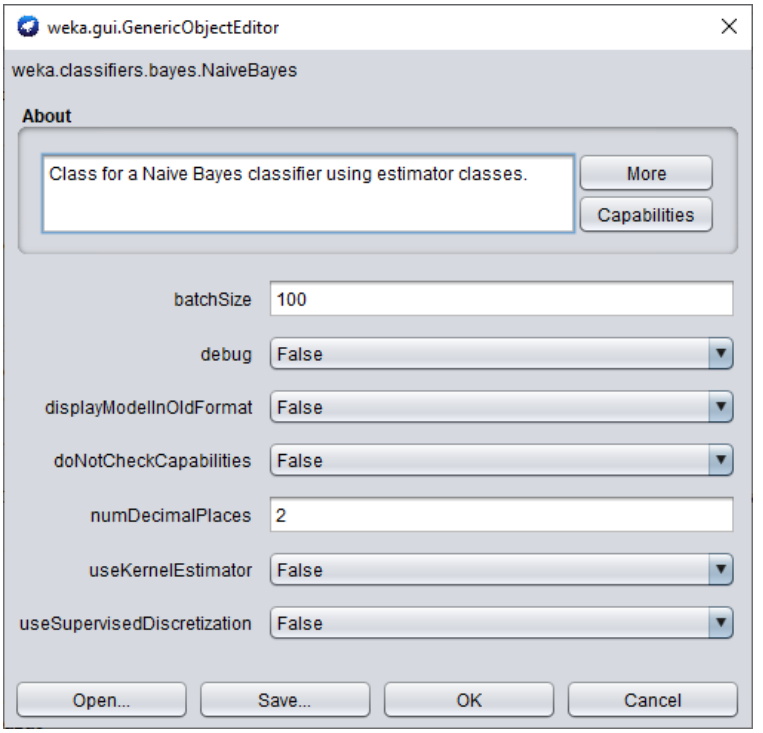


Figura 6.2: Tela de configurações do WEKA para o algoritmo Naive Bayes.

As configurações *batchSize*: 100, *debug*: *false*, *displayModelInOldFormat*: *False*, *doNotCheckCapabilities*: *False*, *numDecimalPlaces*: 2, *useKernelEstimator*: *False* e *useSupervisedDiscretization*: *False* não tiveram seus valores alterados, assumindo os valores padrão no ambiente.

6.2.2. Configuração do Algoritmo J48

No ambiente WEKA, para a parametrização do algoritmo J48, apenas o parâmetro relacionado à poda da árvore (*pruned*) e sem poda da árvore (un*pruned*) foram aplicados para testar o desempenho do algoritmo J48 em ambos os casos.

O parâmetro de configuração do algoritmo no ambiente WEKA, conforme mencionado acima, é apresentado na Figura 6.3.

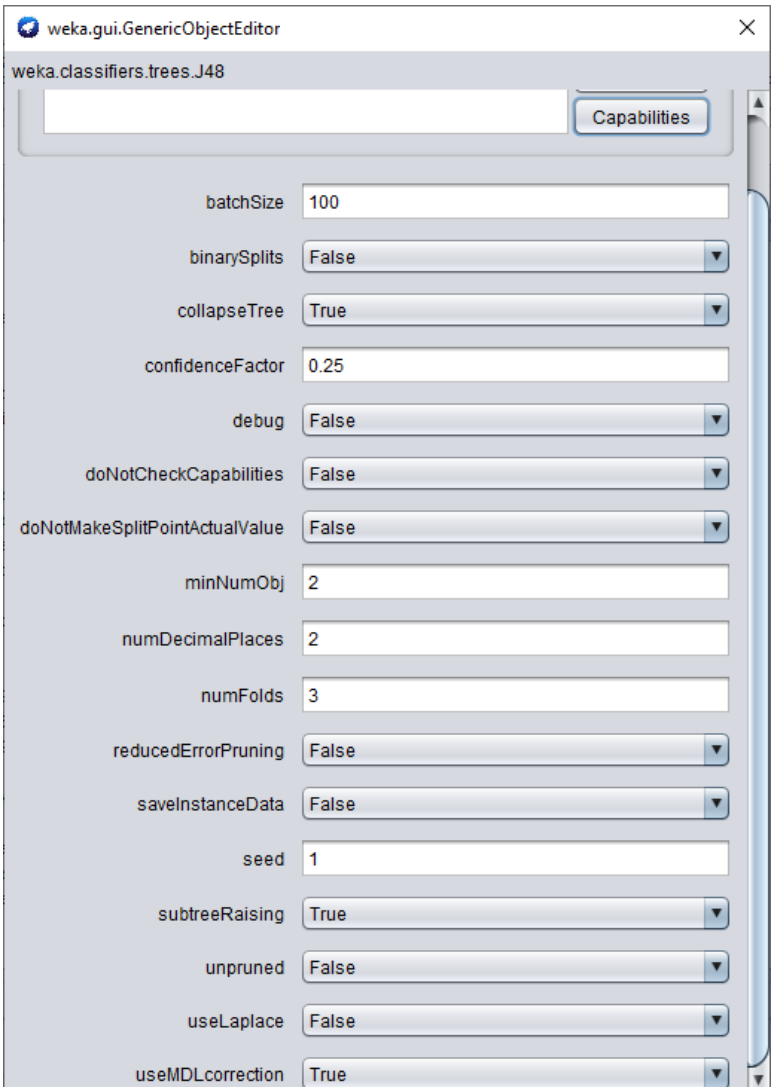


Figura 6.3: Tela de configurações do WEKA para o algoritmo J48.

As configurações *batchSize*: 100, *bynarySlits*: *False*, *collapseTree*: *True*, *confidenceFactor*: 0.25, *debug*: *False*, *doNotCheckCapabilities*: *False*, doNotMakeSplitPointActualValue: *False*, *minNumObj*: 2, *numDecimalPlaces*: 2,

numFolds: 3, *reducedErrorPruning*: *False*, *saveInstanceData*: *False*, *seed*: 1, *subtreeRaising*: *True*, *useLaplace*: *False* e *useMDLCorrection: True* não tiveram seus valores alterados, assumindo os valores padrão no ambiente.

6.2.3. Configuração do Algoritmo IBk

O algoritmo IBk calcula a distância entre a nova instância a ser classificada e as demais instâncias existentes, resgatando os *k* vizinhos mais próximos e atribuindo à nova instância à classe que ocorre com mais frequência entre esses *k* vizinhos. O valor de *k*, a função de pesquisa no algoritmo (NNSearch) e a medida de distância são parâmetros que devem ser configurados no ambiente WEKA e essa configuração está descrita na Tabela 6.1.

Tabela 6.1: Combinações de parâmetros do algoritmo IBk para os experimentos.

k	**NNSEARCH**	**DISTÂNCIA**
1	LINEARNNSEARCH	EUCLIDIANA
3		
5		

Para a configuração do algoritmo IBk no ambiente WEKA foram utilizados os parâmetros KNN: *1, 3* e *5*, *crossValidate*: *false*, *debug*: *false*, *distanceWeighting*: *no distance weighting*, *doNotCheckCapabilities*: *false* e *nearestNeighbourSearchAlgorithm*: *LinearNNSearch*.

Os parâmetros de configuração são apresentados na Figura 6.4 que representa a tela de configuração do algoritmo IBk.

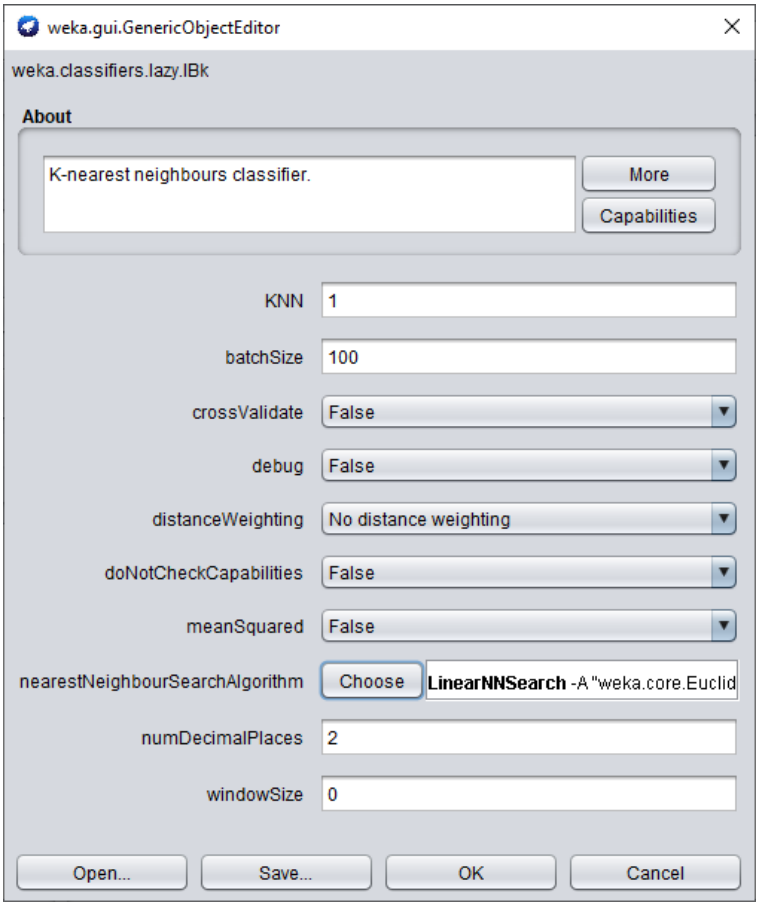


Figura 6.4: Tela de configurações do WEKA para o algoritmo IBk para o valor do parâmetro k = 1.

6.2.4. Uso de Validação Cruzada

Após estabelecer o conjunto de dados a serem utilizados nos experimentos e estabelecer a configuração geral para os algoritmos Naive Bayes, J48 e IBk, para obter estimativas confiáveis de desempenho do classificador é necessário aplicar uma técnica conhecida como validação cruzada. Essa técnica tem como objetivo validar o desempenho dos algoritmos aplicados na mineração de dados, estimando o erro dos algoritmos preditivos. Uma forma comum de validação cruzada é a chamada validação cruzada k-conjuntos (*k-fold cross-validation*). Esta técnica particiona aleatoriamente os dados iniciais em k subconjuntos, sendo que $k - 1$ subconjuntos são utilizados no treinamento e 1 é utilizado para teste. O valor $k = 10$ tem se tornado um valor padrão em termos práticos e foi o utilizado neste trabalho.

Um aspecto importante desse particionamento da base de dados entre treinamento e teste é se os subconjuntos são estratificados ou não. Na validação cruzada estratificada, os subconjuntos são estratificados de modo que a distribuição de classe das tuplas em cada subconjunto seja aproximadamente a mesma que nos dados iniciais, ou seja, distribui as classes das amostras uniformemente entre as os subconjuntos.

Em geral, a validação cruzada estratificada é recomendada para estimar o desempenho devido ao seu viés e variância relativamente baixos (Han *et al.*, 2011). Estudos empíricos também demonstram que a estratificação melhora significativamente os resultados (Maimon *et al.*, 2010).

6.3. Avaliação de Desempenho

Uma vez aplicados os classificadores, deve ser avaliado o desempenho de cada um deles. Existem diferentes métricas para avaliar os resultados obtidos e neste trabalho sendo elas a acurácia, a precisão, a sensibilidade ou revocação, a Medida-F e a área abaixo da curva ROC (AUC), como foi apresentado na Seção 4.5 do Capítulo 4. Para calcular essas métricas é necessário dispor dos valores para os VP (verdadeiro positivos), VN (verdadeiros negativos), FP (falsos positivos) e FN (falsos negativos) em função das instâncias classificadas usadas na fase de teste. Esses valores são representados por meio de uma matriz de confusão como mencionado no Capítulo 5. A seguir será apresentada a matriz de confusão com sua interpretação e as equações que definem as métricas aplicadas na avaliação dos resultados que foram obtidos nos experimentos.

6.3.1. Matriz de Confusão

No ambiente WEKA os resultados dos experimentos são exibidos com *layouts* específicos. Na Figura 6.5 é apresentado um *layout* de exibição da matriz de confusão para um experimento nesse ambiente.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	**VP (aprovado)**	FP (aprov.|reprov.)	FP (aprov.|pend.)	FP (aprov.|evad.)
	reprovado	FP (reprov.|aprov.)	**VP (reprovado)**	FP (reprov.|pend.)	FP (reprov.|evad.)
	pendente	FP (pend.|aprov.)	FP (pend.|reprov.)	**VP (pendente)**	FP (pend.|evad)
	evadido	FP (evad.|aprov.)	FP (evad.|reprov.)	FP (evad.|pend.)	**VP (evadido)**

Figura 6.5: *Layout* da matriz de confusão apresentada no ambiente WEKA para um experimento.

Na interpretação dessa matriz de confusão, no momento que é feita a análise de uma das classes, por exemplo a classe "aprovado", o VP para a classe "aprovado" é a quantidade de verdadeiros positivos para essa classe, e a soma de VP (reprovado), VP (pendente) e VP (evadido) tornam-se os verdadeiros negativos (VN) para fazer os cálculos quando é considerada a classe "aprovado" versus todas as outras classes.

6.3.2 Equações Usadas nos Experimentos

Para avaliar os resultados obtidos nos experimentos foram utilizadas as seguintes métricas:

- Média ponderada da precisão ($MP_{precisão}$), calcula a precisão para cada uma das classes e usa a média ponderada com base no número de amostras por classe.
- Média ponderada da sensibilidade ($MP_{sensibilidade}$), calcula a sensibilidade para cada uma das classes e usa a média ponderada com base no número de amostras por classe.

- Média ponderada da medida-F (MP_F), calcula a medida-F para cada uma das classes e usa a média ponderada com base no número de amostras por classe.
- Área abaixo da curva ROC (MP_{AUC}), calcula a probabilidade de que a curva ROC irá classificar um aluno corretamente.

A média ponderada da precisão é calculada, conforme apresentada na Equação 6.1, pelo somatório das multiplicações entre valores da precisão das classes e o número de instâncias da classe, divididos pela somatória da quantidade de instâncias de cada classe.

$$MP_{precisão} = \frac{(P_{aprovado} * c1) + (P_{reprovado} * c2) + (P_{pendente} * c3) + (P_{evadido} * c4)}{c1 + c2 + c3 + c4} \quad (6.1)$$

Onde:

$P_{aprovado}$ representa a precisão para a classe "aprovado", $P_{reprovado}$ representa a precisão para a classe "reprovado", $P_{pendente}$ representa a precisão para a classe "pendente", $P_{evadido}$ representa a precisão para a classe "evadido", $c1$ representa o número de instâncias da classe "aprovado", $c2$ representa o número de instâncias da classe "reprovado", $c3$ representa o número de instâncias da classe "pendente" e $c4$ representa o número de instâncias da classe "evadido".

As precisões são calculadas pelas Equações 6.2 a 6.5.

$$P_{aprovado} = \frac{VP_{aprovado}}{VP_{aprovado} + Total\ de\ FP_{reprovado,pendente,evadido}} \quad (6.2)$$

Onde:

$Total\ de\ FP_{reprovado,pendente,evadido}$ representa a soma dos valores da linha "aprovado" da matriz de confusão menos o valor presente na primeira coluna da linha.

$$P_{reprovado} = \frac{VP_{reprovado}}{VP_{reprovado} + Total\ de\ FP_{aprovado,pendente,evadido}} \quad (6.3)$$

Onde:

$Total\ de\ FP_{aprovado,pendente,evadido}$ representa a soma dos valores da linha "reprovado" da matriz de confusão menos o valor da coluna "reprovado" dessa linha.

$$P_{pendente} = \frac{VP_{pendente}}{VP_{pendente} + Total\ de\ FP_{aprovado,reprovado,evadido}} \quad (6.4)$$

Onde:

$Total\ de\ FP_{aprovado,reprovado,evadido}$ representa a soma da linha da matriz de confusão menos o valor da coluna "pendente" dessa linha.

$$P_{evadido} = \frac{VP_{evadido}}{VP_{evadido} + Total\ de\ FP_{aprovado,reprovado,pendente}} \quad (6.5)$$

Onde:

$Total\ de\ FP_{aprovado,reprovado,pendente}$ representa a soma dos valores da coluna "evadido" da matriz de confusão menos o valor da coluna "evadido" dessa linha.

A sensibilidade representa a taxa de acerto da classe positiva, ou seja, avalia a capacidade do método de detectar com sucesso resultados classificados como positivos. Neste trabalho consideramos a média ponderada da sensibilidade, calculada de maneira similar a média ponderada da precisão, em função das sensibilidades calculada para cada uma das 4 classes com relação ao resto das classes.

A média ponderada da medida-F (MP_F), também é calculada em função de cada uma das classes com relação ao resto das classes.

Para se obter a área abaixo da curva ROC (AUC) é necessário, antes, obter a curva ROC (Curva Característica de Operação do Receptor**)**, ela é calculada conforme Equação 6.8.

$$ROC = \frac{RPV}{RPF} \quad (6.6)$$

Onde RPV é representado pela sensibilidade (ou taxa de verdadeiros positivos), e RPF = 1 – especificidade ou taxa de falsos positivos, sendo que a especificidade também é conhecida como taxa de verdadeiros negativos. A área abaixo da curva ROC (AUC)

produz valores entre 0 e 1, quanto mais perto de 1 melhor e é obtida conforme apresentado na Figura 6.6.

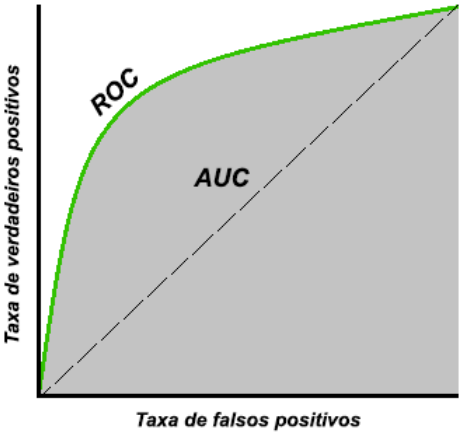


Figura 6.6: Gráfico da curva ROC e AUC.

6.4. Configuração da Técnica SMOTE e Modelo Penalizado no WEKA

Esta seção trata as configurações que devem ser feitas para aplicar a técnica SMOTE e o modelo penalizado no ambiente WEKA para cada conjunto de dados (primeiro, segundo e terceiro ano).

No ambiente WEKA a configuração do SMOTE é aplicada como uma etapa do pré-processamento após a Integração, Limpeza, Redução e Transformação de Dados, podendo ser encontrada no ambiente através do filtro "*weka/filters/supervised/instance/*SMOTE", porém a técnica não é nativa no WEKA, sendo necessário fazer a instalação de um pacote para sua utilização, conforme ilustrado na Figura 6.7.

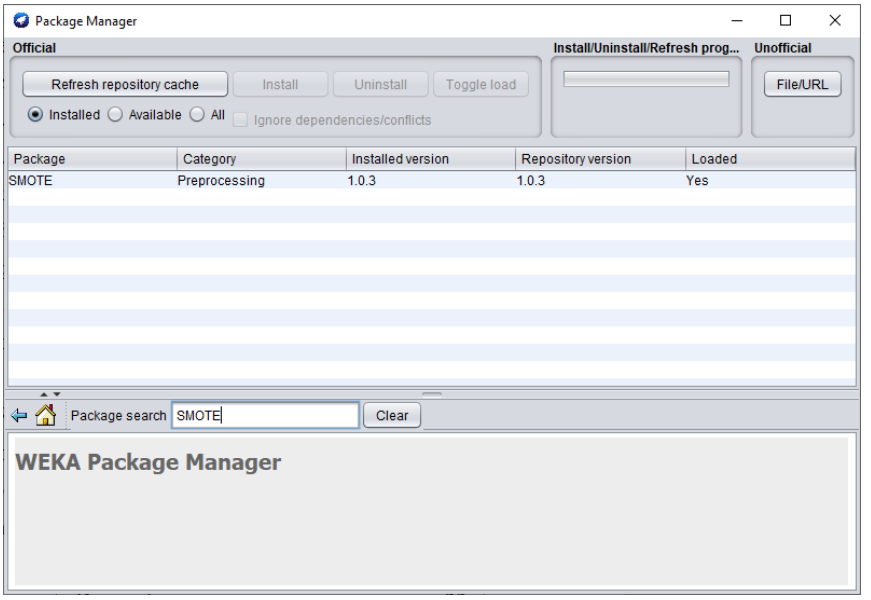


Figura 6.7: Gerenciador de pacotes do WEKA.

Após a instalação do pacote, a técnica SMOTE fica habilitada para ser utilizada, e a técnica pode ser aplicada a cada uma das classes presentes no conjunto de dados. Os parâmetros padrões da técnica são *classValue*: 0, *nearestNeighbors*: 5, *percentage*: 100.0 e *randomSeed*: 1, e a tela de configuração pode ser observada na Figura 6.8.

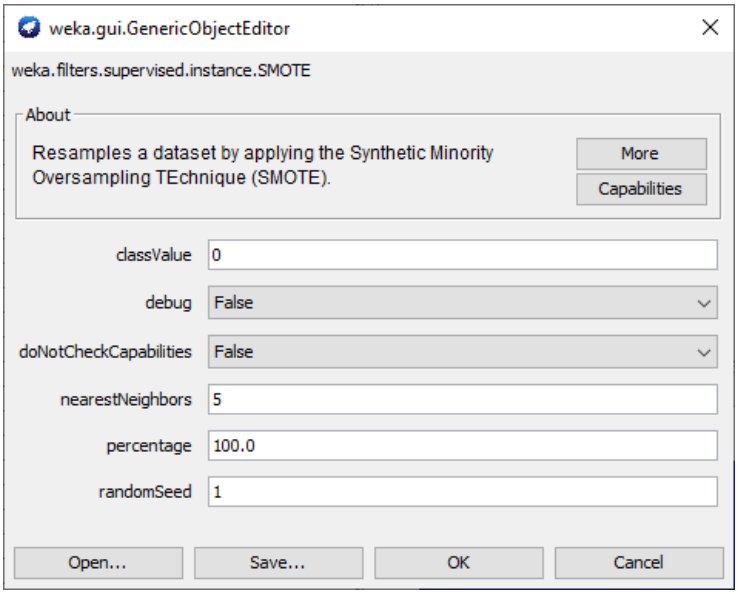


Figura 6.8: Tela de configuração padrão da técnica SMOTE no WEKA.

Porém, como os conjuntos de dados são diferentes para cada ano, os experimentos foram realizados com os seguintes parâmetros:

- Para o primeiro ano a técnica é aplicada apenas a classe “Pendente” (43 registros), onde o conjunto de dados para a classe “Pendente” é multiplicado por 5 (*percentage*: 500.0) resultando em 258 registros (43 registros reais e 215 registros criados sinteticamente), assim, a quantidade de registros da classe “Pendente” está mais próxima da classe “Aprovado” (506 registros). Porém, para as classes “Reprovado” (2 registros) e “Evadido” (1 registro), a técnica não pôde ser aplicada, pois existem poucos registros no conjunto de dados para ambas as classes.
- Para o segundo ano a técnica é aplicada às classes “Pendente” (28 registros) e “Reprovado” (22 registros), nesta etapa o conjunto de dados para ambas as classes também é multiplicado por 5 (*percentage*: 500.0), resultando em 168 registros para a classe “Pendente” (28 registros reais e 140 registros criados sinteticamente), e 132 registros para a classe “Reprovado” (22 registros reais e 110 registros criados sinteticamente). Assim, a quantidade de registros de ambas as classe “`Pendente” e “Reprovado” também estão mais próximas da classe “Aprovado” (434 registros). Porém, neste conjunto de dados, para a classe “Evadido” (2 registros), a técnica não pôde ser aplicada pois existem poucos registros no conjunto de dados para essa classe.
- Para o terceiros ano a técnica é aplicada as classes “Reprovado” (7 registros) e “Evadido” (6 registros), nesta etapa o conjunto de dados para ambas as classes também é multiplicado por 10 (percentage: 1000.0), resultando em 77 registros para a classe “Reprovado” (7 registros reais e 70 registros criados sinteticamente), e 66 registros para a classe “Evadido” (6 registros reais e 60 registros criados sinteticamente) assim, a quantidade de registros de ambas as classe “`Reprovado” e “Evadido” também estão mais próximas da classe “Aprovado” (374 registros). No conjunto de dados dos Terceiros Anos vale relembrar que não há a classe “Pendente”.

Por fim, a técnica de modelo penalizado (*Cost-sensitive Learning*) é aplicada no ambiente WEKA, por meio do classificador “*CostSensitiveClassifier*”, como apresentado na Figura 6.9.

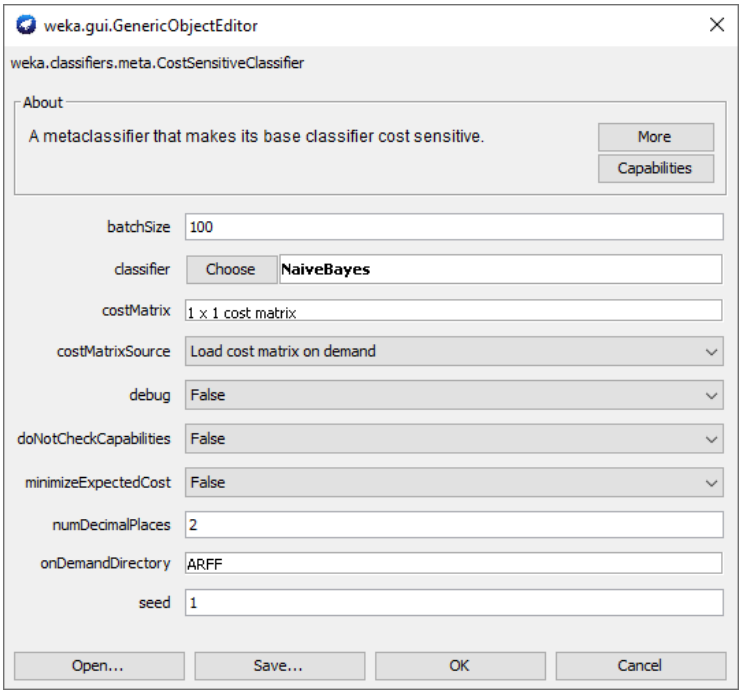


Figura 6.9: Tela de configuração padrão do classificador *CostSensitiveClassifier*.

A técnica *Cost-sensitive Learning* consiste em atribuir diferentes custos de classificação incorreta para as diferentes classes. (Ling *et al.,* 2010).

Para os dados dos alunos do primeiro e segundo ano, a matriz de custos é representada em WEKA por uma matriz de custos de 4 x 4 (correspondente às classes aprovado, reprovado, pendente e evadido), conforme apresentada na Figura 6.10, e para os dados dos alunos do terceiro ano é representada por uma matriz de 3 x 3 (correspondente às classes aprovado, reprovado e evadido), conforme apresentada na Figura 6.11, já o parâmetro *classifier* permite selecionar qual algoritmo será aplicado ao experimento, no caso os algoritmos J48, NaiveBayes e IBk. Os outros parâmetros não foram alterados.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	**CT (aprovado)**	CT (aprov.|reprov.)	CT (aprov.|pend.)	CT (apro.|evad.)
	reprovado	CT (reprov.|aprov.)	**CT (reprovado)**	CT (reprov.|pend.)	CT (reprov.|evad.)
	pendente	CT (pend.|aprov.)	CT (pend.|reprov.)	**CT (pendente)**	CT (pend.|evad.)
	evadido	CT (evad.|aprov.)	CT (evad.|reprov.)	CT (evad.|pend.)	**CT (evadido)**

Figura 6.10: Matriz de custos para as classes do primeiro e segundo ano.

		CLASSE PREDITA		
		aprovado	**reprovado**	**evadido**
CLASSE ORIGINAL	**aprovado**	**CT (aprovado)**	CT (aprov.|reprov.)	CT (apro.|evad.)
	reprovado	CT (reprov.|aprov.)	**CT (reprovado)**	CT (reprov.|evad.)
	evadido	CT (evad.|aprov.)	CT (evad.|reprov.)	**CT (evadido)**

Figura 6.11: Matriz de custos para as classes do terceiro ano.

Nas tabelas acima, a sigla CT representa o custo ou penalidade atribuido. Na diagonal principal de ambas as matrizes , CT representa os custos quando o algoritmo classifica corretamente a classe original, geralmente o valor para essas posições é 0.0. Os outos valores CT que não estão na diagonal representam o custo ou penalidade atribuidos quando o algoritmo classifica incorretamente a classe original.

A matriz de custo, pode ser acessada no ambiente WEKA pela opção *costMatrix*, ao clicar no botão *Choose*, conforme apresentado na Figura 6.12. Ao clicar no botão WEKA é exibido uma caixa de configuração padrão (Figura 6.13).

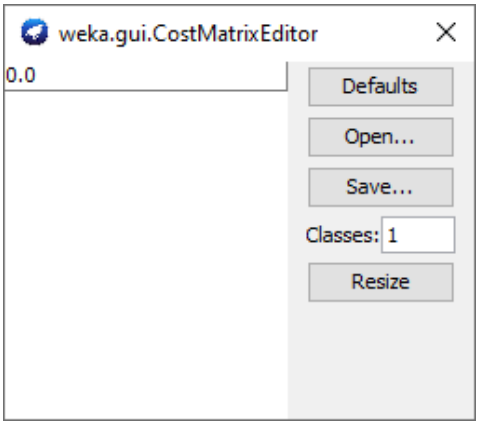


Figura 6.12: Caixa de configuração padrão da matriz de custo.

Por padrão, na Figura 6.12, a caixa de configuração apresenta a quantidade de classes em 1 (*Classes*), para modificar a quantidade de classes o valor deve ser alterado para 3, que representa o conjunto de dados dos terceiros anos, ou 4, para o conjunto de dados do primeiro e segundo ano.

Após selecionado um dos valores para os experimentos é necessesário redimensionar a nova matriz, este redimensionamento é realizado ao clicar no botão *Resize* na Figura 6.12. Na Figura 6.13 é apresentado um exemplo de uma matriz 3x3 com os valores já modificados com seus respectivos custos.

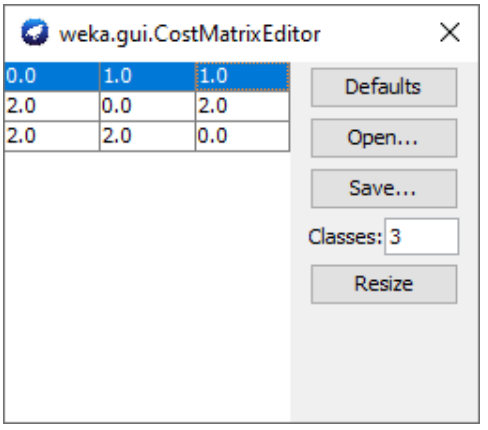


Figura 6.13: Caixa de configuração da matriz com valores de custos modificados.

6.5. Resultados dos Experimentos

Após o estabelecimento da configuração geral para os algoritmos foram feitos os experimentos conforme apresentados na Seção 6.1 para os dados de cada um dos anos da instituição e em seguida foram tabuladas as medidas para avaliar o desempenho dos

classificadores a partir das médias ponderadas. As Tabelas 6.2, 6.3 e 6.4 apresentam os resultados com os valores obtidos para os alunos do primeiro ano, respectivamente para os dados desbalanceados, com a técnica SMOTE e a técnica SMOTE e custos, sendo estas duas técnicas também aplicadas aos dados dos alunos de segundo e terceiro ano. As tabelas referentes as matrizes de custos utilizadas nos experimentos são apresentadas no Apêndice A deste trabalho.

Para calcular as diferentes medidas utilizadas para avaliar o desempenho dos classificadores, em alguns casos foi possível obter resultados, porém em outros não foi possível calcular a medida devido a que o denominador da expressão usada para o cálculo foi zero, sendo assim, nas tabelas a seguir quando não foi possível fazer o cálculo foi colocado um sinal de interrogação.

Tabela 6.2: Resultados para os alunos do primeiro ano utilizando os algoritmos com dados desbalanceados.

	Naive Bayes	**J48 sem poda**	**J48 com poda**	**IBk, k=1**	**IBk, k=3**	**IBk, k=5**
Instâncias corretamente classificadas	0,898551	0,942029	0,947464	0,938406	0,925725	0,934783
Instâncias incorretamente classificadas	0,101449	0,057971	0,052536	0,061594	0,074275	0,065217
Taxa de verdadeiros positivos: aprovado	0,909	0,978	0,990	0,982	0,990	0,996
Taxa de verdadeiros positivos: reprovado	0,000	0,500	0,000	1,000	0,000	0,000
Taxa de verdadeiros positivos: pendente	0,837	0,558	0,512	0,442	0,233	0,279
Taxa de verdadeiros positivos: evadido	0,000	0,000	0,000	0,000	0,000	0,000
Média Ponderada	0,899	0,942	0,947	0,938	0,926	0,935
Taxa de falsos positivos: aprovado	0,152	0,413	0,478	0,543	0,739	0,696
Taxa de falsos positivos: reprovado	0,000	0,002	0,000	0,000	0,000	0,000
Taxa de falsos positivos: pendente	0,096	0,024	0,014	0,018	0,014	0,008
Taxa de falsos positivos: evadido	0,000	0,000	0,000	0,000	0,000	0,000
Média Ponderada	0,147	0,380	0,439	0,500	0,679	0,638
Precisão: aprovado	0,985	0,963	0,958	0,952	0,936	0,940
Precisão: reprovado	?	0,500	?	1,000	?	?
Precisão: pendente	0,424	0,667	0,759	0,679	0,588	0,750
Precisão: evadido	?	?	?	?	?	?
Média Ponderada	?	?	?	?	?	?
Sensibilidade: aprovado	0,909	0,978	0,990	0,982	0,990	0,996
Sensibilidade: reprovado	0,000	0,500	0,000	1,000	0,000	0,000
Sensibilidade: pendente	0,837	0,558	0,512	0,442	0,233	0,279
Sensibilidade: evadido	0,000	0,000	0,000	0,000	0,000	0,000
Média Ponderada	0,899	0,942	0,947	0,938	0,926	0,935
Medida-F: aprovado	0,946	0,971	0,974	0,967	0,963	0,967

Medida-F: reprovado	?	0,500	?	1,000	?	?
Medida-F: pendente	0,563	0,608	0,611	0,535	0,333	0,407
Medida-F: evadido	?	?	?	?	?	?
Média Ponderada	?	?	?	?	?	?
MCC: aprovado	0,580	0,617	0,634	0,535	0,402	0,495
MCC: reprovado	?	0,498	?	1,000	?	?
MCC: pendente	0,550	0,580	0,598	0,518	0,339	0,433
MCC: evadido	?	?	?	?	?	?
Média Ponderada	?	?	?	?	?	?
Área ROC: aprovado	0,956	0,819	0,764	0,730	0,785	0,787
Área ROC: reprovado	0,415	0,749	0,995	1,000	1,000	1,000
Área ROC: pendente	0,947	0,806	0,769	0,714	0,773	0,774
Área ROC: evadido	0,051	0,499	0,254	0,899	0,900	0,898
Média Ponderada	0,952	0,817	0,764	0,730	0,785	0,787
Área PRC: aprovado	0,996	0,968	0,952	0,953	0,963	0,964
Área PRC: reprovado	0,005	0,252	0,325	1,000	1,000	1,000
Área PRC: pendente	0,491	0,475	0,535	0,361	0,390	0,426
Área PRC: evadido	0,002	0,002	0,002	0,009	0,009	0,009
Média Ponderada	0,951	0,926	0,915	0,906	0,917	0,920

Tabela 6.3: Resultados para os alunos do primeiro ano utilizando a técnica SMOTE.

	Naive Bayes	**J48 sem poda**	**J48 com poda**	**IBk, k=1**	**IBk, k=3**	**IBk, k=5**
Instâncias corretamente classificadas	0,89309	0,96219	0,95176	0,95176	0,934811	0,925684
Instâncias incorretamente classificadas	0,10691	0,03781	0,04824	0,04824	0,065189	0,074316
Taxa de verdadeiros positivos: aprovado	0,862	0,972	0,974	0,955	0,925	0,909
Taxa de verdadeiros positivos: reprovado	0,000	1,000	0,000	1,000	0,000	0,000
Taxa de verdadeiros positivos: pendente	0,965	0,946	0,919	0,950	0,965	0,969
Taxa de verdadeiros positivos: evadido	0,000	0,000	0,000	0,000	0,000	0,000
Média Ponderada	0,893	0,962	0,952	0,952	0,935	0,926
Taxa de falsos positivos: aprovado	0,034	0,054	0,080	0,054	0,038	0,034
Taxa de falsos positivos: reprovado	0,001	0,000	0,000	0,000	0,000	0,000
Taxa de falsos positivos: pendente	0,141	0,028	0,029	0,045	0,079	0,094
Taxa de falsos positivos: evadido	0,000	0,001	0,001	0,000	0,000	0,000
Média Ponderada	0,070	0,045	0,063	0,051	0,052	0,054
Precisão: aprovado	0,980	0,972	0,959	0,972	0,979	0,981
Precisão: reprovado	0,000	1,000	?	1,000	?	?
Precisão: pendente	0,776	0,946	0,940	0,914	0,862	0,839
Precisão: evadido	?	0,000	0,000	?	?	?
Média Ponderada	?	0,962	?	?	?	?
Sensibilidade: aprovado	0,862	0,972	0,974	0,955	0,925	0,909
Sensibilidade: reprovado	0,000	1,000	0,000	1,000	0,000	0,000

Sensibilidade: pendente	0,965	0,946	0,919	0,950	0,965	0,969
Sensibilidade: evadido	0,000	0,000	0,000	0,000	0,000	0,000
Média Ponderada	0,893	0,962	0,952	0,952	0,935	0,926
Medida-F: aprovado	0,917	0,972	0,967	0,963	0,951	0,944
Medida-F: reprovado	0,000	1,000	?	1,000	?	?
Medida-F: pendente	0,860	0,946	0,929	0,932	0,910	0,899
Medida-F: evadido	?	0,000	0,000	?	?	?
Média Ponderada	?	0,962	?	?	?	?
MCC: aprovado	0,794	0,919	0,901	0,894	0,867	0,850
MCC: reprovado	-0,002	1,000	?	1,000	?	?
MCC: pendente	0,789	0,918	0,894	0,896	0,864	0,848
MCC: evadido	?	-0,001	-0,001	?	?	?
Média Ponderada	?	0,918	?	?	?	?
Área ROC: aprovado	0,967	0,967	0,957	0,953	0,966	0,975
Área ROC: reprovado	0,382	1,000	0,868	1,000	1,000	1,000
Área ROC: pendente	0,964	0,965	0,948	0,955	0,967	0,975
Área ROC: evadido	0,475	0,496	0,496	0,668	0,668	0,684
Média Ponderada	0,964	0,966	0,953	0,954	0,966	0,975
Área PRC: aprovado	0,984	0,972	0,956	0,961	0,977	0,981
Área PRC: reprovado	0,003	1,000	0,014	1,000	1,000	1,000
Área PRC: pendente	0,896	0,925	0,899	0,905	0,911	0,949
Área PRC: evadido	0,002	0,001	0,001	0,002	0,002	0,002
Média Ponderada	0,951	0,955	0,933	0,941	0,953	0,969

Tabela 6.4: Resultados para os alunos do primeiro ano utilizando a técnica SMOTE e custos.

	Naive Bayes	**J48 sem poda**	**J48 com poda**	**IBk, k=1**	**IBk, k=3**	**IBk, k=5**
Instâncias corretamente classificadas	0,876141	0,96219	0,937419	0,95176	0,912647	0,898305
Instâncias incorretamente classificadas	0,123859	0,03781	0,062581	0,04824	0,087353	0,101695
Taxa de verdadeiros positivos: aprovado	0,836	0,972	0,931	0,955	0,883	0,864
Taxa de verdadeiros positivos: reprovado	0,000	1,000	0,500	1,000	0,000	0,000
Taxa de verdadeiros positivos: pendente	0,965	0,946	0,957	0,950	0,981	0,977
Taxa de verdadeiros positivos: evadido	0,000	0,000	0,000	0,000	0,000	0,000
Média Ponderada	0,876	0,962	0,937	0,952	0,913	0,898
Taxa de falsos positivos: aprovado	0,034	0,054	0,046	0,054	0,023	0,027
Taxa de falsos positivos: reprovado	0,001	0,000	0,000	0,000	0,000	0,000
Taxa de falsos positivos: pendente	0,167	0,028	0,069	0,045	0,120	0,139
Taxa de falsos positivos: evadido	0,000	0,001	0,001	0,000	0,000	0,000
Média Ponderada	0,079	0,045	0,053	0,051	0,055	0,065
Precisão: aprovado	0,979	0,972	0,975	0,972	0,987	0,984
Precisão: reprovado	0,000	1,000	1,000	1,000	?	?
Precisão: pendente	0,746	0,946	0,876	0,914	0,806	0,780

Precisão: evadido	?	0,000	0,000	?	?	?
Média Ponderada	?	0,962	0,941	?	?	?
Sensibilidade: aprovado	0,836	0,972	0,931	0,955	0,883	0,864
Sensibilidade: reprovado	0,000	1,000	0,500	1,000	0,000	0,000
Sensibilidade: pendente	0,965	0,946	0,957	0,950	0,981	0,977
Sensibilidade: evadido	0,000	0,000	0,000	0,000	0,000	0,000
Média Ponderada	0,876	0,962	0,937	0,952	0,913	0,898
Medida-F: aprovado	0,902	0,972	0,952	0,963	0,932	0,920
Medida-F: reprovado	0,000	1,000	0,667	1,000	?	?
Medida-F: pendente	0,841	0,946	0,915	0,932	0,885	0,867
Medida-F: evadido	?	0,000	0,000	?	?	?
Média Ponderada	?	0,962	0,938	?	?	?
MCC: aprovado	0,766	0,919	0,868	0,894	0,829	0,803
MCC: reprovado	-0,002	1,000	0,707	1,000	?	?
MCC: pendente	0,761	0,918	0,871	0,896	0,827	0,801
MCC: evadido	?	-0,001	-0,001	?	?	?
Média Ponderada	?	0,918	0,868	?	?	?
Área ROC: aprovado	0,965	0,967	0,950	0,951	0,962	0,973
Área ROC: reprovado	0,382	1,000	1,000	1,000	1,000	1,000
Área ROC: pendente	0,962	0,965	0,954	0,953	0,964	0,974
Área ROC: evadido	0,473	0,496	0,497	0,508	0,555	0,431
Média Ponderada	0,962	0,966	0,951	0,951	0,962	0,973
Área PRC: aprovado	0,983	0,972	0,951	0,962	0,974	0,981
Área PRC: reprovado	0,003	1,000	1,000	1,000	1,000	1,000
Área PRC: pendente	0,894	0,925	0,928	0,905	0,899	0,948
Área PRC: evadido	0,002	0,001	0,001	0,002	0,002	0,002
Média Ponderada	0,950	0,955	0,942	0,942	0,947	0,968

Para os dados dos alunos do primeiro ano, referente a precisão e com a utilização do algoritmo J48 sem poda da árvore, os dados os quais foram aplicados a técnica SMOTE (0,962) e SMOTE e custos (0,962) obtiveram resultados, e com a utilização do algoritmo J48 com poda da árvore apenas a técnica SMOTE e custos (0,941) obteve resultado, os outros algoritmos não conseguiram obter resultados para comparação.

Nos experimentos, quando considerada a sensibilidade, o algoritmo Naive Bayes obteve 0,899 nos dados desbalanceados, 0,893 na técnica SMOTE e 0,876 na técnica SMOTE e custos. Os algoritmos IBk para $k = 3$ e $k = 5$ obtiveram, respectivamente, 0,926 e 0,935 nos dados desbalanceados, 0,935 e 0,926 na técnica SMOTE e por fim 0,913 e 0,898 na técnica SMOTE e custos. O algoritmo J48 com poda obteve 0,947 nos dados

desbalanceados, 0,952 na técnica SMOTE e 0,937 na técnica SMOTE e custos. Nesses experimentos os algoritmos obtiveram resultados próximos, porém com diminuição na sensibilidade. Já os experimentos com os algoritmos J48 sem poda e IBk ($k = 1$) apresentam um aumento da sensibilidade nas técnicas SMOTE e SMOTE e custos em comparação com os dados desbalanceados, os algoritmos obtiveram, respectivamente, 0,942 e 0,938 nos dados desbalanceados, 0,962 e 0,952 na técnica SMOTE e por fim 0,962 e 0,952 na técnica SMOTE e custos.

Referente a Medida-F, o algoritmo J48 sem poda da árvore obteve resultados, sendo 0,962 na técnica SMOTE e 0,960 na técnica SMOTE e custos, e o algoritmo J48 com poda da árvore obteve resultado, sendo 0,938 na técnica SMOTE e custos, porém para o experimento com os dados desbalanceados e com a aplicação das técnicas os demais algoritmos não conseguiram obter resultados.

Por fim, os experimentos referentes a área abaixo da curva ROC, o algoritmo Naive Bayes obteve uma ligeira melhora nas técnicas SMOTE (0,964) e SMOTE e custos (0,962), em comparação com o experimento com os dados desbalanceados (0,952). Por outro lado, todos os experimentos realizados com os outros algoritmos obtiveram uma melhora significativa nas técnicas SMOTE e SMOTE e custos em comparação aos resultados utilizando os dados desbalanceados.

Portanto, para os experimentos realizados com os dados referentes ao primeiro ano da instituição, pode-se observar que a aplicação das técnicas SMOTE e SMOTE e custos obtêm, na maioria dos experimentos, um melhor desempenho em comparação com os dados desbalanceados.

Já as Tabelas 6.5, 6.6 e 6.7 apresentam os resultados com os valores obtidos para os segundos anos, respectivamente para os dados desbalanceados, com a técnica SMOTE e a técnica SMOTE e custos.

Tabela 6.5: Resultados para os alunos do segundo ano utilizando os algoritmos com dados desbalanceados.

	Naive Bayes	**J48 sem poda**	**J48 com poda**	**IBk, k=1**	**IBk, k=3**	**IBk, k=5**
Instâncias corretamente classificadas	0,884774	0,954733	0,952675	0,925926	0,932099	0,932099
Instâncias incorretamente classificadas	0,115226	0,045267	0,047325	0,074074	0,067901	0,067901
Taxa de verdadeiros positivos: aprovado	0,896	0,988	0,993	0,984	0,995	0,995

Taxa de verdadeiros positivos: reprovado	0,864	0,909	0,909	0,727	0,773	0,864
Taxa de verdadeiros positivos: pendente	0,786	0,536	0,429	0,179	0,143	0,071
Taxa de verdadeiros positivos: evadido	0,000	0,000	0,000	1,000	0,000	0,000
Média Ponderada	0,885	0,955	0,953	0,926	0,932	0,932
Taxa de falsos positivos: aprovado	0,096	0,135	0,192	0,462	0,500	0,538
Taxa de falsos positivos: reprovado	0,006	0,017	0,017	0,002	0,004	0,004
Taxa de falsos positivos: pendente	0,105	0,015	0,011	0,017	0,007	0,007
Taxa de falsos positivos: evadido	0,000	0,000	0,000	0,006	0,004	0,000
Média Ponderada	0,092	0,122	0,173	0,413	0,447	0,481
Precisão: aprovado	0,987	0,984	0,977	0,947	0,943	0,939
Precisão: reprovado	0,864	0,714	0,714	0,941	0,895	0,905
Precisão: pendente	0,314	0,682	0,706	0,385	0,571	0,400
Precisão: evadido	?	?	?	0,400	0,000	?
Média Ponderada	?	?	?	0,912	0,916	?
Sensibilidade: aprovado	0,896	0,988	0,993	0,984	0,995	0,995
Sensibilidade: reprovado	0,864	0,909	0,909	0,727	0,773	0,864
Sensibilidade: pendente	0,786	0,536	0,429	0,179	0,143	0,071
Sensibilidade: evadido	0,000	0,000	0,000	1,000	0,000	0,000
Média Ponderada	0,885	0,955	0,953	0,926	0,932	0,932
Medida-F: aprovado	0,940	0,986	0,985	0,965	0,969	0,966
Medida-F: reprovado	0,864	0,800	0,800	0,821	0,829	0,884
Medida-F: pendente	0,449	0,600	0,533	0,244	0,229	0,121
Medida-F: evadido	?	?	?	0,571	0,000	?
Média Ponderada	?	?	?	0,915	0,916	?
MCC: aprovado	0,631	0,869	0,854	0,625	0,657	0,628
MCC: reprovado	0,857	0,796	0,796	0,820	0,824	0,879
MCC: pendente	0,452	0,583	0,530	0,233	0,267	0,150
MCC: evadido	?	?	?	0,630	-0,004	?
Média Ponderada	?	?	?	0,611	0,640	?
Área ROC: aprovado	0,969	0,929	0,934	0,777	0,809	0,841
Área ROC: reprovado	0,989	0,947	0,949	0,917	0,957	0,956
Área ROC: pendente	0,939	0,792	0,865	0,602	0,635	0,712
Área ROC: evadido	0,696	0,991	0,976	0,997	0,993	0,992
Média Ponderada	0,967	0,922	0,931	0,775	0,807	0,840
Área PRC: aprovado	0,996	0,983	0,985	0,950	0,957	0,964
Área PRC: reprovado	0,787	0,684	0,781	0,726	0,831	0,803
Área PRC: pendente	0,474	0,438	0,448	0,161	0,179	0,206
Área PRC: evadido	0,010	0,226	0,110	0,400	0,286	0,250
Média Ponderada	0,953	0,935	0,941	0,892	0,903	0,910

Tabela 6.6: Resultados para os alunos do segundo ano utilizando a técnica SMOTE.

	Naive Bayes	J48 sem poda	J48 com poda	IBk, k=1	IBk, k=3	IBk, k=5
Instâncias corretamente classificadas	0,925272	0,934783	0,92663	0,94837	0,945652	0,942935
Instâncias incorretamente classificadas	0,074728	0,065217	0,07337	0,05163	0,054348	0,057065
Taxa de verdadeiros positivos: aprovado	0,912	0,949	0,945	0,961	0,949	0,935
Taxa de verdadeiros positivos: reprovado	0,970	0,977	0,977	0,947	0,962	0,992
Taxa de verdadeiros positivos: pendente	0,935	0,875	0,851	0,923	0,935	0,935
Taxa de verdadeiros positivos: evadido	0,000	0,000	0,000	0,500	0,000	0,000
Média Ponderada	0,925	0,935	0,927	0,948	0,946	0,943
Taxa de falsos positivos: aprovado	0,026	0,060	0,073	0,040	0,036	0,030
Taxa de falsos positivos: reprovado	0,015	0,013	0,013	0,007	0,007	0,008
Taxa de falsos positivos: pendente	0,065	0,037	0,040	0,032	0,039	0,049
Taxa de falsos positivos: evadido	0,001	0,001	0,001	0,005	0,004	0,000
Média Ponderada	0,033	0,046	0,055	0,032	0,032	0,030
Precisão: aprovado	0,980	0,958	0,949	0,972	0,974	0,978
Precisão: reprovado	0,934	0,942	0,942	0,969	0,969	0,963
Precisão: pendente	0,809	0,875	0,861	0,896	0,877	0,849
Precisão: evadido	0,000	0,000	0,000	0,200	0,000	?
Média Ponderada	0,930	0,934	0,925	0,952	0,948	?
Sensibilidade: aprovado	0,912	0,949	0,945	0,961	0,949	0,935
Sensibilidade: reprovado	0,970	0,977	0,977	0,947	0,962	0,992
Sensibilidade: pendente	0,935	0,875	0,851	0,923	0,935	0,935
Sensibilidade: evadido	0,000	0,000	0,000	0,500	0,000	0,000
Média Ponderada	0,925	0,935	0,927	0,948	0,946	0,943
Medida-F: aprovado	0,945	0,954	0,947	0,966	0,961	0,956
Medida-F: reprovado	0,952	0,959	0,959	0,958	0,966	0,978
Medida-F: pendente	0,867	0,875	0,856	0,909	0,905	0,890
Medida-F: evadido	0,000	0,000	0,000	0,286	0,000	?
Média Ponderada	0,926	0,934	0,926	0,950	0,947	?
MCC: aprovado	0,876	0,888	0,871	0,919	0,908	0,898
MCC: reprovado	0,941	0,950	0,950	0,949	0,958	0,973
MCC: pendente	0,828	0,838	0,814	0,882	0,876	0,856
MCC: evadido	-0,002	-0,002	-0,002	0,313	-0,003	?
Média Ponderada	0,874	0,885	0,870	0,914	0,907	?
Área ROC: aprovado	0,985	0,950	0,953	0,966	0,981	0,979
Área ROC: reprovado	0,991	0,985	0,982	0,975	0,996	0,992
Área ROC: pendente	0,985	0,924	0,916	0,949	0,961	0,959
Área ROC: evadido	0,635	0,996	0,996	0,861	0,994	0,994
Média Ponderada	0,985	0,950	0,950	0,963	0,979	0,977
Área PRC: aprovado	0,991	0,945	0,948	0,963	0,980	0,980
Área PRC: reprovado	0,909	0,955	0,909	0,934	0,984	0,984

Área PRC: pendente	0,955	0,818	0,815	0,879	0,920	0,914
Área PRC: evadido	0,006	0,400	0,400	0,102	0,286	0,286
Média Ponderada	0,965	0,916	0,909	0,937	0,965	0,964

Tabela 6.7: Resultados para os alunos do segundo ano utilizando a técnica SMOTE e custos.

	Naive Bayes	**J48 sem poda**	**J48 com poda**	**IBk, k=1**	**IBk, k=3**	**IBk, k=5**
Instâncias corretamente classificadas	0,923913	0,944293	0,945652	0,94837	0,945652	0,9375
Instâncias incorretamente classificadas	0,076087	0,055707	0,054348	0,05163	0,054348	0,0625
Taxa de verdadeiros positivos: aprovado	0,908	0,961	0,975	0,961	0,949	0,935
Taxa de verdadeiros positivos: reprovado	0,970	0,962	0,962	0,947	0,947	0,962
Taxa de verdadeiros positivos: pendente	0,940	0,887	0,869	0,923	0,935	0,935
Taxa de verdadeiros positivos: evadido	0,000	1,000	0,000	0,500	1,000	0,000
Média Ponderada	0,924	0,944	0,946	0,948	0,946	0,938
Taxa de falsos positivos: aprovado	0,026	0,056	0,070	0,040	0,036	0,026
Taxa de falsos positivos: reprovado	0,013	0,005	0,007	0,007	0,003	0,010
Taxa de falsos positivos: pendente	0,069	0,032	0,021	0,032	0,039	0,049
Taxa de falsos positivos: evadido	0,001	0,004	0,004	0,005	0,007	0,005
Média Ponderada	0,034	0,041	0,047	0,032	0,031	0,029
Precisão: aprovado	0,980	0,961	0,953	0,972	0,974	0,981
Precisão: reprovado	0,941	0,977	0,969	0,969	0,984	0,955
Precisão: pendente	0,802	0,892	0,924	0,896	0,877	0,849
Precisão: evadido	0,000	0,400	0,000	0,200	0,286	0,000
Média Ponderada	0,930	0,947	0,947	0,952	0,952	0,943
Sensibilidade: aprovado	0,908	0,961	0,975	0,961	0,949	0,935
Sensibilidade: reprovado	0,970	0,962	0,962	0,947	0,947	0,962
Sensibilidade: pendente	0,940	0,887	0,869	0,923	0,935	0,935
Sensibilidade: evadido	0,000	1,000	0,000	0,500	1,000	0,000
Média Ponderada	0,924	0,944	0,946	0,948	0,946	0,938
Medida-F: aprovado	0,943	0,961	0,964	0,966	0,961	0,958
Medida-F: reprovado	0,955	0,969	0,966	0,958	0,965	0,958
Medida-F: pendente	0,866	0,890	0,896	0,909	0,905	0,890
Medida-F: evadido	0,000	0,571	0,000	0,286	0,444	0,000
Média Ponderada	0,925	0,945	0,946	0,950	0,948	0,940
MCC: aprovado	0,871	0,905	0,910	0,919	0,908	0,901
MCC: reprovado	0,945	0,963	0,958	0,949	0,958	0,949
MCC: pendente	0,826	0,857	0,867	0,882	0,876	0,856
MCC: evadido	-0,002	0,631	-0,003	0,313	0,533	-0,004
Média Ponderada	0,872	0,903	0,906	0,914	0,909	0,897
Área ROC: aprovado	0,984	0,952	0,958	0,965	0,975	0,979
Área ROC: reprovado	0,991	0,980	0,985	0,961	0,997	0,992
Área ROC: pendente	0,984	0,925	0,933	0,950	0,956	0,956

Área ROC: evadido	0,635	0,996	0,996	0,627	0,996	0,995
Média Ponderada	0,984	0,951	0,957	0,960	0,975	0,976
Área PRC: aprovado	0,991	0,938	0,945	0,964	0,973	0,975
Área PRC: reprovado	0,913	0,959	0,948	0,932	0,987	0,986
Área PRC: pendente	0,951	0,863	0,847	0,886	0,915	0,910
Área PRC: evadido	0,006	0,400	0,400	0,252	0,325	0,267
Média Ponderada	0,965	0,923	0,922	0,939	0,961	0,960

Referente à precisão para os dados do segundo ano, o algoritmo IBk, para $k = 1$ e $k = 3$, respectivamente obtiveram 0,912 e 0,916 para os dados desbalanceados, 0,952 e 0,948 com a técnica SMOTE e 0,952 e 0,952 com a técnica SMOTE e custos. Em contrapartida, o algoritmo IBk ($k = 5$) obteve 0,943 apenas com a técnica SMOTE e custos. Também pode-se observar que, para o algoritmo Naive Bayes, com os dados desbalanceados, não foi obtido resultado, porém para as demais técnicas obtiveram resultados, sendo que ambas as técnicas SMOTE e SMOTE e custos obtiveram 0,930 nestes experimentos. Contudo, para o algoritmo J48 com e sem a poda da árvore, para os dados desbalanceados, não foram obtidos resultados, entretanto com a aplicação da técnica SMOTE e SMOTE e custos, o algoritmo obteve, respectivamente com a poda da árvore 0,934 e 0,947 e sem a poda da árvore 0,925 e 0,947.

Por sua vez, os experimentos com a sensibilidade, com os dados desbalanceados, o algoritmo J48 com poda da árvore obteve melhor desempenho (0,953) em comparação a técnica SMOTE (0,927) e SMOTE e custos (0,946). O experimento com o algoritmo J48 sem poda da árvore também obteve melhor desempenho (0,955) com os dados desbalanceados em comparação a técnica SMOTE (0,935) e SMOTE e custos (0,944). Porém os demais algoritmos com os dados desbalanceados obtiveram desempenhos inferiores em comparação as técnicas SMOTE e SMOTE e custos.

Nos experimentos relacionados a Medida-F, assim como na precisão, o algoritmo IBk, para $k = 1$ e $k = 3$, respectivamente obtiveram 0,915 e 0,916 para os dados desbalanceados, 0,950 e 0,947 com a técnica SMOTE e 0,950 e 0,948 com a técnica SMOTE e custos. Em contrapartida, o algoritmo IBk ($k = 5$) obteve 0,940 apenas com a técnica SMOTE e custos. Já para o algoritmo Naive Bayes, com os dados desbalanceados, não foi obtido resultado, porém para as demais técnicas foram obtidos resultados, sendo que a técnica SMOTE, neste experimento, obteve um desempenho superior de 0,926 em

comparação com a técnica de SMOTE e custos de 0,925. Contudo, para o algoritmo J48 com e sem a poda da árvore, para dados desbalanceados, não foram obtidos resultados válidos, entretanto com a aplicação das técnicas SMOTE e SMOTE e custos, os algoritmos obtiveram, respectivamente com a poda da árvore 0,926 e 0,946 e sem a poda da árvore 0,934 e 0,945.

Referente a área abaixo da curva ROC, as técnicas SMOTE e SMOTE e custos obtiveram, em todos os experimentos, resultados superiores em comparação aos experimentos realizados com dados desbalanceados.

Portanto, assim como nos experimentos realizados com os dados dos alunos do primeiro ano. fica evidente que a aplicação das técnicas SMOTE e SMOTE e custos para os experimentos dos dados do segundo ano obtêm, na maioria dos experimentos, um melhor desempenho em comparação com os dados desbalanceados e também pode-se observar que em alguns casos a técnica SMOTE com custos foi ligeiramente melhor que a técnica SMOTE.

Por fim, as Tabelas 6.8, 6.9 e 6.10 apresentam os resultados com os valores obtidos para os alunos do terceiro ano, respectivamente para os dados desbalanceados, com a técnica SMOTE e a técnica SMOTE e custos.

Tabela 6.8: Resultados para os alunos do terceiro ano utilizando os algoritmos com os dados desbalanceados.

	Naive Bayes	**J48 sem poda**	**J48 com poda**	**IBk, k=1**	**IBk, k=3**	**IBk, k=5**
Instâncias corretamente classificadas	0,981912	0,979328	0,97416	0,98708	0,98708	0,968992
Instâncias incorretamente classificadas	0,018088	0,020672	0,02584	0,01292	0,01292	0,031008
Taxa de verdadeiros positivos: aprovado	0,995	1,000	1,000	1,000	1,000	1,000
Taxa de verdadeiros positivos: reprovado	0,857	0,714	0,429	0,714	0,857	0,143
Taxa de verdadeiros positivos: evadido	0,333	0,000	0,000	0,500	0,333	0,000
Média Ponderada	0,982	0,979	0,974	0,987	0,987	0,969
Taxa de falsos positivos: aprovado	0,077	0,154	0,231	0,077	0,154	0,385
Taxa de falsos positivos: reprovado	0,011	0,011	0,011	0,005	0,005	0,011
Taxa de falsos positivos: evadido	0,005	0,005	0,008	0,005	0,003	0,008
Média Ponderada	0,075	0,149	0,223	0,075	0,149	0,372
Precisão: aprovado	0,997	0,995	0,992	0,997	0,995	0,987
Precisão: reprovado	0,600	0,556	0,429	0,714	0,750	0,200
Precisão: evadido	0,500	0,000	0,000	0,600	0,667	0,000
Média Ponderada	0,982	0,971	0,966	0,986	0,985	0,957

Sensibilidade: aprovado	0,995	1,000	1,000	1,000	1,000	1,000
Sensibilidade: reprovado	0,857	0,714	0,429	0,714	0,857	0,143
Sensibilidade: evadido	0,333	0,000	0,000	0,500	0,333	0,000
Média Ponderada	0,982	0,979	0,974	0,987	0,987	0,969
Medida-F: aprovado	0,996	0,997	0,996	0,999	0,997	0,993
Medida-F: reprovado	0,706	0,625	0,429	0,714	0,800	0,167
Medida-F: evadido	0,400	0,000	0,000	0,545	0,444	0,000
Média Ponderada	0,981	0,975	0,970	0,986	0,985	0,963
MCC: aprovado	0,886	0,917	0,874	0,959	0,917	0,779
MCC: reprovado	0,711	0,622	0,418	0,709	0,798	0,156
MCC: evadido	0,401	-0,009	-0,011	0,541	0,466	-0,011
Média Ponderada	0,875	0,898	0,852	0,948	0,908	0,756
Área ROC: aprovado	0,995	0,923	0,939	0,973	0,973	0,973
Área ROC: reprovado	0,989	0,852	0,836	0,896	0,994	0,990
Área ROC: evadido	0,806	0,741	0,863	0,738	0,737	0,793
Média Ponderada	0,992	0,919	0,936	0,968	0,970	0,971
Área PRC: aprovado	1,000	0,995	0,996	0,998	0,998	0,998
Área PRC: reprovado	0,472	0,432	0,340	0,518	0,726	0,574
Área PRC: evadido	0,220	0,181	0,224	0,442	0,411	0,257
Média Ponderada	0,978	0,972	0,972	0,981	0,984	0,979

Tabela 6.9: Resultados para os alunos do terceiro ano utilizando a técnica SMOTE.

	Naive Bayes	**J48 sem poda**	**J48 com poda**	**IBk, k=1**	**IBk, k=3**	**IBk, k=5**
Instâncias corretamente classificadas	0,988395	0,982592	0,980658	0,992263	0,990329	0,990329
Instâncias incorretamente classificadas	0,011605	0,017408	0,019342	0,007737	0,009671	0,009671
Taxa de verdadeiros positivos: aprovado	0,997	0,992	0,989	1,000	1,000	1,000
Taxa de verdadeiros positivos: reprovado	0,961	0,974	0,974	0,987	0,961	0,961
Taxa de verdadeiros positivos: evadido	0,970	0,939	0,939	0,955	0,970	0,970
Média Ponderada	0,988	0,983	0,981	0,992	0,990	0,990
Taxa de falsos positivos: aprovado	0,007	0,014	0,014	0,007	0,007	0,007
Taxa de falsos positivos: reprovado	0,002	0,011	0,014	0,005	0,002	0,002
Taxa de falsos positivos: evadido	0,009	0,004	0,004	0,002	0,007	0,007
Média Ponderada	0,007	0,012	0,013	0,006	0,006	0,006
Precisão: aprovado	0,997	0,995	0,995	0,997	0,997	0,997
Precisão: reprovado	0,987	0,938	0,926	0,974	0,987	0,987
Precisão: evadido	0,941	0,969	0,969	0,984	0,955	0,955
Média Ponderada	0,989	0,983	0,981	0,992	0,990	0,990
Sensibilidade: aprovado	0,997	0,992	0,989	1,000	1,000	1,000
Sensibilidade: reprovado	0,961	0,974	0,974	0,987	0,961	0,961
Sensibilidade: evadido	0,970	0,939	0,939	0,955	0,970	0,970
Média Ponderada	0,988	0,983	0,981	0,992	0,990	0,990

Medida-F: aprovado	0,997	0,993	0,992	0,999	0,999	0,999
Medida-F: reprovado	0,974	0,955	0,949	0,981	0,974	0,974
Medida-F: evadido	0,955	0,954	0,954	0,969	0,962	0,962
Média Ponderada	0,988	0,983	0,981	0,992	0,990	0,990
MCC: aprovado	0,990	0,976	0,971	0,995	0,995	0,995
MCC: reprovado	0,969	0,948	0,941	0,977	0,969	0,969
MCC: evadido	0,949	0,947	0,947	0,965	0,957	0,957
Média Ponderada	0,982	0,968	0,964	0,989	0,986	0,986
Área ROC: aprovado	0,995	0,995	0,994	0,998	0,998	0,998
Área ROC: reprovado	0,999	0,992	0,993	0,986	0,987	0,988
Área ROC: evadido	0,995	0,971	0,962	0,968	0,980	0,981
Média Ponderada	0,995	0,991	0,990	0,992	0,994	0,994
Área PRC: aprovado	0,997	0,996	0,995	0,998	0,998	0,998
Área PRC: reprovado	0,996	0,929	0,936	0,961	0,973	0,976
Área PRC: evadido	0,944	0,931	0,931	0,938	0,950	0,950
Média Ponderada	0,990	0,978	0,978	0,985	0,988	0,989

Tabela 6.10: Resultados para os alunos do terceiro ano utilizando a técnica SMOTE e custos.

	Naive Bayes	**J48 sem poda**	**J48 com poda**	**IBk, k=1**	**IBk, k=3**	**IBk, k=5**
Instâncias corretamente classificadas	0,988395	0,98646	0,98646	0,992263	0,992263	0,992263
Instâncias incorretamente classificadas	0,011605	0,01354	0,01354	0,007737	0,007737	0,007737
Taxa de verdadeiros positivos: aprovado	0,997	0,995	0,995	1,000	1,000	1,000
Taxa de verdadeiros positivos: reprovado	0,961	0,987	0,987	0,987	0,987	0,974
Taxa de verdadeiros positivos: evadido	0,970	0,939	0,939	0,955	0,955	0,970
Média Ponderada	0,988	0,986	0,986	0,992	0,992	0,992
Taxa de falsos positivos: aprovado	0,007	0,007	0,014	0,007	0,007	0,007
Taxa de falsos positivos: reprovado	0,002	0,011	0,011	0,005	0,005	0,002
Taxa de falsos positivos: evadido	0,009	0,002	0,000	0,002	0,002	0,004
Média Ponderada	0,007	0,007	0,012	0,006	0,006	0,006
Precisão: aprovado	0,997	0,997	0,995	0,997	0,997	0,997
Precisão: reprovado	0,987	0,938	0,938	0,974	0,974	0,987
Precisão: evadido	0,941	0,984	1,000	0,984	0,984	0,970
Média Ponderada	0,989	0,987	0,987	0,992	0,992	0,992
Sensibilidade: aprovado	0,997	0,995	0,995	1,000	1,000	1,000
Sensibilidade: reprovado	0,961	0,987	0,987	0,987	0,987	0,974
Sensibilidade: evadido	0,970	0,939	0,939	0,955	0,955	0,970
Média Ponderada	0,988	0,986	0,986	0,992	0,992	0,992
Medida-F: aprovado	0,997	0,996	0,995	0,999	0,999	0,999
Medida-F: reprovado	0,974	0,962	0,962	0,981	0,981	0,980
Medida-F: evadido	0,955	0,961	0,969	0,969	0,969	0,970
Média Ponderada	0,988	0,986	0,986	0,992	0,992	0,992

MCC: aprovado	0,990	0,986	0,981	0,995	0,995	0,995
MCC: reprovado	0,969	0,956	0,956	0,977	0,977	0,977
MCC: evadido	0,949	0,956	0,965	0,965	0,965	0,965
Média Ponderada	0,982	0,977	0,975	0,989	0,989	0,989
Área ROC: aprovado	0,995	0,996	0,992	0,999	0,999	0,999
Área ROC: reprovado	0,999	0,990	0,989	0,986	0,986	0,988
Área ROC: evadido	0,995	0,973	0,965	0,969	0,982	0,982
Média Ponderada	0,995	0,992	0,988	0,993	0,995	0,995
Área PRC: aprovado	0,997	0,997	0,994	0,999	0,999	0,999
Área PRC: reprovado	0,996	0,942	0,937	0,972	0,971	0,975
Área PRC: evadido	0,944	0,944	0,959	0,951	0,950	0,950
Média Ponderada	0,990	0,982	0,981	0,989	0,989	0,989

Referente à precisão, sensibilidade, medida-F e área abaixo da curva ROC, para os dados dos alunos do terceiro ano, todos os algoritmos obtiveram desempenhos superiores com a aplicação das técnicas SMOTE e SMOTE e custos em comparação com os resultados obtidos com os dados desbalanceados.

Portanto, para os experimentos realizados com os dados referentes ao terceiro ano da instituição, fica evidente que a aplicação das técnicas SMOTE e SMOTE e custos obtêm, para todos os algoritmos utilizados na pesquisa, uma melhora no desempenho dos algoritmos em comparação com os resultados obtidos com dados desbalanceados e também pode-se observar que na maioria dos casos a técnica SMOTE e custos foi ligeiramente melhor que os experimentos somente com utilização da técnica SMOTE.

Analisando os resultados para todos os anos da instituição observa-se que são semelhantes e que nenhum dos algoritmos tem predominância com relação aos outros, obtendo-se, em geral, bom desempenho para a previsão e, com a aplicação das técnicas SMOTE e SMOTE e custos, é possível observar que as técnicas obtêm resultados superiores em comparação com os obtidos com os dados desbalanceados.

Vale salientar que as classes minoritárias (pendente, reprovado e evadido) possuem uma quantidade significativamente menor de registros nos dados do primeiro ano em comparação com a quantidade do segundo ano. O mesmo ocorre entre o segundo ano em comparação com o terceiro ano. Isto, indica que, quanto mais balanceada a proporção das classes com a aplicação das técnicas SMOTE e SMOTE e custos, melhores resultados podem ser obtidos pelos algoritmos de classificação.

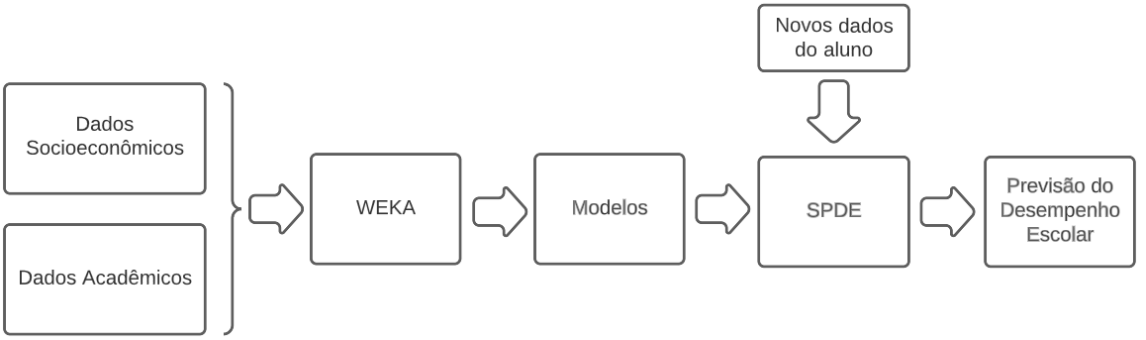
7. Protótipo de um Sistema para Previsão de Desempenho

Neste capítulo é apresentado um sistema denominado “sistema previsor de desempenho escolar” (SPDE), que foi desenvolvido com auxílio da linguagem de programação Java e a ferramenta WEKA. Como já mencionado, a plataforma WEKA contém as técnicas, algoritmos e funcionalidades necessárias e fundamentais para implementar software voltado à previsões dos alunos que possam vir a apresentar baixo desempenho no curso da instituição.

O sistema tem como objetivo auxiliar os gestores escolares de um curso médio-técnico de informática para Internet de uma ETEC, por meio de três algoritmos de mineração de dados (J48, Naive Bayes e KNN), identificando o aluno em uma das quatro possíveis classes, que possa vir a ser aprovado, o aluno aprovado, porém com pendências para o próximo ano, o reprovado ou então evadido do curso, antes que uma das situações citadas anteriormente venham a acontecer. Vale salientar que os alunos que são aprovados com pendências para o próximo ano são aqueles que recebem menção insuficiente na disciplina, tendo que refazer os trabalhos e avaliações posteriormente, os alunos que são reprovados são aqueles que recebem quatro ou mais menções insuficientes, tendo que refazer o ano letivo novamente no próximo ano e por fim os alunos que evadem, são aqueles que, na maioria dos casos, apresentam pendências no ano letivo na instituição e que acabam não retornando no ano seguinte, sem informar aos gestores os motivos da evasão, não dando continuidade aos estudos.

Conforme apresentado na Figura 7.1, o sistema baseado em dados da instituição utiliza as técnicas de *machine learning* apresentadas no Capítulo 4 e no Capítulo 6 para compor um índice final que classifica um aluno com sendo um possível aluno aprovado, reprovado, pendente ou evadido nos cursos da instituição.

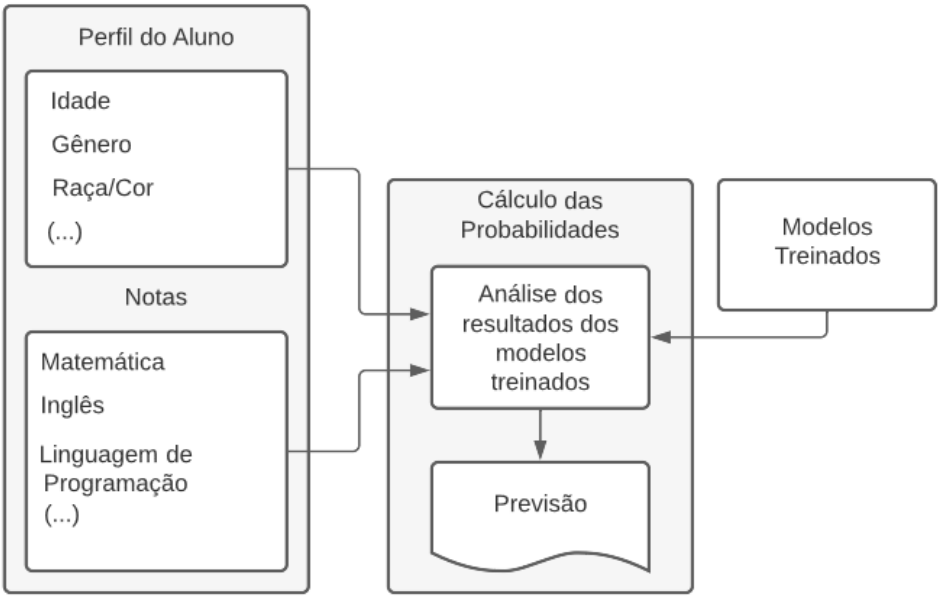
Figura 7.1: Processo geral do SPDE para previsão de desempenho escolar.



O sistema SPDE utiliza os modelos de classificação gerados a partir dos dados fornecidos pela instituição. Com esses modelos e os dados de um aluno específico, o sistema pode predizer o desempenho escolar do mesmo.

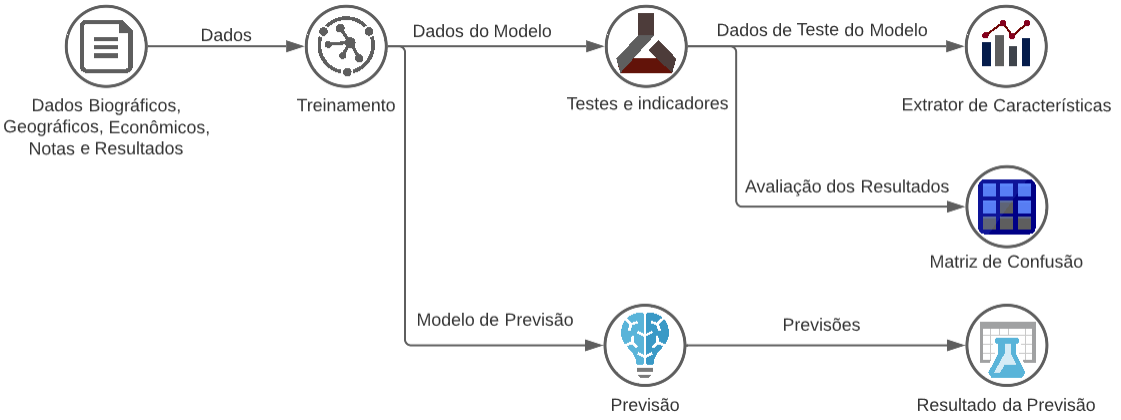
A Figura 7.2 apresenta a arquitetura conceitual do sistema, com os respectivos atributos e o fluxo de dados no SPDE, assim como as ferramentas utilizadas em cada uma das análises de cada algoritmo aplicado à pesquisa. As subseções deste capítulo detalham os componentes da arquitetura conceitual apresentada.

Figura 7.2: Arquitetura conceitual do SPDE.



Com auxílio da ferramenta WEKA, os dados são utilizados para treinar os modelos de classificação aplicados na pesquisa (Figura 7.3), sendo submetidos aos modelos que retornam o rótulo da classe obtendo então um resultado da situação do aluno.

Figura 7.3: Modelo de classificação.



7.1 Cálculo da Previsão

Após a análise dos perfis e notas dos alunos é calculada, usando a função "distributionForInstance()" presente na biblioteca do WEKA, a probabilidade para cada uma das quatro classes em cada um dos algoritmos.

A Figura 7.4 demonstra um trecho da codificação de como o cálculo da previsão é obtido pela função da biblioteca do WEKA para o algoritmo J48 e como a lista de previsão para cada rótulo da classe é gerado.

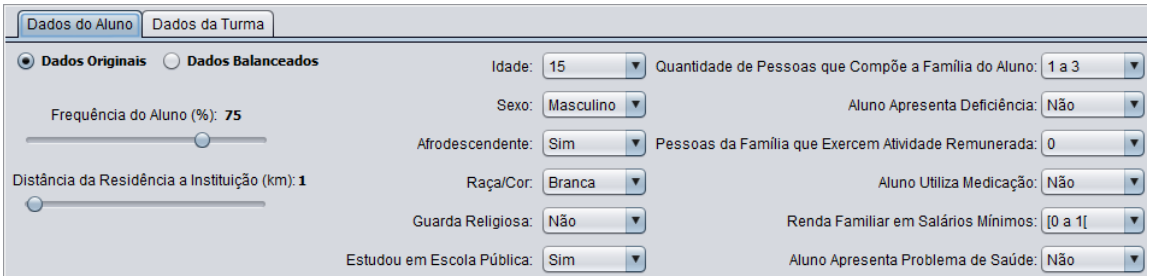
Figura 7.4: Cálculo da previsão utilizando a biblioteca do WEKA para o algoritmo J48.

```
J48 j48 = (J48) modeloJ48.readObject();
modeloJ48.close();

// Criação de novo registro
carregaBaseWeka(i);

Instance selecao = null;
switch (i) {
    case "j48-1":
        selecao = primeiroAnoParaARFF();
        break;
    case "j48-2":
        selecao = segundoAnoParaARFF();
        break;
    default:
        selecao = terceiroAnoParaARFF();
        break;
}
double resultado[] = j48.distributionForInstance(selecao);
DecimalFormat df = new DecimalFormat("###.##");
if (i == "j48-3") {
    txtAprovadoJ48.setText(df.format(resultado[0] * 100) + "%");
    txtReprovadoJ48.setText(df.format(resultado[1] * 100) + "%");
```

7.2 Apresentação do Uso do Sistema

Com o intuito de auxiliar os gestores da instituição de um curso médio-técnico de informática para internet de uma ETEC, foram desenvolvidos dois módulos para o sistema SPDE, com auxílio da linguagem de programação Java e da ferramenta WEKA. O sistema, desenvolvido para ser utilizado por um gestor (usuário), é executado localmente no equipamento e não necessita de conexão à internet e apresenta dois módulos principais (Figura 7.5), o primeiro denominado “Dados do Aluno” é o módulo referente a previsão a partir dos dados de apenas um aluno por vez, e o segundo módulo denominado “Dados da Turma” é responsável por carregar um conjunto de dados de uma fonte externa, podendo ser analisados os dados de diversos alunos simultaneamente.

Figura 7.5: *Layout* da janela do sistema para previsão de desempenho.

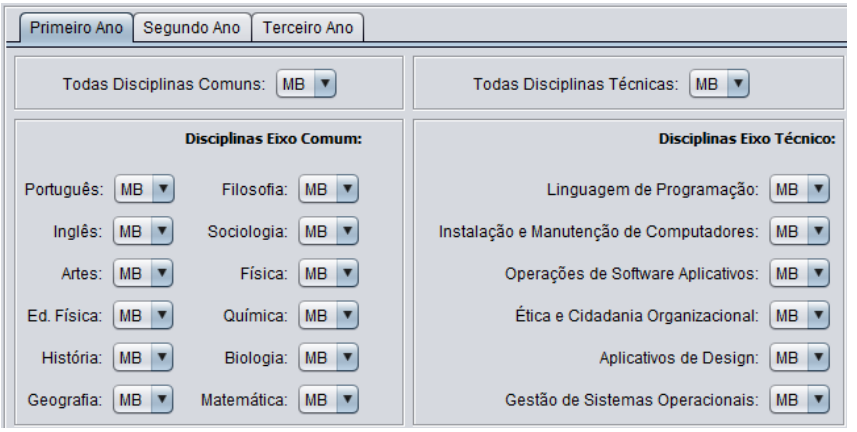


O primeiro módulo realiza a previsão para um aluno específico a partir do preenchimento dos dados do aluno necessários para se obter a previsão do resultado. O usuário tem a possibilidade de escolher a previsão do resultado a partir de modelos gerados com os dados originais ou com dados balanceados (Figura 7.5). A primeira opção denominada “Dados Originais” é a opção padrão, nela os dados utilizados são os originais, ou seja, os dados que foram submetidos apenas ao pré-processamento então, neste caso, as técnicas SMOTE e as funções de custos não foram aplicados. A segunda opção denominada “Dados Balanceados” utiliza as técnicas SMOTE e custos, tornando o arquivo de dados mais balanceado para as classes minoritárias.

O usuário deve alterar os dados do aluno através dos campos que fazem referência aos campos socioeconômicos dos alunos (biográficos, geográficos e econômicos) e dos *sliders* que permitem selecionar a frequência do aluno e a distância da residência do aluno até a instituição, conforme apresentados na Figura 7.5.

Os próximos campos que devem ser selecionados, através das abas “Primeiro Ano”, “Segundo Ano” ou “Terceiro Ano”, são referentes ao ano em que o aluno será analisado pelo sistema, por padrão a aba “Primeiro Ano” está ativa (Figura 7.6), porém o usuário pode selecionar outra aba e, quando uma das abas estiver ativa o sistema automaticamente seleciona o modelo que ele utilizará.

Figura 7.6: Aba selecionada para análise do primeiro ano.



Esses campos possuem as notas dos alunos divididas por eixo comum e eixo técnico e, para o segundo ano (Figura 7.7) e terceiro ano (Figura 7.8) apresentam respectivamente as pendências do primeiro ano e do segundo ano do aluno que está sendo analisado.

Figura 7.7: Aba selecionada para análise do segundo ano.

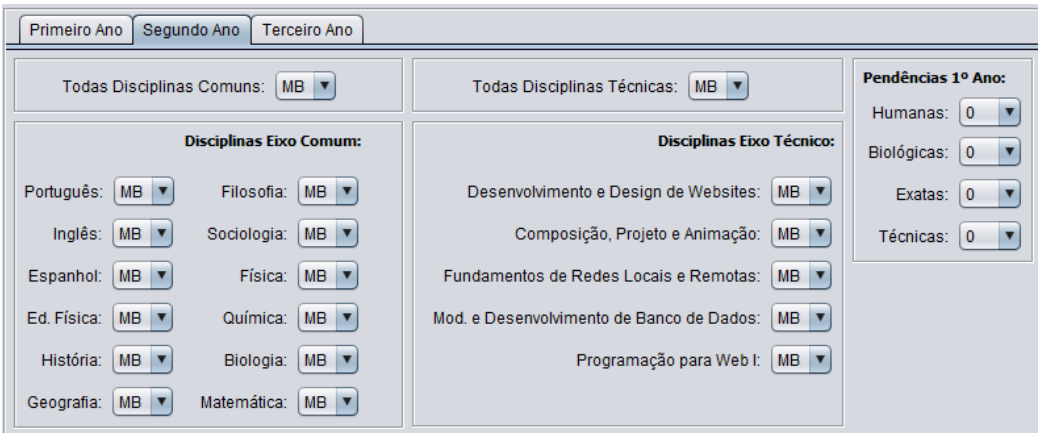
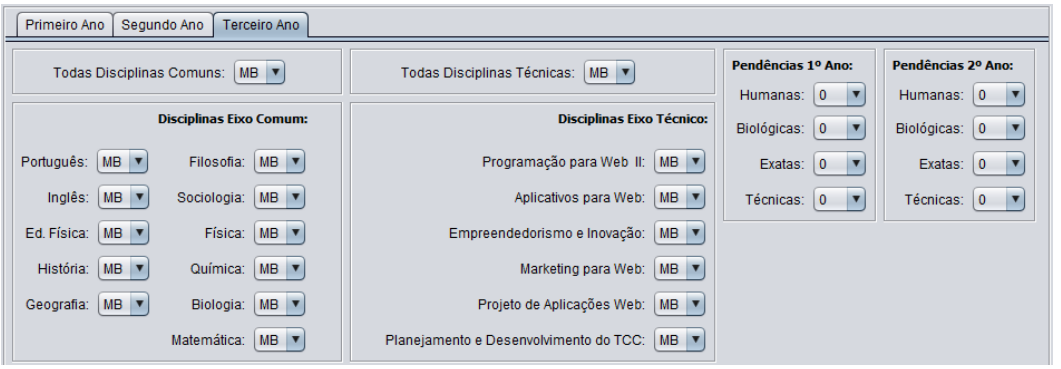
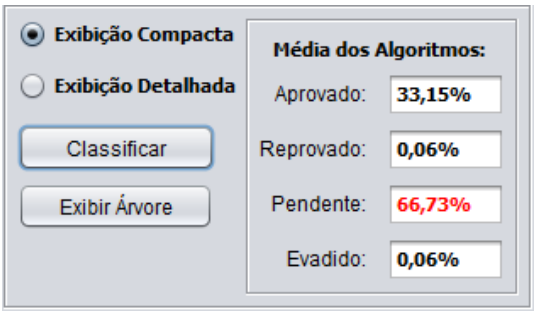


Figura 7.8: Aba selecionada para análise do terceiro ano.



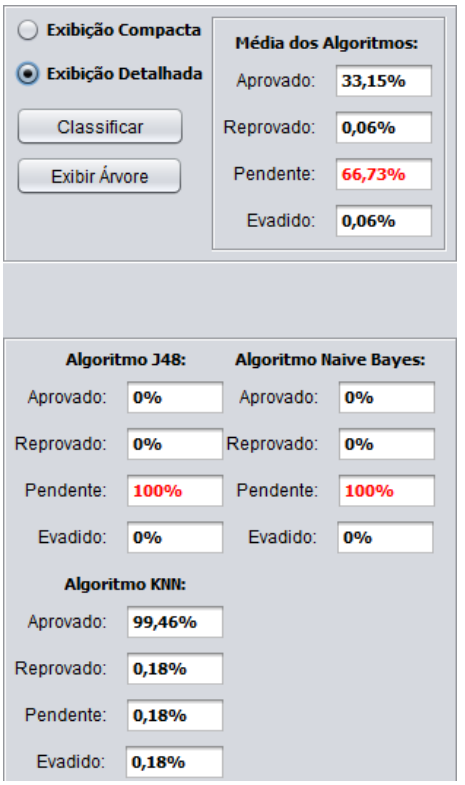
Após todos os campos serem ajustados o usuário pode calcular a média das probabilidades das predições obtidas pelos algoritmos Naive Bayes, J48 e KNN, essas são exibidas no campo denominado “Média dos Algoritmos”, conforme apresentada na Figura 7.9, quando a opção “Exibição Compacta” é selecionada apenas a média dos algoritmos fica habilitada para a visualização do usuário e é possível observar neste exemplo que o resultado exibido na opção “Pendente” fica identificado em vermelho, indicando ao usuário que o aluno tem uma grande probabilidade de ter pendências durante o ano letivo da instituição. Os resultados para “Reprovado” e “Evadido” também ficariam marcados em vermelho caso as probabilidades fossem, cada uma delas, maiores e iguais aos valores pré-determinados pelo sistema.

Figura 7.9: Cálculo da média dos resultados dos algoritmos.



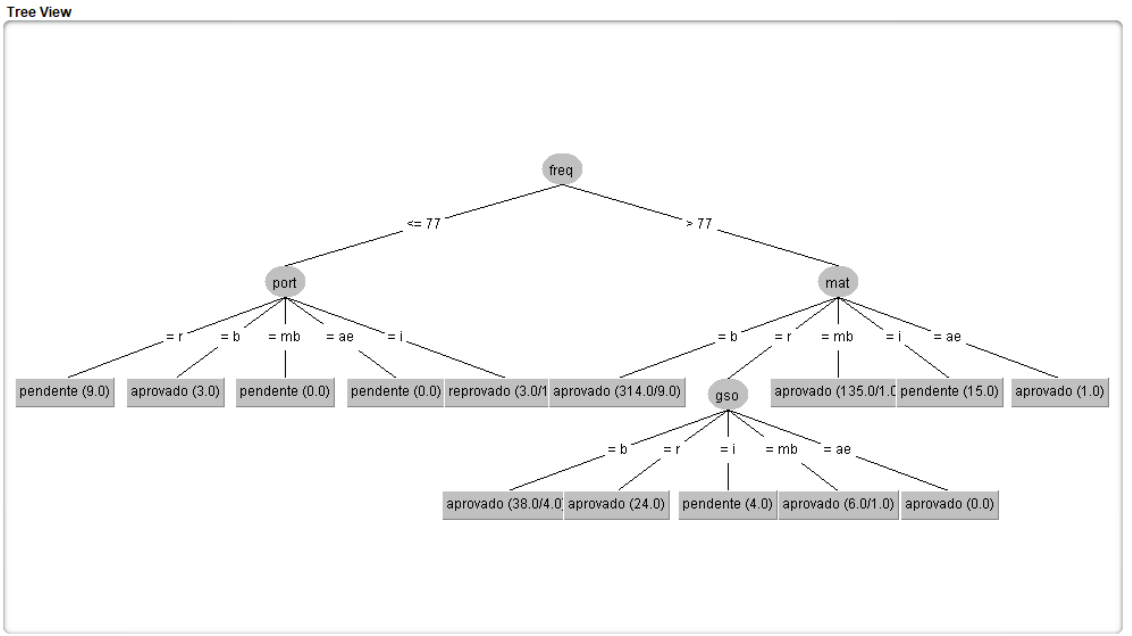
Referente a opção “Exibição Detalhada”, quando selecionada, um novo painel é exibido (Figura 7.10), então é possível observar os resultados obtidos para cada um dos algoritmos.

Figura 7.10: Opção "Exibição Detalhada" exibindo os algoritmos aplicados no sistema.



Um botão "Exibir Árvore" é exibido logo abaixo do botão "Classificar", se o usuário clicar neste botão uma nova janela com a árvore de decisão será aberta, a árvore exibida corresponde à arvore gerada pelo algoritmo J48 e é a árvore específica referente a qual aba ("Primeiro Ano", "Segundo Ano" ou "Terceiro Ano") e qual tipo de dados ("Dados Originais" ou "Dados Balanceados") foram selecionados pelo usuário. A exibição é feita na forma de árvore hierárquica, porém a distribuição visual dos nós a partir da raiz é feita automaticamente de modo a aproveitar da melhor maneira o espaço do navegador, podendo apresentar grandes quantidades de informações que podem ser analisadas pelo usuário. Os campos representados por uma elipse representam os atributos, e a partir da raiz são exibidas as arestas que representam os valores possíveis dos atributos, e as folhas representam as classes da previsão do desempenho dos alunos. A visualização da árvore pode ser controlada com o *mouse*, sendo possível movimentar a árvore com o ponteiro do *mouse* e aproximar/afastar a visualização. A Figura 7.11 apresenta de maneira geral o layout da página de visualização.

Figura 7.11: Árvore de decisão do primeiro ano para os dados desbalanceados.



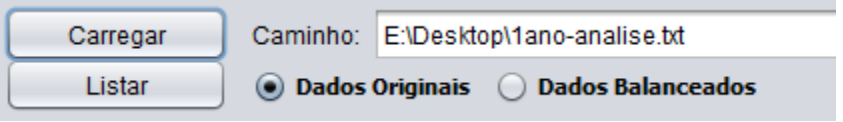
É possível observar que, a árvore de decisão exibe graficamente os atributos mais representativos na previsão dos resultados, assim facilitando a análise de uma forma mais direta e visual para aquele que irá fazer uso do sistema.

Por fim, o segundo módulo representado pela aba “Dados da Turma”, conforme citado anteriormente, é responsável por carregar dados externos para o sistema, a vantagem deste módulo está no fato de poder analisar dados de diversos alunos (conjunto de dados) simultaneamente. Neste módulo, assim como o modulo anterior, é possível selecionar que tipo de dados será utilizado na predição (“Dados Originais” ou “Dados Balanceados”) e que tipo de exibição de resultados será utilizada pelo usuário do sistema (“Exibição Compacta” ou “Exibição Detalhada”). Neste módulo, assim como no módulo anterior, são analisadas as opções biográficas, geográficas, econômicas e de notas dos alunos (disciplinas do eixo comum e técnico).

Para carregar os dados no sistema o usuário deve clicar no botão “Carregar”, uma nova caixa de diálogo do sistema operacional é exibida e o usuário pode localizar e selecionar o arquivo específico, o caminho do arquivo que foi carregado no sistema é exibido no campo “Caminho”, conforme observado na Figura 7.12. Para o usuário este

campo serve para facilitar a visualização do caminho e o nome do arquivo utilizado na previsão, assim evitando que o usuário fique confuso no carregamento.

Figura 7.12: Botão carregar e caminho do arquivo com o nome carregado no sistema para previsão.



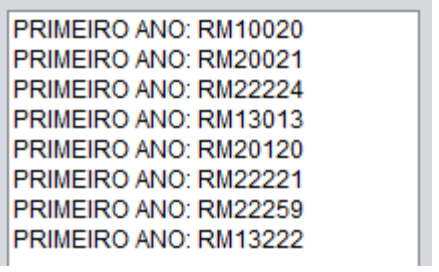
Os tipos de arquivos que são aceitos pelo sistema para previsão de desempenho devem ser obrigatoriamente de extensão ".txt", um exemplo pode ser observado na Figura 7.13, este conjunto de dados deve conter valores separados por vírgula, e o primeiro campo de cada registro de dados deve conter o RM de cada aluno que será analisado pelo sistema, neste exemplo são exibidos os dados referentes a uma situação hipotética referente ao primeiro ano da instituição.

Figura 7.13: Exemplo do arquivo ".txt" com diversos dados para análise.

```
RM10020,16,17,feminino,sim,sim,parda,[1a3],2,[1a2[,nao,nao,nao,nao,98,mb,r,r,r,b,b,i,mb,b,b,i,b,r,b,b,b,r,i,?
RM20021,15,20,masculino,sim,sim,branca,[4a6],1,[0a1[,nao,nao,nao,nao,90,r,mb,b,b,b,b,b,b,r,b,r,r,r,b,b,b,r,b,?
RM22224,15,2,feminino,sim,sim,parda,[4a6],0,[2a3[,nao,nao,nao,nao,80,r,b,b,b,b,r,i,mb,r,b,r,b,b,b,b,b,i,i,?
RM13013,20,7,feminino,sim,sim,preta,[7a9],2,[1a2[,nao,nao,nao,nao,78,b,r,r,r,r,i,b,b,r,r,i,r,r,r,i,b,r,r,?
RM20120,18,50,feminino,sim,sim,parda,[1a3],1,[1a2[,nao,nao,nao,nao,79,mb,r,r,r,b,b,i,mb,b,b,i,b,r,b,b,b,r,i,?
RM22221,15,29,masculino,sim,sim,indigena,[4a6],1,[0a1[,nao,nao,nao,nao,99,r,mb,b,b,b,b,b,b,r,b,r,r,r,b,b,b,r,b,?
RM22259,15,5,feminino,sim,sim,amarela,[1a3],0,[2a3[,nao,nao,nao,nao,67,r,b,i,b,b,r,r,r,r,b,r,i,b,b,b,b,r,i,?
RM13222,19,9,feminino,sim,sim,preta,[4a6],3,[1a2[,nao,nao,nao,nao,50,b,i,r,r,r,i,b,b,i,r,r,r,r,r,i,r,r,?
```

Após o caminho do arquivo que será analisado ser listado no sistema, é necessário selecionar a qual ano ("Primeiro Ano", "Segundo Ano" ou "Terceiro Ano") pertence esses dados e clicar no botão "Listar" que irá exibir a lista de todos os registros contidos no arquivo ".txt" para o ano selecionado e o RM do aluno que será analisado (Figura 7.14).

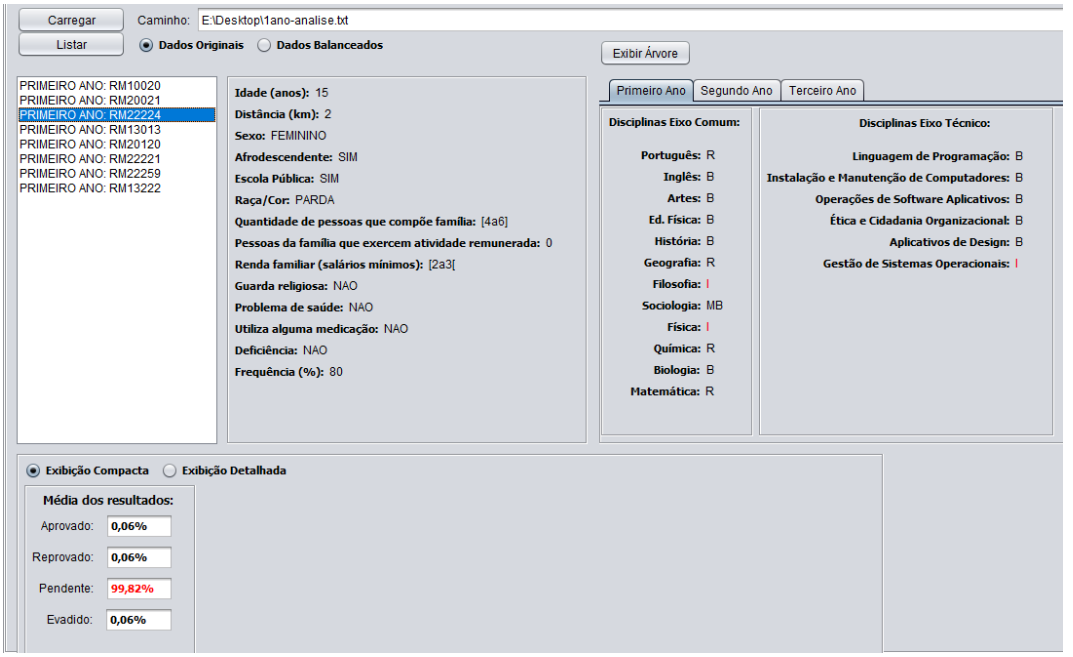
Figura 7.14: Lista do arquivo ".txt" com os alunos para análise.



Quando o usuário clicar em um dos registros exibidos na lista serão apresentados os detalhes daquele aluno e os resultados da média dos algoritmos de forma automática

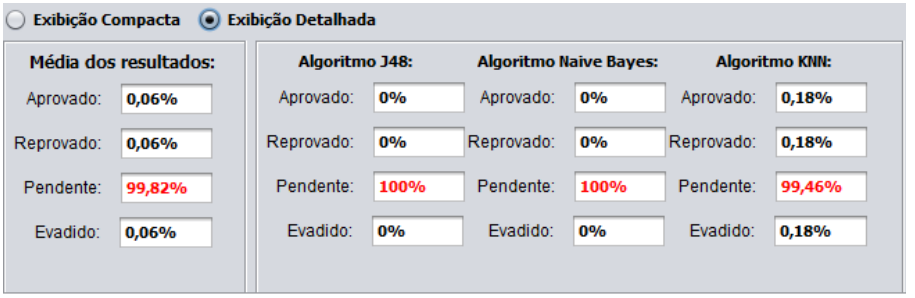
(Figura 7.15). Assim como no módulo anterior, há um botão que exibe a árvore de decisão, para isso o usuário deverá selecionar para qual ano gostaria de exibir a árvore.

Figura 7.15: Detalhes dos dados do aluno e média dos resultados dos algoritmos.



Para que o usuário possa realizar uma análise mais detalhada dos resultados dos algoritmos, este deverá selecionar a opção denominada “Exibição Detalhada” e o sistema exibirá um novo painel, onde serão exibidos os detalhes para os algoritmos (Figura 7.16).

Figura 7.16: Exibição detalhada dos algoritmos e média dos seus resultados.



Em ambos os módulos fica evidente que após realizar as análises pode-se obter uma previsão que não é auspiciosa (reprovado, pendente, evadido) nem para o aluno nem para a instituição e indica que uma análise mais detalhada deve ser realizada para aquele aluno específico, portanto, esse aluno deve receber uma atenção especial e mais

aprofundada, assim evitando que ele perca o interesse pelos estudos e possa ser resgatado pela instituição.

O sistema ainda não foi avaliado formalmente pela instituição, só por alguns funcionários e professores e o seu uso parece ser promissor, embora sejam necessários ajustes, com relação aos parâmetros disponíveis na ferramenta, atributos mais relevantes, entre outros.

Esses ajustes podem ser estruturados a partir de dois métodos, o *Focus Group* (grupo focal) e o Delphi.

Kitzinger (2000) explana o método *Focus Group* como a busca de informações que possam proporcionar a compreensão sobre um tema. O método tem como principal objetivo reunir informações detalhadas referente a esse tema em específico, a partir um grupo de participantes selecionados e pode ser sugerido por um pesquisador, coordenador ou moderador de um grupo.

Outro método conhecido é o Delphi onde, segundo Linstone & Turoff (2002), é utilizado para estruturar um processo de comunicação coletiva de modo que este seja efetivo, ao permitir a um grupo de indivíduos, como um todo, lidar com um problema complexo.

8. Conclusões e Trabalhos Futuros

Este trabalho pesquisou o uso de três algoritmos de MD para predizer o desempenho dos alunos do curso de Informática para Internet Integrado ao Ensino Médio da Escola Técnica Estadual (ETEC) Bartolomeu Bueno da Silva – Anhanguera. Os algoritmos Naive Bayes, J48 com e sem poda e o IBk para k = 1, 3 e 5 foram avaliados considerando três experimentos com dados desbalanceados (que foram fornecidos pela instituição em planilhas do Microsoft Excel) e a aplicação das técnicas SMOTE e SMOTE com Modelos Penalizados (custos), as duas últimas com a finalidade de melhorar os resultados inicialmente obtidos em comparação aos experimentos com os algoritmos e os dados desbalanceados. Considerando os resultados para o primeiro, segundo e terceiro ano, em geral, os melhores resultados foram obtidos, após realização desses experimentos, com a aplicação da técnica SMOTE e custos.

Acredita-se que os resultados obtidos neste estudo possam ajudar os educadores e administradores educacionais, uma vez que é possível obter estimativas sobre o desempenho dos alunos, que podem servir de base para o planejamento de estratégias e políticas que visem diminuir o número de pendências e reprovações, reduzindo, como consequência, a evasão dos alunos do curso. Levando em consideração os resultados obtidos, é possível observar que quanto mais balanceados forem os dados, melhores são os resultados obtidos na previsão pelos algoritmos de classificação e as técnicas que são aplicadas a esses algoritmos, que tem a finalidade de melhorar o desempenho dos algoritmos utilizados na pesquisa.

Ainda considerando os resultados obtidos, foi desenvolvido um sistema que possibilita a integração do processo de geração dos diferentes modelos em uma ferramenta que pode ser utilizada por educadores e administradores para acompanhar os alunos ao longo do curso e assim tomar decisões para melhorar a situação desses alunos, em particular, e da instituição, em geral.

Como trabalhos futuros, considera-se também realizar uma análise mais detalhada do processo de seleção dos atributos de entrada, investigando a existência de outros atributos que possam influenciar no desempenho do aluno regularmente matriculado na instituição, assim como estudar a relação do desempenho dos alunos ao longo do curso.

Um outro ponto a ser considerado é referente ao desbalanceamento dos dados, estudando outras abordagens que permitam obter, eventualmente, melhores resultados em comparação a essa abordagem. Além disso, acredita-se que é possível realizar uma pesquisa semelhante para alunos de outros cursos da instituição, verificando se os resultados são similares para outras disciplinas.

Um outro trabalho futuro é continuar testando o protótipo para a implementação final do sistema de previsão a ser utilizado pela instituição que forneceu os dados.

Referências

Arun, D. K., Namratha, V., Ramyashree, B. V., Jain, Y. P., & Roy Choudhury, A., (2021). *Student Academic Performance Prediction using Educational Data Mining*, *2021, International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, pp. 1-9, doi: 10.1109/ICCCI50826.2021.9457021.

Baker, R., Inventado, P., (2014). *Educational Data Mining and Learning Analytics*. Larusson, J., White, B. (eds) Learning Analytics, pp 61-75, Springer, New York, doi 10.1007/978-1-4614-3305-7_4.

Biolchini, J., Mian, P. G., Natali, A. C. C., & Travassos, G. H. (2005). *Systematic review in software engineering. System Engineering and Computer Science Department* COPPE/UFRJ, *Technical Report ES*, 679(05), pp. 45.

Blum, A. and Mitchell, T. (1998). *Combining labeled and unlabeled data with co-training. In Proc. 11th Annu. Conf. on Comput. Learning Theory, pp.* 92–100. ACM *Press, New York, NY*.

Brito, D., Júnior, I., Queiroga, E., Rêgo, T., (2014). Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina, doi: 882. 10.5753/cbie.sbie.2014.882.

Bujang, S. D. A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., Ghani, N. R., (2021). *Multiclass Prediction Model for Student Grade Prediction Using Machine Learning*, in *IEEE Access*, vol. 9, pp. 95608-95621, doi: 10.1109/ACCESS.2021.3093563.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., (2002). Smote: *Synthetic Minority Over-sampling Technique*, vol. 16, pp. 321–357.

Cover, T. M., Hart, P. E., (1967). *Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory*, pp. 21-27.

Domingos, P. (1999). *Metacost: A General Method for Making Classifiers Cost-sensitive. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–164 San Diego, CA. ACM *Press*.

ETEC (2022). Descrição do Técnico em Informática para Internet. https://www.vestibulinhoetec.com.br/unidades-cursos/curso.asp?c=703 (accessed on 20 June 2022).

Faria, M. M., (2016). Detecção de intrusões em redes de computadores com base nos algoritmos KNN, K-Means++ e J48. Dissertação (Programa de Mestrado em Ciência da Computação) – Faculdade Campo Limpo Paulista – FACCAMP. São Paulo.

Fayyad, U. M., Haussler, D., & Stolorz, P. E., (1996). KDD *for Science Data Analysis: Issues and Examples*. *In* KDD (pp. 50-56).

Gil, P.D., da Cruz Martins, S., Moro, S. et al., (2021). *A data-driven approach to predict first-year students' academic success in higher education institutions*. *Educ Inf Technol* 26, pp. 2165–2190. https://doi.org/10.1007/s10639-020-10346-6.

Han, J., Kamber, M. & Pei, J., (2011). Data Mining Concepts Techniques. 3° ed. São Francisco: Morgan Kaufmann.

Hasib, K. M., Rahman, F., Hasnat, R., Alam, M. G. R., (2022). "*A Machine Learning and Explainable AI Approach for Predicting Secondary School Student Performance*," *IEEE* 12*th Annual Computing and Communication Workshop and Conference* (CCWC*)*, Las Vegas, NV, EUA, pp. 0399-0405, doi: 10.1109/CCWC54503.2022.9720806.

He, H., Bai, Y., Garcia, E. A. & Li, S., (2008). "ADASYN: Adaptive synthetic sampling approach for imbalanced learning", *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

Japkowicz, N., (2000). *The Class Imbalance Problem: Significance and Strategies. In Proceedings of the 2000 International Conference on Artificial Intelligence* (IC-AI'2000): *Special Track on Inductive Learning*, vol. 56, pp. 111-117, *Las Vegas, Nevada*.

Kitchenham, B., Charters, S., (2007). *Guidelines for performing systematic literature reviews in software engineering (version 2.3)*. *Technical report, Keele University and University of Durham*.

KITZINGER, J., (2000). Focus groups with users and providers of health care. In: POPE, C.; MAYS, N. (Org.). Qualitative research in health care. 2. ed. London: BMJ Books.

Kubat, M. & Matwin, S., (1997). *Addressing the Curse of Imbalanced Training Sets: One Sided Selection. In Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 179–186, *Nashville, Tennesse. Morgan Kaufmann*.

Lewis, D., & Catlett, J., (1994). *Heterogeneous Uncertainity Sampling for Supervised Learning*. In *Proceedings of the Eleventh International Conference of Machine Learning*, pp. 148–156, San Francisco, CA. Morgan Kaufmann.

Ling, C., & Li, C., (1998). Data Mining for Direct Marketing Problems and Solutions. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), vol. 98, pp. 73-79, New York, NY. AAAI Press.

Linstone, H. A., & Turoff, M., (2002). The Delphi method: Techniques and applications. Addison Wesley Newark, NJ: New Jersey Institute of Technology. Disponível em <http://www.foresight.pl/assets/downloads/publications/Turoff_Linstone.pdf >.

Little, R. J. A. & Rubin D. B., (2002). Statistical analysis with missing data. 2ª ed. Nova York: Wiley & Sons.

Matsubara, E. T., (2004). O algoritmo de aprendizado semi-supervisionado CO-TRAINING e sua aplicação na rotulação de documentos. Dissertação de Mestrado, ICMC-USP. http://www.teses.usp.br/teses/disponiveis/55/55134/tde-19082004-092311/.

Meedech, P., Iam-On, N., Boongoen, T., (2016). *Prediction of Student Dropout Using Personal Profile and Data Mining Approach*. *In*: Lavangnananda, K., Phon-Amnuaisuk, S., Engchuan, W., Chan, J. (eds) *Intelligent and Evolutionary Systems*. *Proceedings in Adaptation, Learning and Optimization*, vol 5. Springer, Cham. https://doi.org/10.1007/978-3-319-27000-5_12.

Monard, M. C., Baranauskas, J. A., (2003). Conceitos sobre Aprendizado de Máquina. Sistemas Inteligentes Fundamentos e Aplicações. 1 ed. Barueri-SP: Manole Ltda, pp. 89-114. ISBN 85-204-168.

Nahar, K., Shova, B. I., Ria, T. et al., (2021). *Mining educational data to predict students performance*. *Educ Inf Technol* 6, pp. 6051–6067. https://doi.org/10.1007/s10639-021-10575-3.

Neiva, F., Silva, R., (2016). Revisão Sistemática da Literatura em Ciência da Computação - Um Guia Prático. 10.13140/RG.2.1.1445.3361.

Paganatto, R.A, Monteiro, A.M., (2021). Previsão de evasão de alunos utilizando mineração de dados: uma revisão da literatura. pp. 1-8. XVII Workshop de Computação da UNIFACCAMP.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T. & Brunk, C., (1994). *Reducing Misclassification Costs*. *In Proceedings of the Eleventh International Conference on Machine Learning San Francisco, CA. Morgan Kauffmann.*

Pyle, D. *Data preparation for data mining*. São Francisco: Morgan Kaufmann, 1999.

Quinlan J. R., "*Induction of decision tree*", *Machine Learning*, 1986, vol. 1, pp.81-106.

Ramaphosa, K. I. M., Zuva, T., Kwuimi, R., (2018). "*Educational Data Mining to Improve Learner Performance in Gauteng Primary Schools*," *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa*, pp. 1-6, doi: 10.1109/ICABCD.2018.8465478.

Romero, C., Ventura, S. (2013). *Data Mining in Education.* Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery*, vol 3, pp. 12-27. https://doi.org/10.1002/widm.1075.

Saa, A. A., Al-Emran, M., Shaalan, K., (2020). *Mining Student Information System Records to Predict Students' Academic Performance*. *In*: Hassanien, A., Azar, A., Gaber, T., Bhatnagar, R., F. Tolba, M. (eds) *The International Conference on Advanced Machine Learning Technologies and Applications* (AMLTA2019). AMLTA 2019. *Advances in Intelligent Systems and Computing*, vol 921, pp. 229-239. Springer, Cham. https://doi.org/10.1007/978-3-030-14118-9_23.

WEKA (2022). *Data Mining Software in Java*. Available at https://www.cs.waikato.ac.nz/ml/weka/ (accessed on 15 June 2022).

Witten, I. H., Frank, E. & Hall, M. A., (2011). *Data Mining Practical Machine Learning Tools and Techniques. 3° ed. Burlington: Morgam Kaufman & Elsevier*.

Wu, X. et al., (2008). *Top 10 algorithms in data mining. Knowledge and Information Systems*, vol. 14, pp. 1-37. https://doi.org/10.1007/s10115-007-0114-2.

Yağcı, M., (2022). *Educational data mining: prediction of students' academic performance using machine learning algorithms*. *Smart Learn. Environ*. 9, pp. 11. https://doi.org/10.1186/s40561-022-00192-z.

Apêndice A - Matrizes de custos

Nessa seção são apresentas as matrizes de custos, estas matrizes são configuradas no ambiente WEKA para a utilização da técnica SMOTE com custos, conhecida como *Cost-sensitive Learning* (Aprendizado Sensível ao Custo), a técnica aplicada a pesquisa é utilizada em conjunto com sobreamostragem e usa modelos que penalizam os erros gerados pelos algoritmos na classificação, a técnica aplica diferentes funções de custos, podendo penalizar as classes de maneira igual ou diferente para cada classe.

A seguir são apresentadas as matrizes de custos para os algoritmos Naive Bayes, J48 com e sem poda da árvore e IBk com $k = 1, 3$ e 5 para as classes do primeiro ano da instituição.

Tabela A1: Matriz de custos para as classes do primeiro ano para o algoritmo Naive Bayes.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	2.0	0.0	2.0	2.0
	pendente	2.0	2.0	0.0	2.0
	evadido	2.0	2.0	2.0	0.0

Tabela A2: Matriz de custos para as classes do primeiro ano para o algoritmo J48 sem poda.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	0.5	0.0	0.5	0.5
	pendente	0.5	0.5	0.0	0.5
	evadido	0.5	0.5	0.5	0.0

Tabela A3: Matriz de custos para as classes do primeiro ano para o algoritmo J48 com poda.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	2.0	0.0	2.0	1.5
	pendente	2.0	1.0	0.0	0.5
	evadido	2.0	2.0	1.0	0.0

Tabela A4: Matriz de custos para as classes do primeiro ano para o algoritmo IBk com $k = 1$.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	2.0	0.0	2.0	2.0
	pendente	2.0	2.0	0.0	2.0
	evadido	2.0	2.0	2.0	0.0

Tabela A5: Matriz de custos para as classes do primeiro ano para o algoritmo IBk com $k = 3$.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	2.0	0.0	2.0	2.0
	pendente	2.0	2.0	0.0	2.0
	evadido	2.0	2.0	2.0	0.0

Tabela A6: Matriz de custos para as classes do primeiro ano para o algoritmo IBk com $k = 5$.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	2.0	0.0	2.0	2.0
	pendente	2.0	2.0	0.0	2.0
	evadido	2.0	2.0	2.0	0.0

Nas tabelas a seguir são apresentadas as matrizes de custos para os algoritmos Naive Bayes, J48 com e sem poda da árvore e IBk com $k = 1, 3$ e 5 para as classes do segundo ano da instituição.

Tabela A7: Matriz de custos para as classes do segundo ano para o algoritmo Naive Bayes.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	2.0	0.0	1.5	0.5
	pendente	0.5	0.5	0.0	0.5
	evadido	2.0	2.0	2.0	0.0

Tabela A8: Matriz de custos para as classes do segundo ano para o algoritmo J48 sem poda.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	2.0	0.0	2.0	0.5
	pendente	2.0	0.5	0.0	0.5
	evadido	2.0	2.0	2.0	0.0

Tabela A9: Matriz de custos para as classes do segundo ano para o algoritmo J48 com poda.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	1.0	0.0	0.5	0.5
	pendente	1.0	0.5	0.0	0.5
	evadido	1.5	0.5	0.5	0.0

Tabela A10: Matriz de custos para as classes do segundo ano para o algoritmo IBk com $k = 1$.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	2.0	0.0	2.0	2.0
	pendente	2.0	2.0	0.0	2.0
	evadido	2.0	2.0	2.0	0.0

Tabela A11: Matriz de custos para as classes do segundo ano para o algoritmo IBk com $k = 3$.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	0.5	0.0	0.5	0.5
	pendente	0.5	0.5	0.0	0.5
	evadido	2.0	2.0	2.0	0.0

Tabela A12: Matriz de custos para as classes do segundo ano para o algoritmo IBk com $k = 5$.

		CLASSE PREDITA			
		aprovado	**reprovado**	**pendente**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5	0.5
	reprovado	2.0	0.0	1.0	0.5
	pendente	0.5	0.5	0.0	0.5
	evadido	2.0	2.0	2.0	0.0

Por fim, nas tabelas a seguir são apresentadas as matrizes de custos para os algoritmos Naive Bayes, J48 com e sem poda da árvore e IBk com $k = 1, 3$ e 5 para as classes do terceiro ano da instituição.

Tabela A13: Matriz de custos para as classes do terceiro ano para o algoritmo Naive Bayes.

		CLASSE PREDITA		
		aprovado	**reprovado**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5
	reprovado	2.0	0.0	2.0
	evadido	2.0	2.0	0.0

Tabela A14: Matriz de custos para as classes do terceiro ano para o algoritmo J48 sem poda.

		CLASSE PREDITA		
		aprovado	**reprovado**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5
	reprovado	1.5	0.0	0.5
	evadido	0.5	0.5	0.0

Tabela A15: Matriz de custos para as classes do terceiro ano para o algoritmo J48 com poda.

		CLASSE PREDITA		
		aprovado	**reprovado**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5
	reprovado	1.5	0.0	0.5
	evadido	0.5	0.5	0.0

Tabela A16: Matriz de custos para as classes do terceiro ano para o algoritmo IBk com $k = 1$.

		CLASSE PREDITA		
		aprovado	**reprovado**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5
	reprovado	2.0	0.0	2.0
	evadido	2.0	2.0	0.0

Tabela A17: Matriz de custos para as classes do terceiro ano para o algoritmo IBk com $k = 3$.

		CLASSE PREDITA		
		aprovado	**reprovado**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5
	reprovado	2.0	0.0	2.0
	evadido	1.5	0.5	0.0

Tabela A18: Matriz de custos para as classes do terceiro ano para o algoritmo IBk com $k = 5$.

		CLASSE PREDITA		
		aprovado	**reprovado**	**evadido**
CLASSE ORIGINAL	**aprovado**	0.0	0.5	0.5
	reprovado	2.0	0.0	2.0
	evadido	0.5	0.5	0.0