UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO DE JOINVILLE
ENGENHARIA MECATRÔNICA

EMANUEL TÚRMINA TORRES

ESTUDO DE CASO PARA A PREDIÇÃO DE REPROVAÇÃO ESTUDANTIL NA DISCIPLINA DE PROGRAMAÇÃO I: UMA ANÁLISE EXPLORATÓRIA E APLICAÇÃO DE APRENDIZADO DE MÁQUINA

Joinville
2024

EMANUEL TÚRMINA TORRES

ESTUDO DE CASO PARA A PREDIÇÃO DE REPROVAÇÃO ESTUDANTIL NA DISCIPLINA DE PROGRAMAÇÃO I: UMA ANÁLISE EXPLORATÓRIA E APLICAÇÃO DE APRENDIZADO DE MÁQUINA

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

Orientador: Dr. Benjamin Grando Moreira

Joinville
2024

Dedico este trabalho aos meus queridos pais e à minha amada família.

AGRADECIMENTOS

Agradeço a minha família por todo o apoio e encorajamento constantes ao longo da minha trajetória estudantil. A presença e o carinho de vocês sempre me incentivaram a perseguir meus sonhos, e só cheguei onde estou graças a vocês. Um agradecimento especial à minha mãe, Diane, ao meu pai, Jeferson, e à minha irmã, Polyana. Reencontrar vocês durante as férias sempre renovava minhas energias para enfrentar cada novo semestre.

Agradeço a meu orientador, Dr. Benjamin Grando Moreira, por todo o conhecimento compartilhado para condução deste trabalho. Obrigado por me acompanhar neste projeto, pela paciência conforme os problemas foram surgindo e por acreditar em meu potencial. Através de conversas, conselhos e desafios, sua orientação contribuiu muito para meu desenvolvimento profissional e acadêmico.

Agradeço a meus amigos por deixaram essa jornada muito mais divertida. Mesmo diante dos períodos mais complicados, o fato de poder contar com a companhia e o apoio de vocês sempre me trouxe tranquilidade. Guardo com carinho todos os momentos que compartilhamos, desde as celebrações natalinas ao final de cada ano até as viagens espontâneas durante um semestre caótico. Muito obrigado; esses anos ao lado de vocês foram maravilhosos.

"Life's but a walking shadow, a poor player, that struts and frets his hour upon the stage, and then is heard no more."

William Shakespeare, Macbeth

RESUMO

A retenção estudantil é uma preocupação crítica no ambiente acadêmico, com desafios complexos que afetam a trajetória educacional dos estudantes. A reprovação é uma das principais causas, e a intervenção precoce é apontada como uma estratégia eficaz na solução desse problema. Nesse contexto, propõe-se nesse trabalho a implementação de um modelo preditivo classificativo baseado em aprendizado de máquina destinado a antecipar o desempenho dos estudantes. O trabalho faz uso da mineração de dados educacionais a partir de relatórios obtidos do ambiente Moodle e do sistema acadêmico da UFSC. O método foi avaliado em uma disciplina de Programação I do CTJ, onde antecipou estudantes reprovados para 25%, 50%, 75% e 95% do semestre letivo, obtendo sensibilidade de 90,8%, 92,3%, 95,4% e 100% com LR, SVM, RNA e AD, respectivamente. Visando-se criar um algoritmo genérico, RNA foi selecionado por demonstrar uma sensibilidade média de 91,13% para os três primeiros períodos, conjunto de tempo no qual a intervenção por parte do professor é efetiva.

Palavra-chave: modelo preditivo; mineração de dados educacionais; reprovação escolar.

ABSTRACT

Student retention is a critical concern in the academic environment, with complex challenges affecting students' educational trajectories. Failure is one of the main causes, and early intervention is identified as an effective strategy in solving this problem. In this context, this paper proposes the implementation of a predictive classification model based on machine learning aimed at anticipating student performance. The work utilizes educational data mining from reports obtained from the Moodle environment and the academic system of UFSC. The method was evaluated in a Programming I course at CTJ, where it anticipated failing students for 25%, 50%, 75%, and 95% of the academic semester, achieving sensitivities of 90.8%, 92.3%, 95.4%, and 100% with LR, SVM, RNA, and AD, respectively. Aiming to create a generic algorithm, RNA was selected for demonstrating an average sensitivity of 91.13% for the first three periods, a time frame in which teacher intervention is effective.

Keywords: predictive model; educational data mining; academic failure.

LISTA DE FIGURAS

Figura 1 – Aprendizado Supervisionado por Classificação 17
Figura 2 – Matriz de Confusao Genérica 22
Figura 3 – Parte de um Relatório de Dados Cadastrais do Sistema Acadêmico 27
Figura 4 – Parte de um Relatório de Presenças do Moodle 31
Figura 5 – Parte de um Relatório de Notas do Moodle 32
Figura 6 – Exemplo de Fluxo na Ferramenta Orange 37
Figura 7 – Widgets *Discretize* e *Test and Score* do Orange 38
Figura 8 – Widgets Confusion Matrix e Data Table do Orange 38
Figura 9 – Análise Descritiva por Curso 42
Figura 10 – Distribuição dos Alunos por Ano de Ingresso 42
Figura 11 – Distribuição do Tipo de Ensino por Ano de Ingresso 43
Figura 12 – Tendência do IAP e IAA dos Estudantes ao Longo dos Anos 44
Figura 13 – Ferramenta EduMap 45
Figura 14 – Distinção Regional entre Engenharia Civil e Engenharia Naval para População Branca 46
Figura 15 – Distinção Regional entre Engenharia Mecatrônica e Engenharia Automotiva para População Parda 47
Figura 16 – Distribuição da População Parda de Engenharia Naval por Gênero . 48
Figura 17 – Distribuição dos Grupos de Valor Alto e Médio-Baixo para População Parda 50
Figura 18 – Fluxo de Dados Integrados 58

LISTA DE TABELAS

Tabela 1 – Dicionário de Índices Cadastrais 29
Tabela 2 – Dicionário de Índices Tratados 34
Tabela 3 – Distribuição Percentual do IAA por Raça e Gênero 49
Tabela 4 – Distribuição Percentual do IAA por Curso e Gênero 51
Tabela 5 – Métricas de Classificação de Reprovados com Dados Cadastrais e Diferentes Discretizações 52
Tabela 6 – Distribuição de Aprovados e Reprovados Para as Categorias de Índices com Maior Impacto 53
Tabela 7 – Métricas de Classificação de Reprovados com Índices de Presença por Período Letivo 54
Tabela 8 – Métricas de Classificação de Aprovados com Índices de Presença por Período Letivo 55
Tabela 9 – Métricas de Classificação de Reprovados com Índices Cadastrais e de Presença por Período Letivo 56
Tabela 10 – Métricas de Classificação de Aprovados com Índices Cadastrais e de Presença por Período Letivo 56
Tabela 11 – Métricas de Classificação de Reprovados com Índices de Nota por Período Letivo 57
Tabela 12 – Métricas de Classificação de Reprovados por Período Letivo Através da Integração dos Dados 59
Tabela 13 – Métricas de Sensibilidade de Reprovados por Período Letivo para Cada Conjunto de Dados 59
Tabela 14 – Média de Sensibilidade para Cada Modelo Classificativo Considerando 75% do Período Letivo 60

LISTA DE SIGLAS

AD	Árvore de Decisão
AM	Aprendizado de Máquina
AVA	Ambientes Virtuais de Aprendizado
CTJ	Centro Tecnológico de Joinville
CMD	Comprimento Mínimo de Descrição
FN	Falso Negativo
FP	Falso Positivo
GB	Gradient Boosting
IAA	Índice de Aproveitamento Acumulado
IM	Índice de Matrícula
IAP	Índice de Aproveitamento Probatório
KNN	K-Nearest Neighbors
LR	Regressão Logística
LVP	Laboratórios Virtuais de Programação
MDE	Mineração de Dados Educacionais
NB	Naive Bayes
PR	Paraná
RF	Random Forest
RNA	Redes Neurais Artificiais
SC	Santa Catarina
SP	São Paulo
SVM	Máquinas de Vetores de Suporte
UFSC	Universidade Federal de Santa Catarina
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO 12
1.1 OBJETIVOS 13
1.1.1 Objetivo geral 13
1.1.2 Objetivos Específicos 13
2 FUNDAMENTAÇÃO TEÓRICA 14
2.1 MINERAÇÃO DE DADOS EDUCACIONAIS 14
2.1.1 Tipos de dados obtidos com a MDE 14
2.2 APRENDIZADO DE MÁQUINA 15
2.2.1 Aprendizado supervisionado 16
2.2.2 Pré-processamento dos dados 17
2.2.2.1 Limpeza 17
2.2.2.2 Manipulação Direta 18
2.2.2.3 Discretização 19
2.2.2.4 Normalização 20
2.2.2.5 Seleção de Características 20
2.2.3 Validação dos Modelos 21
2.2.3.1 Divisão dos Dados por Amostragem Aleatória Estratificada 21
2.2.3.2 Matriz de Confusão 21
2.2.3.2.1 Sensibilidade 22
2.2.3.2.2 Precisão 23
2.2.3.2.3 Pontuação F1 23
2.2.3.2.4 Acurácia 24
2.3 TRABALHOS RELACIONADOS 24
3 METODOLOGIA 27
3.1 CONJUNTO DE DADOS 27
3.1.1 Dados Cadastrais 27
3.1.1.1 Dados Acadêmicos 28
3.1.1.2 Dados Sociodemográficos 30
3.1.2 Relatório de Presenças 31
3.1.3 Relatório de Notas 32
3.2 ANÁLISES ESTATÍSTICAS 33
3.3 TRATAMENTO DOS DADOS 33
3.3.1 Limpeza dos dados 35
3.3.2 Manipulação dos dados 35
3.3.3 Segmentação dos Dados 36
3.4 DESENVOLVIMENTO DE FLUXOS PARAS ANÁLISE PREDITIVA . 37

3.5 ANÁLISES PREDITIVAS 38
4 MODELOS E RESULTADOS 40
4.1 ANÁLISES ESTATÍSTICAS 40
4.1.1 Análise Descritiva 40
4.1.2 Análise Temporal 42
4.1.3 Software para Mapeamento e Análise Geográfica 44
4.1.4 Análise Exploratória 46
4.1.4.1 Distribuição Geográfica 46
4.1.4.2 Distribuição por Desempenho 48
4.2 ANÁLISES PREDITIVAS APLICADAS PARA TURMA DE PROGRAMAÇÃO I 51
4.2.1 Fluxo Cadastral 51
4.2.2 Fluxo de Presenças 54
4.2.3 Integração dos Fluxos de Dados Cadastrais e de Presenças . . . 55
4.2.4 Fluxo de notas 56
4.2.5 Fluxo Final: Integração de Todos os Dados 57
5 CONCLUSÃO 61
6 PERSPECTIVAS FUTURAS 62
6.1 DADOS ACADÊMICOS 62
6.2 DADOS SOCIODEMOGRÁFICOS 62
REFERÊNCIAS 63
APÊNDICE A – MATRIZES DE CONFUSÃO PARA ANÁLISES PREDITIVAS 66

1 INTRODUÇÃO

A retenção estudantil é questão complexa que envolve as motivações e desafios enfrentados pelos alunos. De acordo com o National Center for Educational Statistics (NCES, 2021), o baixo desempenho acadêmico e a reprovação estão entre as principais razões que levam os estudantes a abandonar os estudos.

Em um contexto específico das engenharias, cerca de 40% dos estudantes não conseguem progredir além do primeiro ano de estudos. Além disso, dentre aqueles que conseguem, aproximadamente 30% enfrentam dificuldades significativas em disciplinas fundamentais, como cálculo e programação (Boylan-Ashraf; Haughery, 2018).

A intervenção precoce, por parte das instituições de ensino, com estudantes de risco, tem demonstrado ser eficaz na redução das taxas de reprovação, com uma redução média de 13% (Ramos *et al.*, 2015). Portanto, conseguir identificar de forma antecipada padrões e fatores que afetam o desempenho dos estudantes se tornou um campo de interesse no setor educacional (Romero; Ventura, 2010).

Nesse cenário, a Mineração de Dados Educacionais (MDE) é a área que se concentra na conversão de dados estudantis em informação útil para identificar e resolver desafios acadêmicos. Utilizando técnicas de Aprendizado de Máquina e Deep Learning para previsão de desempenho, a MDE tem estimulado um rápido crescimento nos últimos anos (Mclaren; Sheuer, 2011).

Dada a importância de identificar antecipadamente estudantes em risco de reprovação, propõe-se neste trabalho o desenvolvimento de um modelo preditivo baseado em aprendizado de máquina fundamentado por dados educacionais dos estudantes.

A fim de validar este modelo e especificar fatores que afetem a possibilidade de reprovação, é realizado um estudo de caso da disciplina de Programação I, ofertada pelo Centro Tecnológico de Joinville (CTJ) da Universidade Federal de Santa Catarina (UFSC). Os dados utilizados são coletados do ambiente virtual de aprendizado Moodle e do sistema acadêmico, abrangendo variáveis sociodemográficas, frequência de participação nas aulas e notas obtidas.

Inicialmente, são realizadas avaliações estatísticas sobre as informações dos alunos, objetivando esboçar as características dos estudantes, detectar tendências no processo de ingresso e investigar a dispersão geodemográfica. Finalmente, é conduzida uma análise implícita, na qual diferentes algoritmos de aprendizado de máquina são comparados em quatro períodos distintos do semestre letivo: 25%, 50%, 75% e 95%. Para cada intervalo, examinou-se a influência dos diferentes tipos de dados, isolados e em conjunto, na antecipação correta dos estudantes reprovados.

1.1 OBJETIVOS

Para contribuir com a identificação de fatores que afetam o desempenho dos estudantes, propõem-se neste trabalho os seguintes objetivos.

1.1.1 Objetivo geral

Desenvolver e avaliar um modelo preditivo baseado em aprendizado de máquina para identificar o risco de reprovação na disciplina de Programação I, utilizando dados educacionais extraídos de sistema acadêmico e do ambiente Moodle.

1.1.2 Objetivos Específicos

- Desenvolver uma ferramenta para facilitar a análise exploratória dos dados;
- Identificar as principais variáveis relacionadas com o risco de reprovação que permitam a predição;
- Manipular os dados, incluindo limpeza, transformação e seleção de atributos relevantes a fim de serem processados;
- Desenvolver modelos preditivos utilizando diferentes algoritmos de aprendizado de máquina;
- Avaliar o desempenho dos modelos preditivos.

2 FUNDAMENTAÇÃO TEÓRICA

Visando identificar variáveis prognósticas do risco de reprovação de estudantes universitários, é necessário compreender os conceitos utilizados para a construção de um modelo preditivo, neste caso, baseado na aprendizagem de máquina. Neste capítulo, inicialmente, é explorada a abordagem de mineração de dados e as categorias de informações que poderiam ser utilizadas. Em seguida, são apresentadas as tecnologias de aprendizado de máquina e seu subtipo supervisionado, como os dados devem ser tratados e como os modelos são avaliados. Por fim, é feita a contextualização de trabalhos similares que auxiliaram na escolha dos recursos empregados.

2.1 MINERAÇÃO DE DADOS EDUCACIONAIS

A Mineração de Dados estabelece que o processo de descoberta de padrões em grandes conjuntos de informações é passo essencial para a obtenção de conhecimento por meio de bancos de dados (Han *et al.*, 2011). Sua vertente, a MDE, é método que possibilita a transformação desses padrões em informação potencialmente útil para a melhora do desempenho estudantil (Romero; Ventura, 2010).

Romero e Ventura (2007) classificam o trabalho em MDE como uma fusão das categorias de Estatísticas e Mineração Web, resultando em uma abordagem integrada que busca combinar os princípios e as capacidades preditivas de ambos os campos. Essa integração baseia e expande áreas correlatas, como Aprendizado de Máquina (AM) e Deep Learning, com o propósito de desenvolver modelos capazes de deduzir aspectos específicos dos dados educacionais, tais como a probabilidade de reprovação e a construção de perfis educacionais (Mclaren; Sheuer, 2011).

Nesse sentido, as predições em MDE podem assumir diferentes formas principais: classificação, regressão e estimação de densidade. Essas predições operam através da análise e fusão dos diversos aspectos encontrados nos dados, e necessitam que uma certa quantidade desses seja manualmente codificada para viabilizar a correta identificação e padronização (Baker; Yacef, 2017).

2.1.1 Tipos de dados obtidos com a MDE

No contexto da MDE, onde a integração de diversas abordagens visa aprimorar a compreensão e predição de padrões, é crucial considerar não apenas as técnicas analíticas, mas também os tipos de dados envolvidos nesse processo (González *et al.*, 2021).

Embora tenha sido reconhecida recentemente como um campo científico, a MDE é uma das aplicações mais populares e antigas envolvendo mineração de dados.

Sua crescente relevância nos últimos anos está diretamente ligada ao surgimento de Ambientes Virtuais de Aprendizado (AVA), como o sistema Moodle (Mclaren; Sheuer, 2011).

Os AVA oferecem uma rica fonte de informações educacionais, como notas, frequência estudantil, atividades realizadas e interações dos usuários com a plataforma. Esses dados são coletados em arquivos de log e são conhecidos como dados virtuais (Mclaren; Sheuer, 2011).

No entanto, para uma compreensão abrangente e precisa, é importante considerar também dados demográficos, como etnia, nível de escolaridade, idade e gênero. Essas variáveis são obtidas por meio de pesquisas, registros acadêmicos e outras fontes externas e são consideradas como possíveis preditores para determinar o êxito acadêmico dos estudantes (Kovaĉić, 2010).

Essa abordagem holística, podendo incorporar tanto dados virtuais provenientes de AVA quanto dados demográficos, enriquece a análise e a aplicação preditiva da MDE (Cortez; Silva, 2010).

2.2 APRENDIZADO DE MÁQUINA

O AM é um ramo da área de Inteligência Artificial que se diferencia por sua capacidade de permitir que as máquinas aprendam padrões e façam decisões sem serem explicitamente programadas. Em vez de seguir instruções específicas, os algoritmos de AM são alimentados com dados e, a partir desses, aprendem a realizar tarefas específicas (Dhankar; Gupta, 2021).

Um dos primeiros experimentos notáveis envolvendo AM é o Perceptron, desenvolvido por Frank Rosenblatt em 1957. O Perceptron foi alimentado com entradas binárias e ajustou seus pesos sinápticos através de um processo de aprendizado supervisionado. O objetivo era que o Perceptron pudesse aprender a distinguir entre duas classes de padrões (Dreyfus, 1990).

No entanto, Minsky e Papert (1969) demonstraram que o Perceptron não poderia aprender a resolver problemas mais complexos que não fossem linearmente separáveis. Apesar dessas limitações, o experimento é considerado um marco importante na história do AM, e as técnicas e conceitos desenvolvidos desde então levaram a avanços significativos nas décadas seguintes (Shawe-Taylor, 2009).

Em essência, o AM se configura como um sistema capaz de ajustar autonomamente seu comportamento com base em experiências passadas. Essa adaptação envolve a formulação de regras lógicas, geradas pelo reconhecimento de padrões nos dados analisados, com o intuito de aprimorar o desempenho em tarefas específicas ou tomar decisões apropriadas para classificar dados futuros, assim, derivando novo conhecimento (Semeraro *et al.*, 2023).

O fluxo constante de dados, contanto que tenham sido devidamente tratados, desempenha um papel relevante no AM, facilitando a identificação de novos padrões. Modelos e algoritmos precisam ser capazes de lidar com mudanças contínuas nos dados para manterem seu desempenho e utilidade ao longo do tempo (Semeraro *et al.*, 2023).

Além disso, por se concentrar em replicar o processo de aprendizado humano, Dhankar e Gupta (2021) destacam esse campo como área crucial da ciência de dados. Através do uso de métodos estatísticos, os algoritmos são treinados para realizar classificações ou previsões, revelando insights valiosos em projetos de mineração de dados (Dhankar; Gupta, 2021).

Como explorado por Walker (2018), existem várias abordagens no AM, incluindo aprendizado supervisionado (onde o modelo é treinado em pares de entrada e saída) e não supervisionado (onde o modelo encontra padrões nos dados sem rótulos de saída). Por conta da natureza dos dados explorados neste trabalho, a categoria de aprendizado supervisionado é utilizada e será discutida em detalhes a seguir.

2.2.1 Aprendizado supervisionado

O aprendizado supervisionado é uma abordagem essencial no campo do AM, destacando-se por seu foco em conjuntos de dados rotulados. Nesse método, cada exemplo do conjunto inclui uma entrada (variável independente) associada a uma saída (variável dependente) conhecida, possibilitando ao modelo aprender as relações entre as características da entrada e as saídas desejadas (Osisanwo *et al.*, 2017).

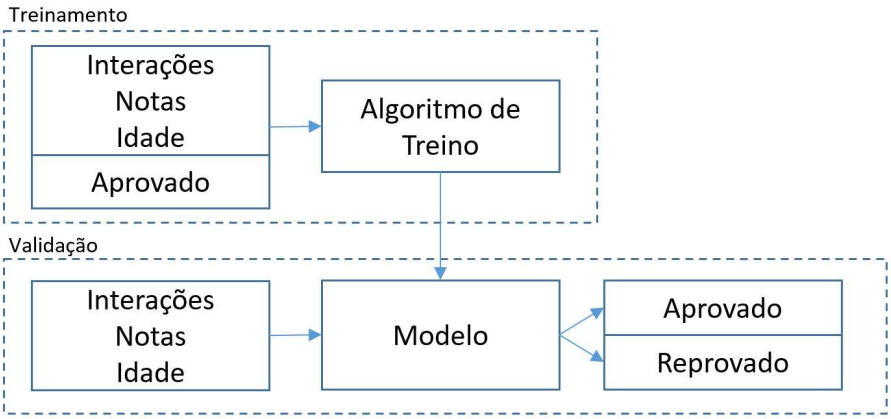
As tarefas comuns no aprendizado supervisionado incluem previsão e classificação. Na previsão, o modelo é treinado para antecipar valores contínuos, como preços de ações, enquanto que na classificação, o objetivo é categorizar as entradas em classes distintas, como identificar se um e-mail é spam ou não. A escolha do tipo de tarefa depende do contexto específico da aplicação (Osisanwo *et al.*, 2017).

A Figura 1 utiliza o cenário classificativo proposto para este trabalho como exemplo. Um algoritmo é exposto a perfis estudantis rotulados, onde a quantidade de interações, notas e idade atuam como variáveis independentes, e a aprovação ou reprovação como variável dependente.

Durante a fase de treinamento, o algoritmo irá ajustar retroativamente seus parâmetros para minimizar a diferença entre suas previsões e as saídas reais. Esse ajuste é necessário para que o algoritmo generalize bem e seja capaz de fazer previsões precisas em situações não vistas durante o ensaio. Quando o algoritmo decifra a relação entre os dados, um modelo é criado e precisa ser validado (Burkart; Huber, 2021).

A avaliação do modelo ocorre posteriormente, utilizando um segundo conjunto de dados, ainda não analisados, a fim de verificar a eficácia do modelo em cenários

Figura 1 – Aprendizado Supervisionado por Classificação



Fonte: Elaborado pelo Autor (2024).

reais e garantir sua aplicabilidade prática. (Burkart; Huber, 2021)

Diversos algoritmos são empregados no aprendizado supervisionado, cada um adequado a variados tipos de tarefas. Para tarefas de previsão, por exemplo, a regressão linear pode ser utilizada, enquanto em tarefas de classificação, algoritmos como Máquinas de Vetores de Suporte, do inglês *Support Vector Machines* (SVM), Árvores de Decisão (AD) e Redes Neurais Artificiais (RNA) são comuns. A seleção do algoritmo depende do tipo de dados disponíveis. (Alloghani *et al.*, 2020).

2.2.2 Pré-processamento dos dados

Segundo García *et al.* (2014), o pré-processamento de dados é uma etapa no desenvolvimento de modelos de aprendizado que transforma os dados brutos em um formato adequado para a análise e treinamento. Este processo envolve técnicas que visam melhorar a qualidade dos dados, remover inconsistências, reduzir a dimensionalidade e prepará-los para que os algoritmos possam aprender de forma eficaz (García *et al.*, 2014).

Entre as principais técnicas de pré-processamento estão a limpeza dos dados, que trata da remoção de valores ausentes e correção de inconsistências; a manipulação direta dos dados, que inclui a criação de novos índices e a alteração de atributos; a discretização, que transforma variáveis contínuas em discretas; a normalização, que ajusta a escala dos atributos; e a seleção de características, que elege as variáveis com maior relevância (García *et al.*, 2014).

2.2.2.1 Limpeza

A limpeza dos dados é etapa inicial no pré-processamento. Este método envolve o uso de técnicas para identificar e corrigir problemas nos dados, assegurando a integridade e a consistência dos mesmos (García *et al.*, 2014).

Um dos primeiros passos na limpeza é o tratamento de valores ausentes. Dados faltantes são comuns em conjuntos de dados reais, podendo ocorrer devido a coleta incompleta. Técnicas para lidar com valores ausentes incluem a remoção de registros incompletos, o que pode ser adequado quando a quantidade de dados ausentes é pequena, ou a imputação, onde valores ausentes são substituídos por estimativas, como a média, mediana ou valores preditivos obtidos a partir de outros dados (García *et al.*, 2014).

Além dos valores ausentes, existe a correção de valores inconsistentes. Inconsistências podem surgir de diversas fontes, como duplicação de registros, erros de entrada manual ou discrepâncias entre diferentes fontes de dados. Técnicas de padronização asseguram que os dados estejam em um formato uniforme. Por exemplo, padronizar unidades de medida ou formatos de data pode prevenir erros de análise (García *et al.*, 2014).

Outro aspecto importante na limpeza de dados é a remoção de dados irrelevantes que têm pouca, ou nenhuma, contribuição para a análise ou modelagem. A presença de variáveis irrelevantes pode aumentar a complexidade do modelo e reduzir sua eficiência. Técnicas como análise de correlação são usadas para identificar e eliminar esses dados. Ao remover informações redundantes ou desnecessárias, o modelo pode focar nos atributos mais relevantes, melhorando sua interpretabilidade (García *et al.*, 2014).

Em resumo, a limpeza de dados é um processo iterativo e contínuo, frequentemente repetido durante o desenvolvimento do modelo. Novos problemas podem ser identificados conforme o entendimento do conjunto de dados se aprofunda e o modelo evolui. É uma etapa que influencia diretamente a qualidade dos modelos, onde dados limpos e bem preparados resultam em previsões mais precisas (García *et al.*, 2014).

2.2.2.2 Manipulação Direta

A manipulação direta dos dados envolve técnicas que permitem sua transformação e reestruturação para que se ajustem às necessidades da análise. Através da criação de novos índices, concatenação de informações e alteração do significado de atributos, os dados podem ser refinados para melhor suportar a tomada de decisões (Zheng; Casari, 2018).

Zheng e Casari (2018) definem a criação de novos índices, ou atributos derivados, como uma das principais técnicas da manipulação direta. Isso pode envolver cálculos ou transformações matemáticas simples, como a padronização de dados, ou operações mais complexas, como a criação de índices compostos que agregam múltiplas variáveis (Zheng; Casari, 2018).

A concatenação de informações é outra técnica importante, que envolve a junção de dados de diferentes fontes ou a combinação de múltiplos atributos em um

único. Isso é realizado através de operações de mesclagem e junção, que permitem integrar dados de diferentes tabelas ou conjuntos de dados com base em uma chave comum (Kazil; Jarmul, 2016).

Além disso, a alteração direta do significado de atributos, ou engenharia de atributos, envolve modificar ou recodificar variáveis existentes para melhor refletir a realidade ou facilitar a análise. Isso pode incluir a transformação de dados categóricos em indicadores binários (one-hot encoding) e a reclassificação de categorias (Zheng; Casari, 2018).

Por fim, a manipulação direta pode incluir a agregação de dados para criar resumos ou estatísticas derivadas. Isso é particularmente útil em grandes conjuntos de dados, onde sumarizar a informação pode revelar tendências que não são aparentes em seu formato bruto. Operações de agregação comuns incluem médias, somas, contagens e cálculos de mediana ou percentis, que podem ser aplicados a diferentes subgrupos de dados para melhorar sua percepção (Kazil; Jarmul, 2016).

2.2.2.3 Discretização

A discretização é o processo de converter variáveis contínuas em variáveis discretas, dividindo o domínio dos atributos em intervalos finitos. Esta técnica é particularmente útil para algoritmos de aprendizado que funcionam melhor com dados categóricos, como AD e Naive Bayes (NB) (Dougherty *et al*., 1997).

Existem duas abordagens principais para a discretização: supervisionada e não supervisionada. A discretização supervisionada utiliza a variável alvo para guiar a criação dos intervalos. Métodos como a discretização baseada em entropia minimizam a incerteza dentro dos intervalos, alinhando-os de modo que a informação da variável alvo seja maximizada. Outro exemplo é a discretização por Comprimento Mínimo de Descrição (CMD), que busca dividir os dados em intervalos de maneira a minimizar o comprimento total da descrição, que inclui tanto a descrição dos intervalos quanto a da variável alvo dentro desses intervalos (Fayyad; Irani, 1993).

Por outro lado, a discretização não supervisionada não considera a variável alvo e se baseia apenas nas propriedades intrínsecas dos dados. A discretização por igual amplitude divide o intervalo de valores contínuos em segmentos de tamanho igual, enquanto a discretização por igual frequência cria intervalos de forma que cada um contenha aproximadamente o mesmo número de amostras. Esta última é particularmente útil quando os dados são altamente assimétricos, garantindo que cada intervalo tenha uma representação adequada dos dados (Fayyad; Irani, 1993).

Embora a discretização simplifique a estrutura dos dados e possa melhorar o desempenho dos algoritmos, ela apresenta desafios. A transformação de dados contínuos em discretos pode resultar na perda de informação, especialmente se os intervalos não forem cuidadosamente escolhidos. A determinação do número e dos

limites dos intervalos requer experimentação e validação para evitar a criação de intervalos que não capturem bem a variabilidade dos dados (Liu *et al.*, 2002).

A aplicação prática da discretização geralmente envolve a experimentação com diferentes métodos e parâmetros para determinar a configuração que melhor se ajusta aos dados específicos e ao problema em questão (Dougherty *et al.*, 1997).

2.2.2.4 Normalização

A normalização é uma técnica de pré-processamento que visa ajustar a escala dos atributos para que eles contribuam de forma equilibrada no treinamento dos modelos (Zheng; Casari, 2018).

Uma das abordagens mais comum para a normalização é a *Min-Max Normalization*, onde os valores dos atributos são transformados para uma escala específica, geralmente entre 0 e 1. Esta técnica ajusta os valores com base no valor mínimo e máximo do atributo, conforme Equação 1, onde X é o valor original e X_{min} e X_{max} são os valores mínimo e máximo, respectivamente (Zheng; Casari, 2018).

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{1}$$

A normalização é essencial quando os dados possuem diferentes unidades de medida ou variações de escala, pois ajuda a melhorar o desempenho e a estabilidade dos algoritmos de aprendizado. Para exemplificar, em um conjunto de dados que inclui a altura e o peso de indivíduos, as diferenças de grandeza entre centímetros e quilogramas podem causar problemas. A normalização ajusta esses valores para que ambos contribuam igualmente no processo de aprendizado, evitando que atributos com valores maiores dominem os cálculos de distância ou a influência nos modelos (Zheng; Casari, 2018).

2.2.2.5 Seleção de Características

A seleção de características é uma etapa que visa identificar e selecionar as variáveis mais relevantes para os modelo. Este processo ajuda a reduzir a dimensionalidade dos dados, eliminando atributos redundantes ou irrelevantes, o que pode melhorar o desempenho e a interpretabilidade (Zheng; Casari, 2018).

Uma das técnicas utilizadas para a seleção de características é o ranqueamento das variáveis pela proporção de ganho de informação. O ganho de informação é uma métrica que quantifica a redução da incerteza sobre a variável alvo ao conhecer o valor de uma variável preditora. A proporção é calculada com base na entropia da variável alvo antes e depois de observar a variável preditora. Variáveis com maior

ganho de informação são consideradas mais relevantes para a construção do modelo (Zheng; Casari, 2018).

2.2.3 Validação dos Modelos

Segundo Witten *et al.* (2016), a validação de modelos visa assegurar que o modelo treinado não apenas se ajuste bem aos dados de treinamento, mas também generalize adequadamente para novos dados. A validação envolve uma série de técnicas e métricas que permitem avaliar de forma abrangente o desempenho do algoritmo. Este processo inclui a divisão dos dados em conjuntos de treinamento e teste, a construção e interpretação da matriz de confusão e o cálculo de métricas de desempenho específicas, como sensibilidade, precisão, F1-score e acurácia (Witten *et al.*, 2016).

2.2.3.1 Divisão dos Dados por Amostragem Aleatória Estratificada

A amostragem aleatória estratificada é uma técnica para divisão de dados em conjuntos de treinamento e teste, fazendo com que cada subconjunto, conhecido como estrato, seja representativo da população total. A amostragem estratificada é empregada para evidenciar as diferenças entre os grupos de uma população, diferentemente da amostragem aleatória simples, que considera todos os membros da população como iguais e com a mesma probabilidade de serem selecionados (Taherdoost, 2016).

O processo de amostragem começa com a identificação das diferentes classes ou estratos no conjunto de dados. Cada estrato representa uma subpopulação com características semelhantes. A seguir, uma amostra aleatória é extraída de cada estrato, proporcional ao tamanho da classe na população total. Isso assegura que a distribuição das classes no conjunto de treinamento e no conjunto de teste reflete a distribuição original dos dados (Taherdoost, 2016).

Uma vantagem dessa técnica é a mitigação de vieses que poderiam surgir de uma simples divisão aleatória, que pode não preservar a proporção das classes. Por exemplo, em um problema de classificação com classes desbalanceadas, uma divisão aleatória simples pode resultar em um conjunto de treinamento ou teste com muito poucos exemplos da classe minoritária. A amostragem estratificada evita este problema, melhorando a generalização do modelo treinado (Taherdoost, 2016).

2.2.3.2 Matriz de Confusão

A matriz de confusão oferece uma representação do desempenho do modelo, permitindo a visualização das previsões corretas e incorretas. Através dela, é possível entender não apenas a acurácia geral, mas também como as previsões se distribuem entre as diferentes classes (Witten *et al.*, 2016).

Para um problema de classificação binária, a matriz de confusão é estruturada conforme Figura 2 e é aplicada de acordo com as seguintes quatro unidades de avaliação (Witten *et al.*, 2016):

- **Verdadeiro Positivo (VP):** São as instâncias que o modelo previu corretamente como pertencentes à classe positiva. Por exemplo, para classificação dos alunos, um verdadeiro positivo seria um estudante aprovado que foi corretamente identificada pelo modelo.
- **Verdadeiro Negativo (VN):** São as instâncias que o modelo previu corretamente como pertencentes à classe negativa. Usando o mesmo exemplo, um verdadeiro negativo seria um estudante reprovado que foi corretamente identificado.
- **Falso Positivo (FP):** São as instâncias que o modelo previu incorretamente como pertencentes à classe positiva. No contexto da aprovação, um falso positivo seria um estudante reprovado que foi incorretamente identificada como aprovado.
- **Falsos Negativos (FN):** São as instâncias que o modelo previu incorretamente como pertencentes à classe negativa. Novamente, no exemplo de aprovação, um falso negativo seria um estudante aprovado que foi incorretamente identificada como reprovado.

Figura 2 – Matriz de Confusao Genérica

Fonte: Elaborado pelo Autor (2024).

Além disso, a matriz de confusão e as unidades básicas de avaliação são usadas para calcular as métricas de desempenho do modelo em diferentes situações, como sensibilidade, precisão, pontuação F1 e acurácia (Witten *et al.*, 2016).

2.2.3.2.1 Sensibilidade

A sensibilidade, também conhecida como recall, é uma métrica especialmente útil em problemas de classificação binária. Esta métrica quantifica a capacidade do modelo de identificar corretamente todas as instâncias positivas, calculando a proporção de VP em relação ao total de instâncias que realmente pertencem à classe

positiva. Entretanto, vale ressaltar que a métrica não identifica o número de elementos negativos classificados como FP. (Witten *et al.*, 2016).

A Equação 2 é usada para calcular a sensibilidade. É importante notar que sua aplicação não se limita aos casos de VP. É possível também medir a proporção de VN corretamente identificados pelo modelo ao substituir VP por VN, e FN por FP na equação (Witten *et al.*, 2016).

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

De acordo com Witten *et al.* (2016), esta métrica assume particular importância em contextos onde a identificação precisa de instâncias positivas ou negativas é crítica . No contexto deste trabalho, uma alta sensibilidade na identificação de verdadeiros negativos é essencial para garantir que a maioria dos estudantes em risco de reprovação seja corretamente identificada, permitindo que recebam a ajuda necessária em tempo hábil.

2.2.3.2.2 Precisão

Complementando a sensibilidade, a precisão é uma métrica que mede a exatidão com que o modelo prevê instâncias positivas, ou seja, calcula a proporção de VP entre todas as instâncias classificadas como positivas, conforme ilustrado na Equação 3 (Witten *et al.*, 2016).

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

A precisão é particularmente relevante em cenários onde o custo de uma classificação incorreta é elevado (Witten *et al.*, 2016). Retomando o contexto deste trabalho, embora classificar incorretamente estudantes aprovados como reprovados não seja tão prejudicial, visto que o auxílio extra pode contribuir para que obtenham médias ainda melhores, a precisão ainda se mantém como uma métrica de importância para garantir a adequada avaliação e intervenção.

2.2.3.2.3 Pontuação F1

A pontuação F1, ou F1-score, combina a precisão e a sensibilidade de um modelo de em um único valor, oferecendo um equilíbrio entre essas duas métricas. Representada pela Equação 4, esta métrica é útil em cenários onde é necessário mediar entre a precisão e a sensibilidade, especialmente quando a distribuição das classes é desequilibrada (Witten *et al.*, 2016).

$$\text{F1-score} = 2 \times \frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \tag{4}$$

A pontuação F1 é a média harmônica da precisão e da sensibilidade, o que significa que ela favorece modelos que têm um bom equilíbrio entre as duas. Um valor alto de F1 indica que o modelo não só é preciso, mas também eficiente em identificar corretamente as instâncias positivas (Witten *et al.*, 2016).

Esta métrica é especialmente importante em problemas onde tanto falsos positivos quanto falsos negativos têm consequências significativas (Witten *et al.*, 2016).

2.2.3.2.4 Acurácia

Finalmente, a acurácia, representada pela Equação 5, mede a proporção de previsões corretas realizadas pelo modelo em relação ao total de previsões, fornecendo uma visão geral da performance do modelo (Witten *et al.*, 2016).

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{5}$$

Em cenários onde as classes são balanceadas, a acurácia pode ser uma métrica útil para fornecer uma visão geral do desempenho do modelo. É importante, no entanto, complementá-la com outras métricas descritas, como precisão, sensibilidade e F1-score, para obter uma avaliação mais completa, especialmente em conjuntos de dados desbalanceados (Witten *et al.*, 2016).

2.3 TRABALHOS RELACIONADOS

A revisão de trabalhos similares tem como objetivo fornecer uma base sólida para o desenvolvimento deste estudo, identificando lacunas no conhecimento existente e destacando metodologias eficazes. Ao compreender as pesquisas anteriores, busca-se avançar na análise da predição de reprovação estudantil, e contribuir para a expansão do corpo de conhecimento da área.

Na tentativa de explorar a usabilidade de dados sociodemográficos, Kovaĉić (2010) aplicou o método de Árvores de Classificação e Regressão para a predição de um conjunto de 435 estudantes. No entanto, com uma acurácia máxima de 60%, o autor destaca que as informações não fornecem dados suficientes para uma predição precisa entre indivíduos bem-sucedidos e malsucedidos (Kovaĉić, 2010).

Trabalhos subsequentes também se dedicaram em explorar esse modelo de dados, empregando, no entanto, técnicas distintas, como NB, RNA e Regressão Logística, do inglês Logistic Regression (LR). A constância nos resultados, mantendo-se dentro de uma faixa de 50% a 60% de acurácia, ressalta a incompatibilidade de uso

dessa vertente de informação nos problemas relacionados à reprovação estudantil (Cheewaprakobkit, 2013; Cortez; Silva, 2010).

Buscando uma abordagem diferente, Almarabeh (2017) utilizou dados educacionais virtuais, como notas, presença em seminários e participação em laboratórios ao longo de um semestre como fonte de análise. Após o processamento desses dados com cinco classificadores diferentes, concluiu que a Rede Bayesiana apresentou o melhor desempenho, alcançando uma acurácia de 91%. Os resultados sublinham que dados virtuais podem oferecer retornos promissores na predição de desfechos acadêmicos (Almarabeh, 2017).

Com uma variável distinta dos dados virtuais, Barrozo (2022) utilizou a técnica de classificação AdaBoost para analisar as interações dos estudantes com o ambiente virtual Moodle. Durante as primeiras cinco semanas do curso, uma acurácia média de 56,2% foi alcançada, e esse desempenho aumentou para 70,5% nas três semanas finais. Notavelmente, os usuários que interagiam mais frequentemente com o ambiente demonstraram uma probabilidade menor de reprovação.

De forma similar, Kaensar e Wongnin (2023) utilizaram do Moodle para conduzir uma pesquisa em quatro estágios distintos (25%, 50%, 75% e 100%) do progresso acadêmico, e obtiveram um F1-score de 81,1% com AD ao final do curso. No entanto, um aspecto notável surgiu durante a análise: o desempenho dos algoritmos varia significativamente em diferentes estágios do período letivo, especificamente, SVM obteve o melhor desempenho para o estágio inicial, RL foi mais eficaz para os estágios intermediários, e as AD demonstraram superioridade no estágio final (Kaensar; Wongnin, 2023).

Finalmente, em um contexto similar ao deste trabalho, Lomba (2023) analisou a influência dos diferentes tipos de atividades avaliativas na probabilidade de aprovação dos estudantes de duas turmas de Programação I no CTJ da UFSC. Entretanto, ao invés de separar a análise por períodos de tempo, foi feita a segmentação por tipo de atividade. Para a primeira turma, obteve uma sensibilidade de 90,48% para as provas e questionários quando analisados tanto individualmente quanto em conjunto. Para a segunda turma, os resultados foram melhores, alcançando uma sensibilidade de 95% para questionários e provas enquanto isolados, e 97,5% quando analisados em conjunto. O modelo utilizado am ambas as turmas foi o de RNA (Lomba, 2023).

Ainda com dados de turmas de Programação I do CTJ, Dalfovo (2023) buscou identificar fatores que apontem a possibilidade de desistência dos alunos nas disciplinas. A partir da manipulação e segmentação por período letivos das presenças e atividades avaliativas, obteve com NB para a primeira turma sensibilidades de 87%, 82%, 82% e 80% para os estágios de 25%, 50%, 75% e 100%, respectivamente. Para a segunda turma, entretanto, os resultados foram mais promissores. Ainda com NB, obteve sensibilidades de 95,31%, 99,38%, 100,00% e 100,00% para a mesma distri-

buição de períodos. Embora apresenta alta variabilidade para cada base de dados analisada, a abordagem se provou eficiente.

3 METODOLOGIA

Neste capitulo, são descritos os métodos utilizados para alcançar o objetivo de identificar a relação das variáveis com o desempenho dos estudantes e sua influência na predição dos reprovados. Inicialmente, são especificados os conjuntos de dados analisados neste estudo. Após esta introdução, as análises estatísticas realizadas em um subconjunto dos dados são explicadas. Segue-se uma descrição do tratamento aplicado aos dados, visando aprimorar a interpretação pelos modelos de AM. Posteriormente, o processo de criação de fluxos por meio do software Orange Data Mining é delineado, facilitando as análises preditivas. Finalmente, a metodologia das análises preditivas é abordada, consolidando o propósito deste estudo.

3.1 CONJUNTO DE DADOS

As análises empreendidas neste estudo fundamentaram-se em relatórios extraídos do Moodle e do Sistema de Controle Acadêmico da UFSC. Estes sistemas forneceram um conjunto abrangente de dados, abarcando variáveis cadastrais, registros de presença e distribuição de notas dos estudantes.

3.1.1 Dados Cadastrais

O relatório de dados cadastrais é composto por informações sobre o curso, situação acadêmica, dados sociodemográficos, e histórico escolar dos estudantes, delineando um panorama do perfil discente e fornecendo substrato para análise preditiva. Conforme ilustrado na Figura 3, alguns índices do relatório são expostos, evidenciando os atributos com maior impacto para com este estudo.

Figura 3 – Parte de um Relatório de Dados Cadastrais do Sistema Acadêmico

Curso	Sexo	racaCor	dataNascimento	Naturalidade	anoSemestreIngresso	ensinoPublico?	IAP	IAA	IM	UFSG	AnoConclusãoSG
605	M	branca	28-04-1995	SC - Joinville	20131		7137	5038	4576	SC	2012
602	M	branca	16-09-1996	RS - Caxias do Sul	20151	Sim	7230	6202	5426	RS	2013
607	M	branca	22-11-1995	PR - Curitiba	20151	Sim	7839	6113	5837	PR	2012
608	M	branca	11-03-1997	SC - Joinville	20151		7853	6163	5598	SC	2014
607	M	branca	31-12-1995	SC - Joinville	20151		7707	6959	6873	SC	2012
603	M	parda	14-05-1997	SP - Campinas	20151		7310	5915	5471	SP	2014
603	M	branca	13-03-1991	PR - Maringa	20151		7044	3883	3446	PR	2010
603	M	branca	08-07-1970	SP - Sao Jose do Rio	20151		8260	4241	1219	PR	2003
602	M	branca	19-01-1996	SC - Joinville	20151		7114	5784	5061	SC	2013
602	M	branca	25-11-1996	SC - Chapecó	20151		7047	4434	2937	SC	2013

Fonte: Elaborado pelo Autor (2024).

Vale ressaltar que o documento em questão não aprofunda nos tipos específicos de informações coletadas e em como elas se correlacionam. A fim de melhorar isso, a Tabela 1 apresenta um dicionário dos índices presentes no arquivo. Esse co-

nhecimento se prova importante uma vez que novos indicadores podem ser criados para expandir o escopo do relatório através da manipulação desses dados.

Por possuir informação sensível dos estudantes, o acesso ao relatório não é automaticamente disponibilizado a todos os interessados, requerendo uma solicitação formal à administração do campus.

3.1.1.1 Dados Acadêmicos

No contexto da análise acadêmica da UFSC, utilizam-se índices como o Índice de Aproveitamento Acumulado (IAA) e o Índice de Aproveitamento Probatório (IAP) para avaliar o desempenho dos estudantes. Tais índices contribuem para o entendimento de diferentes aspectos de performance, proporcionando uma análise das diferentes trajetórias tomadas pelos estudantes (UFSC, 2024).

O IAA é uma métrica que reflete o desempenho ao longo do curso. Diferentemente de índices que consideram apenas o resultado de um semestre ou de disciplinas individuais, o IAA leva em conta todas as notas obtidas em todas as disciplinas cursadas até o momento, ajustadas pelo peso de créditos de cada uma. Isso significa que disciplinas com maior carga horária têm um impacto proporcionalmente maior no cálculo do índice, tornando o IAA um reflexo ponderado e abrangente do desempenho acadêmico (UFSC, 2024).

De maneira complementar, o IAP avalia o desempenho com base unicamente na eficiência das aprovações. Este índice calcula a proporção de disciplinas em que o estudante foi aprovado em relação ao total de disciplinas em que se matriculou, oferecendo uma medida direta da capacidade do aluno em concluir com êxito suas obrigações acadêmicas. Tal análise permite uma avaliação da progressão dos estudantes e auxilia na identificação daqueles que podem necessitar de suporte adicional (UFSC, 2024).

A transição entre o conceito de eficácia imediata, representado pelo IAP, e a avaliação da carga acadêmica, introduz o Índice de Matrícula (IM). O IM é uma métrica educacional projetada para avaliar a carga acadêmica de um individuo em um determinado período. Ele utiliza o IAA do aluno, sua Carga Horária Cursada (CHC), e a Carga Horária Total (CHT) possível ou recomendada para o período em questão, conforme demonstrado pela Equação (6) (UFSC, 2024).

$$\mathrm{IM} = \frac{\mathrm{IAA} \cdot \mathrm{CHC}}{\mathrm{CHT}} \tag{6}$$

O IM tem o propósito de oferecer um panorama sobre como os estudantes distribuem seu tempo e esforço entre as disciplinas, possibilitando a identificação de situações de sobrecarga ou de subaproveitamento. Um IM elevado pode indicar uma carga de trabalho excessiva, potencialmente levando a estresse e diminuição na qualidade do aprendizado. Por outro lado, um IM reduzido pode sugerir que o estudante

Tabela 1 – Dicionário de Índices Cadastrais

Termo	**Descrição**
Curso	Código numérico do curso.
nomeCurso	Nome completo do curso frequentado pelo estudante.
Situacao	Situação acadêmica atual do estudante (ex: regular, trancado).
semestreSituacao	Semestre correspondente à situação acadêmica registrada.
Sexo	Gênero do estudante.
racaCor	Categoria de raça ou cor do estudante.
dataNascimento	Data de nascimento do estudante.
estadoCivil	Estado civil do estudante.
Nacionalidade	Nacionalidade do estudante.
Naturalidade	Local de nascimento do estudante.
estadoMatriculaAtual	Estado atual da matrícula (ex: ativa, trancada, cancelada).
Curriculo	Currículo do curso seguido pelo estudante.
anoSemestreIngresso	Ano e semestre de ingresso do estudante na instituição.
provavelFormatura	Ano previsto para a formatura do estudante.
Polo	Polo de ensino ao qual o estudante está vinculado (se aplicável).
formaIngresso	Forma de ingresso na instituição (ex: vestibular, ENEM, transferência).
categoriaIngresso	Categoria de ingresso do estudante (ex: ampla concorrência, cotas).
ensinoPublico?	Indica se o estudante veio de escola pública (sim ou não).
IAP	Índice de aproveitamento baseado em aprovações.
IAA	Índice de aproveitamento acadêmico acumulado.
IM	Índice de matrícula, utilizado para fins administrativos.
ClassificacaoVestibular	Classificação do estudante no processo seletivo de ingresso.
OrdemClassificacao	Ordem de classificação dentro da categoria de ingresso.
Usuario	Identificador do usuário no sistema acadêmico.
EscolaSG	Nome da escola de ensino médio do estudante.
CursoSG	Curso de ensino médio frequentado pelo estudante.
MunicipioSG	Município onde se localiza a escola de ensino médio do estudante.
UFSG	Unidade federativa da escola de ensino médio do estudante.
AnoConclusãoSG	Ano de conclusão do ensino médio pelo estudante.

Fonte: Elaborado pelo Autor (2024).

não esteja aproveitando plenamente as oportunidades acadêmicas disponíveis, impactando negativamente em sua formação (UFSC, 2024).

A análise desses índices e a aplicação de suas métricas auxiliam para uma gestão educacional efetiva, facilitando a identificação de áreas que necessitam de melhoria e a implementação de estratégias direcionadas ao sucesso dos estudantes.

3.1.1.2 Dados Sociodemográficos

Dentro do âmbito sociodemográfico, informações sobre sexo, raça/cor, data de nascimento, estado de origem e semestre de ingresso fornecem uma base para o entendimento das diversas facetas que compõem o corpo discente. A análise do perfil dos estudantes permite a compreensão da diversidade e das dinâmicas sociais presentes dentro da instituição acadêmica (Thornton *et al.*, 2010).

Além disso, a exploração de padrões e correlações entre as características demográficas e os resultados acadêmicos também facilita a identificação de possíveis fatores de risco que afetem o desempenho estudantil. A investigação da relação entre o histórico educacional e as notas obtidas, por exemplo, pode elucidar a influência do tipo de ensino médio frequentado — público ou privado — no sucesso acadêmico.

Adicionalmente, a criação de novos índices a partir desses dados oferece uma maneira de aprofundar a compreensão acerca da influência que essas variáveis exercem. Entre os criados para este trabalho, estão:

- **Intervalo Pré-Graduação:** este índice mede o tempo entre a conclusão do ensino médio e o ingresso na universidade, apontando para intervalos nos estudos que podem impactar a readaptação ao ambiente acadêmico.
- **Idade de Ingresso:** é calculada através da comparação entre a data de nascimento do estudante e a data de ingresso na instituição. Oferece uma métrica para examinar o impacto da idade no desempenho do estudante nos primeiros semestres.
- **Idade Cursando:** é determinada pela diferença entre a data de nascimento do estudante e a data em que o estudante está matriculado na disciplina. Similar à idade de ingresso, oferece uma métrica sobre o impacto da idade no desempenho do estudante na disciplina.
- **Diferença entre Idade de Ingresso e Cursando:** é calculada pela diferença, em anos, entre a idade do estudante no momento do ingresso na instituição e a idade enquanto cursa a disciplina. Este índice permite analisar como o momento em que o estudante decide cursar a disciplina, seja logo no primeiro ano de curso ou após alguns anos, afeta seu desempenho.
- **Reprovação:** é obtido através da comparação entre IAA e IAP de cada estudante. Caso sejam diferentes, significa que o estudante já teve reprovações em

seu histórico acadêmico. É importante destacar que isso significa uma reprovação em qualquer disciplina de seu histórico, e não necessariamente na de programação I. Este índice ajuda a entender se reprovações passadas afetam a chance de reprovação na disciplina atual.

A utilização desses dados e dos índices derivados facilita uma análise mais detalhada, auxiliando na identificação de estudantes com maior probabilidade de enfrentar dificuldades. Além disso, é importante reconhecer a variedade e a complexidade dentro de cada categoria sociodemográfica para evitar generalizações e permitir uma abordagem mais minuciosa.

3.1.2 Relatório de Presenças

O relatório de presenças registra a frequência dos alunos às aulas, detalhando a participação ou ausência em cada sessão ao longo do período letivo. Assim como o relatório de dados cadastrais, o formato deste documento é padronizado para todas as disciplinas, facilitando a uniformidade e a compreensão dos dados registrados. A Figura 4 apresenta uma seção de um relatório extraído do Moodle.

Figura 4 – Parte de um Relatório de Presenças do Moodle

Aluno	10/08/2023 10:30 T	15/08/2023 10:30 T	17/08/2023 10:30 T	22/08/2023 10:30 T	24/08/2023 10:30 T
NOME1	Au (0/2)	Au (0/2)	Au (0/2)	Au (0/2)	?
NOME2	Inscrição de usuári	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)
NOME3	Au (0/2)	Pr (2/2)	Pr (2/2)	Au (0/2)	?
NOME4	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	?
NOME5	Au (0/2)	Ju (1/2)	Pr (2/2)	Au (0/2)	?

Fonte: Elaborado pelo Autor (2024).

No relatório, cada coluna representa um dia de aula, listando-os incrementalmente desde o primeiro até o último dia registrado. Os registros de presença operam através de um sistema de pontuação, e podem assumir diferentes valores indicativos da situação do aluno em cada aula:

- **Pr (2/2):** Indica que o aluno participou integralmente da aula.
- **At (1/2):** Indica que o aluno atrasou e participou parcialmente da aula.
- **Ju (1/2):** Indica que o aluno justificou sua ausência.
- **Au (0/2):** Indica que o aluno esteve ausente ou faltou à aula.
- **"?":** Denota que o aluno esteve ausente, mas a pontuação ainda não foi registrada para a aula.
- **"Inscrição de usuários inicia xx.xx.xxxx":** Indica que o aluno não iniciou regularmente na disciplina no começo da oferta dessa, e sim na data especificada.

O relatório é obtido através da aba de exportação do módulo de presenças do Moodle, e pode ser exportado em formatos compatíveis com Microsoft Excel e OpenOffice.

3.1.3 Relatório de Notas

O relatório de notas detalha as avaliações aplicadas ao longo da disciplina que constituem a nota final do aluno. Ao contrário dos dados cadastrais e do relatório de presenças, que seguem um modelo padrão, o formato do relatório de notas é específico para cada disciplina. A Figura 5 apresenta algumas das atividades avaliativas aplicadas na turma de programação I analisada neste estudo.

Figura 5 – Parte de um Relatório de Notas do Moodle

Aluno	Nome do Curso	LVP 0	LVP 1	LVP 2	Questionário 1	Questionário 2	Avaliações Parciais	LVP 3	Avaliação Final	Recuperação	Presença	Total do Curso
ALUNO1	Engenharia Naval [t	-	0	-	7	6,9	4,3	-	0,5	0	7,8	2
ALUNO2	Engenharia Autom	-	-	-	8	-	0,8	-	0	0	5,6	0,5
ALUNO3	Engenharia Civil de	-	10	-	10	-	7,3	-	0	0	5,2	3
ALUNO4	Engenharia Civil de	-	10	-	9	5,5	3,5	-	0	0	4,8	1,5
ALUNO5	Ciência e Tecnolog	-	2,5	1,7	8	-	5,9	-	1,8	4	7,1	4
ALUNO6	Engenharia Ferrovi	-	-	-	9	-	0,5	-	0	0	2,5	0
ALUNO7	Engenharia de Trar	-	10	10	10	-	10	-	4	0	9,3	6,5
ALUNO8	Engenharia Naval [t	-	-	-	-	4,1	0,6	0	1,5	0	7,8	1
ALUNO9	Engenharia Aeroes	-	-	10	10	10	6,4	10	3,5	5,5	8,3	5
ALUNO10	Engenharia Naval [t	-	-	-	10	-	7	-	2	0	9	4

Fonte: Elaborado pelo Autor (2024).

O relatório em questão apresenta atividades como Laboratórios Virtuais de Programação (LVP), questionários, conteúdo interativo e as avaliações parciais, final e de recuperação da disciplina. Caso um aluno não tenha realizado uma das atividades, sua nota é indicada por um hífen (-). Para os demais casos, as notas podem ter sido determinadas por atribuição automática ou avaliadas manualmente.

A disposição das colunas é definida pela estrutura que o professor define e não indica os pesos relativos de cada atividade na nota final. Entretanto, para a disciplina em questão, os LVP e os questionários constituem a categoria de *Avaliações Parciais* e representam 40% da nota, e a *Avaliação Final*, avaliada através de quatro atividades aplicadas ao final do curso, representa os 60% restantes. Além disso, há atividades que podem estar no sistema sem terem sido disponibilizadas e devem ser desconsideradas, como é o caso do LVP 0.

Assim como o relatório de presenças, o relatório pode ser exportado pela aba do módulo de notas do Moodle. O formato do arquivo pode ser ajustado para compatibilidade com Microsoft Excel ou OpenOffice.

3.2 ANÁLISES ESTATÍSTICAS

Após a coleta dos dados, realizaram-se análises estatísticas preliminares dos conjuntos de dados cadastrais. O objetivo principal dessas análises é delinear o perfil dos estudantes, identificar padrões de ingresso e explorar a distribuição geográfica e demográfica dos alunos. Para alcançar esses objetivos, utilizaram-se técnicas estatísticas e ferramentas de análise de dados.

A análise descritiva foi a primeira etapa e teve como objetivo fornecer uma visão geral das principais características dos estudantes, tanto de forma isolada quanto por curso. Para isso, foram utilizadas distribuições de frequência para descrever características como idade de ingresso, distribuição por gênero, raça, tipo de ensino (público ou privado) e região de origem.

Em seguida, foi realizada uma análise temporal para identificar tendências e padrões no ingresso de estudantes ao longo dos anos, destacando variações significativas, especialmente em períodos de impacto externo, como a pandemia de COVID-19 entre 2020 e 2022. Para isso, foram utilizadas séries temporais e visualizações de dados, como gráficos de linha e de barras, que ilustram as flutuações no ingresso anual, a evolução do tipo de ensino (público vs privado) e a variação de IAA e IAP ao longo dos anos.

Finalmente, foi realizado o desenvolvimento de uma ferramenta em Python para a análise exploratória dos dados. Esta ferramenta permite mapear e analisar a distribuição geográfica dos alunos, visualizando a disposição dos dados em mapas de calor e facilitando a análise de padrões regionais. Com isso, a análise exploratória se centralizou na distribuição dos estudantes por raça, gênero e curso em diferentes regiões, utilizando mapas temáticos para identificar concentrações e padrões. Além disso, foi investigada a correlação entre características demográficas e o IAA, categorizando os estudantes em grupos de desempenho e explorando possíveis desigualdades.

3.3 TRATAMENTO DOS DADOS

Os relatórios obtidos fornecem uma quantidade considerável de informações sobre os estudantes. No entanto, esses dados, em seu estado bruto, apresentam desafios para serem diretamente aplicados em modelos de AM. Dentre esses desafios, destacam-se a presença de dados redundantes, valores ausentes e a variabilidade na qualidade e formato dos dados.

Para superar esses desafios e tornar os dados mais adequados para análises preditivas, foram realizados processos de tratamento que visam melhorar sua interpretação e utilidade. O resultado dessas manipulações, e versão final que será usada na análise, é o dicionário de índices apresentado na Tabela 2.

Tabela 2 – Dicionário de Índices Tratados

Índices Cadastrais	
nomeCurso	Nome completo do curso frequentado pelo estudante.
Sexo	Gênero do estudante.
racaCor	Categoria de raça ou cor do estudante.
anoIngresso	Ano de ingresso do aluno na instituição.
ensinoPublico?	Indica se o estudante veio de escola pública (sim ou não).
IAA	Índice de aproveitamento acadêmico acumulado.
reprovou?	Indica se o aluno já reprovou em alguma disciplina (sim ou não).
origemRegiao	Região de origem do aluno. Exceção para SC, onde são especificados os estudantes de Joinville.
intervaloGrad	Intervalo entre a graduação do ensino médio e ingresso do estudante na instituição.
idadeCursando	Idade do estudante durante a disciplina.
idadeIngresso	Idade do estudante no momento do ingresso na instituição.
idadeDif	Diferença em anos entre o ingresso e o momento atual.
Índices Presenças	
presenca_XX	Percentual de presença nas aulas, subdividido em quartis (25, 50, 75, 95).
faltas_Consecutivas_XX	Número de faltas consecutivas, subdividido em quartis.
Índices Notas	
LPV_XX	Média dos LPV completados, subdividido em quartis.
LPV_incompletos_XX	Número de LPV não completados, subdividido em quartis.
Quest_XX	Média dos questionários respondidos, subdividido em quartis.
Quest_incompletos_XX	Número de questionários não completados, subdividido em quartis.
Atv_AF	Média das atividades avaliativas finais completadas, para o quartil de 95%.
Atv_AF_incompletas	Número de atividades avaliativas finais não completadas, para o quartil de 95%.

Fonte: Elaborado pelo Autor (2024).

Para se chegar a esses índices, foram seguidos três passos principais: a limpeza, manipulação e a segmentação dos dados. Cada um desses processos é descrito nas subseções a seguir.

3.3.1 Limpeza dos dados

A etapa inicial na limpeza dos dados envolveu a anonimização dos estudantes, removendo dos relatórios quaisquer índices relacionados à sua identificação, como nome, sobrenome, endereço de e-mail e ID UFSC. Para permitir a integração futura entre os relatórios, foi criada uma nova coluna para categorizar os estudantes de maneira sequencial utilizando identificadores numéricos.

Em seguida, foram tratados os valores ausentes. No relatório de dados cadastrais, os valores faltantes dos índices numéricos foram substituídos pela média, enquanto os índices de categoria utilizaram o valor mais frequente ou relacional. No relatório de notas, as atividades não aplicadas foram removidas. O relatório de presenças não apresentou valores vazios.

Para concluir, foram excluídos índices irrelevantes para a análise com base em sua redundância, baixa variação grupal ou falta de influência discernível nos parâmetros de interesse. Para os dados cadastrais, por exemplo, índices como *Curso* e *nomeCurso* classificam o mesmo tipo de informação, *Nacionalidade* e *estadoCivil* possuem um valor que representa mais de 98% de seus casos e, finalizando, o nome da escola em que um estudante concluiu o ensino médio, representado por *EscolaSG*, não contribui significativamente com a análise devido à sua ampla quantidade de opções.

Para o relatório de presenças, a manipulação subsequente dos dados resultou na redundância e remoção de índices como o total de ausências, sessões anotadas e porcentagem de presenças. Apenas as pontuações por sessão foram mantidas. Além disso, sessões em que todos os alunos estiveram presentes ou ausentes foram consideradas inválidas e excluídas.

No relatório de notas, índices como *Curso* e *Presença* foram eliminados por já estarem presentes nos outros relatórios. Ademais, os conteúdos interativos foram removidos por conta da baixa participação por parte dos estudantes com essas atividades. No fim, restaram apenas as atividades avaliativas relacionadas aos LVP, questionários e avaliação final.

3.3.2 Manipulação dos dados

Após a limpeza, os dados foram submetidos a uma série de manipulações. Essa etapa envolveu a criação de novos índices, a concatenação de informação de alguns e alteração direta do significado de outros.

Para o relatório de dados cadastrais, foram criados índices como *idadeIngresso* e *reprovou?*, conforme introduzido na Subseção 3.1.1.2. Esses índices substituíram alguns dos índices originais, permitindo seu uso como variáveis contínuas nos modelos. Em seguida, *UFSG* foi adaptado junto de *MunicipioSG* para discernir quais estudantes de SC são de Joinville e quais são de outros municípios. Esse passo se fez importante pois há um grande número de estudantes oriundos dessa cidade, conforme será apresentado futuramente na Seção 4.1.1. Além disso, foi renomeado para *origemRegiao* e unificou os estados em suas regiões. Por fim, *anoSemestreIngresso* foi modificado para representar apenas o ano de ingresso, sem a presença do semestre.

No relatório de presenças, foi incorporado um novo índice para indicar o número de faltas consecutivas de cada estudante. Adicionalmente, a porcentagem de presença foi reformulada para refletir de forma mais precisa a situação de cada aluno. A pontuação total exclui os dias em que o aluno ainda não estava inscrito na disciplina. Por exemplo, um estudante que participou de quatro aulas e esteve inscrito desde o primeiro dia terá uma taxa de presença de 100% com oito pontos. Em contraste, um estudante que ingressou na disciplina no segundo dia terá uma taxa de presença de 100% com seis pontos, já que o primeiro dia é desconsiderado por ainda não estar inscrito.

No relatório de notas, foram adicionados novos índices para indicar o número de LVP, questionários e atividades que compõem a avaliação final da disciplina que o estudante não realizou.

3.3.3 Segmentação dos Dados

Para viabilizar o objetivo de identificar alunos que estão em risco de reprovação com antecedência suficiente para permitir a intervenção do professor, o conjunto de dados referente às presenças e notas foi segmentado em porções de 25%, 50%, 75% e 95%, desconsiderando os últimos 5% referentes à avaliação de recuperação.

Para o relatório de presenças, por seguir uma sequência temporal, a redução foi realizada de forma a considerar os registros acumulados em cada um dos quatro intervalos mencionados. Os índices de faltas consecutivas e a porcentagem de presença foram adaptados para refletir isso. Vale ressaltar que o primeiro período será mais instável por conta da presença de alunos que são matriculados com atraso — devido a chamadas de matrícula realizadas após o início do semestre — e que possuem um peso maior para cada dia.

A redução do relatório de notas buscou considerar apenas os LVP e questionários, já que as quatro atividades que compõem a avaliação final são todas aplicadas simultaneamente ao final do curso, conforme explicado na seção 3.1.3. Para cada intervalo, foram selecionadas quantias condizentes de cada atividade e sua média foi

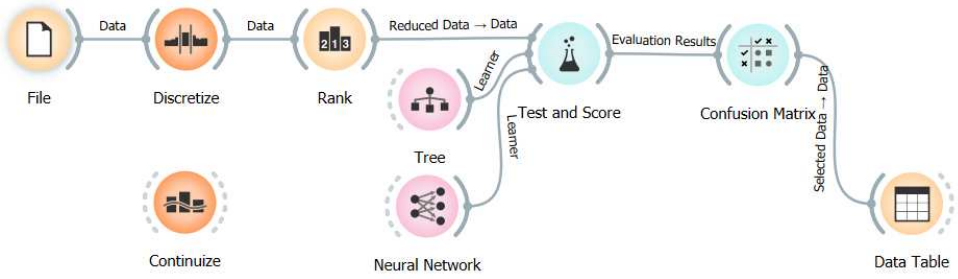
tomada. Por fim, novos índices foram criados com essas métricas e o índice de atividades não realizadas foi adaptado para também refletir essa distribuição.

Os dados cadastrais não foram segmentados pois já possibilitam uma previsão da chance de reprovação dos estudantes antes mesmo que ingressem na disciplina.

3.4 DESENVOLVIMENTO DE FLUXOS PARAS ANÁLISE PREDITIVA

Para realizar a análise preditiva dos dados tratados, foi utilizado o software Orange Data Mining, uma ferramenta open-source de mineração de dados que oferece uma interface gráfica intuitiva. Orange permite a criação de fluxos de trabalho interativos para pré-processamento, modelagem, visualização e análise de dados. Sua flexibilidade e variedade de widgets tornam-no adequado para a implementação dos modelos preditivos empregados neste trabalho. Um exemplo de fluxo similar ao implementado neste estudo pode ser observado na Figura 6, e nesta subseção é descrito como este pode ser configurado.

Figura 6 – Exemplo de Fluxo na Ferramenta Orange



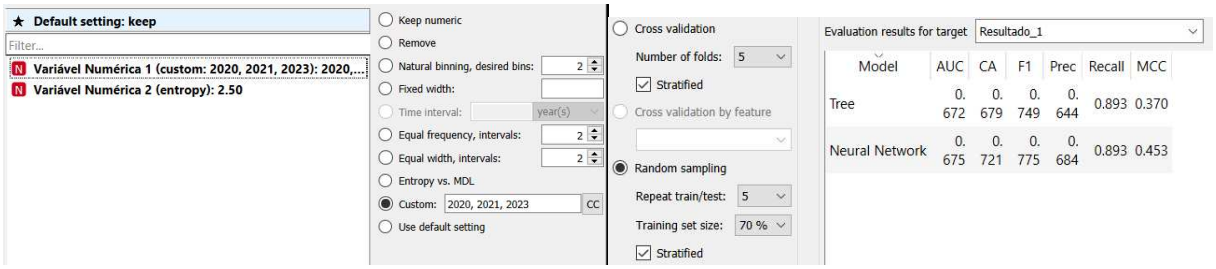
Fonte: Elaborado pelo Autor (2024).

A etapa inicial envolve a importação dos dados, processo simplificado pela interface do Orange, que permite a seleção da variável alvo, o ajuste de meta-informações e a reclassificação de índices como categóricos ou numéricos. A identificação dos alunos, por exemplo, é uma meta-informação utilizada apenas na integração dos diferentes dados e, ao ser classificada como tal, é excluída das análises subsequentes.

Após a importação dos dados, segue-se o processo de pré-processamento. Utilizando o widget *Discretize*, é possível discretizar dados numéricos em intervalos definidos por igual frequência, por entropia, ou ainda em intervalos definidos manualmente. A Figura 7a mostra um exemplo de variável discretizada com intervalos customizados, além de outra variável processada pela opção de entropia disponível no Orange. Seguindo o pré-processamento, o widget *Continuize* é utilizado para normalizar os valores numéricos. Já o widget *Rank* é utilizado para selecionar os atributos mais informativos, empregando métodos como a Proporção de Ganho de Informação.

Em seguida, os dados já processados são integrados ao widget *Test and Score*, que é responsável pela aplicação dos modelos de AM sobre os dados. Este widget permite a divisão dos dados e facilita a visualização das métricas de desempenho obtidas por cada técnica, tais como acurácia, precisão e recall, conforme ilustrado na Figura 7b. É importante destacar que, nesta etapa, a ferramenta oferece opções de customização para as técnicas utilizadas em cada modelo, permitindo, por exemplo, limitar o número de iterações no treinamento da RNA. Para os fluxos desenvolvidos neste trabalho, as configurações de cada técnica foram ajustadas experimentalmente para otimizar os resultados.

Figura 7 – Widgets *Discretize* e *Test and Score* do Orange

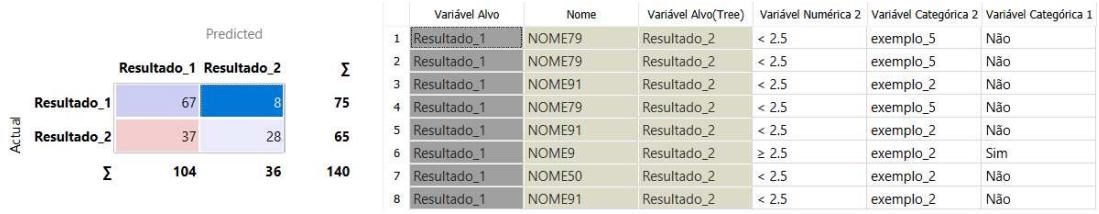


(a) Discretize (b) Test and Score

Fonte: Elaborado pelo Autor (2024).

Por fim, os resultados podem ser visualizados através do widget *Confusion Matrix*, que exibe uma matriz de confusão, facilitando a análise dos acertos e erros de cada modelo. Adicionalmente, ao incorporar o widget *Data Table*, torna-se possível examinar especificamente os dados que foram classificados incorretamente. A Figura 8 exemplifica como os falsos positivos, identificados pelo algoritmo de AD, são analisados. Esta análise ajuda a entender as causas dos erros de classificação e refinar os modelos preditivos e tratamento dos dados.

Figura 8 – Widgets Confusion Matrix e Data Table do Orange



Fonte: Elaborado pelo Autor (2024).

3.5 ANÁLISES PREDITIVAS

Com os dados devidamente tratados e o método de desenvolvimento de fluxos estabelecido, procedeu-se às análises preditivas. Com estas análises, busca-se

avaliar a influência de variáveis cadastrais, de presenças e notas na reprovação dos estudantes.

Inicialmente, foram elaborados fluxos distintos para cada tipo de dado, adaptando as técnicas de tratamento e análise para maximizar a extração de informações relevantes. Dentre as técnicas empregadas, destacam-se a discretização e ranqueamento por proporção de ganho para os dados cadastrais e a normalização para as presenças e notas.

Em seguida, para cada fluxo desenvolvido, a divisão dos dados foi guiada por técnicas de amostragem aleatória estratificada, na qual 70% dos dados foram alocados para treinamento, com o processo sendo repetido cinco vezes. Esta técnica foi selecionada para que os conjuntos de treinamento e teste refletissem a diversidade do corpo discente.

Posteriormente, foram avaliadas as métricas de sensibilidade, precisão, pontuação F1 e acurácia para cada caso, utilizando os classificadores LR, AD, RNA, SVM, NB, *Random Forest* (RF), *K-Nearest Neighbors* (KNN) e *Gradient Boosting* (GB). Além disso, foram analisadas também as matrizes de confusão e tabelas de dados para identificar os valores responsáveis pela classificação correta e incorreta dos dados como VP, VN, FP e FN.

Por fim, após a execução e avaliação inicial de cada modelo dentro de seus respectivos fluxos, procedeu-se com a integração dos modelos para explorar sinergias entre diferentes tipos de dados. Esta etapa visou ampliar a compreensão sobre como as variáveis interagem entre si e qual a sua influência combinada na predição dos resultados acadêmicos.

4 MODELOS E RESULTADOS

Neste capitulo, os resultados obtidos durante a análise dos diferentes conjuntos de dados são discorridos. Inicialmente, são apresentadas as análises estatísticas realizadas nos dados cadastrais antes de seu tratamento, com o objetivo de compreender a distribuição de perfis dos estudantes e sua influência no desempenho acadêmico. Finalmente, os fluxos preditivos criados para os três agrupamentos de dados, com o objetivo de classificar os estudantes como aprovados ou reprovados, são descritos e as análises realizadas sobre estes são discutidas.

4.1 ANÁLISES ESTATÍSTICAS

A compreensão das dinâmicas educacionais requer uma análise atenta dos dados acadêmicos, com um enfoque às camadas de informação que delineiam o cenário. Nessa seção, serão expostas análises estatísticas realizadas sobre os dados cadastrais dos 1399 estudantes que ingressaram entre os semestres 2015.1 e 2023.2 e ainda estão matriculados, ou seja, desconsiderando desistentes, graduados e aqueles com matrícula trancada.

Inicialmente, é apresentada uma análise descritiva das variáveis-chave que moldam o perfil dos estudantes. Em seguida, a análise temporal explora a evolução do ingresso dos estudantes ao longo dos anos. Finalmente, é introduzida uma ferramenta desenvolvida para facilitar a análise exploratória. Esta análise investiga padrões na distribuição geográfica e demográfica dos estudantes, e é dividida em duas partes: a primeira aborda a origem regional dos estudantes por raça e gênero, enquanto a segunda explora a relação entre essas características e o desempenho acadêmico (IAA).

As análises nesta seção abordem exclusivamente os dados cadastrais dos estudantes, e é importante destacar que os resultados alcançados não são limitados a qualquer disciplina específica, incluindo a de Programação I. Por explorar facetas inerentes a cada estudante, independentemente da disciplina cursada, as informações apresentadas podem ser aplicadas a diferentes contextos, contanto que sejam apropriadamente tratadas.

4.1.1 Análise Descritiva

A análise descritiva aqui apresentada visa fornecer uma descrição breve e direta das distribuições das principais características dos estudantes, como idade de ingresso, distribuição por gênero, e origem geográfica.

Distribuição da Idade de Ingresso: A média da idade de ingresso na universidade é de 20,6 anos. A maioria dos estudantes (75%) tem entre 18 e 21 anos. A idade mínima é 16 anos e a máxima é 71, indicando que muitos iniciam imediatamente após o ensino médio, enquanto outros retomam os estudos depois de passarem um tempo trabalhando.

Intervalo entre Ensino Médio e Graduação: O tempo médio entre o término do ensino médio e o ingresso na graduação é de 1,56 anos. Cerca de 75% dos estudantes entram na universidade até dois anos após a conclusão da escola. O maior intervalo registrado é de 37 anos.

Distribuição por Gênero (Geral): A predominância masculina é marcante, com 1.024 homens (72,5%) e 388 mulheres (27,5%). A disparidade de gênero é uma tendência comum em cursos de engenharia e tecnologia, destacando a necessidade de incentivar a participação feminina.

Distribuição por Gênero (Por Curso): Engenharia Automotiva e Engenharia Mecatrônica têm maioria masculina. Já em cursos como Engenharia Civil, Engenharia Naval e Ciência e Tecnologia, a proporção de estudantes femininas é mais equilibrada, ou até mesmo superior, a de masculinos. A distribuição de cada curso pode ser visualizada na Figura 9b.

Distribuição por Raça/Cor: Estudantes que se identificam como brancos são a maioria, com 76,6%. Pardos representam 15,1%, pretos 3,3%, amarelos 3,0%, indígenas 0,07% e não declarados 1,4%.

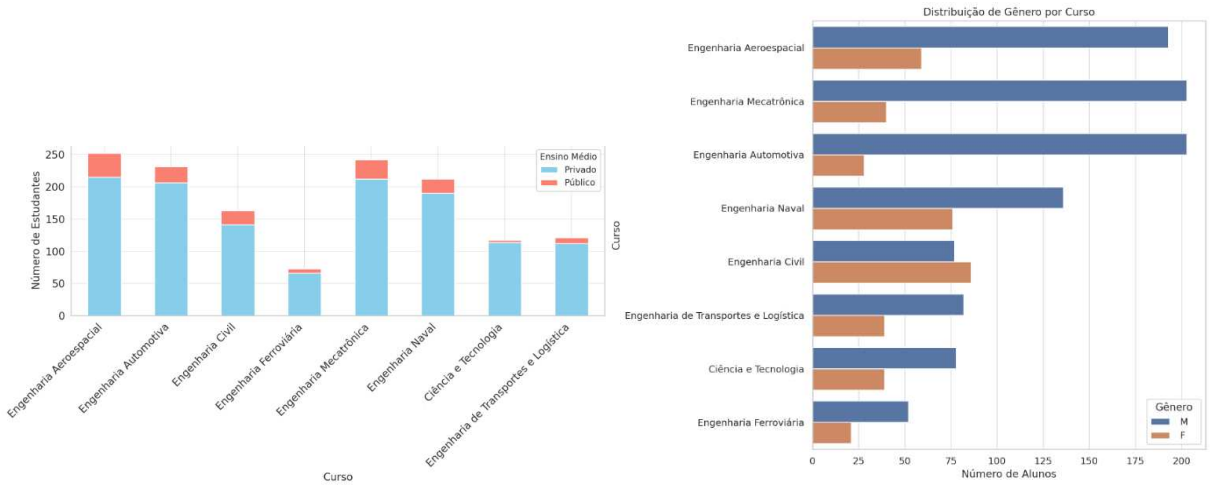
Distribuição por Tipo de Ensino: A maioria dos estudantes (88,9%) vem de escolas particulares, enquanto apenas 11,1% são de escolas públicas, conforme Figura 9a.

Distribuição por Região: Estudantes do Sul são maioria (62,3%), seguidos pelo Sudeste (18,1%), Centro-Oeste (5,6%), Norte (5,4%) e Nordeste (3,3%).

Distribuição por Unidade Federativa: A maioria dos estudantes é da região de Santa Catarina (42,4%). Outros estados relevantes incluem Paraná (14,7%), São Paulo (12,5%), Rio Grande do Sul (4,8%) e Minas Gerais (2,8%).

Distribuição de Joinvilenses em Relação a SC: Dos ingressantes de Santa Catarina, 48,16% são de Joinville e 51,84% são de outras cidades, mostrando que a proximidade da universidade influência na seleção dos estudantes.

Figura 9 – Análise Descritiva por Curso



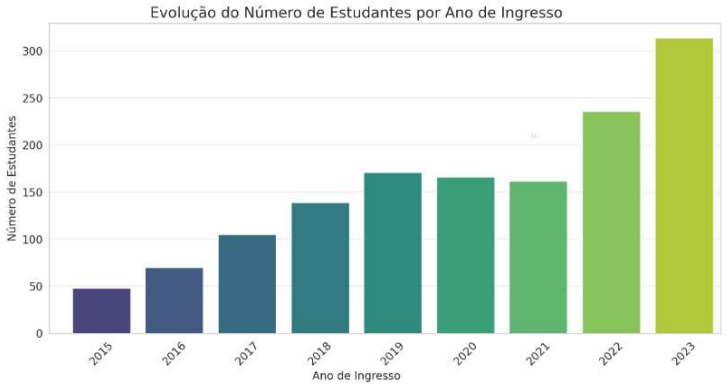
(a) Distribuição de Ensino Público vs Privado (b) Distribuição por Gênero

Fonte: Elaborado pelo Autor (2024).

4.1.2 Análise Temporal

A análise temporal dos dados revela padrões no comportamento estudantil ao longo dos anos, com um olhar particular sobre o impacto da pandemia de COVID-19 nos anos recentes. A Figura 10 ilustra essa evolução na admissão de estudantes, delineando as flutuações na quantidade de alunos ingressantes por ano na universidade.

Figura 10 – Distribuição dos Alunos por Ano de Ingresso



Fonte: Elaborado pelo Autor (2024).

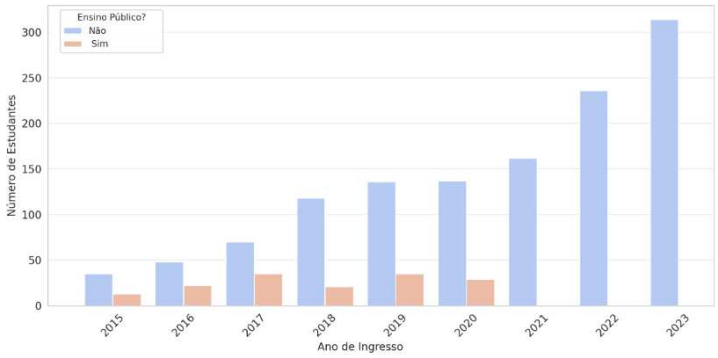
De 2015 a 2019, a instituição experienciou um aumento progressivo no número de novos estudantes, refletindo, possivelmente, a atratividade crescente do campus e da oferta acadêmica. Contudo, ao comparar com o ano de 2020, observa-se uma inversão dessa tendência: o número de alunos sofre uma redução. Esta diminuição, que persistiu até 2021, coincide com o período da pandemia, sugerindo que os desafios impostos pelo contexto global podem ter influenciado a decisão de ingresso dos estudantes.

Em 2022, com o arrefecimento da crise, nota-se um fenômeno que poderia ser interpretado como um 'estouro de ingresso': uma recuperação expressiva que supera os números pré-pandemia. Este pico pode ser atribuído à normalização das condições de vida e à retomada das atividades presenciais. De maneira similar, esse ímpeto se mantém em 2023, indicando uma possível estabilização no interesse pelo ensino superior e uma recuperação do setor educacional.

Uma camada adicional de análise emerge ao se investigar as origens educacionais dos alunos admitidos anualmente. No entanto, é importante considerar a possibilidade de inconsistências ou erros de registro no sistema de dados, que podem ter falhado em capturar adequadamente as informações de admissão de alunos oriundos do ensino público a partir de 2021. Se os dados forem precisos, tal fenômeno demanda uma investigação imediata e a implementação de medidas corretivas.

A Figura 11 oferece um panorama dessa distribuição ao longo dos anos estudados, evidenciando as tendências de ingresso de estudantes provenientes de escolas públicas e privadas.

Figura 11 – Distribuição do Tipo de Ensino por Ano de Ingresso

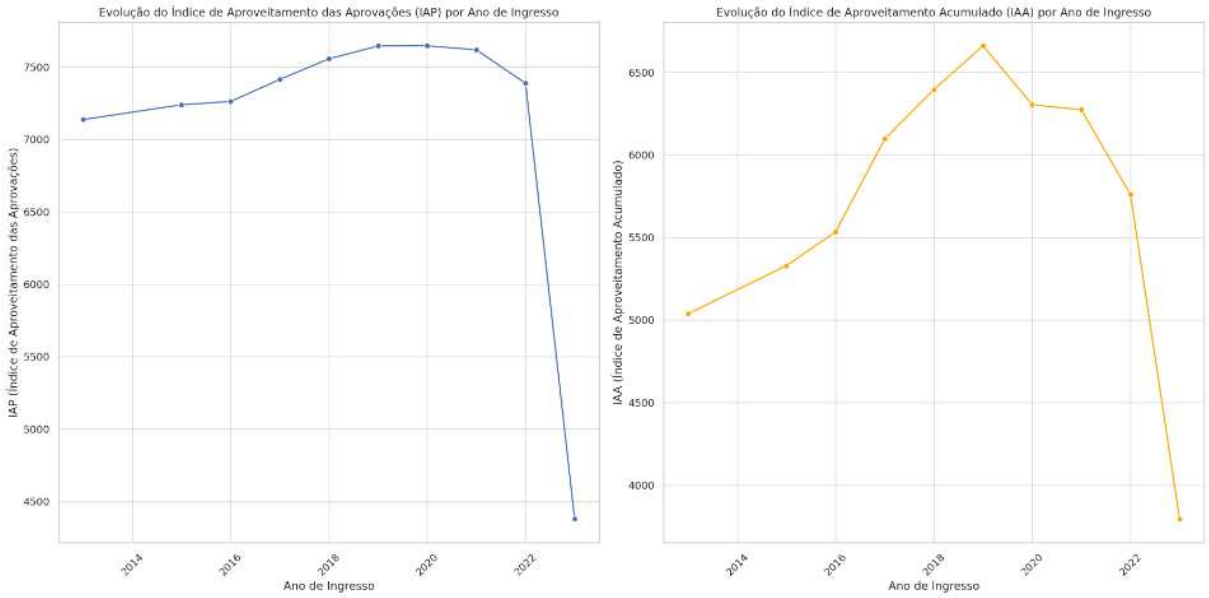


Fonte: Elaborado pelo Autor (2024).

De 2015 a 2020, o contingente de alunos oriundos do ensino privado mostrou um incremento consistente. Em contraste, o número de alunos do ensino público demonstrou uma variação mais errática. O cenário fica ainda mais extremo a partir de 2021, quando o impacto da pandemia no sistema educacional se torna evidente. Durante esse período, o ingresso de alunos de escolas públicas é aproximadamente nulo, possível reflexo das disparidades ampliadas pelo ensino à distância emergencial e os desafios socioeconômicos enfrentados.

Por fim, a análise do desempenho acadêmico, capturado pelo IAA e pelo IAP, sugere uma leve influência da pandemia nesses parâmetros. A Figura 12 mostra que os alunos que ingressaram pouco antes da pandemia experimentaram um incremento em seus índices de aproveitamento. É plausível considerar que essa melhoria nos indicadores reflita uma diminuição na rigidez avaliativa durante o período de transição para o ensino remoto.

Figura 12 – Tendência do IAP e IAA dos Estudantes ao Longo dos Anos



Fonte: Elaborado pelo Autor (2024).

Em contraste, observa-se que os estudantes admitidos durante e logo após a pandemia tiveram seu IAA médio reduzido. Tal fenômeno pode ser interpretado como um desafio na readaptação às rotinas de estudo tradicionais com a retomada das atividades presenciais. Essa queda nos índices pode também sinalizar dificuldades residuais decorrentes do ensino à distância, que, apesar do retorno ao presencial, ainda impactam o desempenho dos alunos.

Para uma análise mais precisa, seria necessário avaliar o índice de aproveitamento por semestre ao longo do período pandêmico - informação essa não disponível no relatório de dados cadastrais.

4.1.3 Software para Mapeamento e Análise Geográfica

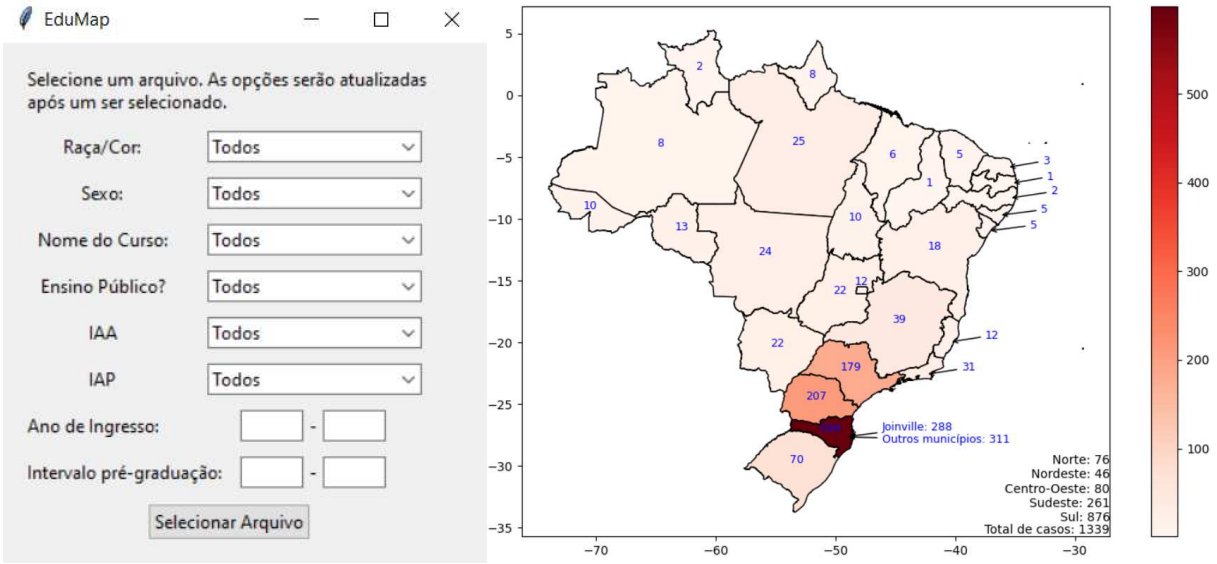
O desenvolvimento de uma ferramenta para mapeamento e análise geográfica, nomeada EduMap, iniciou-se com a identificação da necessidade de um meio eficiente para visualizar e interpretar os conjuntos de dados cadastrais gerados pelo sistema acadêmico. Para atender a esse requisito, optou-se por uma abordagem modular em Python, integrando tecnologias de front-end para visualização e bibliotecas de back-end para processamento e cálculo.

A EduMap permite ao usuário selecionar e carregar diretamente o arquivo contendo o relatório, sem a necessidade de manipulação prévia do documento. Esta característica é particularmente eficaz devido ao fato de que o relatório possui um padrão estabelecido, garantindo assim a consistência e a precisão na análise das informações importadas. Este recurso não apenas melhora a usabilidade e a flexibilidade da ferramenta, mas também assegura que os usuários possam realizar análises

customizadas com base em dados atualizados.

Adicionalmente, a concepção da interface gráfica da ferramenta foi planejada para garantir uma experiência de usuário fluida e acessível. Conforme Figura 13a, os usuários são capazes de navegar pelo sistema e selecionar variáveis de interesse. A escolha de concentrar-se nessas características foi motivada pela concepção de relevância que tais dados têm na compreensão da análise preditiva.

Figura 13 – Ferramenta EduMap



(a) Menu da Ferramenta (b) Exemplo de Resultado

Fonte: Elaborado pelo Autor (2024).

Como resultado, ao aplicar os filtros selecionados, o programa carrega um mapa temático (heatmap) do Brasil. Este mapa mostra a distribuição dos estudantes, colorindo os estados com intensidades diferentes com base no número de ocorrências. O mapa também apresenta numericamente o volume de casos por estado, com adendo especial para Santa Catarina, onde os estudantes, devido a seu alto número, são divididos entre residentes de Joinville e residentes de outros municípios. Além disso, é exibida a distribuição por região e o número total de casos para a combinação de características selecionadas.

O usuário pode refazer o processo de filtragem e visualização quantas vezes desejar, permitindo uma análise flexível de diferentes conjuntos. Um exemplo de mapa gerado sem quaisquer filtros aplicados é ilustrado na Figura 13b.

O processo iterativo de desenvolvimento enfatizou a prototipagem rápida e o teste contínuo de novas funcionalidades, assegurando que a ferramenta atenda às necessidades de análise de padrões demográficos, suportando, assim, o planejamento e a pesquisa no contexto educacional do presente trabalho. O código-fonte da ferra-

menta está disponível no repositório do Github[1].

4.1.4 Análise Exploratória

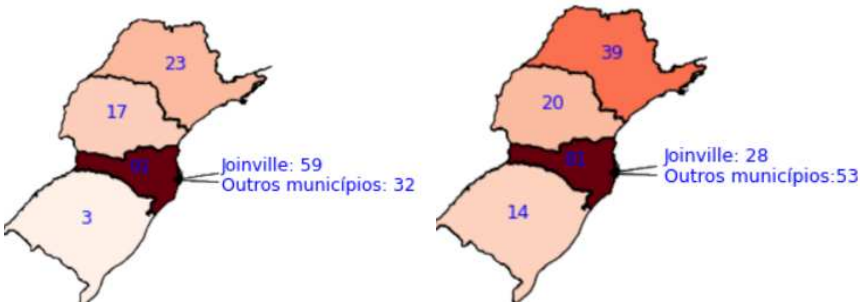
A seguir são apresentados os resultados da análise exploratória realizada através da ferramenta apresentada. A análise foi dividida em duas etapas e revelou padrões na distribuição geográfica e demográfica dos estudantes. Inicialmente, é apresentada a análise que visou identificar a origem regional dos cursos com base na raça e gênero dos estudantes, e, em seguida, a que aborda a distribuição de cada grupo com base no IAA.

É importante destacar que as considerações realizadas em relação a gênero e cor consideraram apenas esses elementos de forma isolada. Outros fatores que poderiam justificar os aspectos apresentados seriam se o aluno estudou em escola pública ou privada, a renda familiar e se o aluno apenas estuda ou precisa trabalhar. Entretanto, tais dados não apresentaram influência significativa ou não estão presentes no relatório de dados cadastrais.

4.1.4.1 Distribuição Geográfica

Ao separar as amostras por raça, os grupos de população branca e amarela apresentaram focos de origem nos estados de Santa Catarina (SC), Paraná (PR) e São Paulo (SP). Além disso, a concentração em cada estado varia entre os cursos. Para os cursos de engenharia mecatrônica e aeroespacial, por exemplo, há uma alta quantidade de estudantes oriundos do PR. Já nos cursos de engenharia naval e civil, a origem dos estudantes está predominantemente em SC e SP, com uma distinção notável em SC, conforme Figura 14: a maior parte dos alunos de engenharia civil são de Joinville, enquanto os de engenharia naval vêm de outras cidades do estado.

Figura 14 – Distinção Regional entre Engenharia Civil e Engenharia Naval para População Branca



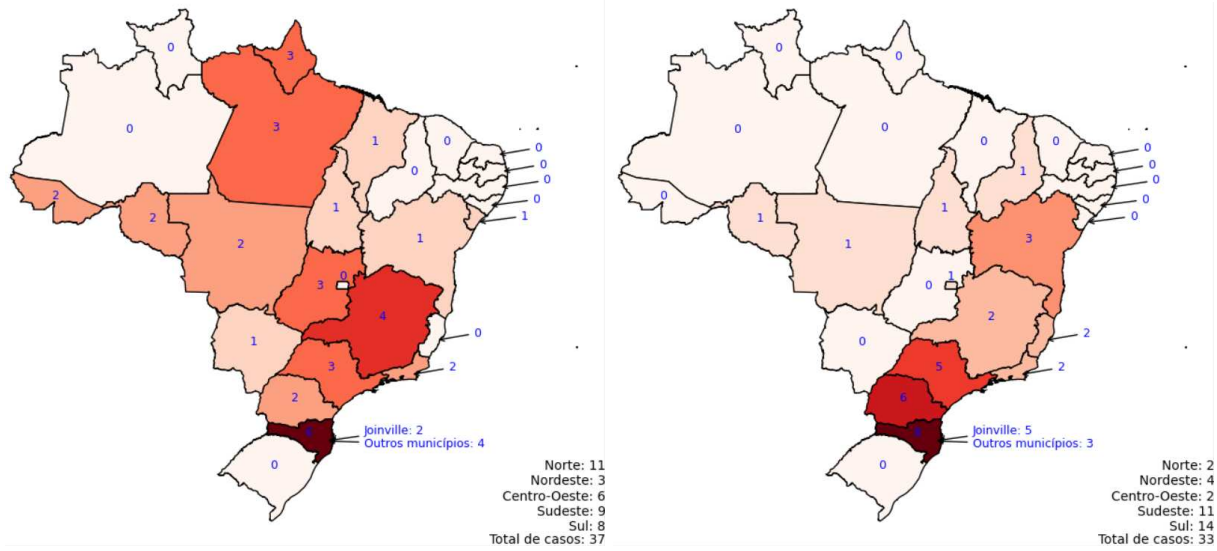
(a) Engenharia Civil (b) Engenharia Naval

Fonte: Elaborado pelo Autor (2024).

[1] https://github.com/emanueltt/Heatmap_AE

Para a população parda e preta, observa-se uma distribuição mais equilibrada e distinta por curso em todo o território nacional. O curso de engenharia mecatrônica apresenta uma distribuição uniforme, enquanto os cursos de engenharia automotiva e aeroespacial têm concentração nos estados litorâneos. Já os cursos de engenharia de transporte e logística, civil e naval têm foco no Sul e no Pará. Um exemplo dessa distribuição está presente na Figura 15, onde as populações dos cursos de engenharia mecatrônica e automotiva são comparadas.

Figura 15 – Distinção Regional entre Engenharia Mecatrônica e Engenharia Automotiva para População Parda



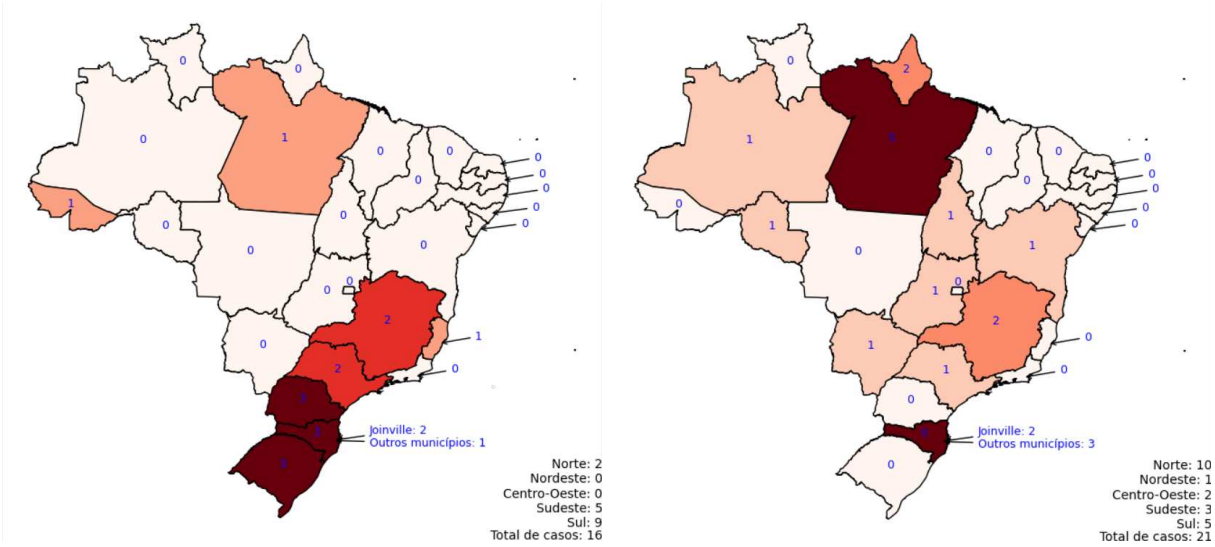
(a) Engenharia Mecatrônica (b) Engenharia Automotiva

Fonte: Elaborado pelo Autor (2024).

É possível aprofundar a análise da distribuição ao incluir a variável de gênero, revelando particularidades interessantes em alguns casos. Na engenharia naval, por exemplo, a população feminina parda está concentrada nas regiões Sul e Sudeste, enquanto a população masculina está mais distribuída pelo Norte, conforme a Figura 16. No curso de engenharia mecatrônica, por outro lado, as mulheres brancas estão predominantemente concentradas em SC, enquanto os homens têm uma distribuição mais equilibrada entre SC e PR.

As análises das outras combinações de raça, curso e gênero seguiram padrões semelhantes aos discutidos, mostrando distribuições comparáveis em termos regionais e demográficos. Embora não tenham sido abordadas detalhadamente neste capítulo, essas combinações também oferecem informações valiosas sobre a diversidade e a distribuição dos estudantes.

Figura 16 – Distribuição da População Parda de Engenharia Naval por Gênero



(a) População Feminina (b) População Masculina

Fonte: Elaborado pelo Autor (2024).

4.1.4.2 Distribuição por Desempenho

Para a análise por desempenho, são utilizadas as informações previamente analisadas sobre a distribuição de estudantes por raça, curso e gênero, combinando-as com o IAA para investigar possíveis relações entre essas características e o desempenho acadêmico. A análise categoriza os estudantes em quatro grupos de desempenho: baixo ($0.1 \leq IAA < 5$), médio-baixo ($5 \leq IAA < 6$), médio-alto ($6 \leq IAA < 7.5$) e alto ($IAA > 7.5$). Vale destacar que o grupo de valores inferiores a 0.1 não está incluído, pois pode conter alunos que acabaram de ingressar na universidade e ainda não possuem IAA definido.

Inicialmente, ao fazer a análise isolada por regiões, observa-se que o Nordeste e Centro-Oeste possuem 32% dos estudantes nas categorias de valor baixo e médio-baixo. A região Norte apresenta uma porcentagem significativamente maior, com 60% dos estudantes pertencendo a esses grupos. Por fim, as regiões Sudeste e Sul possuem valores intermediários de 50% e 42%, respectivamente. Particularmente no Sul, nota-se uma concentração crescente de estudantes no PR e em outros municípios de SC para os grupos de valor médio-alto e alto. Conforme os valores começam a diminuir para os grupos de valor médio-baixo e baixo, a concentração de estudantes em Joinville começa a crescer e os outros estados da região passam a ter uma distribuição mais homogênea.

Em seguida, de maneira similar à distribuição geográfica, é feita a inclusão das variáveis de raça e gênero na análise do IAA. A Tabela 3 apresenta a distribuição dos grupos para as populações parda, branca, amarela e preta de acordo com o gênero.

Embora não presentes na tabela, a análise englobou as regiões em que cada grupo pertence.

Tabela 3 – Distribuição Percentual do IAA por Raça e Gênero

	Parda F	Parda M	Branca F	Branca M
Alto (> 7.5)	16,07%	16,26%	23,68%	24,79%
Médio-alto (6 - 7.5)	39,29%	34,15%	40,23%	38,24%
Médio-baixo (5 - 6)	19,64%	18,70%	19,17%	18,63%
Baixo (< 5)	25,00%	30,89%	16,92%	18,35%
	Amarela F	Amarela M	Preta F	Preta M
Alto (> 7.5)	33,33%	24,00%	0,00%	3,57%
Médio-alto (6 - 7.5)	50,00%	40,00%	33,33%	32,14%
Médio-baixo (5 - 6)	0,00%	12,00%	22,22%	17,86%
Baixo (< 5)	16,67%	24,00%	44,44%	46,43%

Fonte: Elaborado pelo Autor (2024).

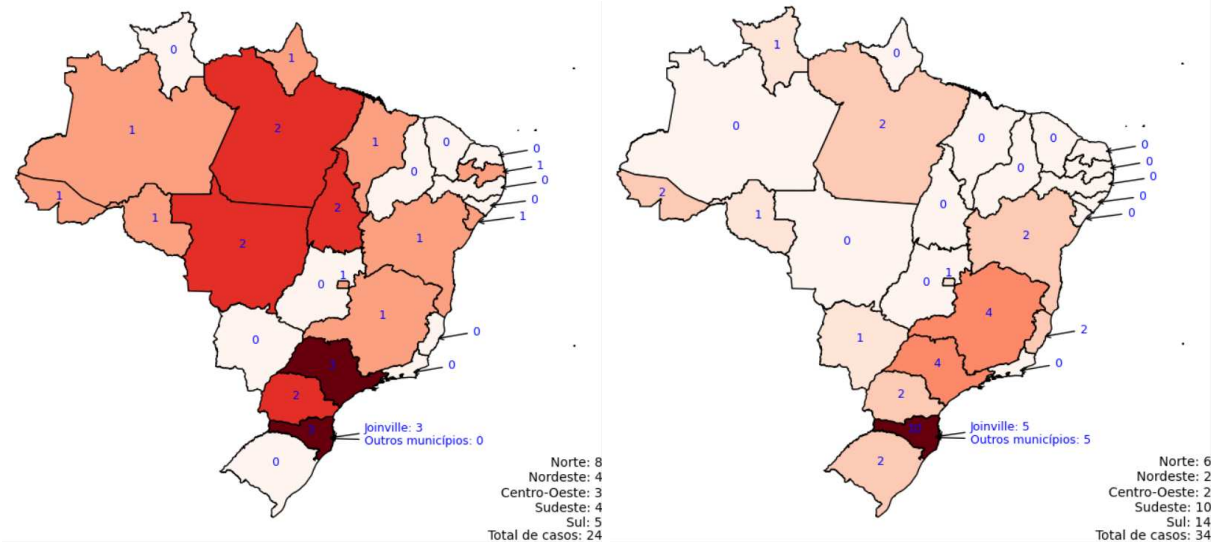
Para a população parda, a distribuição entre os gêneros é similar no grupo de notas altas, onde aproximadamente 16% da população feminina e masculina pertence, sendo majoritariamente da região Norte. A mesma similaridade é observada no grupo de notas médio-baixas, com cerca de 18% de ambos os gêneros, porém com concentração nas regiões Sudeste e Sul. A principal diferença surge nos grupos de notas médio-altas e baixas: as mulheres têm uma proporção 5% maior no grupo de notas médio-altas (39% contra 34%), sem uma concentração específica em qualquer região. Já no grupo de notas baixas, os homens apresentam uma proporção maior (30,89% contra 25%), e a distribuição é semelhante à das notas médio-baixas, com maior presença nas regiões Sudeste e Sul. A Figura 17 demonstra a comparação entre a distribuição para os grupos de valor alto e médio-baixo da população parda.

A população branca apresenta uma distribuição similar para ambos os gêneros em todas as categorias: 24% para notas altas, 40% para notas médio-altas, 19% para notas médio-baixas e 17% para notas baixas. No entanto, além de SC, as notas mais baixas possuem uma concentração maior em SP, enquanto as notas mais altas vem do PR.

A população amarela feminina representa o grupo com o melhor índice acadêmico. Cerca de 83% das mulheres pertencem aos grupos de notas médio-altas (50%) e altas (33%) e estão distribuídas entre os estados do PR e SC. A população masculina amarela apresenta uma distribuição similar à população branca, porém com um percentual maior de alunos pertencentes ao grupo de valor baixo. A maior parte da população com notas inferiores à média pertence ao estado de SP.

Em contraste, a população preta é a que apresenta maiores dificuldades. Tanto para homens quanto para mulheres, cerca de 45% da população tem notas de

Figura 17 – Distribuição dos Grupos de Valor Alto e Médio-Baixo para População Parda



(a) Valor Alto (b) Valor Médio-Baixo

Fonte: Elaborado pelo Autor (2024).

valor baixo, e apenas 2,5% da população geral consegue alcançar notas altas. Para notas médio-altas, 33% representa um valor similar ao das outras raças, embora um pouco inferior.

Por fim, a análise do desempenho acadêmico também foi realizada através da distribuição por curso e gênero, conforme Tabela 4. Destaca-se que ao separar os dados por curso, observou-se que as populações resultantes tornaram-se reduzidas a ponto de serem insuficientes para identificar padrões na distribuição regional. Devido a esse obstáculo, a análise focará na relação entre o desempenho acadêmico em cada curso e os gêneros.

Para a maioria dos cursos, uma maior parte da população do gênero feminino pertence aos grupos de valor alto e médio-alto, com uma diferença média de 6% em relação ao gênero masculino. Notadamente, os cursos de engenharia de transportes e logística e ferroviária apresentam as maiores disparidades, com diferenças de 15,53% e 9,21%, respectivamente. Curiosamente, o curso de engenharia civil é uma exceção, onde a população masculina supera a feminina por 1,05% nesses grupos de desempenho.

Além disso, a maioria dos cursos apresenta uma distribuição equilibrada entre os grupos de valor superior e inferior à média de aprovação da UFSC (6), com aproximadamente 52% dos estudantes acima e 48% abaixo dessa média. Entre as exceções notáveis, destacam-se os cursos de engenharia mecatrônica e aeroespacial, nos quais 71% dos estudantes estão nos grupos de desempenho médio-alto e alto. Em contraste, o curso de engenharia ferroviária tem apenas 33% de sua população

Tabela 4 – Distribuição Percentual do IAA por Curso e Gênero

	Cientec F	Cientec M	Aero F	Aero M	Auto F	Auto M
Alto (> 7.5)	26,47%	12,50%	28,57%	36,81%	25,93%	12,89%
Medio alto (6 - 7.5)	27,78%	34,72%	42,86%	32,97%	33,33%	43,30%
Medio baixo (5 - 6)	13,89%	9,72%	8,93%	12,64%	18,52%	19,59%
Baixo (< 5)	19,44%	37,50%	10,71%	13,19%	14,81%	16,49%

	Civil F	Civil M	Transp F	Transp M	Ferro F	Ferro M
Alto (> 7.5)	10,98%	9,72%	15,79%	12,50%	9,52%	4,44%
Medio alto (6 - 7.5)	36,59%	38,89%	44,74%	32,50%	28,57%	24,44%
Medio baixo (5 - 6)	23,17%	22,22%	21,05%	17,50%	19,05%	26,67%
Baixo (< 5)	23,17%	27,78%	15,79%	25,00%	33,33%	42,22%

	Meca F	Meca M	Naval F	Naval M
Alto (> 7.5)	37,50%	32,46%	16,22%	18,46%
Medio alto (6 - 7.5)	37,50%	34,55%	41,89%	32,31%
Medio baixo (5 - 6)	12,50%	14,14%	20,27%	22,31%
Baixo (< 5)	5,00%	14,66%	17,57%	17,69%

Fonte: Elaborado pelo Autor (2024).

nesses grupos de desempenho.

4.2 ANÁLISES PREDITIVAS APLICADAS PARA TURMA DE PROGRAMAÇÃO I

Com o objetivo de compreender a influência dos diferentes tipos de dados na reprovação dos estudantes, esta seção detalha os fluxos de análise preditiva desenvolvidos e os resultados obtidos, conforme o método descrito na Seção 3.4. A análise incide sobre os 93 estudantes da disciplina de Programação I do segundo semestre de 2023, buscando entender a influência de cada índice sobre os casos classificados correta e incorretamente.

Inicialmente, é descrito o desenvolvimento do fluxo de dados cadastrais. Em seguida, aborda-se o fluxo de presenças e a concatenação de ambos. Posteriormente, detalha-se o fluxo de notas e, por fim, a integração de todos os fluxos mencionados. As matrizes de confusão analisadas em cada fluxo podem ser encontradas no Apêndice A.

4.2.1 Fluxo Cadastral

Para o fluxo de dados cadastrais, as variáveis numéricas foram inicialmente discretizadas com o intuito de criar grupos com distribuições homogêneas. Cada variável foi segmentada de modo que cada grupo resultante contivesse um número similar de estudantes, promovendo uma análise equilibrada.

Posteriormente, adotou-se uma abordagem diferente na discretização para maximizar o ganho de informação, baseando-se nas análises detalhadas na seção anterior. O índice *anoIngresso*, por exemplo, foi segmentado nos períodos de 2020, 2022 e 2023, refletindo o impacto que a pandemia teve sobre o desempenho dos estudantes. Além disso, para variáveis ainda não exploradas, métodos baseados em entropia e no critério de CMD foram utilizados para identificar os pontos de corte.

Após a discretização, os índices foram ranqueados com base na proporção de ganho de informação. Embora tenham permanecido majoritariamente os mesmos para ambos os casos, o valor das informações para o segundo foi 80% maior. Os oito índices mais relevantes foram então utilizados nos algoritmos de aprendizado de máquina, e as métricas de classificação obtidas para cada abordagem estão apresentadas na Tabela 5.

Tabela 5 – Métricas de Classificação de Reprovados com Dados Cadastrais e Diferentes Discretizações

Discretização	Algoritmo	Sensibilidade (%)	Precisão (%)	F1 (%)	Acurácia (%)
Homogênea	NB	75,0%	72,1%	73,7%	75,0%
Por Ganho	GB	80,0%	85,2%	82,5%	84,3%

Fonte: Elaborado pelo Autor (2024).

A análise dos resultados demonstra que a discretização focada em ganho de informação é mais eficiente para identificar e classificar corretamente os dados. Portanto, essa abordagem será adotada como padrão para as análises futuras, assegurando uma maior eficácia nas previsões.

Além disso, é importante destacar que os algoritmos selecionados foram aqueles com maior sensibilidade aos resultados associados à reprovação dos estudantes. Para esse mesmo caso, por exemplo, utilizando a discretização focada em ganho de informação, o modelo KNN obteve uma sensibilidade de 93,3% para os aprovados, porém apenas 58,5% para os reprovados.

Se o modelo classifica corretamente os aprovados, mas apresenta um alto número de falso positivos, muitos estudantes que necessitam de suporte não serão identificados e, consequentemente, não receberão a ajuda necessária a tempo. Portanto, um algoritmo com métricas gerais menos favoráveis, mas com alta sensibilidade para identificar os reprovados, é mais vantajoso.

Dando sequência ao fluxo, a tabela de dados foi estudada para entender a influência dos índices sobre o desempenho dos estudantes, de maneira similar às análises realizadas na Seção 4.1. Na Tabela 6 estão os valores dos índices com maior impacto na aprovação dos estudantes. Os índices restantes, apresentando uma influência menos significativa em torno de 50%, não foram incluídos, porém sua relação com o desempenho dos estudantes é discutida.

Tabela 6 – Distribuição de Aprovados e Reprovados Para as Categorias de Índices com Maior Impacto

Índices	Aprovados	Reprovados	Aprovados (%)
Não Reprovou	22	04	84,61%
idadeCursando entre 17 e 19 anos	19	06	76,00%
É de SC, mas veio de outros municípios	16	06	72,72%
IAA superior ou igual a 3.527	49	21	70,00%
anoIngresso durante 2023	36	16	69,23%
idadeIngresso superior a 22 anos	03	12	20,00%
anoIngresso anterior a 2022	03	15	16,67%
idadeCursando de ou superior a 24 anos	01	10	9,09%
idadeDif superior a 3 anos	01	13	7,14%
IAA entre 0.1 e 3.527	01	20	4,76%

Fonte: Elaborado pelo Autor (2024).

A análise longitudinal dos dados de ingresso revela que estudantes que ingressaram durante ou antes da pandemia apresentaram taxas de aprovação baixas, de 16,67%. No entanto, observou-se uma melhoria progressiva: um aumento para 52,00% no ano subsequente ao término da pandemia, culminando em uma taxa de aprovação de 69,23% para os que ingressaram em 2023.

De maneira relacional ao ano de ingresso, embora não mutuamente exclusivos, as faixas etárias também demonstram uma influência significativa nas taxas de aprovação. Estudantes entre 17 e 19 anos apresentam as maiores taxas de sucesso, com 76,00%, que diminuem gradualmente com o aumento da idade: 54,38% para aqueles entre 19 e 24 anos e apenas 9,09% para estudantes de 24 anos ou mais.

O IAA e o histórico de reprovações são outras variáveis que apresentam uma forte correlação com o desempenho dos estudantes. Estudantes com IAA acima de 3.527 e sem reprovações alcançam taxas de aprovação de 70% e 84,61%, respectivamente. Em contraste, aqueles com IAA abaixo de 3.527 ou com reprovações exibem taxas substancialmente menores, sendo de apenas 4,76% para o primeiro grupo e 43,93% para o segundo.

Finalmente, apesar de a maioria das regiões de origem apresentar taxas de aprovação em torno de 50%, os estudantes oriundos de Santa Catarina, excluindo-se Joinville, destacam-se com uma taxa de 72,72%.

Para concluir a análise deste fluxo, a matriz de confusão foi examinada em detalhe, juntamente das percentagens associadas, com o objetivo de identificar as causas subjacentes às classificações incorretas pelos algoritmos. Nos casos de falso negativos, a maioria das ocorrências envolveu valores atípicos, onde estudantes com quatro ou mais índices associados a baixas taxas de aprovação acabaram sendo aprovados. Quanto aos falso positivos, valores atípicos representaram aproximadamente

50% das ocorrências. Para o restante, não foi possível determinar um motivo claro que justificasse a classificação incorreta.

4.2.2 Fluxo de Presenças

Para o fluxo de presenças, diferente do de dados cadastrais, os índices não foram discretizados. O resultado da análise foi melhor quando mantidos em seu formato numérico. Contudo, os valores foram submetidos a um processo de normalização, de modo que a amplitude dos mesmos se estabelecesse entre os limites de 0 e 1.

Além disso, não foi necessário ranquear os índices, pois cada segmento de tempo conta com apenas dois. Todos os índices foram, portanto, utilizados simultaneamente na análise. As métricas específicas alcançadas em cada intervalo estão detalhadas na Tabela 7.

Tabela 7 – Métricas de Classificação de Reprovados com Índices de Presença por Período Letivo

Período	Algoritmo	Sensibilidade (%)	Precisão (%)	F1 (%)	Acurácia (%)
25%	NB	75,4%	65,3%	70,0%	70,0%
50%	GB	73,8%	77,4%	75,6%	77,9%
75%	KNN	75,4%	86,0%	80,3%	82,9%
95%	GB	76,9%	92,6%	84,0%	86,4%

Fonte: Elaborado pelo Autor (2024).

De forma similar aos resultados encontrados por Kaensar e Wongnin (2023), observou-se uma variação no desempenho dos algoritmos conforme o período do curso. Especificamente, o método NB se mostrou mais eficaz no primeiro período, o GB destacou-se no segundo e quarto período, e o KNN foi superior no terceiro.

Ademais, foi possível encontrar perfis para cada caso ao se analisar a matriz de confusão e os dados para o período de 50% do curso. No caso dos reprovados corretamente classificados, 97,5% possuem uma taxa de presença inferior a 60% e mais que duas faltas consecutivas. De maneira similar, 88,1% dos aprovados corretamente classificados possuem presença acima de 80% e, no máximo, uma falta.

Em seguida, foram analisados os casos de falso negativos, que apresentaram apenas valores atípicos. Esse fenômeno ocorre com estudantes que registram um alto índice de faltas no início do curso, mas que mostram recuperação nos períodos subsequentes. Para exemplificar, o estudante denominado *Aluno37* teve uma taxa de presença de 60% no segundo período e registrou quatro faltas consecutivas, sendo inicialmente classificado como reprovado. Porém, nos terceiro e quarto períodos, ele estabilizou sua taxa de presença em 77%, e o número de faltas consecutivas tornou-se menos relevante em comparação com outros estudantes. Assim, ele foi posteriormente classificado como aprovado.

No outro extremo, os casos de falso positivos também revelaram apenas valores atípicos. Nessa situação, entretanto, trata-se de estudantes que compareceram à maioria das aulas, porém acabaram reprovando devido a fatores externos que não podem ser analisados aqui, como as notas.

Em resumo, as informações encontradas nos casos de falso positivos e falso negativos também explicam por que os algoritmos demonstram uma sensibilidade muito mais alta para os aprovados, conforme indicado na Tabela 8. Um aluno com alta percentagem de presença ainda pode ser reprovado por baixas notas; no entanto, um aluno com baixa presença será reprovado por Frequência Insuficiente (FI) ou terá desistido do curso, o que deixa pouca margem para erro.

Tabela 8 – Métricas de Classificação de Aprovados com Índices de Presença por Período Letivo

Período	Algoritmo	Sensibilidade (%)	Precisão (%)	F1 (%)	Acurácia (%)
25%	LR	85,3%	71,1%	77,6%	73,6%
50%	LR	93,3%	74,5%	82,8%	79,3%
75%	LR	97,3%	76,0%	85,4%	82,1%
95%	LR	98,7%	74,0%	84,6%	80,7%

Fonte: Elaborado pelo Autor (2024).

Interessantemente, para este caso, não houve variação do algoritmo com melhor desempenho para cada período; a LR se sobressaiu em todos.

4.2.3 Integração dos Fluxos de Dados Cadastrais e de Presenças

Após a análise individual, a integração dos fluxos de dados cadastrais e de presenças foi explorada, buscando avaliar os benefícios dessa concatenação para a qualidade dos modelos preditivos. A hipótese subjacente é que essa abordagem possa aprimorar as métricas de classificação dos estudantes reprovados ao permitir a análise de interações entre variáveis que poderiam passar despercebidas quando examinadas isoladamente.

De maneira similar aos fluxos anteriores, os dados cadastrais foram discretizados, enquanto os índices de presença foram normalizados. Após a preparação, esses dados foram combinados e ranqueados de acordo com as oito variáveis de maior relevância. A Tabela 9 apresenta as métricas de classificação obtidas.

O primeiro período apresenta métricas similares às encontradas pelo fluxo de dados cadastrais, indicando pouca influência das presenças. Entretanto, a partir do segundo período, as métricas começaram a melhorar significativamente. Quando comparadas com as métricas do fluxo de presenças para 50%, 75% e 95% dos dados, observou-se um aumento de 7,7%, 6,1% e 7,7% em sensibilidade e de 12,4%, 5,4%

Tabela 9 – Métricas de Classificação de Reprovados com Índices Cadastrais e de Presença por Período Letivo

Período	Algoritmo	Sensibilidade (%)	Precisão (%)	F1 (%)	Acurácia (%)
25%	RNA	80,0%	83,9%	81,9%	83,6%
50%	NB	81,5%	89,8%	85,5%	87,1%
75%	RF	81,5%	91,4%	86,2%	87,9%
95%	RF	84,6%	94,8%	89,4%	90,7%

Fonte: Elaborado pelo Autor (2024).

e 2,2% em precisão, respectivamente. Além disso, a mesclagem dos dados alterou o algoritmo de melhor desempenho para cada segmento.

Para a tabela de dados e matriz de confusão, os padrões e perfis relacionados aos índices encontrados em cada fluxo se mantiveram consistentes. No entanto, a mesclagem dos dados melhorou a classificação dos valores atípicos que anteriormente eram problemáticos. Retomando o caso do *Aluno37*, que estava sendo erroneamente classificado no segundo período devido ao número de faltas consecutivas, a inclusão dos índices cadastrais permitiu que fosse corretamente classificado.

Por fim, o fato de ambos os conjuntos de dados já apresentarem melhores resultados na classificação dos estudantes aprovados foi ainda mais intensificado com a mesclagem dos mesmos, conforme demonstrado na Tabela 10.

Tabela 10 – Métricas de Classificação de Aprovados com Índices Cadastrais e de Presença por Período Letivo

Período	Algoritmo	Sensibilidade (%)	Precisão (%)	F1 (%)	Acurácia (%)
25%	KNN	92,0%	75,0%	82,6%	79,3%
50%	KNN	98,7%	76,3%	86,0%	82,9%
75%	LR	98,7%	84,1%	90,8%	89,3%
95%	LR	98,7%	85,2%	92,1%	90,1%

Fonte: Elaborado pelo Autor (2024).

4.2.4 Fluxo de notas

O último conjunto de dados analisado é o de notas. De maneira similar ao fluxo de presenças, os índices foram normalizados entre os limites de 0 e 1 e não houve necessidade de ranqueamento. As métricas de desempenho foram classificadas por segmento de tempo e são detalhadas na Tabela 11.

Em contraste com os outros fluxos, nos quais as métricas se destacaram na classificação dos aprovados, as sensibilidades alcançadas pelo fluxo de notas foram similares tanto para os casos de reprovados quanto de aprovados. A LR manteve-se

Tabela 11 – Métricas de Classificação de Reprovados com Índices de Nota por Período Letivo

Período	Algoritmo	Sensibilidade (%)	Precisão (%)	F1 (%)	Acurácia (%)
25%	NB	89,2%	85,3%	87,2%	87,9%
50%	RF	92,3%	80,0%	85,7%	85,7%
75%	SVM	93,8%	93,8%	93,8%	94,3%
95%	RNA	98,5%	95,5%	97,0%	97,1%

Fonte: Elaborado pelo Autor (2024).

como o melhor algoritmo para a classificação dos aprovados em todos os períodos, similar aos resultados obtidos pelo fluxo de presenças.

Além disso, quando comparadas com os outros fluxos, as métricas para o fluxo de notas são superiores, pois refletem diretamente o desempenho acadêmico do estudante, estabelecendo uma relação causal com os critérios de aprovação. As notas são indicadores consistentes do progresso do estudante e são menos suscetíveis a variações externas em comparação com presenças ou dados sociodemográficos, que são fatores mais indiretos.

Dando sequência ao fluxo, a análise da tabela de dados para os estudantes na metade do curso revelou pontos de corte específicos para a classificação de aprovados e reprovados. Entre os reprovados, 94,58% possuíam dois ou mais LVP incompletos, com média igual ou inferior a 4. Por outro lado, 94,11% dos aprovados completaram todos os seus LVP, com média superior a 9. Quanto aos questionários, ambos os grupos apresentaram em média cerca de um questionário incompleto; as médias foram em torno de 6 para os reprovados e superiores a 8 para os aprovados.

Na análise dos falso negativos, todos os casos envolveram estudantes que obtiveram médias boas em um formato de atividade, seja LVP ou questionário, mas apresentaram desempenho inferior no outro, com valores abaixo de 6. Para os falso positivos, a maioria dos casos foi de valores atípicos, envolvendo estudantes que completaram todas as atividades e alcançaram médias boas em ambas, mas foram reprovados. Esses casos são explicados ao expandir a análise para o quarto período, onde os mesmos estudantes obtiveram notas ruins na avaliação final, que compõe 60% da nota do curso. Por exemplo, os estudantes denominados *Aluno44*, *Aluno66* e *Aluno83* tiveram médias superiores a 9 nos LVP e questionários, porém inferiores a 1 na avaliação final.

4.2.5 Fluxo Final: Integração de Todos os Dados

De maneira similar ao que foi feito na Subseção 4.2.3, o último fluxo criado, ilustrado na Figura 18, resultou da integração dos três conjuntos de dados.

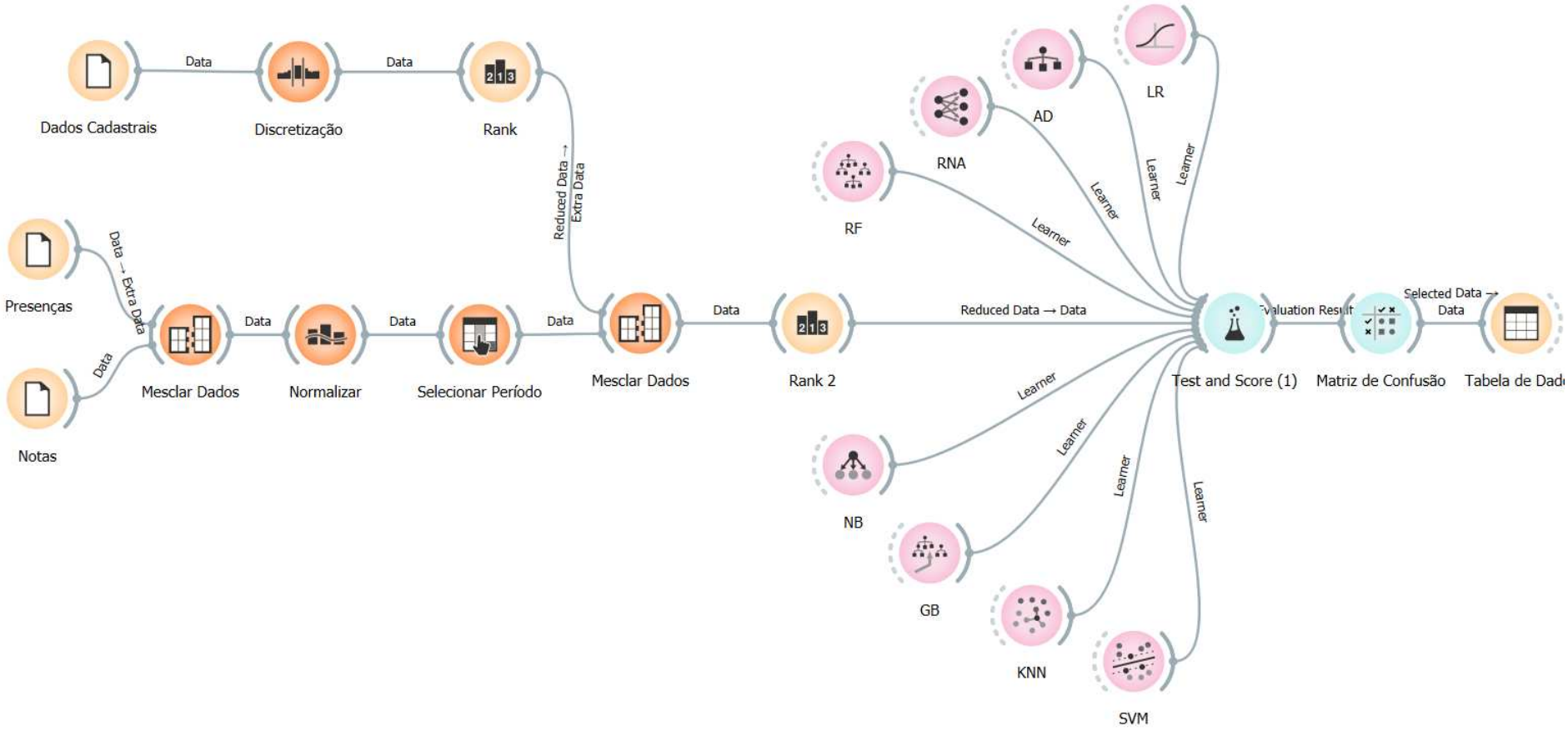


Figura 18 – Fluxo de Dados Integrados

Fonte: Elaborado pelo Autor (2024).

Neste caso, o pré-processamento de cada tipo de variável foi realizado conforme os respectivos fluxos individuais. No entanto, o ranqueamento foi ajustado para incluir os nove melhores índices, em vez dos oito anteriormente considerados. As métricas resultantes dessa abordagem estão apresentadas na Tabela 12.

Tabela 12 – Métricas de Classificação de Reprovados por Período Letivo Através da Integração dos Dados

Período	Algoritmo	Sensibilidade (%)	Precisão (%)	F1 (%)	Acurácia (%)
25%	LR	90,8%	92,2%	91,5%	92,1%
50%	SVM	92,3%	90,9%	91,6%	92,1%
75%	RNA	95,4%	92,5%	93,9%	94,3%
95%	AD	100,0%	92,9%	96,3%	96,4%

Fonte: Elaborado pelo Autor (2024).

Os resultados, quando comparados com os modelos que utilizaram os dados isoladamente, revelam que a concatenação proporcionou uma melhoria notável na sensibilidade e, principalmente, nas demais métricas de desempenho. A abordagem integrada, ao capturar a inter-relação das variáveis, apresenta-se como uma interpretação superior para a predição.

Essa melhora fica mais evidente ao se analisar a Tabela 13, onde a sensibilidade de cada conjunto de dados é comparada. Por si só, a sensibilidade obtida considerando apenas o índice de notas é superior aos outros formatos de dados. Entretanto, ao serem integrados, houve uma melhora de 1,6%, 0%, 1,6% e 1,5% para 25%, 50%, 75% e 95% do período letivo.

Tabela 13 – Métricas de Sensibilidade de Reprovados por Período Letivo para Cada Conjunto de Dados

Período	Integrado	Notas	Presença & Cadastrais	Presença	Cadastrais
25%	90,8%	89,2%	80,0%	75,4%	80,0%
50%	92,3%	92,3%	81,5%	73,8%	-
75%	95,4%	93,8%	81,5%	75,4%	-
95%	100,0%	98,5%	84,6%	76,9%	-

Fonte: Elaborado pelo Autor (2024).

Finalmente, ao se considerar a aplicação dos modelos em um cenário real, o uso de diferentes algoritmos para períodos distintos pode ser problemático. Para exemplificar, um professor pode querer prever as probabilidades de reprovação dos seus estudantes após 33% do curso ter decorrido. Nesse contexto, torna-se difícil determinar qual algoritmo seria o mais adequado, já que não é possível identificar com precisão em que momento o SVM supera a LR em desempenho. Para resolver isto,

a Tabela 14 apresenta a sensibilidade média dos melhores modelos, considerando os três primeiros períodos do curso.

Tabela 14 – Média de Sensibilidade para Cada Modelo Classificativo Considerando 75% do Período Letivo

	LR	SVM	RNA	AD
Sensibilidade Média (%)	90,26%	90,76%	91,13%	85,14%

Fonte: Elaborado pelo Autor (2024).

O algoritmo de RNA se mostrou superior aos outros três, com um diferença de +0,37% quando comparado com o segundo melhor modelo. Senso assim, a fim de atender às características gerais que podem surgir da análise de diferentes turmas, propõe-se a criação de um fluxo genérico com seleção dos nove melhores atributos por proporção de ganho, método de validação por amostragem aleatória estratificada e classificação por RNA.

5 CONCLUSÃO

A retenção estudantil permanece como um desafio para instituições educacionais, onde o baixo desempenho acadêmico e as taxas elevadas de reprovação emergem como fatores críticos que contribuem para o abandono escolar. Intervenções precoces, tais como tutoria personalizada, aconselhamento acadêmico e suporte psicológico, são estratégias proativas eficazes na melhoria do desempenho dos estudantes.

Neste contexto, o presente trabalho explorou o potencial das técnicas de mineração de dados e aprendizado de máquina na predição do sucesso de estudantes na disciplina de Programação I. A motivação por trás desta investigação reside na possibilidade de identificar estudantes em risco de reprovação durante as fases iniciais da disciplina, permitindo a realização de intervenções especializadas de forma oportuna.

Para isso, utilizando relatórios cadastrais, de presença e notas do sistema acadêmico e do Moodle, aplicou-se uma metodologia quantitativa para desenvolver e validar modelos preditivos. Considerando que a identificação precisa de instâncias verdadeiro negativas (reprovados) é crítica, optou-se pela métrica de sensibilidade como principal critério de avaliação.

O modelo com RNA demonstrou melhores resultados, alcançando uma sensibilidade média de 91,13% na identificação de alunos em risco de reprovação nos períodos iniciais do semestre letivo. Este desempenho não só evidencia a adequação das técnicas utilizadas como também a qualidade dos dados empregados no modelo, confirmando a viabilidade de aplicação prática das predições para intervenções.

Ademais, ressalta-se a necessidade contínua de explorar e integrar novas fontes de dados e métodos analíticos para aprimorar tanto as capacidades preditivas dos modelos quanto o entendimento das dinâmicas educacionais presentes na instituição. As perspectivas para pesquisas futuras incluem a expansão do modelo para outras disciplinas e a inclusão de variáveis psicossociais e socioeconômicas, buscando uma análise mais abrangente.

Em conclusão, este trabalho alcançou seus objetivos propostos e estabelece uma fundação para futuras investigações que poderão contribuir para o sucesso dos estudantes e para a eficácia das práticas educacionais.

6 PERSPECTIVAS FUTURAS

Ao abordar a complexidade inerente à predição do risco de reprovação utilizando modelos baseados em aprendizagem de máquina, alguns desafios técnicos e conceituais foram encontrados. A compreensão desses obstáculos é útil para a evolução contínua das pesquisas na área e para o aprimoramento das técnicas empregadas. Este capítulo visa discutir as principais dificuldades encontradas durante o desenvolvimento do projeto e sugerir possíveis caminhos que futuras investigações poderiam explorar. Desta forma, espera-se contribuir não apenas para o avanço do conhecimento científico, mas também para a melhoria prática na aplicação de modelos preditivos em contextos educacionais.

6.1 DADOS ACADÊMICOS

O IAA considerado para este estudo reflete exclusivamente o desempenho acadêmico dos estudantes durante o semestre específico de coleta dos dados. Esta limitação restringe a possibilidade de realizar análises longitudinais, o que impede a compreensão das variações e tendências do desempenho acadêmico ao longo do tempo.

No sistema de controle acadêmico da UFSC, já é possível acessar informações sobre o Índice de Aproveitamento (IA) semestral e a evolução do IAA ao longo do curso. Entretanto, a elaboração de um relatório detalhado, nos moldes do relatório de dados cadastrais que foi analisado, ainda não foi implementada. A criação de tal documento permitira identificar períodos críticos nos quais os estudantes enfrentam maiores dificuldades e enriqueceria a capacidade preditiva de modelos futuros.

6.2 DADOS SOCIODEMOGRÁFICOS

Como mencionado na Subseção 4.1.2, a ausência de dados sobre alunos oriundos de escolas públicas a partir de 2021 constitui uma limitação para a análise e a compreensão da diversidade do corpo discente. A verificação e a validação da precisão na coleta desses dados são importantes para a integridade das análises e para permitir uma expansão das investigações exploratórias.

Além disso, as análises exploratórias realizadas focaram em variáveis limitadas, como gênero, cor/raça e IAA. Contudo, a inclusão de dados socioeconômicos adicionais, como a renda familiar dos estudantes e se eles possuem atividades laborais concomitantes aos estudos expandiria o estudo. Esses dados poderiam oferecer percepções sobre as interações entre o status socioeconômico e o desempenho acadêmico, facilitando a identificação de padrões mais subtis.

REFERÊNCIAS

ALLOGHANI, M.; AL-JUMEILY, D.; MUSTAFINA, J.; HUSSAIN, A.; ALJAAF, A. J. A systematic review on supervised and unsupervised machine learning algorithms for data science. **Supervised and unsupervised learning for data science**, 2020. Disponível em: https://nces.ed.gov/fastfacts/display.asp?id=16. Acesso em: 21 out. 2023.

ALMARABEH, H. Analysis of students' performance by using different data mining classifiers. **International Journal of Modern Education and Computer Science**, v. 9, n. 8, p. 9–15, 2017. Disponível em: https://doi.org/10.5815/ijmecs.2017.08.02.

BAKER, R.; YACEF, K. The state of educational data mining in 2009: A review and future visions. **Journal of Educational Data Mining**, v. 1, n. 1, p. 3–17, 2017. Disponível em: https://doi.org/10.5281/zenodo.3554657.

BOYLAN-ASHRAF, P. C.; HAUGHERY, J. R. Failure rates in engineering: Does it have to do with class size? *In*: ASEE ANNUAL CONFERENCE & EXPOSITION, Salt Lake City. 2018. Disponível em: https://peer.asee.org/failure-rates-in-engineering-does-it-have-to-do-with-class-size. Acesso em: 21 out. 2023.

BURKART, N.; HUBER, M. F. A survey on the explainability of supervised machine learning. **Journal of Artificial Intelligence Research**, v. 70, p. 245–317, 2021.

CHEEWAPRAKOBKIT, P. Study of factor analysis affecting achievements of undergraduate. *In*: INTERNATIONAL MULTI CONFERENCE OF ENGINEERS AND COMPUTER SCIENTISTS, Hong Kong. 2013. Disponível em: https://www.iaeng.org/publication/IMECS2013/IMECS2013_pp332-336.pdf. Acesso em: 26 set. 2023.

CORTEZ, P.; SILVA, A. Using data mining to predict secondary school student performance. **5th Annual Future Business Technology Conference**, Porto, Portugal, 2010. Disponível em: http://www3.dsi.uminho.pt/pcortez/student.pdf. Acesso em: 02 set. 2023.

S. DALFOVO, B. de. **Estudo de Caso com Análise de Dados para a Detecção da Desistência de Estudantes em Disciplinas Ofertadas com Apoio do Ambiente Moodle**. 2023. Trabalho de Conclusão de Curso (Graduação em Engenharia Mecatrônica) — Centro Tecnológico de Joinville, Universidade Federal de Santa Catarina, Joinville, 2023.

DHANKAR, A.; GUPTA, N. A systematic review of techniques, tools and applications of machine learning. *In*: THIRD INTERNATIONAL CONFERENCE ON INTELLIGENT COMMUNICATION TECHNOLOGIES AND VIRTUAL MOBILE NETWORKS (ICICV), India. 2021. p. 764–768. Disponível em: https://ieeexplore.ieee.org/document/9388637. Acesso em: 23 nov. 2023.

DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and unsupervised discretization of continuous features. **ICML**, v. 1, set. 1997. Disponível em: https://doi.org/10.1016/B978-1-55860-377-6.50032-3.

DREYFUS, S. Artificial neural networks, back propagation, and the kelley-bryson gradient procedure. **Journal of Guidance Control and Dynamics**, v. 13, p. 926–928, 1990. Disponível em: https://doi.org/10.2514/3.25422.

FAYYAD, U. M.; IRANI, K. B. Multi-interval discretization of continuous-valued attributes for classification learning. *In*: **International Joint Conference on Artificial Intelligence**. [s.n.], 1993. Disponível em: https://api.semanticscholar.org/CorpusID:18718011. Acesso em: 07 jun. 2024.

GARCÍA, S.; LUENGO, J.; HERRERA, F. **Data Preprocessing in Data Mining**. Springer International Publishing, 2014. (Intelligent Systems Reference Library). ISBN 9783319102474. Disponível em: https://books.google.com.br/books?id=SbFkBAAAQBAJ. Acesso em: 19 mai. 2024.

GONZÁLEZ, M.; RUÍZ, M.; ORTIN, F. Massive lms log data analysis for the early prediction of course-agnostic student performance. **Computers Education**, v. 163, p. 104–108, 2021. Disponível em: https://doi.org/10.1016/j.compedu.2020.104108.

KAENSAR, C.; WONGNIN, W. Analysis and prediction of student performance based on moodle log data using machine learning techniques. **Journal of Emerging Technologies in Learning**, v. 18, n. 10, p. 184–203, 2023. Disponível em: https://doi.org/10.3991/ijet.v18i10.35841.

KAZIL, J.; JARMUL, K. **Data Wrangling with Python: Tips and Tools to Make Your Life Easier**. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2016. ISBN 1491948817.

KOVAĈIĆ, Z. J. Early prediction of student success: Mining students enrolment data. **Informing Science and IT Education**, jun 2010. Disponível em: https://api.semanticscholar.org/CorpusID:14191196. Acesso em: 26 set. 2023.

LIU, H.; HUSSAIN, F.; TAN, C. L.; DASH, M. Discretization: An enabling technique. **Data Min. Knowl. Discov.**, v. 6, p. 393–423, out. 2002. Disponível em: https://doi.org/10.1023/A:1016304305535.

LOMBA, A. G. **Predição de Reprovação em Turmas da Disciplina de Programação I na Universidade Federal de Santa Catarina - Campus Joinville**. 2023. Trabalho de Conclusão de Curso (Graduação em Engenharia Mecatrônica) — Centro Tecnológico de Joinville, Universidade Federal de Santa Catarina, Joinville, 2023.

MCLAREN, B. M.; SHEUER, O. Educational data mining. **Encyclopedia of the Sciences of Learning**, 2011. Disponível em: https://link.springer.com/referenceworkentry/10.1007/978-1-4419-1428-6_618. Acesso em: 14 nov. 2023.

MINSKY, M.; PAPERT, S. **Perceptrons**. Cambridge, MA, USA: MIT Press, 1969.

NATIONAL CENTER FOR EDUCATION STATISTICS. **College Dropout Trends in the United States: An Analysis of NCES Data**. U.S., 2021. Disponível em: https://nces.ed.gov/fastfacts/display.asp?id=16. Acesso em: 17 nov. 2023.

OSISANWO, F. *et al.* Supervised machine learning algorithms: classification and comparison. **Journal of Computer Trends and Technology (JCTT)**, v. 48, n. 3, p. 128–138, 2017.

RAMOS, V.; WAZLAWICK, R.; GALIMBERTI, M.; FREITAS, M.; MARIANI, A. C. A comparação da realidade mundial do ensino de programação para iniciantes com a realidade nacional: Revisão sistemática da literatura em eventos brasileiros. **Simpósio Brasileiro de Informática na Educação**, v. 26, n. 1, p. 318, 2015. Disponível em: https://doi.org/10.5753/cbie.sbie.2015.318.

ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. **Expert Systems with Applications**, v. 33, p. 135–146, 2007. Disponível em: https://doi.org/10.1016/j.eswa.2006.04.005.

ROMERO, C.; VENTURA, S. Educational data mining: A review of the state of the art. **Systems, Man, and Cybernetics**, v. 40, p. 601–618, 2010. Disponível em: https://doi.org/10.1109/TSMCC.2010.2053532.

SEMERARO, F.; GRIFFITHS, A.; CANGELOSI, A. Human–robot collaboration and machine learning: A systematic review of recent research. **Robotics and Computer-Integrated Manufacturing**, v. 79, p. 102–432, 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S0736584522001156. Acesso em: 17 nov. 2023.

SHAWE-TAYLOR, J. Technical perspective: Machine learning for complex predictions. **Communications of the ACM**, v. 52, p. 96–96, 2009. Disponível em: https://doi.org/10.1145/1592761.1592782.

TAHERDOOST, H. Sampling methods in research methodology; how to choose a sampling technique for research. **International Journal of Academic Research in Management**, v. 5, p. 18–27, 01 2016. Disponível em: https://doi.org/10.2139/ssrn.3205035.

THORNTON, M. *et al.* Diversity and social integration on higher education campuses in india and the uk: Student and staff perspectives. **Research in Post-Compulsory Education**, v. 15, p. 159–176, 06 2010. Disponível em: https://doi.org/10.1080/13596741003790682.

UNIVERSIDADE FEDERAL DE SANTA CATARINA. **O que é IA, IAA, IAP e IM?** Florianópolis, 2024. Disponível em: https://cenq.paginas.ufsc.br/o-que-e-ia-iaa-iap-e-im-e-como-se-calcula/. Acesso em: 06 abr. 2024.

WALKER, J. S. **Machine Learning for Beginners**. North Charleston, SC, USA: CreateSpace Independent Publishing Platform, 2018.

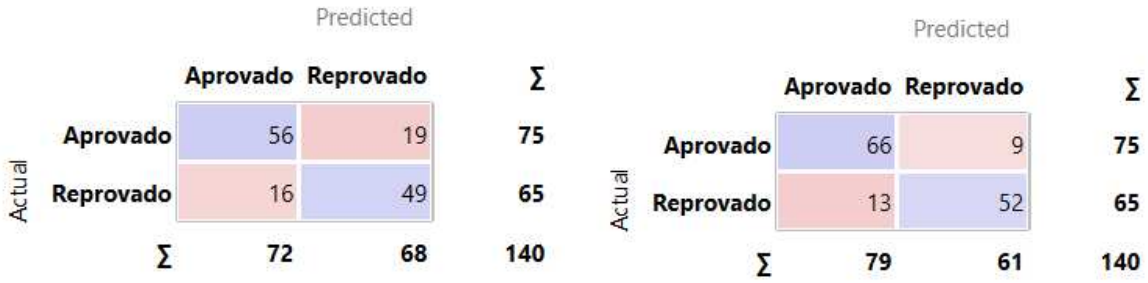
WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2016. ISBN 0123748569.

ZHENG, A.; CASARI, A. **Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists**. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2018. ISBN 1491953241.

APÊNDICE A – MATRIZES DE CONFUSÃO PARA ANÁLISES PREDITIVAS

As Figuras 1, 2, 3, 4 e 5 apresentam as matrizes de confusão obtidas e utilizadas na análise preditiva de cada fluxo.

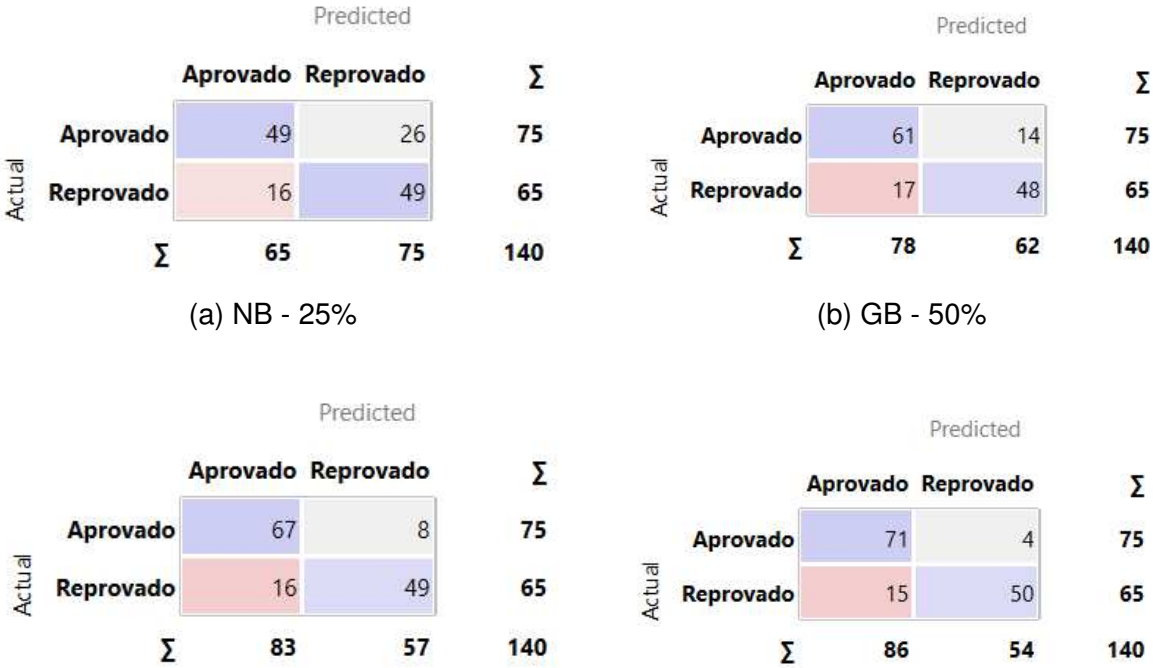
Figura 1 – Matrizes de Confusão dos Melhores Algoritmos por Método de Discretização para o Fluxo Cadastral



(a) NB - Homogênea (b) GB - Por Ganho

Fonte: Elaborado pelo Autor (2024).

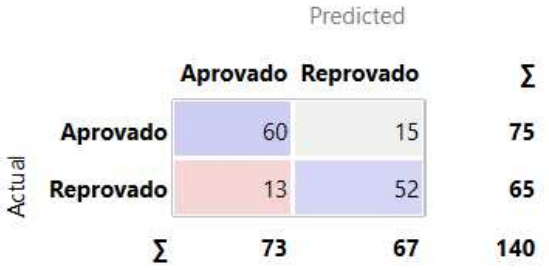
Figura 2 – Matrizes de Confusão dos Melhores Algoritmos por Período Letivo para o Fluxo de Presenças



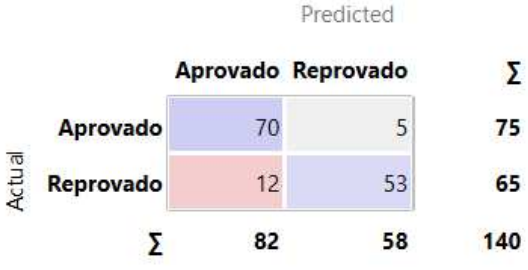
(c) KNN - 75% (d) GB - 95%

Fonte: Elaborado pelo Autor (2024).

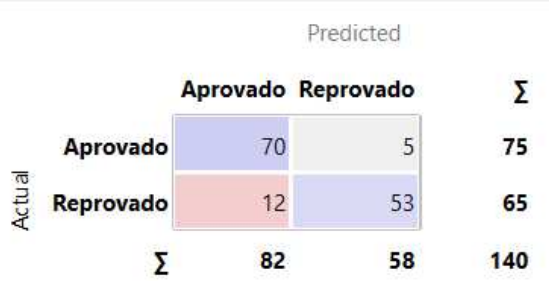
Figura 3 – Matrizes de Confusão dos Melhores Algoritmos por Período Letivo para o Fluxo Concatenado de Índices Cadastrais e de Presenças



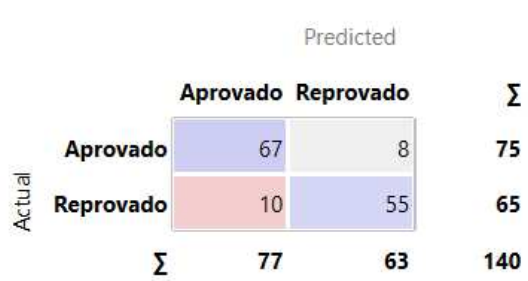
(a) RNA - 25%



(b) NB - 50%



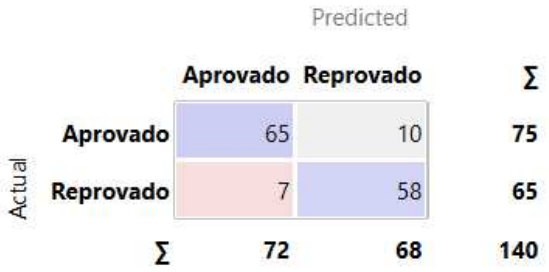
(c) RF - 75%



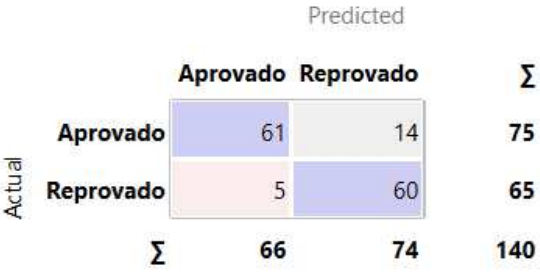
(d) RF - 95%

Fonte: Elaborado pelo Autor (2024).

Figura 4 – Matrizes de Confusão dos Melhores Algoritmos por Período Letivo para o Fluxo de Notas

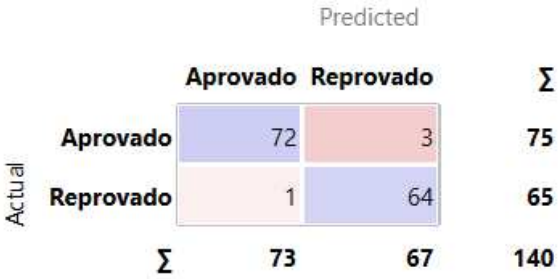


(a) NB - 25%



(b) RF - 50%

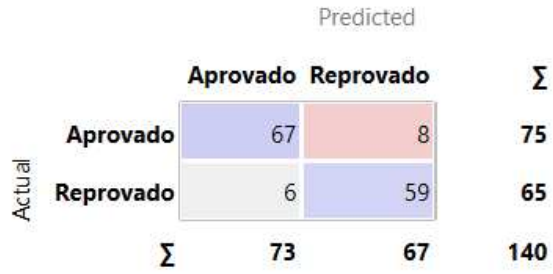
(c) SVM - 75%



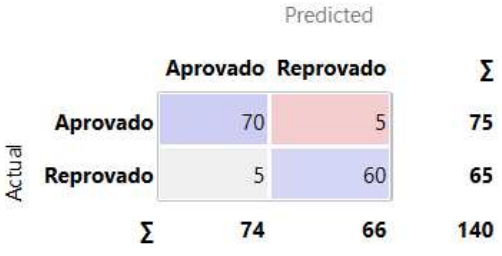
(d) RNA - 95%

Fonte: Elaborado pelo Autor (2024).

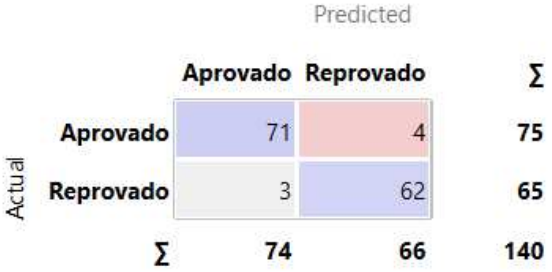
Figura 5 – Matrizes de Confusão dos Melhores Algoritmos por Período Letivo para o Fluxo Integrado



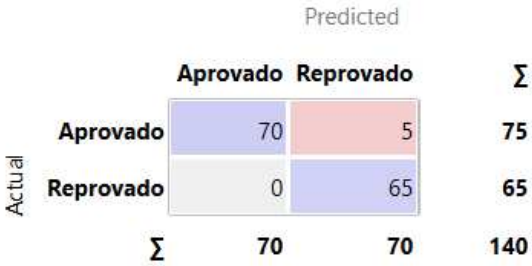
(a) LR - 25%



(b) SVM - 50%



(c) RNA - 75%



(d) AD - 95%

Fonte: Elaborado pelo Autor (2024).