

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

ANDRÉ GOMES LOMBA JÚNIOR

PREDIÇÃO DE REPROVAÇÃO EM TURMAS DA DISCIPLINA DE PROGRAMAÇÃO I  
NA UNIVERSIDADE FEDERAL DE SANTA CATARINA - CAMPUS JOINVILLE

Joinville  
2023

ANDRÉ GOMES LOMBA JÚNIOR

PREDIÇÃO DE REPROVAÇÃO EM TURMAS DA DISCIPLINA DE PROGRAMAÇÃO I  
NA UNIVERSIDADE FEDERAL DE SANTA CATARINA - CAMPUS JOINVILLE

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

Orientador: Dr. Benjamin Grando Moreira

Joinville  
2023

Dedico este trabalho à minha família

## **AGRADECIMENTOS**

Agradeço a Deus por tornar isso possível.

Agradeço à minha família por todo o apoio e suporte oferecido.

Agradeço ao meu orientador por acreditar e me ajudar nesse desafio.

Agradeço aos meus amigos por tornar esta jornada mais leve e divertida.

Cavalheiros, vocês precisam colocar a placa de blindagem onde os furos das balas não estão, porque é onde os furos ocorreram nos aviões que não retornaram (Abraham Wald - 1942).

## RESUMO

O presente trabalho tem como objetivo estabelecer relações entre as notas obtidas em atividades avaliativas e o desempenho final dos alunos matriculados na disciplina de Programação I da Universidade Federal de Santa Catarina, Campus de Joinville. A análise dessas relações será conduzida por meio de uma abordagem sistemática, envolvendo a coleta e mineração de dados provenientes dos relatórios disponíveis no Moodle. A mineração de dados proporcionou a identificação de padrões, por meio dos quais foram gerados novos conhecimentos sobre as diferentes trajetórias acadêmicas dos alunos, visando uma compreensão das relações de desempenho ao longo da disciplina, considerando variáveis como atividades práticas, questionários e frequência de acesso. Este estudo busca não apenas compreender o cenário atual, mas contribuir para uma abordagem informada no processo educacional, identificando padrões e tendências para fornecer subsídios à implementação de estratégias pedagógicas eficazes.

**Palavras-chave:** mineração de dados educacionais; reprovação acadêmica; classificação.

## **ABSTRACT**

The present work aims to establish relationships between the grades obtained in evaluative activities and the final performance of students enrolled in the Programming I course at the Federal University of Santa Catarina, Joinville Campus. The analysis of these relationships will be conducted through a systematic approach, involving the collection and mining of data from the available reports in Moodle. Data mining has provided the identification of patterns, through which new knowledge has been generated about the different academic trajectories of students, aiming to understand performance relationships throughout the course, considering variables such as practical activities, quizzes, and access frequency. This study seeks not only to understand the current scenario but also to contribute to an informed approach in the educational process, identifying patterns and trends to provide support for the implementation of effective pedagogical strategies.

**Keywords:** educational data mining; academic failure; classification.

## LISTA DE FIGURAS

Figura 1 – Processo de mineração de dados . . . . .	15
Figura 2 – Ciclo dos dados educacionais . . . . .	15
Figura 3 – Ciclo da Análise de Aprendizagem . . . . .	16
Figura 4 – Diferenças entre coleta ativa e passiva de dados . . . . .	17
Figura 5 – Matriz de Confusão . . . . .	20
Figura 6 – Tabela de atributos . . . . .	22
Figura 7 – Agrupamento obtido . . . . .	23
Figura 8 – Impacto das ferramentas analisadas . . . . .	24
Figura 9 – Gráfico correlacionando número de acessos e notas obtidas . . . . .	25
Figura 10 – Número de acessos por semana . . . . .	25
Figura 11 – Ambiente do Moodle na disciplina de Programação I . . . . .	27
Figura 12 – Exemplo de tópicos de assuntos na disciplina de Programação I . . . . .	27
Figura 13 – Desempenho geral da turma base na disciplina . . . . .	29
Figura 14 – Exemplo de tabela com as variáveis utilizadas e a variável alvo . . . . .	29
Figura 15 – Correlação das variáveis com a nota final . . . . .	30
Figura 16 – Relatório de presença . . . . .	31
Figura 17 – Matriz de confusão com base na presença . . . . .	31
Figura 18 – VPLs tratados individualmente . . . . .	32
Figura 19 – Análise da média dos VPLs . . . . .	33
Figura 20 – Questionários tratados individualmente . . . . .	34
Figura 21 – Considerando a média dos Questionários . . . . .	34
Figura 22 – Modelagem da análise por assuntos no Orange . . . . .	35
Figura 23 – Matriz de confusão dos Assuntos Iniciais . . . . .	36
Figura 24 – Matriz de confusão das Estruturas de Repetição . . . . .	36
Figura 25 – Matriz de confusão dos Vetores . . . . .	37
Figura 26 – Matriz de confusão das Matrizes . . . . .	37
Figura 27 – Matriz de confusão dos Ponteiros e Strings . . . . .	38
Figura 28 – Matriz de confusão das Funções . . . . .	38
Figura 29 – Probabilidades de aprovação geradas pela simulação . . . . .	40
Figura 30 – Modelagem com RNA no Orange . . . . .	42
Figura 31 – Modelagem da turma com outra abordagem . . . . .	43
Figura 32 – Desempenho na Turma - Métricas da Prova 1 . . . . .	44
Figura 33 – Desempenho na Turma - Métricas dos Questionários 1 e 2 . . . . .	44
Figura 34 – Desempenho na Turma - Métricas dos Questionários 1 e 2 + Prova 1 . . . . .	45
Figura 35 – Probabilidade de aprovação na turma com outra abordagem . . . . .	46



## LISTA DE TABELAS

Tabela 1 – Tabela Comparativa - Valores das métricas de presença . . . . .	32
Tabela 2 – Tabela Comparativa - VPLs Individuais vs. Média dos VPLs . . . . .	33
Tabela 3 – Tabela Comparativa - Questionários . . . . .	35
Tabela 4 – Tabela Comparativa - Desempenho por Assunto . . . . .	39
Tabela 5 – Tabela Comparativa - Métricas dos Questionários na segunda turma	41
Tabela 6 – Tabela Comparativa - Métricas dos VPLs na segunda turma . . . . .	41
Tabela 7 – Tabela Comparativa - Métricas do Desempenho por Assunto na segunda turma . . . . .	41
Tabela 8 – Tabela Comparativa - Métricas dos Questionários utilizando RNAs .	42
Tabela 9 – Tabela Comparativa - Métricas dos VPLs utilizando RNAs . . . . .	42
Tabela 10 – Tabela Comparativa - Métricas de Desempenho por Assunto usando RNAs . . . . .	43
Tabela 11 – Tabela Comparativa - Métricas de desempenho em turma com abordagem diferente . . . . .	45
Tabela 12 – Tabela Comparativa - Desempenho em outra turma com abordagem diferente . . . . .	46

## LISTA DE ABREVIATURAS E SIGLAS

<i>ARFF</i>	Attribute-Relation File Format
<i>CRISP – DM</i>	Cross-Industry Standard Process for Data Mining
<i>FN</i>	Falso negativo
<i>FP</i>	Falso positivo
<i>RNA</i>	Redes Neurais Artificiais
<i>SQL</i>	Structured Query Language
<i>UFSC</i>	Universidade Federal de Santa Catarina
<i>VPL</i>	Virtual Programming Laboratory
<i>VP</i>	Verdadeiro positivo
<i>VN</i>	Verdadeiro negativo

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>12</b>
1.1	Objetivo	13
<b>1.1.1</b>	<b>Objetivo Geral</b>	<b>13</b>
<b>1.1.2</b>	<b>Objetivos Específicos</b>	<b>13</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>14</b>
2.1	Mineração de Dados (Data Mining)	14
2.2	Mineração de Dados Educacionais	15
2.3	Learning Analytics	16
2.4	Coleta de Dados	16
2.5	Aprendizagem Supervisionada e Não Supervisionada	17
2.6	Clustering	18
2.7	Algoritmos para mineração de dados utilizados	18
<b>2.7.1</b>	<b>Árvores de Decisão</b>	<b>18</b>
<b>2.7.2</b>	<b>Redes Neurais Artificiais</b>	<b>19</b>
2.8	Matriz de Confusão	19
2.9	Orange	20
<b>3</b>	<b>TRABALHOS SIMILARES</b>	<b>22</b>
3.1	Estudo de caso de mineração de dados no Moodle	22
3.2	Mineração de dados educacionais no Moodle	23
3.3	Análise de Sistemas de Aprendizagem Remotos através de EDM	24
3.4	Desempenho do estudante por relatórios do Moodle	24
<b>4</b>	<b>METODOLOGIA</b>	<b>26</b>
4.1	Apresentação Inicial do Ambiente para Mineração de Dados	26
4.2	Coleta de Dados do Moodle	28
4.3	Preparação de Dados, simulações e análises	28
<b>4.3.1</b>	<b>Análise das Relações entre as Variáveis</b>	<b>29</b>
<b>4.3.2</b>	<b>Preparação e simulação de dados relacionados à presença</b>	<b>30</b>
<b>4.3.3</b>	<b>Simulações por VPLs</b>	<b>32</b>
4.3.3.1	VPLs Individuais	32
4.3.3.2	Média dos VPLs	33
4.3.3.3	Tabela Comparativa	33
<b>4.3.4</b>	<b>Simulações por Questionários</b>	<b>33</b>
4.3.4.1	Questionários Individuais	34
4.3.4.2	Média dos Questionários	34

4.3.4.3	Tabela Comparativa - Questionários Individuais vs Média dos Questionários	35
<b>4.3.5</b>	<b>Simulações por Assuntos</b>	<b>35</b>
4.3.5.1	Fundamentos Básicos	36
4.3.5.2	Estruturas de Repetição	36
4.3.5.3	Vetores	37
4.3.5.4	Matrizes	37
4.3.5.5	Strings e Ponteiros	38
4.3.5.6	Funções	38
4.3.5.7	Tabela Comparativa - Desempenho por Assunto	39
<b>4.3.6</b>	<b>Identificação de Falsos Negativos e Falsos Positivos e probabilidades de aprovação</b>	<b>39</b>
<b>4.3.7</b>	<b>Comparação com outra turma</b>	<b>41</b>
<b>4.3.8</b>	<b>Uso de Redes Neurais</b>	<b>41</b>
<b>4.3.9</b>	<b>Simulações em Turmas de Outro Professor(a)</b>	<b>43</b>
4.3.9.1	Tabela Comparativa	45
4.3.9.2	Probabilidade de aprovação	45
4.3.9.3	Utilizando uma segunda turma como teste	46
<b>5</b>	<b>RESULTADOS E DISCUSSÕES</b>	<b>48</b>
5.1	Análise das Variáveis	48
5.2	Análise de Presença	48
5.3	Análise dos VPLs	49
5.4	Análise de Questionários	49
5.5	Análise por Assunto	49
5.6	Identificação de Falsos Negativos e Falsos Positivos e probabilidade de aprovação	50
5.7	Análise em outra turma	50
5.8	Comparação da simulação de RNAs com árvores	51
5.9	Análises na turma de outro professor	51
5.10	Discussão geral dos resultados	52
<b>6</b>	<b>CONCLUSÕES</b>	<b>53</b>
	<b>REFERÊNCIAS</b>	<b>54</b>

## 1 INTRODUÇÃO

A educação é um dos pilares do desenvolvimento de um país, e nessa área há uma gama de universidades no Brasil, com os mais variados cursos; entretanto, são enfrentados desafios significativos no que diz respeito à reprovação de alunos em disciplinas específicas, principalmente em cursos de Engenharia e afins. Nesse contexto, surge a necessidade de abordagens inovadoras e eficazes para entender e resolver problemas de desempenho acadêmico, como por exemplo em (Romero; Ventura; García, 2008).

Assim, este trabalho se propõe a analisar e oferecer soluções para um problema específico: a elevada taxa de reprovação na disciplina de Programação I, presente na matriz curricular de todos os cursos do campus de Joinville da Universidade Federal de Santa Catarina. De acordo com as perspectivas contemporâneas, a ciência de dados emerge como uma ferramenta relevante e necessária para auxiliar no processo educacional, explorando dados para identificar padrões e fornecer informações valiosas sobre o aprendizado dos alunos (Romero; Ventura, 2007).

Visando minimizar o problema em questão, professores da universidade trabalham para promover maior entendimento de como funciona o aprendizado dos alunos. Para tal, são coletados dados de discentes que servirão como base para entender os fatores que possivelmente influenciaram no desempenho acadêmico. Esse processo, já abordado na comunidade acadêmica e científica, é conhecido como mineração de dados educacionais a partir de análises de aprendizagem (Educational Data Mining, 2023).

Ao longo da história, diferentes processos e abordagens de aprendizagem foram testados e aplicados com o intuito de maximizar a aquisição de novos conhecimentos. Dentro desse cenário, de acordo com Hattie e Timperley (2007), o retorno de informações, popularmente conhecido como feedback, é de suma importância para que seja possível avaliar e alterar, caso necessário, as técnicas aplicadas.

Aliado ao cenário citado, tem-se o desenvolvimento de novas tecnologias de forma rápida e incessante, sendo um de seus campos a ciência de dados. Com o objetivo de descobrir informações previamente ocultas e produzir novos conhecimentos que auxiliarão nas tomadas de decisões, pode ser explicada como uma combinação de matemática, estatística, programação, análise, inteligência artificial e aprendizado de máquina (IBM, 2023b).

Dentro da vasta área de ciência de dados, há um domínio que visa explorar seus conceitos e técnicas para aplicá-los na educação, sendo conhecido como ciência

de dados educacionais.

Dado que o avanço da tecnologia de informação ocorre de forma acelerada e ininterrupta, possibilitando cada vez mais a integração entre as mais diversas áreas de conhecimento, este trabalho apresenta uma análise de dados educacionais para uma possível implementação de melhorias no processo de aprendizagem. A crescente disponibilidade de dados provenientes de plataformas de aprendizagem, registros acadêmicos e interações em ambientes virtuais de ensino, oferece uma ampla fonte de informações sobre o comportamento e o desempenho dos alunos.

Quanto à metodologia do trabalho, primeiramente é analisada uma turma base do semestre 2023-1. Com base nessa turma, são analisadas as variáveis selecionadas para verificar a influência de cada uma. Em seguida, são identificados os alunos que não atenderam às predições e, por meio de probabilidades, são feitas sugestões de quais desses alunos teriam maior chance de aprovação com a devida ajuda. Posteriormente, são realizados testes de modelo utilizando outra turma, outro algoritmo e também uma nova turma com uma abordagem diferente, cujo objetivo nesse caso, é agregar valor adicional às análises.

## 1.1 OBJETIVO

Com o intuito de estabelecer critérios de análise, propõe-se os seguintes objetivos.

### 1.1.1 Objetivo Geral

Analisar as informações dos estudantes da disciplina de Programação I do campus Joinville da UFSC, com o propósito de gerar informações que forneçam entendimento do alto nível de reprovação nesta matéria. A partir destas informações, são gerados conhecimentos pertinentes e aplicáveis no campo da ciência de dados educacionais.

### 1.1.2 Objetivos Específicos

- Apresentar conceitos teóricos sobre ciência de dados na educação;
- Desenvolver critérios de análise específicos para avaliar os dados coletados;
- Aplicar técnicas de mineração de dados para realizar uma análise das variáveis que influenciam no desempenho dos alunos;
- Testar o modelo a partir do uso de outro algoritmo e outras turmas;
- Estabelecer relações com base nos resultados obtidos.

## 2 FUNDAMENTAÇÃO TEÓRICA

Para uma compreensão dos aspectos que influenciam na reprovação dos alunos através da ciência de dados, é relevante abordar alguns tópicos específicos. Esta seção apresentará esses tópicos, juntamente com conceitos e explicações, contribuindo assim para o entendimento do trabalho.

### 2.1 MINERAÇÃO DE DADOS (DATA MINING)

A mineração de dados, também conhecida como data mining, figura como um conceito fundamental na análise de dados. Esse processo pode ser executado por modelos que variam em níveis de complexidade, começando pelo uso da Structured Query Language (SQL), considerado mais simples. A complexidade aumenta com a aplicação de técnicas como árvores de decisão e atinge seu ponto mais elevado com o uso de Redes Neurais Artificiais (RNA).

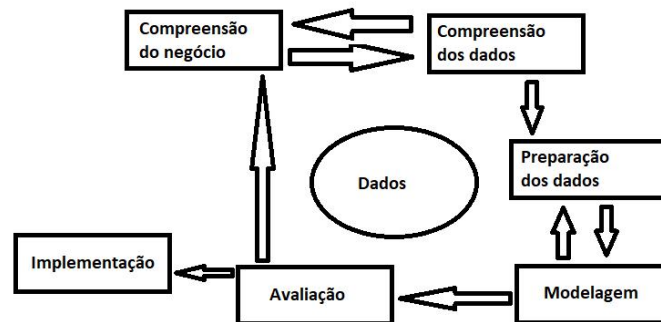
Segundo Provost e Fawcett (2013), a mineração de dados é uma prática que combina arte, ciência e tecnologia. Para tal, utiliza o processo Cross-Industry Standard Process for Data Mining (CRISP-DM) para guiar os profissionais durante a mineração. As fases do CRISP-DM são as seguintes:

1. **Compreensão do Negócio (Business Understanding):** Inicialmente, definem-se os objetivos do projeto a partir do levantamento de requisitos, metas e restrições, elaborando assim uma perspectiva de negócio;
2. **Compreensão dos Dados (Data Understanding):** Nesta etapa, são realizadas explorações nos dados, incluindo coletas, identificação de problemas e um entendimento prévio, buscando compreender sua natureza;
3. **Preparação dos Dados (Data Preparation):** Durante esta etapa, os dados são preparados para análise por meio de tratamentos como filtros para dados ausentes, subdivisões, limpeza e outras transformações necessárias para atender aos requisitos do modelo;
4. **Modelagem (Modeling):** Nesta fase, escolhem-se e aplicam-se diversas técnicas de modelagem aos dados preparados, envolvendo a seleção e implementação de algoritmos de mineração de dados. Adicionalmente, os modelos podem passar por ajustes para otimização, se necessário;
5. **Avaliação (Evaluation):** Neste momento, os resultados obtidos são rigorosamente avaliados para determinar se o modelo está em conformidade com o projeto, podendo ocorrer a reformulação do mesmo;
6. **Implementação (Deployment):** Por fim, se os resultados do modelo estiverem de acordo com o escopo do projeto, eles são implementados em seu ambiente

específico, podendo também ser integrados a outros sistemas;

A Figura 1 apresenta as etapas do processo de mineração de dados.

Figura 1 – Processo de mineração de dados



Fonte: Adaptado de (Shearer, 2000)

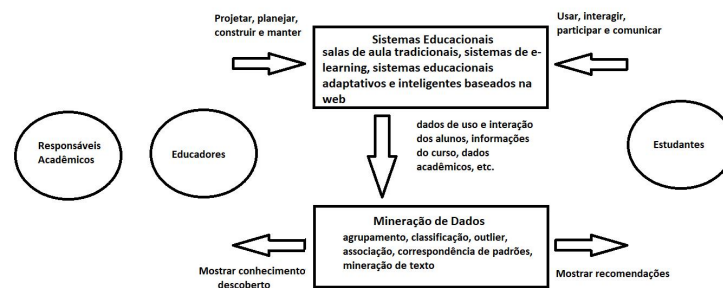
Dentre as características, nota-se que algumas etapas permitem o reinício do ciclo.

## 2.2 MINERAÇÃO DE DADOS EDUCACIONAIS

A Mineração de Dados Educacionais (MDE) é uma ferramenta crucial para compreender os alunos e seus processos de aprendizado por meio da análise de dados coletados durante suas interações educacionais (Educational Data Mining, 2023). Esse campo desempenha papel vital na otimização dos ambientes educacionais.

Conforme delineado por Romero e Ventura (2007), a MDE envolve um ciclo iterativo de formação de hipóteses, realização de testes e aprimoramento contínuo. A Figura 2 esquematiza visualmente esse processo.

Figura 2 – Ciclo dos dados educacionais



Fonte: Adaptado de (Romero; Ventura, 2007)

Como ilustrado na Figura 2, a eficácia desse processo está ligada à interação



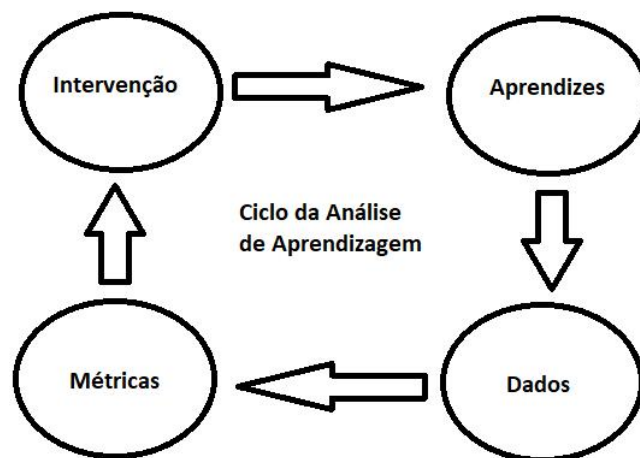
dos estudantes com o sistema educacional. Quanto mais robusta for essa comunicação no ambiente educacional, maior será a quantidade de dados coletados. Isso, por sua vez, amplia a gama de informações disponíveis para análise.

### 2.3 LEARNING ANALYTICS

De acordo com Tsai (2023), learning analytics, ou análise de aprendizagem, refere-se ao entendimento e otimização da aprendizagem por meio da medição, coleta, análise e relatórios de dados sobre os alunos. Conforme mencionado por Clow (2013), a análise de aprendizagem não é uma disciplina sólida com metodologias já estabelecidas, mas sim um campo de investigação que se utiliza das mais variadas ferramentas.

A Figura 3 mostra o ciclo da análise de aprendizagem. De acordo com o que foi apresentado, o ciclo é compreendido como a interação dos alunos com dados que estabelecem métricas pré-definidas, influenciando no processo de aprendizagem desses alunos.

Figura 3 – Ciclo da Análise de Aprendizagem



Fonte: Adaptado de (Clow, 2013)

### 2.4 COLETA DE DADOS

A coleta de dados pode ser classificada em dois modos distintos, os quais serão detalhados a seguir.

1. **Coleta Ativa:** A coleta ativa de dados ocorre quando o participante, cujas informações serão coletadas, fornece ativamente os dados. Nesse método, há uma certa subjetividade associada, uma vez que os dados são fornecidos pelo próprio indivíduo (Kuik, 2018);

2. **Coleta Passiva:** Em contraste, a coleta passiva de dados ocorre sem a participação direta do usuário, podendo até ocorrer sem o seu conhecimento. Esse método é comumente empregado em navegadores e sites na internet. A coleta passiva é considerada objetiva, uma vez que os dados não são fornecidos explicitamente pelos participantes (Kuik, 2018);

De maneira geral, ambos os tipos possuem relevância para coleta de dados, e podem ser diferenciados, como ilustrado na Figura 4.

Figura 4 – Diferenças entre coleta ativa e passiva de dados



Fonte: Adaptado de (Kuik, 2018)

A coleta de dados no Moodle é do tipo passiva e ocorre quando os usuários registram suas atividades, participam de eventos, solicitam informações ou respondem a pesquisas, entre outras interações. Para este trabalho, os dados serão coletados dos estudantes matriculados na disciplina de Programação I, seguindo critérios estabelecidos.

A coleta é realizada por meio de exportação de relatórios. Esses relatórios podem ser gerados pelos administradores da turma, que, neste contexto, são os professores responsáveis pela disciplina (Moodle, 2023). A partir desses dados, são obtidas informações sobre as atividades realizadas pelos estudantes, fundamentais para a análise proposta.

Entre os módulos coletados, destacam-se as informações de notas e presença. O módulo de presença fornece um registro de assiduidade dos estudantes, enquanto os módulos de notas abrangem desde avaliações individuais até módulos compostos, que agrupam notas de diferentes práticas (Moodle, 2023).

## 2.5 APRENDIZAGEM SUPERVISIONADA E NÃO SUPERVISIONADA

No contexto da aprendizagem de máquina, diversas abordagens são empregadas. A aprendizagem supervisionada utiliza rótulos durante o treinamento,

enquanto a aprendizagem não supervisionada não faz uso de rótulos.

No aprendizado supervisionado, o objetivo é realizar uma análise precisa dos dados. Para atingir esse objetivo, os dados de treinamento são previamente rotulados de acordo com as características de interesse. Na mineração de dados, algoritmos de classificação ou regressão podem ser aplicados para essa finalidade (IBM, 2023a).

Em contraste, na aprendizagem não supervisionada, o objetivo principal é realizar o agrupamento de conjuntos de dados por meio de algoritmos. No campo da mineração de dados, a técnica de agrupamento, também conhecida como clustering, é amplamente utilizada, embora existam outras técnicas disponíveis (IBM, 2023a).

## 2.6 CLUSTERING

Uma parte fundamental em uma análise de dados é o agrupamento de informações por semelhanças, também conhecido como clustering. Entre as características essenciais, destaca-se que os exemplares de um determinado agrupamento devem ser o mais similares possível, enquanto instâncias de agrupamentos diferentes devem ser o mais diferentes possível. Para alcançar esse objetivo, a medição dessas similaridades deve ser clara e objetiva (Xu; Tian, 2015).

Ao longo de todo o processo, algumas etapas devem ser seguidas. Inicialmente, são selecionadas as informações mais pertinentes do conjunto de dados a serem analisados. Em seguida, essas informações passam por um algoritmo de agrupamento, escolhido pelo responsável pela análise dentre a vasta gama de algoritmos disponíveis. Ao finalizar, os resultados encontrados pelo algoritmo são analisados, juntamente com uma explicação desses resultados.

## 2.7 ALGORITMOS PARA MINERAÇÃO DE DADOS UTILIZADOS

Nesta seção são abordados os dois principais algoritmos de mineração de dados. Essas técnicas são conhecidas como Árvores de Decisão e Redes Neurais Artificiais.

### 2.7.1 Árvores de Decisão

De acordo com Loh (2011), as árvores, tanto de classificação quanto de regressão, são métodos de aprendizado de máquina. Tais métodos utilizam-se de dados para construir modelos preditivos.

No contexto do aprendizado de máquina, as árvores de decisão representam um dos modelos mais utilizados. Nesse modelo, que tem como base uma estrutura semelhante a uma árvore, cada nó representa uma decisão baseada em um atributo. A partir dessa decisão, a árvore passa por divisões sucessivas em subconjuntos, onde

um nó “pai” gera nós “filhos” que podem, por sua vez, servir como nó pai para outros subconjuntos.

Quanto às aplicações das árvores de decisão, existem as árvores de classificação, que visam determinar a classe à qual a variável pertence, e as árvores de regressão, cujo objetivo é prever um valor. A configuração dessas árvores depende da manipulação dos parâmetros e da variável alvo, podendo também serem utilizadas de forma combinada.

### **2.7.2 Redes Neurais Artificiais**

Com o objetivo de simular e aprimorar o processamento do cérebro humano, surgem as Redes Neurais Artificiais (RNA). Essas redes são compostas por neurônios artificiais, que desempenham um papel fundamental em processos de aprendizado de máquina e inteligência artificial em geral (Nielsen, 2015).

Entre as diversas aplicações dessa ferramenta, destaca-se a mineração de dados educacionais. Considerando a complexidade envolvida na análise e eventual previsão do comportamento dos estudantes para posterior auxílio, o uso de RNAs torna-se extremamente interessante, uma vez que são capazes de fornecer soluções para problemas complexos, como o tratado neste contexto.

Quanto às suas características, é relevante destacar as mais significativas. Como mencionado, os neurônios representam a unidade mais básica de processamento nas redes e atuam em diferentes camadas da RNA. Essas camadas incluem uma camada de entrada, que recebe os dados, camadas intermediárias, que processam esses dados, e uma camada de saída, que fornece o resultado final.

Além das características intrínsecas das Redes Neurais Artificiais, é crucial considerar um fator importante que permeia todo o processo de simulação: a função de ativação. Essa função é responsável por introduzir a não linearidade, permitindo a resolução de problemas mais complexos que dificilmente seriam solucionados com uma abordagem mais simples.

## **2.8 MATRIZ DE CONFUSÃO**

Entre as ferramentas utilizadas para avaliar o resultado de uma análise de dados, destaca-se a matriz de confusão. Geralmente, ao realizar um processo por meio de aprendizado de máquina, essa matriz é empregada para medir o desempenho do procedimento. A Figura 5 exemplifica uma matriz de confusão.

Figura 5 – Matriz de Confusão

		Previsto	
		Positivo	Negativo
Real	Positivo	Verdadeiro positivo	Falso negativo
	Negativo	Falso positivo	Verdadeiro negativo

Fonte: Autor (2023)

Ao observar a matriz, nota-se que ela é dividida em quadrantes. Dentre os valores assumidos, temos: verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN) (James et al., 2013). Com base nesses elementos, é possível calcular algumas características da matriz, sendo elas listadas junto com suas equações a seguir.

**Acurácia (Accuracy):** A acurácia mede o número de elementos calculados corretamente em relação ao total, conforme visto na Equação 1.

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}} \quad (1)$$

**Precisão (Precision):** A precisão calcula os valores verdadeiros positivos corretamente em relação ao total de positivos previstos, conforme visto na Equação 2.

$$\text{Precisão (Precision): } \text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \quad (2)$$

**Sensibilidade (Recall):** A sensibilidade avalia os valores verdadeiros positivos corretamente em relação ao número real de positivos, conforme visto na Equação 3.

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} \quad (3)$$

A matriz de confusão será utilizada como métrica neste trabalho.

## 2.9 ORANGE

Para a condução das análises e visualizações de dados, optou-se pela plataforma Orange, projetada com o propósito de facilitar tais atividades e revelar padrões ocultos (Orange, 2023).

A interface gráfica do Orange é intuitiva, tornando-a acessível mesmo para usuários sem muita experiência em programação. Essa característica permite

desenvolver análises mais complexas sem a necessidade de habilidades avançadas. Destacam-se também os widgets, que são blocos de análise especializados, abrangendo uma variedade de funções, desde importação de arquivos e seleção de dados até a aplicação de algoritmos e métricas de avaliação.

No âmbito específico do aprendizado de máquina, o Orange oferece suporte a diversos algoritmos, incluindo árvores de decisão e redes neurais artificiais. Essa flexibilidade possibilita a criação de modelos preditivos com diferentes níveis de complexidade, adaptando-se às necessidades específicas de cada análise. Destaca-se também a disponibilidade de métricas de análise, como a matriz de confusão, que é apresentada de maneira visual, simplificando a interpretação dos resultados.

A escolha pelo Orange auxilia significativamente no processo de análise de dados, proporcionando uma compreensão mais clara e ampliando as possibilidades de descoberta de padrões nos conjuntos de dados, além de exibir os resultados obtidos.

### 3 TRABALHOS SIMILARES

Neste capítulo, será realizada uma revisão bibliográfica de alguns trabalhos relacionados à temática de relatórios no Moodle e mineração de dados educacionais.

#### 3.1 ESTUDO DE CASO DE MINERAÇÃO DE DADOS NO MOODLE

O estudo de caso elaborado por Romero, Ventura e García (2008) teve como objetivo analisar a mineração de dados no curso de gerenciamento de sistemas. Foram empregadas técnicas de mineração, como estatística, visualização, classificação, agrupamento e regras de associação dos dados.

O processo de mineração foi dividido em coleta, preparação, aplicação e avaliação dos dados. Para a coleta de dados, foram considerados os dados armazenados no banco de dados do sistema Moodle. Dado que o Moodle armazena os dados em bancos de dados relacionais, há maior flexibilidade e facilidade em obter informações em comparação com arquivos de texto, o que é útil para o tratamento dos dados.

No pré-processamento dos dados, a primeira etapa envolveu a seleção dos dados a serem utilizados. Neste caso, foram utilizados dados de 438 estudantes. Com essas informações disponíveis, foi criada uma tabela com atributos que armazenam dados dos estudantes. A Figura 6 exibe os atributos considerados.

Figura 6 – Tabela de atributos

Attributes used for each student	
Name	Description
Course	Identification number of the course
n_assignment	Number of assignments done
n_quiz	Number of quizzes done
n_quiz_a	Number of quizzes passed
n_quiz_s	Number of quizzes failed
n_messages	Number of messages sent to the chat
n_messages_ap	Number of messages sent to the teacher
n_posts	Number of messages sent to the forum
n_read	Number of messages read on the forum
total_time_assignment	Total time spent in assignment
total_time_quiz	Total time used on quizzes
total_time_forum	Total time used on forum
Mark	Final mark the student obtained in the course

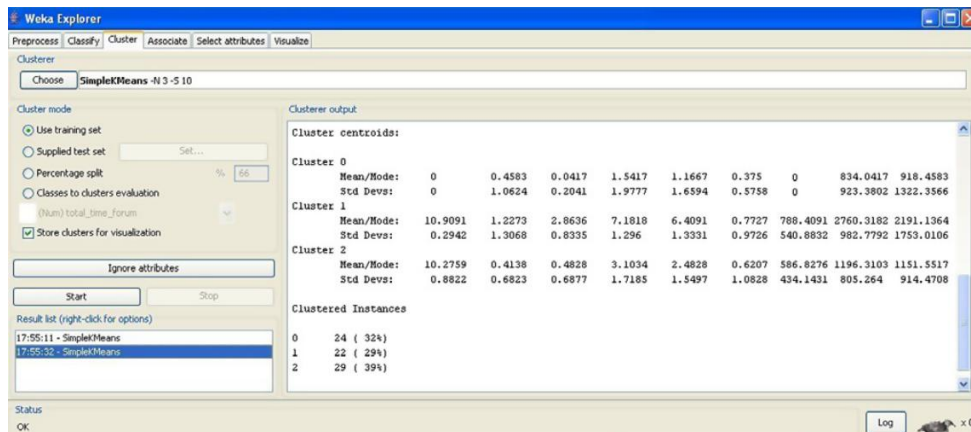
Fonte: (Romero; Ventura; García, 2008)

Na aplicação dos dados, foram utilizados Weka e Keel, sistemas gratuitos implementados em Java que utilizam o formato Attribute-Relation File Format (ARFF). Para o agrupamento, foi empregado o algoritmo KMeans, resultando na divisão dos estudantes em três grupos, conforme pode ser visto na Figura 7 que exibe os resultados obtidos de cada cluster. O cluster 0 representa os alunos não ativos, enquanto o cluster

1 engloba os alunos muito ativos, e o cluster 2 representa os estudantes ativos, situando-se entre os dois extremos.

Com base nesses dados, os estudantes foram classificados como insuficientes, bons ou excelentes. Os resultados obtidos foram utilizados para auxiliar os estudantes classificados como insuficientes, que não estavam obtendo notas suficientes nos questionários.

Figura 7 – Agrupamento obtido



Fonte: (Romero; Ventura; García, 2008)

Diante desses resultados, os professores obtêm entendimentos que podem orientar ações específicas para aprimorar o desempenho dos estudantes.

### 3.2 MINERAÇÃO DE DADOS EDUCACIONAIS NO MOODLE

O estudo conduzido por Shrestha e Pokharel (2021) teve como objetivo identificar as ferramentas que mais influenciam as notas dos alunos com base nos dados do Moodle. Foram consideradas oito variáveis independentes, correspondentes ao número de acessos nas seguintes ferramentas: tarefas, bate-papo (chat), fórum, arquivos, visualizações de curso, links de vídeos e enciclopédia. A nota foi tratada como variável dependente.

No estudo, foram empregados cinco algoritmos de classificação para aprendizado supervisionado: K Nearest Neighbour, Naïve Bayes, Support Vector Machine, Random Forest e Classification and Regression Trees.

Ao concluir a pesquisa em uma sala com oitenta e um estudantes, observou-se que o algoritmo SVM obteve maior precisão e que os acessos ao sistema e arquivos foram os mais impactantes.

A Figura 8 abaixo mostra o impacto significativo de cada variável, com destaque para as duas escolhidas.



Figura 8 – Impacto das ferramentas analisadas

Features	meanImp	decision
Assignment.Click	-0.76998092	Rejected
Chat.Click	0.06111646	Rejected
File.Click	8.45510448	Confirmed
Forum.Click	0.80458942	Rejected
System.Click	5.27863130	Confirmed
Url.Click	3.64773971	Rejected
Wiki.Click	-0.82471889	Rejected

Fonte: (Shrestha; Pokharel, 2021)

As técnicas utilizadas, como clustering e classificação, mostram-se promissoras para identificar padrões e fatores impactantes nas notas, proporcionando aos educadores uma visão mais clara e embasada para aprimorar o processo de ensino.

### 3.3 ANÁLISE DE SISTEMAS DE APRENDIZAGEM REMOTOS ATRAVÉS DE EDM

Com o objetivo de compreender a relação dos estudantes com os sistemas de aprendizagem remotos e antecipar como essa relação impacta o desempenho, o estudo conduzido por Lile (2011) propõe uma abordagem de mineração de dados educacionais empregando diversos métodos.

Inicialmente, foi elaborado um relatório das atividades dos estudantes, englobando o acesso e o envio de tarefas, notas em questionários e participação em fóruns de discussão. Através de algoritmos de clustering, os estudantes foram agrupados conforme padrões de comportamento no Moodle.

Utilizando o algoritmo k-means, foram identificados três grupos. O cluster 0 representa os estudantes mais ativos, enquanto o cluster 1 engloba os estudantes moderadamente ativos, e o cluster 2 inclui os estudantes menos ativos. Em seguida, os estudantes moderadamente ativos apresentaram notas abaixo porém não tão distantes, enquanto que os membros do cluster 2 apresentaram notas significativamente mais baixas. Essa análise revela conexões sobre a relação entre o engajamento dos estudantes e seu desempenho acadêmico em ambientes de aprendizagem remotos.

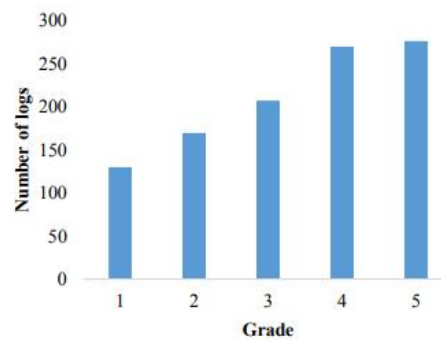
### 3.4 DESEMPENHO DO ESTUDANTE POR RELATÓRIOS DO MOODLE

O estudo realizado por Kadoić e Oreški (2018) propõe uma análise do desempenho do estudante com base nos relatórios gerados pelo Moodle. Para tal,

foram consideradas a frequência de acesso e atividades realizadas como parâmetros para a análise.

Inicialmente, tem-se um gráfico que correlaciona o número de acessos com as notas obtidas, como pode ser visto na Figura 9. Após visualizar, pode-se concluir que os estudantes com mais acessos obtiveram melhores notas.

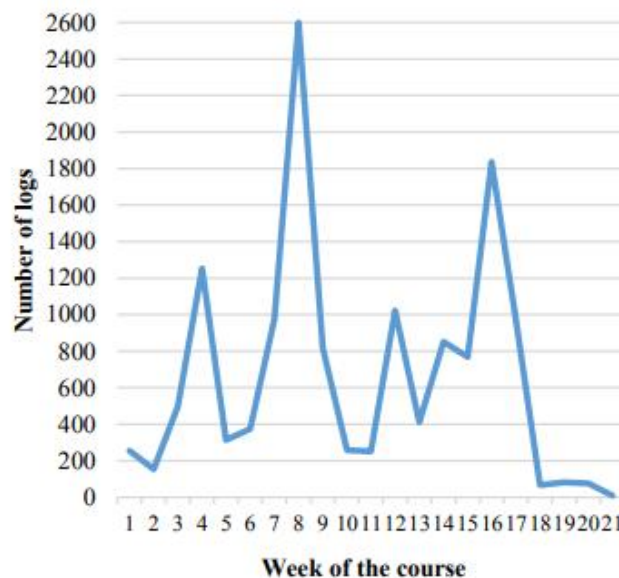
Figura 9 – Gráfico correlacionando número de acessos e notas obtidas



Fonte: (Kadoić; Oreški, 2018)

Além disso, também foram feitas outras comparações. A Figura 10 exibe o número de acessos por semana.

Figura 10 – Número de acessos por semana



Fonte: (Kadoić; Oreški, 2018)

Com base na Figura 10, nota-se que os pontos de pico foram nas semanas 8 e 16. Tal fenômeno acontece, pois, como citado em Kadoić e Oreški (2018), essas semanas se referem às semanas de prova, ou seja, espera-se que os alunos tenham uma alteração considerável nas horas de estudo, comprovando esta ideia.

## 4 METODOLOGIA

Após revisar trabalhos relacionados ao tema, a realização da mineração de dados educacionais é apresentada.

Esta metodologia proporciona uma abordagem abrangente para a mineração de dados educacionais, permitindo a análise profunda e a interpretação significativa dos resultados obtidos.

### 4.1 APRESENTAÇÃO INICIAL DO AMBIENTE PARA MINERAÇÃO DE DADOS

Antes de iniciar o processo de mineração de dados, é importante destacar que o ambiente do Moodle foi previamente organizado para acompanhar melhor o processo de aprendizagem dos estudantes.

As ferramentas disponíveis passaram por uma pré-análise, visando facilitar o processo de análise de aprendizagem e mineração de dados educacionais.

Dentre as ferramentas da disciplina de Programação I, tem-se questionários e Laboratórios Virtuais de Programação (VPL - Virtual Programming Laboratories, do inglês), que são realizados ao longo do semestre.

Enquanto os questionários representam a capacidade do aluno em entender os conceitos teóricos, os VPLs desempenham um papel significativo no processo de aprendizado prático, proporcionando aos estudantes a oportunidade de aplicar os conceitos teóricos em atividades práticas de programação.

Avaliações finais também são aplicadas, integrando a nota total do curso. No entanto, como são realizadas apenas no final, não serão utilizadas como parâmetros para analisar o desempenho dos estudantes ao longo do curso.

Para elucidar o exposto, as Figuras 11 e 12 apresentam o ambiente do Moodle com algumas de suas funcionalidades. Como pode ser observado, há um espaço dedicado aos laboratórios virtuais, que serão fundamentais para grande parte da análise. Além deles, existem questionários relacionados aos tópicos lecionados, que são respondidos ao final do assunto estudado. Há também um espaço dedicado à presença, outra variável de interesse para este trabalho.

Além das ferramentas citadas, também existem outras, tais como avisos, plano de ensino e informações sobre horários e disponibilidade de monitoria.

De forma geral, observa-se que o Moodle pode ser dividido conforme idealização do professor. Além destas características, também é informado o progresso do estudante, ou seja, há espaços de marcação automática que indicam quando o aluno completou determinada atividade.

Figura 11 – Ambiente do Moodle na disciplina de Programação I

## EMB5600-02602/02603/02604... (20232) - Programação I

Painel ▶ Meus cursos ▶ EMB5600-02602/02603/02604/02607 (20232)

Seu progresso

### NAVEGAÇÃO

Painel

- Página inicial do site
- Moodle UFSC
- Meus cursos
  - EMB5636-06605 (20231)
  - EMB5617-07605 (20232)
  - EMB5636-06605 (20232)
  - ESE410022-41010086ME (20232)
  - EMB5045-09605 | EMB5650-09605 (20232)
  - EMB5600-02602/02603/02604/02607 (20232)**
    - Participantes
    - Emblemas
    - Competências
    - Notas
    - Programação I
    - Laboratórios virtuais
    - Material extra
    - Arquitetura computacional
    - Sistemas de numeração
    - Algoritmos / pseudocódigo / portugal
    - Introdução à linguagem C
    - Desvio condicional
    - Estrutura de repetição

### Programação I

- Avisos
- Plano de ensino 1.2Mb Documento PDF
- Presença
- Monitores e horários de monitoria

---

### Laboratórios virtuais

Exercícios autoavaliativos para você fixar o conteúdo aprendido

- Tutorial sobre o VPL [versão que fala do ambiente adicional]
- Tutorial sobre os VPLs
- VPL 0 - Retângulo maior
- VPL 1 - Fatorial
- VPL 2 - Valor de uma série
- VPL 3 - Produto interno
- VPL 4 - Maior valor do vetor

Fonte: Autor (2023)

Figura 12 – Exemplo de tópicos de assuntos na disciplina de Programação I

Lista 1 - Computador simplificado Documento PDF

Gabarito - Lista 1 Documento PDF

**Restrito** Disponível se: A atividade Lista 1 - Computador simplificado esteja marcada como concluída

---

### Sistemas de numeração

Este tópico engloba o estudo dos sistemas de numeração e a conversão de dados entre diferentes sistemas de numeração.

- H-P** Videoaula - Sistemas de numeração
  - Apresentação sobre sistemas de numeração 292.2Kb Documento PDF
- H-P** Videoaula - Conversão de dados em diferentes sistemas de numeração
  - Apresentação sobre a conversão de bases 371.7Kb Documento PDF
- Exercícios resolvidos sobre conversão de base
- Outros vídeos sobre sistemas de numeração
- Lista 2 - Sistemas de numeração 207.9Kb Documento PDF
- Gabarito - Lista 2 297.7Kb Documento PDF
- Restrito** Disponível se: A atividade Lista 2 - Sistemas de numeração esteja marcada como concluída
- Avaliação do conteúdo - sistemas de numeração

Algoritmos / pseudocódigo / portugal

Fonte: Autor (2023)

Após essa apresentação inicial do ambiente virtual, é detalhado a seguir todo o processo de mineração de dados. Para tal, destacam-se as fonte de dados, os relatórios e também o anonimato dos alunos.

## 4.2 COLETA DE DADOS DO MOODLE

A coleta de dados desempenha um papel essencial na mineração de dados. No contexto específico deste estudo, a coleta de dados é realizada por meio da plataforma Moodle, um ambiente virtual de aprendizagem amplamente utilizado. Essa etapa visa obter informações detalhadas sobre as interações dos alunos com o conteúdo do curso, suas participações nas atividades e seu desempenho global.

Quanto às fontes de dados, o Moodle armazena uma riqueza de informações valiosas em seu sistema, que registram uma variedade de dados, incluindo atividades, notas, realização de questionários e laboratórios virtuais. A coleta dessas informações é essencial para compreender o comportamento dos alunos e identificar padrões relevantes para a análise.

Para iniciar a coleta de dados, são utilizados relatórios personalizados fornecidos pelo Moodle. Esses relatórios são configurados para extrair informações específicas, como presença nas aulas, desempenho em questionários e conclusão de laboratórios virtuais. Essa abordagem permite uma visão detalhada e estruturada das atividades dos alunos.

Em respeito à privacidade dos alunos, os dados coletados passam por um processo de anonimização. Os alunos são representados por números ordenados, garantindo que suas identidades individuais sejam protegidas durante todo o processo de análise.

## 4.3 PREPARAÇÃO DE DADOS, SIMULAÇÕES E ANÁLISES

Inicialmente, toma-se como base uma turma de Programação I do primeiro semestre de 2023 da UFSC Joinville. Essa turma, composta por 86 alunos, será fundamental para a maioria dos processos de análise. O número real de aprovados e reprovados pode ser visualizado na Figura 13.

Considerando que serão conduzidas várias análises, cada uma com características distintas, é fundamental realizar uma preparação adequada dos dados. A preparação dos dados visa garantir que estejam em um formato adequado para a análise, incluindo limpeza, tratamento de valores ausentes e conversões necessárias. Nesta seção, serão descritas as etapas envolvidas nesse processo, abrangendo diferentes tipos de análises propostas.

Dado que cada análise será conduzida de forma distinta, o processo de preparação de dados será individual para cada caso, ou seja, cada análise terá sua própria etapa de preparação, seguida pela simulação e, por fim, pela análise. Quanto à etapa de preparação, será conduzida nas planilhas do Google, seguida pela simulação na plataforma Orange e, finalmente, pelas análises propriamente ditas.

Figura 13 – Desempenho geral da turma base na disciplina

		Predicted		$\Sigma$
		Aprovado	Reprovado	
Actual	Aprovado	38	0	38
	Reprovado	0	48	48
$\Sigma$		38	48	86

Fonte: Autor (2023)

Com base na Figura, observa-se que 38 dos 86 alunos foram aprovados, representando aproximadamente 44,19% dos estudantes.

Antes de iniciar as análises específicas, deve-se esclarecer as métricas utilizadas. A precisão refere-se à capacidade do modelo de identificar corretamente os casos positivos em relação ao total de casos previstos como positivos. Em termos mais simples, é a proporção de alunos corretamente previstos como aprovados em relação ao número total de alunos previstos como aprovados. Por outro lado, a sensibilidade, foca na capacidade do modelo de identificar todos os casos positivos em relação ao número total real de casos positivos. Em outras palavras, é a proporção de alunos corretamente previstos como aprovados em relação ao número total real de alunos aprovados.

#### 4.3.1 Análise das Relações entre as Variáveis

Com base nas informações coletadas, o processo de simulação é iniciado. Primeiramente, exporta-se o arquivo com os dados a serem utilizados para o software, visíveis na tabela de dados, conforme mostra a Figura 14.

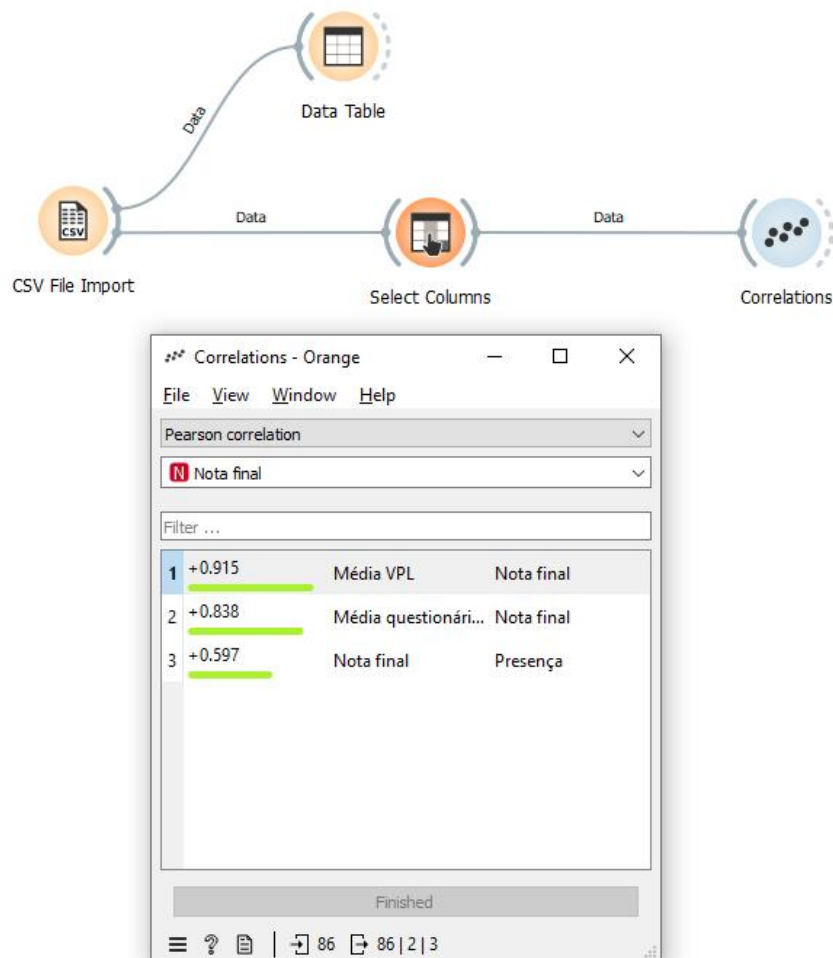
Figura 14 – Exemplo de tabela com as variáveis utilizadas e a variável alvo

	Media VPL	Media Quest	Total do curso	Porcentagem
1	10	9.92	10	96
2	9.05556	8	7	69.2
3	6.66667	9	6.5	90.5
4	9.88889	8.59	9	90
5	1.11111	8.24	1.5	94.7
6	9.22222	9.67	7	95.7
7	3.33333	5.27	1.5	100
8	9.88889	9.83	8	84.2
9	10	9.88	9	100
10	0	4.98	0.5	95.2
11	6.11111	7.52	7	85
12	7.77778	8.94	8	91.3
13	0	0	0	100
14	1.44444	6.71	5.5	27.3

Fonte: Autor (2023)

Após realizar algumas configurações para estabelecer relações entre a nota final (variável alvo) e as demais variáveis independentes, obtém-se a correlação entre elas, como mostra a Figura 15.

Figura 15 – Correlação das variáveis com a nota final



Fonte: Autor (2023)

Dessa forma, constata-se que a variável que mais influenciou na nota final foi a média das notas dos VPLs, seguida pela nota nos questionários e, por fim, pela presença nas aulas. Essas informações servirão como base para as comparações que serão realizadas.

#### 4.3.2 Preparação e simulação de dados relacionados à presença

Nesta subseção, é apresentada a preparação dos relatórios de presença dos alunos, e embora seja a menos influente dentre as variáveis analisadas, sua análise é válida. Para tal, tem-se os seguintes estados possíveis: o aluno estava presente na

aula, sendo representado pelo símbolo Pr(2/2), caso esteve ausente, representa-se por Au(2/2), e, para os casos que há uma interrogação, atribui-se ausência. A Figura 16 exhibe estas características.

Figura 16 – Relatório de presença

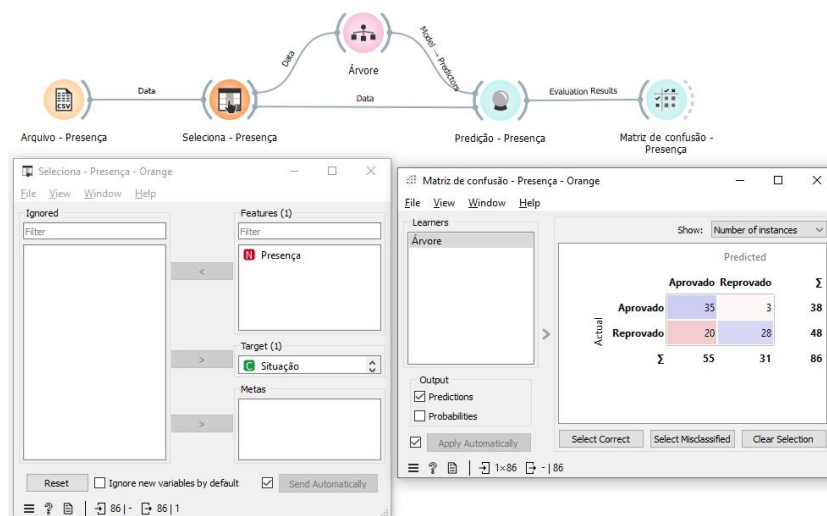
Nome	23/03/2023 13:3	28/03/2023 10:3	30/03/2023 10:2	30/03/2023 13:3	4/04/2023 10:30	6/04/2023 10:20
1	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	Pr (2/2)
2	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	?
3	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	Pr (2/2)
4	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	?
5	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	?
6	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	Pr (2/2)
7	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	?
8	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	?
9	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	?
10	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	?
11	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	?
12	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	Pr (2/2)
13	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	Pr (2/2)
14	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Au (0/2)	?
15	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	Pr (2/2)
16	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	Pr (2/2)
17	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	Pr (2/2)
18	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	?
19	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	Pr (2/2)
20	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Au (0/2)	?
21	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	?
22	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Au (0/2)	?
23	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	Pr (2/2)	?
24	Pr (2/2)	Pr (2/2)	Pr (2/2)	?	Pr (2/2)	Pr (2/2)

Fonte: Autor (2023)

Com base no relatório geral de presença, estabeleceu-se o seguinte critério: o aluno que mais assistiu às aulas foi definido como referência, ou seja, ele obteve 100% de frequência, enquanto que a porcentagem dos outros é relativa com base neste aluno.

A matriz de confusão com base na presença pode ser vista na Figura 17.

Figura 17 – Matriz de confusão com base na presença



Fonte: Autor (2023)



Com base na exposição da matriz de confusão relacionada à presença dos alunos nas aulas, foram obtidas as métricas apresentadas na Tabela 1.

Tabela 1 – Tabela Comparativa - Valores das métricas de presença

Modelo	Acurácia (%)	Precisão (%)	Sensibilidade (%)
Presença	73,26	63,64	92,10

Fonte: Autor (2023)

Essas métricas são fundamentais para avaliar o desempenho do modelo de análise em relação à presença dos alunos.

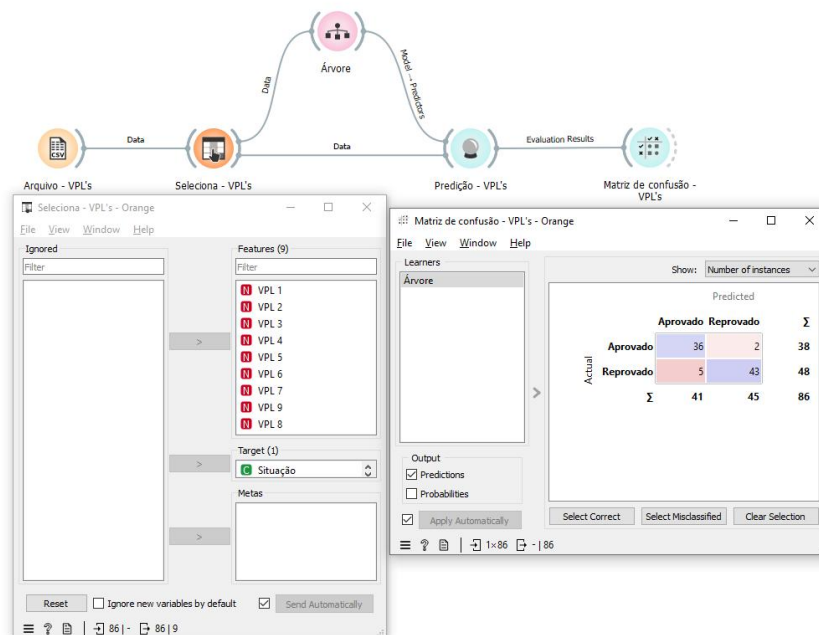
### 4.3.3 Simulações por VPLs

Nesta parte, são apresentadas as análises individuais de cada VPL e também considerando a média total. O objetivo é estabelecer as relações das notas destes laboratórios práticos com o total e confirmar se a média é uma boa aproximação para os casos individuais.

#### 4.3.3.1 VPLs Individuais

Aqui, será apresentada a análise do desempenho dos alunos com base nas notas individuais em cada laboratório virtual de programação. A Figura 18 mostra esse caso.

Figura 18 – VPLs tratados individualmente



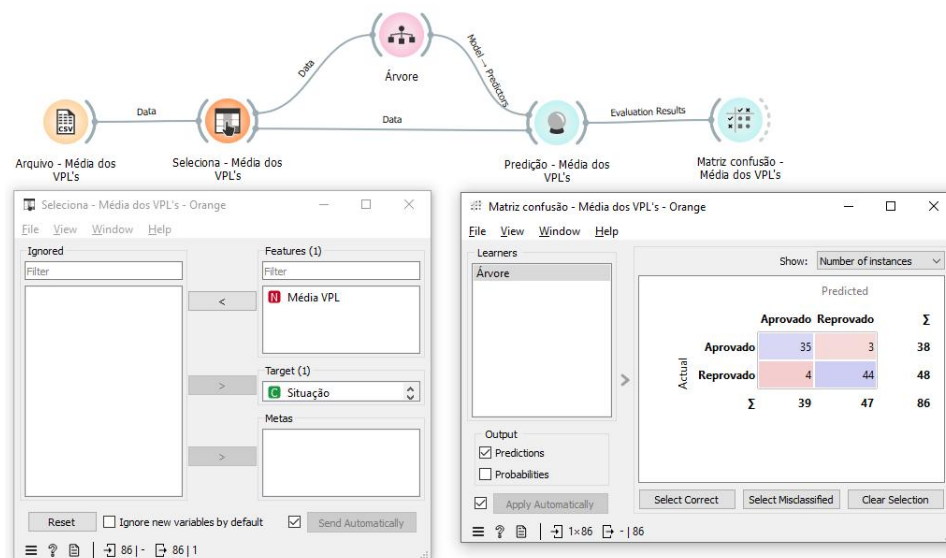
Fonte: Autor (2023)

Os valores encontrados serão utilizados nos cálculos das métricas.

#### 4.3.3.2 Média dos VPLs

Para a análise que relaciona a média das notas dos VPLs com as notas finais, é elaborado o fluxo mostrado na Figura 19, que exibe a matriz de confusão obtida.

Figura 19 – Análise da média dos VPLs



Fonte: Autor (2023)

A partir dessa matriz, juntamente com a matriz exposta na Figura 18, são realizadas comparações.

#### 4.3.3.3 Tabela Comparativa

A Tabela 2 compara as métricas de desempenho entre os VPLs Individuais e a Média dos VPLs.

Tabela 2 – Tabela Comparativa - VPLs Individuais vs. Média dos VPLs

Modelo	Acurácia (%)	Precisão (%)	Sensibilidade (%)
VPLs Individuais	91,86	87,80	94,74
Média dos VPLs	91,86	89,74	92,10

Fonte: Autor (2023)

Os resultados obtidos são discutidos no capítulo 5.

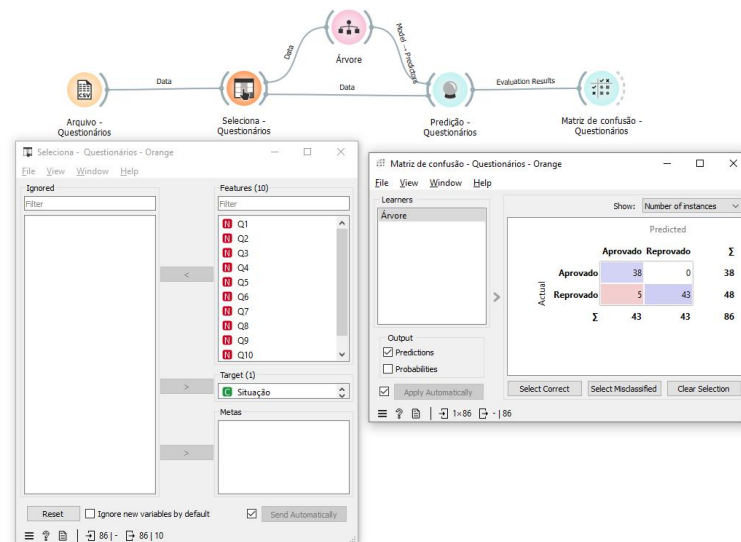
#### 4.3.4 Simulações por Questionários

Nesta subseção, serão apresentadas as análises relacionadas aos questionários, a parte teórica da matéria, explorando tanto a nota de cada questionário individual quanto a média deles. Aqui, também serão analisados e comparados os questionários de forma individual de acordo com a média total.

#### 4.3.4.1 Questionários Individuais

Inicialmente, a análise é direcionada para o desempenho dos alunos em cada questionário. A Figura 20 mostra a metodologia e a previsão das notas nos questionários tratados isoladamente.

Figura 20 – Questionários tratados individualmente

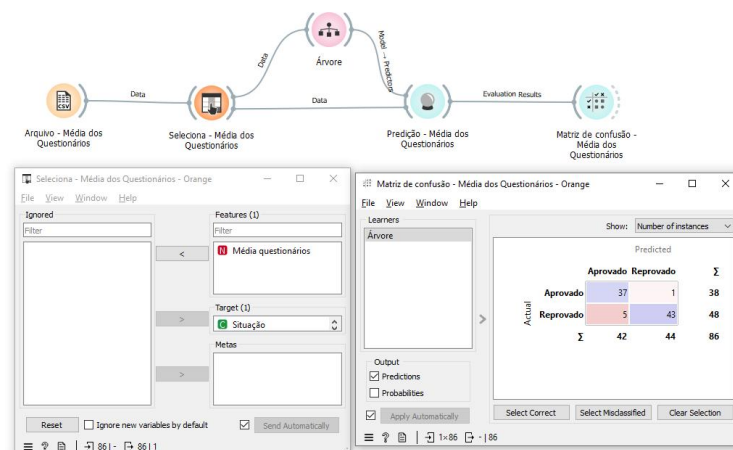


Fonte: Autor (2023)

#### 4.3.4.2 Média dos Questionários

Em seguida, a análise foca na média das notas dos questionários. A Figura 21 apresenta o desenvolvimento e as previsões obtidas.

Figura 21 – Considerando a média dos Questionários



Fonte: Autor (2023)

A partir destes dados, faz-se também uma nova tabela comparativa.

#### 4.3.4.3 Tabela Comparativa - Questionários Individuais vs Média dos Questionários

Além das análises visuais, é útil apresentar uma tabela comparativa das métricas obtidas nas análises dos questionários. A Tabela 3 resume as principais métricas de desempenho.

Tabela 3 – Tabela Comparativa - Questionários

Modelo	Acurácia (%)	Precisão (%)	Sensibilidade (%)
Questionários Individuais	94,19	88,37	100
Média dos Questionários	93,02	88,09	97,37

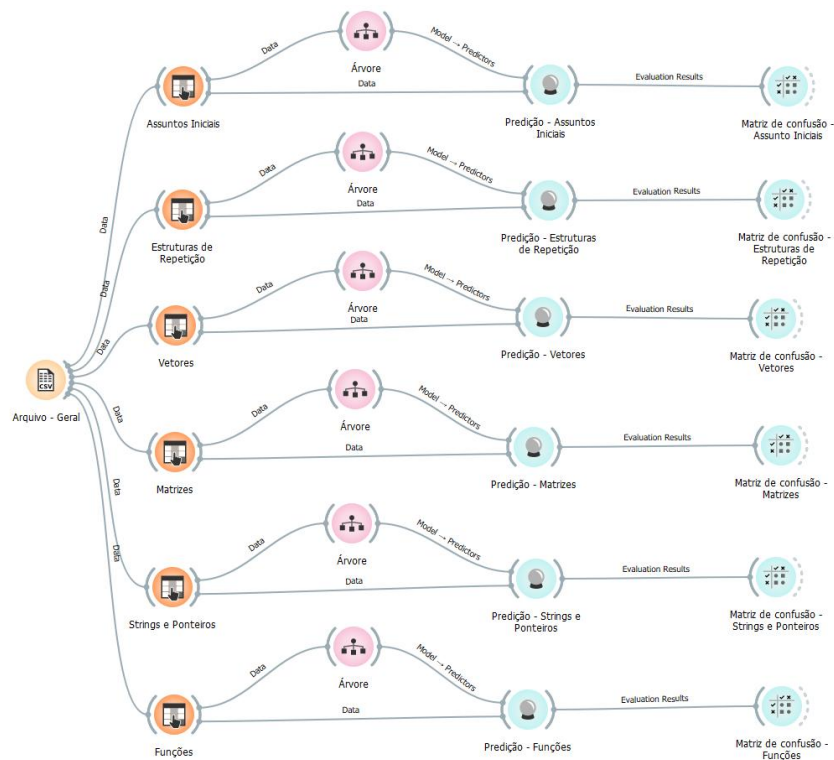
Fonte: Autor (2023)

Os resultados exibidos serão discutidos no capítulo 5.

#### 4.3.5 Simulações por Assuntos

Nesta subseção, é explorada a estruturação do curso de Programação I em seis partes distintas, cada uma representando um conjunto específico de tópicos abordados ao longo do semestre. A divisão do conteúdo foi pensada de forma a aprofundar a compreensão dos alunos em áreas específicas. A Figura 22 apresenta a modelagem baseada nas divisões por assunto.

Figura 22 – Modelagem da análise por assuntos no Orange



Ativa  
Acesse

Fonte: Autor (2023)

A Figura 22 retrata os seis casos de simulação.

#### 4.3.5.1 Fundamentos Básicos

A primeira fase do curso concentrou-se nos fundamentos básicos da programação, avaliados principalmente por meio de questionários. Este período introdutório estabeleceu a base conceitual para tópicos mais avançados nas etapas seguintes. A matriz de confusão resultante pode ser vista na Figura 23.

Figura 23 – Matriz de confusão dos Assuntos Iniciais



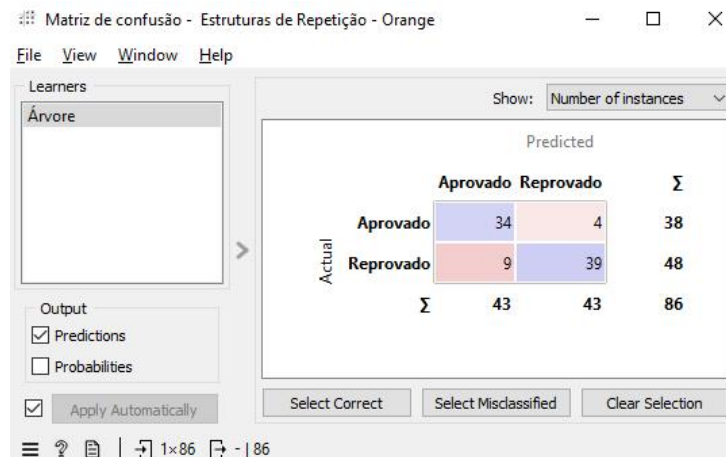
Fonte: Autor (2023)

As métricas obtidas pela matriz estão presentes na Tabela 4.

#### 4.3.5.2 Estruturas de Repetição

Na segunda parte, explora-se a aplicação de estruturas de repetição. A análise revela como os alunos lidaram com conceitos fundamentais relacionados às iterações e loops, destacando a importância dessa habilidade no contexto da programação. A Figura 24 apresenta a matriz de confusão obtida.

Figura 24 – Matriz de confusão das Estruturas de Repetição



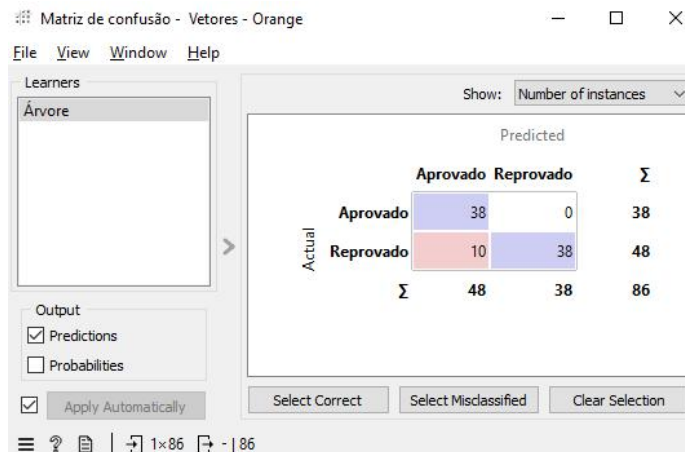
Fonte: Autor (2023)

As métricas obtidas pela matriz estão presentes na Tabela 4.

#### 4.3.5.3 Vetores

A terceira fase do curso introduziu o tema de vetores. A análise desses tópicos oferece entendimentos sobre a capacidade dos alunos em lidar com estruturas de dados lineares, fundamental para a resolução de problemas mais complexos. A matriz de confusão pode ser vista na Figura 25.

Figura 25 – Matriz de confusão dos Vetores



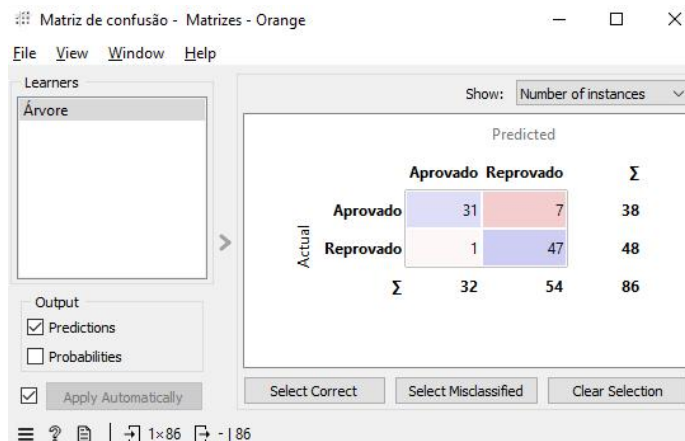
Fonte: Autor (2023)

As métricas obtidas pela matriz estão presentes na Tabela 4.

#### 4.3.5.4 Matrizes

A quarta parte focou na compreensão e aplicação de matrizes. A análise revela como os alunos abordaram desafios relacionados à manipulação e processamento de dados em vetores com mais de uma dimensão. A matriz de confusão é exibida na Figura 26.

Figura 26 – Matriz de confusão das Matrizes



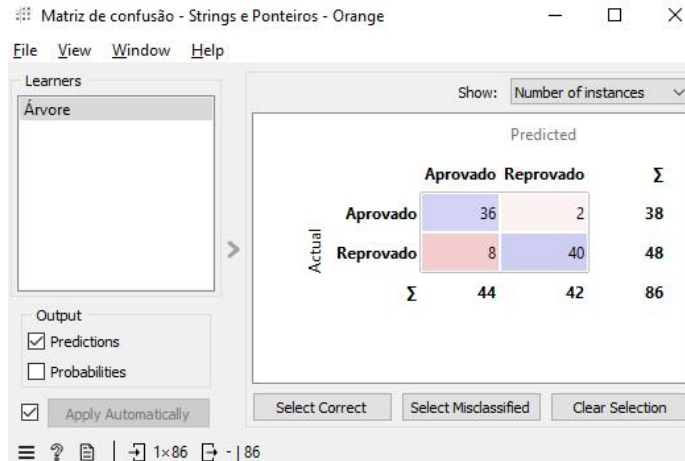
Fonte: Autor (2023)

As métricas obtidas pela matriz estão presentes na Tabela 4.

#### 4.3.5.5 Strings e Ponteiros

A quinta etapa explorou strings e ponteiros, analisando como os alunos trabalham com manipulação de strings e gerenciamento de memória. Neste caso, o assunto ponteiro foi tratado apenas teoricamente. A Figura 27 apresenta a matriz encontrada.

Figura 27 – Matriz de confusão dos Ponteiros e Strings



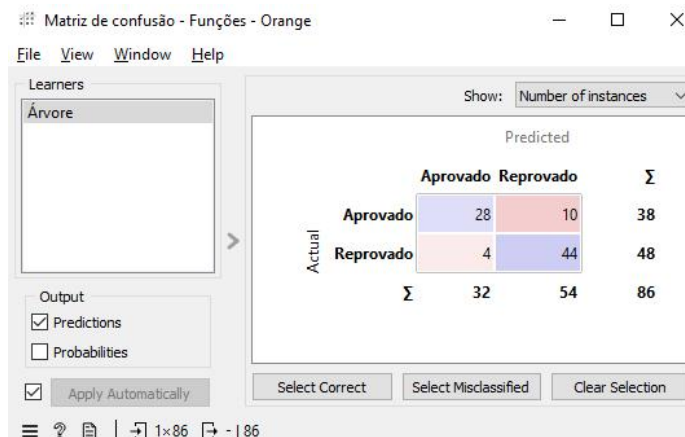
Fonte: Autor (2023)

As métricas obtidas pela matriz estão presentes na Tabela 4.

#### 4.3.5.6 Funções

Na última parte, tem-se o entendimento e aplicação de funções. A análise destaca como os alunos integraram conceitos aprendidos anteriormente para criar e utilizar funções, demonstrando uma compreensão mais avançada dos princípios de programação. A matriz de confusão pode ser vista na Figura 28.

Figura 28 – Matriz de confusão das Funções



Fonte: Autor (2023)

As métricas obtidas pela matriz estão presentes na Tabela 4.

#### 4.3.5.7 Tabela Comparativa - Desempenho por Assunto

Além das análises visuais, é útil apresentar uma tabela comparativa das métricas obtidas em cada fase da análise dos assuntos. A Tabela 4 resume as principais métricas de desempenho.

Tabela 4 – Tabela Comparativa - Desempenho por Assunto

<b>Assunto</b>	<b>Acurácia (%)</b>	<b>Precisão (%)</b>	<b>Sensibilidade (%)</b>
Fundamentos Básicos	86,05	80,95	89,47
Estruturas de Repetição	84,88	79,01	89,47
Vetores	88,37	79,17	100
Matrizes	90,70	96,87	81,58
Strings e Ponteiros	88,37	81,81	94,74
Funções	83,72	87,50	73,68

Fonte: Autor (2023)

A tabela proporciona uma visão mais objetiva e comparativa do desempenho dos alunos em cada assunto abordado no curso, promovendo uma abordagem abrangente do desempenho dos alunos em áreas específicas que será discutida com mais detalhes no capítulo 5.

#### 4.3.6 Identificação de Falsos Negativos e Falsos Positivos e probabilidades de aprovação

Além da análise realizada, busca-se aprimorar a detecção de alguns fatores. O objetivo principal foi identificar com maior precisão quais alunos apresentam uma probabilidade mais significativa de aprovação na disciplina. Esses alunos específicos seriam então selecionados para receber auxílio personalizado.

Na identificação desses alunos, optou-se por analisar um caso com parâmetros fixados para obter respostas de forma simplificada. Primeiramente, são observadas a classificação dos alunos como falsos negativos e falsos positivos com base em cada assunto específico.

Como referência, faz-se a utilização da turma base do trabalho, correspondente ao semestre 2023-1. Os alunos são identificados por meio de números, separados por hífens.

O Quadro 1 apresenta os alunos que foram classificados erroneamente, destacando aqueles que deveriam ter sido reprovados mas foram aprovados, denominados como falsos negativos (FN), e aqueles previstos como aprovados mas que acabaram sendo reprovados, denominados falsos positivos (FP).

Analisando os casos de falsos negativos, destaca-se, por exemplo, o aluno 41, que, apesar de um desempenho inferior em três assuntos, foi aprovado. Quanto aos falsos positivos, o aluno 32 chama atenção, pois foi previsto como aprovado em todos os assuntos, mas acabou sendo reprovado.



Quadro 1 – Quadro com falsos negativos e falsos positivos

	<b>Falsos negativos</b>	<b>Falsos positivos</b>
<b>Fundamentos Básicos</b>	9-41-55-58	6-19-21-30-32-43-56-64
<b>Estruturas de Repetição</b>	25-41-59-71	18-21-30-32-37-43-47-49-69
<b>Vetores</b>	-	8-19-32-37-43-49-56-60-64-67
<b>Matrizes</b>	4-12-16-25-41-58-83	32
<b>Strings e Ponteiros</b>	12-55	27-30-32-43-56-64-67-84
<b>Funções</b>	13-16-17-25-33-45-59-71-81-83	19-32-56-60

Fonte: Autor (2023)

Considerando um cenário em que o professor precisa selecionar alunos para receber ajuda, devido à dificuldade de auxiliar todos, é pertinente estabelecer critérios para essa escolha. Nesse sentido, a ideia é utilizar valores probabilísticos calculados pela simulação para determinar quais alunos têm maior chance de aprovação com a ajuda adequada.

A Figura 29 apresenta, com base no assunto Matrizes, escolhido pela sua alta precisão, a situação real dos alunos, seguida pelo número de identificação e a predição. Ao lado, são exibidos os valores relacionados à probabilidade de aprovação dos alunos, seguidos pela probabilidade de reprovação. Com base nesses números, o professor pode obter parâmetros que o auxiliam a identificar os alunos com maior probabilidade de aprovação.

Figura 29 – Probabilidades de aprovação geradas pela simulação

Aprovado	71	Aprovado	0.966667	0.033333
Aprovado	81	Aprovado	0.966667	0.033333
Aprovado	82	Aprovado	0.966667	0.033333
Aprovado	9	Aprovado	0.966667	0.033333
Aprovado	25	Reprovado	0.285714	0.714286
Reprovado	37	Reprovado	0.285714	0.714286
Aprovado	41	Reprovado	0.285714	0.714286
Reprovado	67	Reprovado	0.285714	0.714286
Reprovado	69	Reprovado	0.285714	0.714286
Reprovado	79	Reprovado	0.285714	0.714286
Reprovado	84	Reprovado	0.285714	0.714286
Reprovado	15	Reprovado	0.25	0.75
Aprovado	16	Reprovado	0.25	0.75
Reprovado	31	Reprovado	0.25	0.75
Reprovado	43	Reprovado	0.25	0.75
Reprovado	49	Reprovado	0.25	0.75
Reprovado	52	Reprovado	0.25	0.75
Reprovado	6	Reprovado	0.25	0.75
Aprovado	83	Reprovado	0.25	0.75
Aprovado	12	Reprovado	0.222222	0.777778

Fonte: Autor (2023)

Com base nesses valores, são realizadas discussões no capítulo 5.

#### 4.3.7 Comparação com outra turma

A incorporação de dados de uma segunda turma auxilia na análise, proporcionando uma perspectiva mais abrangente do desempenho dos alunos em diferentes contextos. As comparações realizadas entre as turmas foram sobre questionários, VPLs e desempenho por assunto.

Os questionários, conforme a Tabela 5 revelaram resultados absolutos na segunda turma, com acurácia, precisão e sensibilidade atingindo 100%.

Tabela 5 – Tabela Comparativa - Métricas dos Questionários na segunda turma

<b>Modelo</b>	<b>Acurácia (%)</b>	<b>Precisão (%)</b>	<b>Sensibilidade (%)</b>
Questionários	100	100	100

Fonte: Autor (2023)

Também foram feitas análises com base nos VPLs, conforme mostra a Tabela 6.

Tabela 6 – Tabela Comparativa - Métricas dos VPLs na segunda turma

<b>Modelo</b>	<b>Acurácia (%)</b>	<b>Precisão (%)</b>	<b>Sensibilidade (%)</b>
VPL's	80,90	82,46	87,04

Fonte: Autor (2023)

A Tabela 7 mostra as métricas para cada assunto nesta nova turma.

Tabela 7 – Tabela Comparativa - Métricas do Desempenho por Assunto na segunda turma

<b>Assunto</b>	<b>Acurácia (%)</b>	<b>Precisão (%)</b>	<b>Sensibilidade (%)</b>
Fundamentos Básicos	91,01	89,65	96,30
Estruturas de Repetição	100	100	100
Vetores	87,64	90,57	88,89
Matrizes	84,27	90	83,33
Strings e Ponteiros	85,39	82,54	96,30
Funções	80,90	87,75	79,63

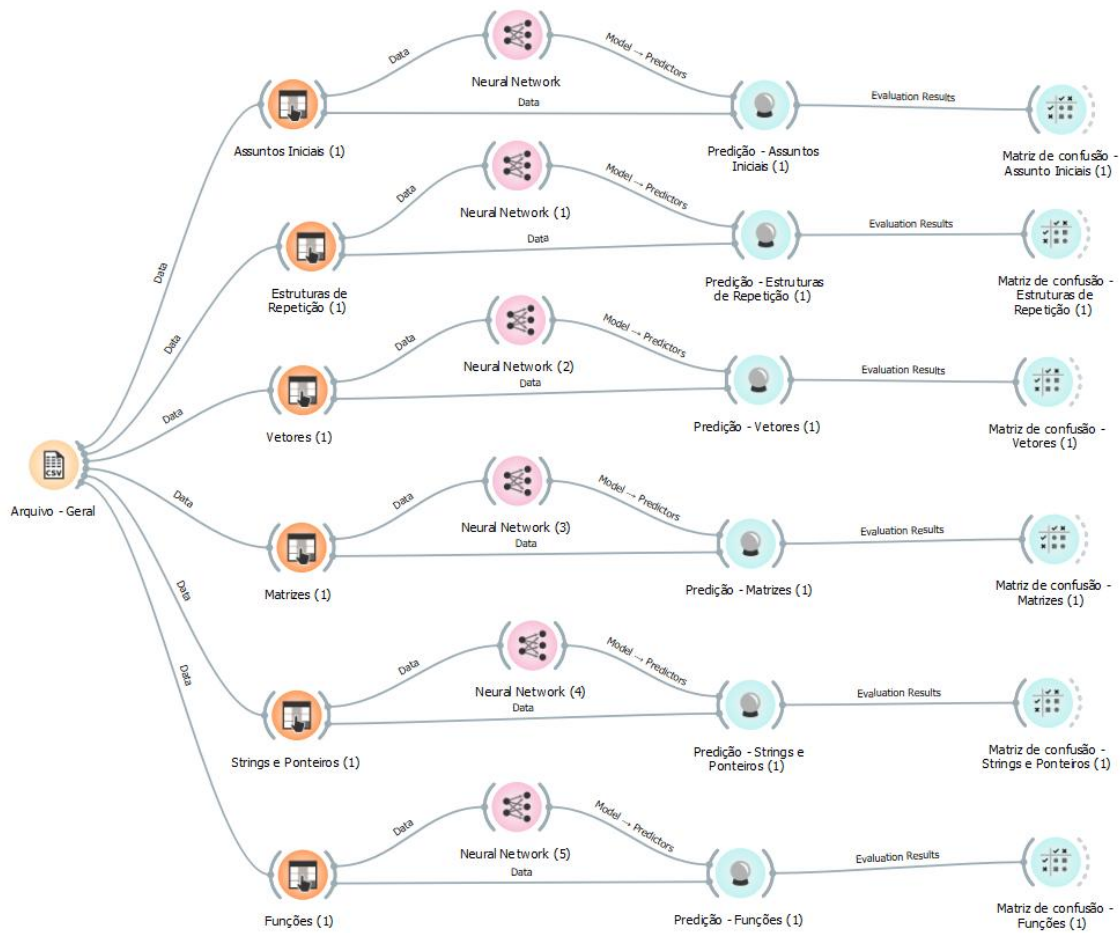
Fonte: Autor (2023)

#### 4.3.8 Uso de Redes Neurais

Após realizar análises utilizando árvores de decisão, foi decidido também realizar sua abordagem a partir de Redes Neurais Artificiais buscando identificar variações causadas por algoritmos. Nesta seção, as análises por questionários, VPLs e assuntos serão abordadas novamente com um novo algoritmo. Para este caso, também foram utilizados dados da turma do primeiro semestre de 2023.

A modelagem no Orange segue o mesmo padrão utilizado para árvores. A Figura 30 exemplifica isso.

Figura 30 – Modelagem com RNA no Orange



Fonte: Autor (2023)

Assim, também é realizada a simulação de questionários. As métricas atingidas são exibidas na Tabela 8.

Tabela 8 – Tabela Comparativa - Métricas dos Questionários utilizando RNAs

Modelo	Acurácia (%)	Precisão (%)	Sensibilidade (%)
Questionários	86,05	80,95	89,47

Fonte: Autor (2023)

Também foram feitas análises com base nos VPLs, conforme mostra a Tabela 9.

Tabela 9 – Tabela Comparativa - Métricas dos VPLs utilizando RNAs

Modelo	Acurácia (%)	Precisão (%)	Sensibilidade (%)
VPL's	88,37	83,33	92,10

Fonte: Autor (2023)

Por fim, são realizadas encontradas as métricas por assuntos, conforme mostra a Tabela 10.

Tabela 10 – Tabela Comparativa - Métricas de Desempenho por Assunto usando RNAs

Assunto	Acurácia (%)	Precisão (%)	Sensibilidade (%)
Fundamentos Básicos	83,72	75	94,74
Estruturas de Repetição	88,37	71,43	78,95
Vetores	90,70	87,50	92,10
Matrizes	89,53	93,94	81,58
Strings e Ponteiros	89,53	80,85	100
Funções	88,37	91,18	81,58

Fonte: Autor (2023)

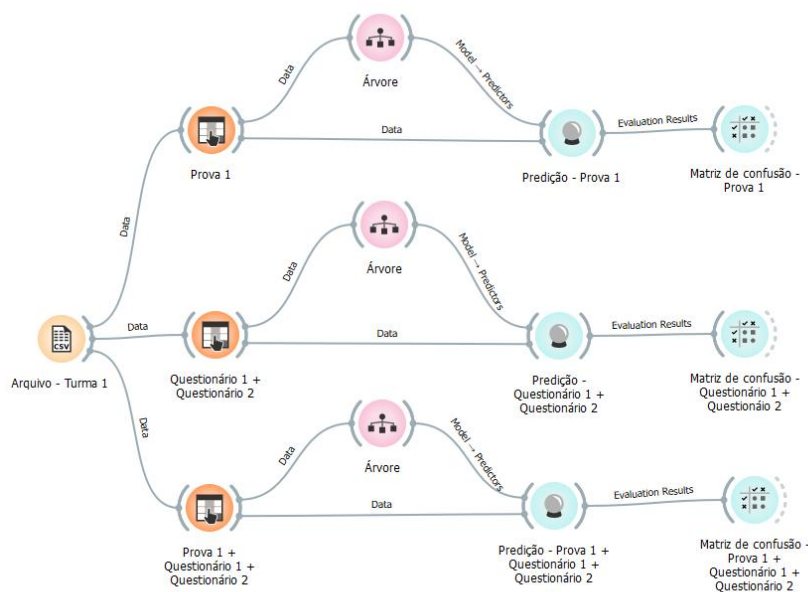
Os valores obtidos são utilizados para discussões no capítulo 5.

### 4.3.9 Simulações em Turmas de Outro Professor(a)

Além das análises realizadas na turma do professor(a) responsável por este estudo, foram conduzidas simulações em uma turma ministrada por outro professor(a). Apesar da compatibilidade na grade curricular, os métodos avaliativos são diferentes, com a aplicação da primeira prova ocorrendo após a realização de dois questionários, quando a turma atinge aproximadamente 30% do conteúdo. Esta abordagem permite uma outra investigação sobre a eficácia das previsões iniciais em relação aos resultados finais.

A Figura 31 apresenta a modelagem geral para esta turma. Observa-se uma divisão entre prova, questionários e prova acrescida de questionários.

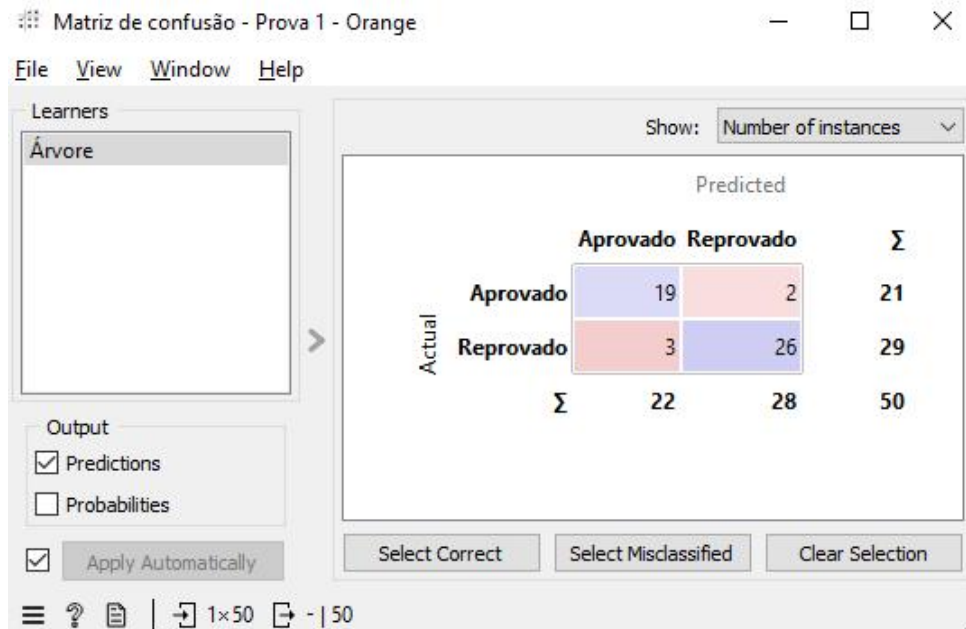
Figura 31 – Modelagem da turma com outra abordagem



Fonte: Autor (2023)

Ao analisar a Figura 32, que destaca o desempenho na primeira prova, é possível avaliar como os alunos lidaram com os conceitos iniciais até o ponto em que a avaliação foi aplicada.

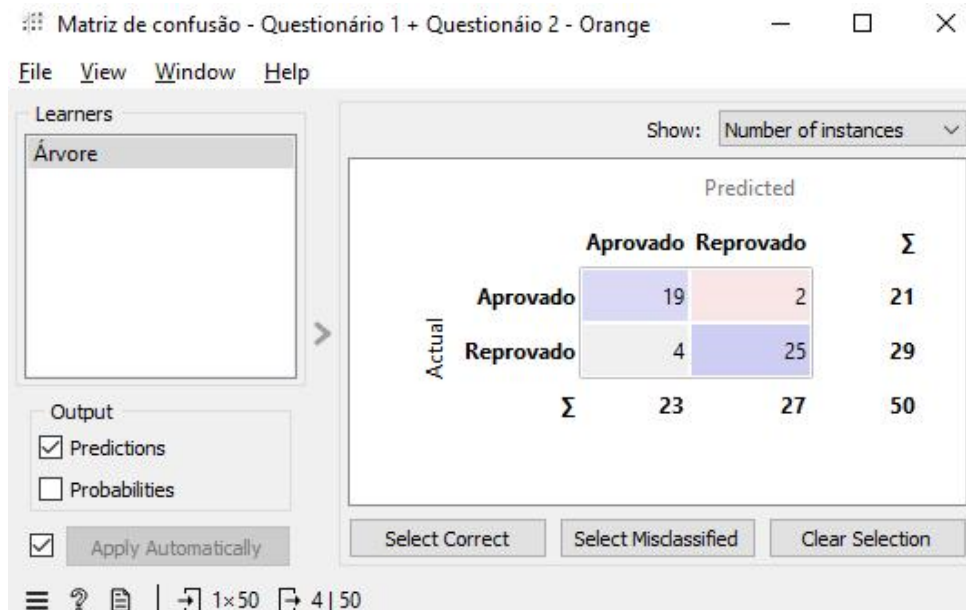
Figura 32 – Desempenho na Turma - Métricas da Prova 1



Fonte: Autor (2023)

A Figura 33 destaca o desempenho combinado dos questionários da matéria.

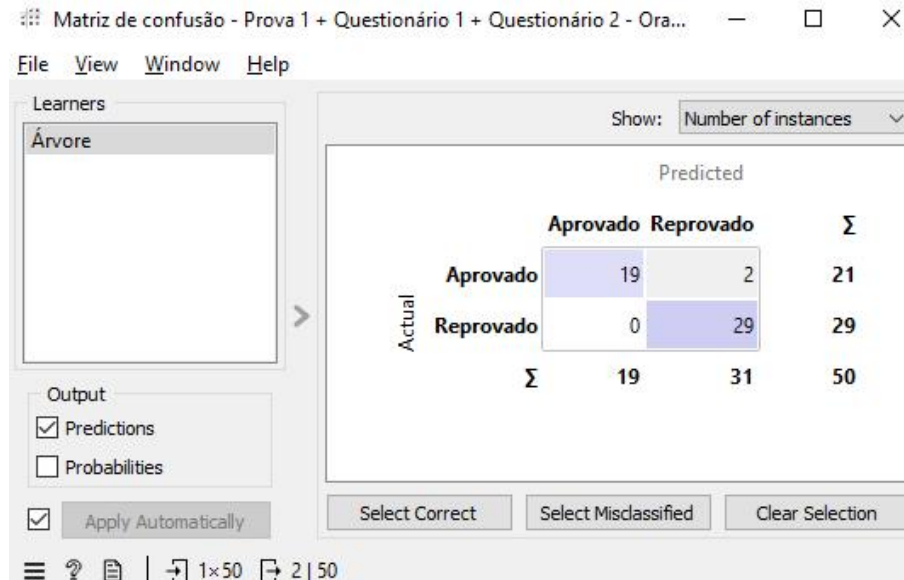
Figura 33 – Desempenho na Turma - Métricas dos Questionários 1 e 2



Fonte: Autor (2023)

Ao integrar os resultados da primeira prova com os questionários, a Figura 34 proporciona uma visão do desempenho dos alunos nesta primeira parte do curso.

Figura 34 – Desempenho na Turma - Métricas dos Questionários 1 e 2 + Prova 1



Fonte: Autor (2023)

Os dados presentes na Figura 34 são utilizados para comparações na Tabela 11.

#### 4.3.9.1 Tabela Comparativa

Além das análises visuais, é útil apresentar uma tabela comparativa das métricas obtidas em cada tipo de avaliação. A Tabela 11 resume as principais métricas de desempenho.

Tabela 11 – Tabela Comparativa - Métricas de desempenho em turma com abordagem diferente

<b>Tipo de Avaliação</b>	<b>Acurácia (%)</b>	<b>Precisão (%)</b>	<b>Sensibilidade (%)</b>
Prova	90	86,36	90,48
Questionários	88	82,61	90,48
Prova + Questionários	96	100	90,48

Fonte: Autor (2023)

A tabela resume as principais métricas de desempenho para cada tipo de avaliação.

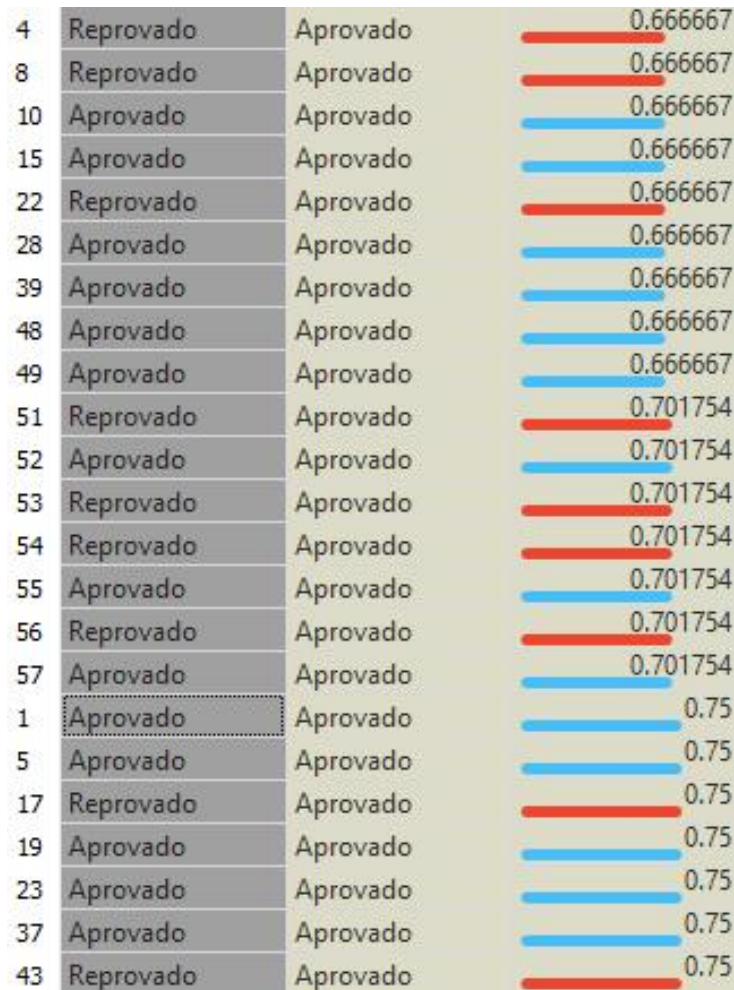
#### 4.3.9.2 Probabilidade de aprovação

Assim como na turma base, nesta também foram realizados cálculos de probabilidade de aprovação que podem ajudar a decidir quais alunos ajudar. Os valores,



obtidos pela análise das provas junto com os questionários, são mostrados na Figura 35.

Figura 35 – Probabilidade de aprovação na turma com outra abordagem



Fonte: Autor (2023)

Os valores obtidos são discutidos no capítulo 5.

#### 4.3.9.3 Utilizando uma segunda turma como teste

Para avaliar o modelo obtido, é utilizada uma segunda turma que realiza as mesmas atividades, para tentar encontrar possíveis individualidades de cada turma. As métricas encontradas são apresentadas na Tabela 12.

Tabela 12 – Tabela Comparativa - Desempenho em outra turma com abordagem diferente

Tipo de Avaliação	Acurácia (%)	Precisão (%)	Sensibilidade (%)
Prova	89,47	90,48	95
Questionários	82,46	82,61	95
Prova + Questionários	87,81	86,67	97,5

Fonte: Autor (2023)

Aqui, encerra-se o capítulo de metodologia. No capítulo 5 são discutidos os resultados encontrados nas simulações.



## 5 RESULTADOS E DISCUSSÕES

Nesta seção, serão apresentados os resultados das análises realizadas anteriormente, seguidos por discussões mais aprofundadas. Essa abordagem visa não apenas verificar se os objetivos propostos foram alcançados, mas também identificar eventuais problemas e possíveis correções.

### 5.1 ANÁLISE DAS VARIÁVEIS

Conforme apresentado em 4.3.1, os resultados indicam que a média das notas dos VPLs apresenta a correlação mais significativa com a nota final, seguida pelas pontuações nos questionários e, por último, pela frequência de presença nas aulas. Essa observação sugere que o desempenho nos VPLs tem um impacto substancial nas notas finais dos alunos.

A tabela de dados, que pode ser visualizada na Figura 15, fornece uma visão detalhada das informações utilizadas nessa análise.

Essas descobertas fornecem uma base para as próximas etapas da pesquisa, permitindo comparações mais aprofundadas.

### 5.2 ANÁLISE DE PRESENÇA

Partindo de 4.3.2, a investigação da presença dos alunos nas aulas, embora menos significativa que os outros, é um aspecto relevante para compreender seu engajamento no curso. Utilizando a matriz de confusão, obtiveram-se as métricas.

No que diz respeito à acurácia de 73,26%, observa-se uma taxa satisfatória de previsões precisas no contexto da presença dos alunos. A sensibilidade de 92,10% destaca a eficácia do modelo em identificar casos reais de alunos presentes, indicando que apenas três dos oitenta e seis alunos foram aprovados sem comparecer a uma quantidade significativa de aulas.

Por fim, a precisão de 63,64% aponta que, das previsões positivas feitas pelo modelo, aproximadamente dois terços estavam corretas. É importante ressaltar que esse resultado pode ser influenciado por alunos que, mesmo desistindo da disciplina, continuam frequentando as aulas para evitar reprovação por frequência insuficiente (FI).

### 5.3 ANÁLISE DOS VPLS

Com base na Tabela 2 apresentada em 4.3.3, observa-se que alguns alunos mantiveram um desempenho consistente em todos os VPLs, evidenciando uma compreensão robusta e aplicação prática dos conceitos abordados nas atividades práticas. Por outro lado, a presença de outliers indica que certos alunos enfrentaram dificuldades específicas em VPLs particulares. Esses resultados sugerem que os VPLs são uma parte fundamental da avaliação dos alunos, contribuindo de maneira significativa para a predição do desempenho global.

### 5.4 ANÁLISE DE QUESTIONÁRIOS

Conforme a Tabela 3 presente em 4.3.4, os resultados indicam um desempenho consistente nos questionários, tanto em análises individuais quanto na média. Isso sugere que os questionários são uma métrica sólida para avaliar o conhecimento teórico dos alunos.

### 5.5 ANÁLISE POR ASSUNTO

Em relação às métricas de desempenho por assunto apresentadas em 4.3.5, nota-se que tanto Fundamentos Básicos quanto Estruturas de Repetição exibem acurácias e sensibilidades semelhantes, com ligeira superioridade nos Fundamentos Básicos em termos de precisão.

Vetores demonstram uma acurácia relativamente alta, embora a precisão seja um pouco inferior. No entanto, eles se destacam pela sensibilidade excepcionalmente alta, atingindo 100%, indicando eficácia do modelo em identificar alunos que dominam este assunto específico.

Matrizes apresentam a maior acurácia e precisão, sinalizando a confiabilidade significativa do modelo na previsão e avaliação do desempenho dos alunos neste tópico crucial. A alta precisão e a baixa sensibilidade sugerem que os alunos que dominam este assunto estão quase unanimemente preparados para a aprovação.

Os tópicos de strings e ponteiros exibem boa acurácia e sensibilidade, com a alta sensibilidade indicando que essas variáveis não são indicadores tão precisos de aprovação.

Por fim, as funções mostram uma acurácia baixa, porém uma precisão notavelmente alta e uma sensibilidade extremamente baixa em comparação com outros tópicos. Essa baixa sensibilidade sugere que o assunto pode ser considerado de maior grau de complexidade, já que muitos alunos previstos como reprovados foram, na verdade, aprovados.

## 5.6 IDENTIFICAÇÃO DE FALSOS NEGATIVOS E FALSOS POSITIVOS E PROBABILIDADE DE APROVAÇÃO

Após analisar os resultados obtidos na identificação de alunos, destacando tanto os falsos negativos (alunos previstos como reprovados, mas que foram aprovados) quanto os falsos positivos (alunos previstos como aprovados, mas que foram reprovados), conforme é mostrado em 4.3.2, surgem alguns pontos a serem discutidos.

A análise dos Falsos Negativos revela casos em que o modelo subestimou o desempenho dos alunos, prevendo sua reprovação, mas esses alunos acabaram sendo aprovados. Um exemplo é o aluno 41, que, apesar de um desempenho abaixo da média em três assuntos, conseguiu ser aprovado. Outros casos, como 9, 55 e 58, apresentam características semelhantes, ou seja, alunos que conseguiram se superar nos momentos finais da disciplina.

Os Falsos Positivos, por sua vez, representam situações em que o modelo previu a aprovação, mas os alunos foram, na verdade, reprovados. Tratando o aluno de número 32 por exemplo, nota-se que esse foi previsto como aprovado em todos os assuntos, mas acabou sendo reprovado. Ao abrir os relatórios gerais, percebe-se que ele obteve desempenhos ruins nas provas que não corresponderam ao desempenho ao longo do curso. Esse caso representa uma exceção ao modelo, ou seja, prever que alguém que foi bem ao longo do semestre será reprovado é difícil. Além dele, também há outros casos de alunos, como 6, 19, 21 e 43, que acabaram falhando nos momentos cruciais, que no caso, seriam as provas.

Além identificar os falsos negativos e falsos positivos, o modelo calcula a probabilidade de aprovação dos alunos. Com esses valores, é possível ter uma ideia de quais alunos poderiam ser aprovados com a ajuda adequada. É importante ressaltar que esses valores foram obtidos com base no assunto Matrizes, que obteve a maior precisão entre os analisados. Ao analisar a porcentagem de aprovação, nota-se que ela é consideravelmente alta para os alunos aprovados, sendo de aproximadamente 96,67%.

Dentre os valores exibidos, destaca-se também os alunos previstos como reprovados, mas que foram aprovados, evidenciando a sensibilidade não tão boa quanto à precisão desse assunto analisado.

De modo geral, cabe ao professor decidir quais parâmetros serão levados em conta para decidir quais alunos deveriam obter maior atenção.

## 5.7 ANÁLISE EM OUTRA TURMA

Na seção 4.3.7 buscou-se analisar uma segunda turma. Essa turma exibiu uma precisão perfeita nos resultados dos questionários, algo que de certa forma era inesperado, não apenas por se diferenciar da primeira mas também por ser improvável

de acontecer, com acurácia, precisão e sensibilidade atingindo 100%. Esses valores sugerem que o aprendizado teórico foi essencial para a aprovação.

Embora tenham ocorrido ligeiras reduções na acurácia, precisão e sensibilidade na segunda turma em comparação com a primeira, os resultados permanecem robustos. Isso indica que a importância dos VPLs são mantidas, embora neste caso seja menor quando comparada com os questionários.

Dentre os fatores observado na análise por assunto, nota-se a alta precisão das estruturas de repetição, atingindo métricas absolutas. Além disso, nota-se também que os assuntos de matrizes são menos precisos quando comparados com a primeira turma, embora seu valor mantenha-se significativo.

## 5.8 COMPARAÇÃO DA SIMULAÇÃO DE RNAS COM ÁRVORES

Após realizar análises dos mais variados tipos utilizando Árvores de Decisão, optou-se por testar as Redes Neurais Artificiais na seção 4.3.8 para identificar eventuais dissonâncias nos valores. Comparando-se somente os questionários, nota-se que para este caso as árvores apresentaram maior precisão. O mesmo é válido para a análise de VPLs, na qual as árvores apresentaram novamente uma maior precisão.

Quanto a simulação por assuntos, nota-se que matrizes novamente foi o assunto com maior precisão, indicando sua generalização independente da técnica utilizada. O oposto também é válido, visto que estruturas de repetição foi o assunto menos preciso. De forma geral, as árvores apresentaram melhores resultados do que as RNAs, e uma possível explicação para tal pode ser os parâmetros utilizados.

## 5.9 ANÁLISES NA TURMA DE OUTRO PROFESSOR

Por fim, ao estender as investigações para uma turma com abordagem diferenciada, na qual a primeira prova ocorre logo no início, representando cerca de 30% do curso, antes de iniciar a realização dos VPLs, destacam-se alguns pontos relevantes.

De acordo com o exposto em 4.3.9, na primeira turma, tanto a prova quanto os questionários apresentam acurácias próximas e sensibilidades idênticas, indicando que, de maneira geral, ambos oferecem boas previsões. Já na segunda, a prova apresentou significativamente maior acurácia e precisão.

No que diz respeito à precisão, observa-se que a prova é um indicativo mais robusto do desempenho final dos estudantes em comparação aos questionários em ambas as turmas. Além disso, constata-se também na primeira turma que os questionários somados à prova atingiram uma precisão de 100%.

Também foram realizados cálculos de probabilidade de aprovação. Nota-se que a parte central, ou seja, aqueles em que os alunos estão com notas aproximadamente

na média, é difícil de prever quais serão realmente aprovados. Quanto a isso, conclui-se que o modelo é eficiente para identificar os casos extremos mas pode não apresentar análises precisas quando analisados os casos que estão na média devido ao alto grau de complexidade dessa análise.

Para concluir, é interessante notar que essas informações foram obtidas considerando apenas cerca de 30% do conteúdo do curso. Mesmo no início do semestre, já era possível ter uma noção significativa de quantos alunos poderiam ser aprovados.

Essas observações ressaltam a importância e a eficácia das avaliações, especialmente das provas, como indicadores confiáveis do desempenho dos alunos, mesmo em fases iniciais do curso.

## 5.10 DISCUSSÃO GERAL DOS RESULTADOS

Apresentadas todas as análises realizadas, juntamente com os resultados obtidos, chega-se a alguns resultados globais. Dentre esses resultados, destaca-se que, apesar de variações entre assuntos, desde o começo da disciplina tem-se fortes indícios de quais alunos serão aprovados e quais serão reprovados. Quanto às identificações e cálculos de probabilidade de aprovação, nota-se que são precisos para casos extremos, mas podem apresentar equívocos em casos medianos devido ao alto grau de complexidade da análise.

A comparação entre turmas revela tanto consistências quanto variações, destacando a importância de considerar as particularidades de cada grupo de estudantes. A introdução de RNAs como alternativa às Árvores de Decisão revela diferenças nos resultados, ressaltando a influência dos parâmetros na precisão das previsões, embora essa diferença não seja tão significativa. Além disso, a análise na turma de outro professor, que segue outra metodologia, reforça a importância das avaliações, especialmente das provas, como indicadores confiáveis do desempenho dos alunos, mesmo nas fases iniciais do curso.

## 6 CONCLUSÕES

Ao longo deste trabalho, foram exploradas as relações entre as notas obtidas em atividades avaliativas ao longo do semestre e o desempenho final dos alunos em turmas da disciplina de Programação I da Universidade Federal de Santa Catarina.

Além disso, a análise detalhada dos dados coletados não apenas destaca áreas específicas passíveis de aprimoramento, mas também fornece uma base sólida para promover melhorias contínuas com base em evidências concretas.

Dentre as análises realizadas, a análise de presença destaca a relevância do engajamento dos alunos, já a análise por assunto oferece uma visão aprofundada do desempenho dos alunos em diferentes áreas do conteúdo do curso. Além dessas análises, a comparação entre turmas destaca a importância de considerar contextos específicos, seja entre turmas com professores diferentes ou até mesmo de um único professor. Vale também destacar a possibilidade de trabalhar com diferentes técnicas de mineração de dados, aumentando o conhecimento na área.

Em síntese, os resultados obtidos neste trabalho marcam a importância da análise de dados educacionais como uma ferramenta poderosa para compreender, aprimorar e inovar nos processos de ensino e aprendizagem.

Diante disso, é evidente que a análise de dados educacionais, quando conduzida de maneira criteriosa, pode ser uma ferramenta poderosa para compreender, aprimorar e inovar nos processos de ensino e aprendizagem.

Para trabalhos futuros, sugere-se a inclusão de outras variáveis que possam afetar o desempenho acadêmico dos estudantes, como a participação em monitorias, por exemplo. Além disso, é recomendado aumentar o número de alunos analisados, a fim de proporcionar resultados mais abrangentes e que sirvam como base para melhorar o desempenho do maior número possível de alunos.

## REFERÊNCIAS

CLOW, D. An overview of learning analytics. **Teaching in Higher Education**, Taylor & Francis, 2013.

Educational Data Mining. **Educational Data Mining**. 2023. Disponível em: <https://educationaldatamining.org/>. Acesso em: 02 set. 2023.

HATTIE, J.; TIMPERLEY, H. The power of feedback. **Review of educational research**, v. 77, n. 1, p. 81–112, 2007.

IBM. **Supervised vs. Unsupervised Learning: What's the Difference?** 2023. Disponível em: <https://www.ibm.com/blog/supervised-vs-unsupervised-learning/>. Acesso em: 04 set. 2023.

IBM. **What is data science?** 2023. Disponível em: <https://www.ibm.com/topics/data-science>. Acesso em: 06 set. 2023.

JAMES, G. et al. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112.

KADOIĆ, N.; OREŠKI, D. **Analysis of Student Behavior and Success Based on Logs in Moodle**. IEEE, 2018. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8400123>.

KUIK, K. **Active data vs Passive data**. 2018. Disponível em: <https://www.lab.one/insights/active-data-vs-passive-data>. Acesso em: 15 set. 2023.

LILE, A. Analyzing e-learning systems using educational data mining techniques. **Mediterranean Journal of Social Sciences**, v. 2, n. 3, p. 403–419, 2011.

LOH, W.-Y. Classification and regression trees. **Wiley interdisciplinary reviews: data mining and knowledge discovery**, Wiley Online Library, v. 1, n. 1, p. 14–23, 2011.

MOODLE. **Moodle plugins**. 2023. Disponível em: <https://moodle.org/plugins/>. Acesso em: 15 set. 2023.

NIELSEN, M. A. **Neural networks and deep learning**. [S.l.]: Determination press San Francisco, CA, USA, 2015. v. 25.

ORANGE. **Orange**. 2023. Disponível em: <https://orangedatamining.com/>. Acesso em: 15 set. 2023.

ROMERO, C.; VENTURA, S. **Educational data mining: A survey from 1995 to 2005**. Elsevier, 2007. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417406001266>.

ROMERO, C.; VENTURA, S.; GARCÍA, E. **Data mining in course management systems: Moodle case study and tutorial**. Elsevier, 2008. Disponível em: [https://www.sciencedirect.com/science/article/pii/S0360131507000590?casa\\_token=kKWl4s3rggYAAAAA:wPaiyO-By1fxU1t0mXf1OHyQvo9tB3L8LZ5s-VR\\_-HN\\_cRWWNI9Of7QuJRJcm-xXh6BLRM9z-l3e](https://www.sciencedirect.com/science/article/pii/S0360131507000590?casa_token=kKWl4s3rggYAAAAA:wPaiyO-By1fxU1t0mXf1OHyQvo9tB3L8LZ5s-VR_-HN_cRWWNI9Of7QuJRJcm-xXh6BLRM9z-l3e).

SHEARER, C. The crisp-dm model: the new blueprint for data mining. **Journal of data warehousing**, THE DATA WAREHOUSE INSTITUTE, v. 5, n. 4, p. 13–22, 2000.

SHRESTHA, S.; POKHAREL, M. Educational data mining in moodle data. **International Journal of Informatics and Communication Technology (IJ-ICT)**, v. 10, n. 1, p. 9–18, 2021.

TSAI, Y.-S. **What is Learning Analytics?** 2023. Disponível em: [https://www.solaresearch.org/about/what-is-learning-analytics/#:~:text=As%20a%20research%20and%20teaching,%2DCentered%20Design%20\(e.g.%20usability%2C](https://www.solaresearch.org/about/what-is-learning-analytics/#:~:text=As%20a%20research%20and%20teaching,%2DCentered%20Design%20(e.g.%20usability%2C). Acesso em: 02 set. 2023.

XU, D.; TIAN, Y. A comprehensive survey of clustering algorithms. **Annals of Data Science**, Springer, p. 165–193, 2015.