

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ  
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO  
ENGENHARIA DE COMPUTAÇÃO

LUCAS RICARDO DE LIMA FRANCO

**ANÁLISE DE DADOS ACADÊMICOS E CONSTRUÇÃO DE  
PLATAFORMA PARA APOIO A REDUÇÃO DA RETENÇÃO NO  
ENSINO SUPERIOR**

TRABALHO DE CONCLUSÃO DE CURSO

CORNÉLIO PROCÓPIO  
2021

LUCAS RICARDO DE LIMA FRANCO

# **ANÁLISE DE DADOS ACADÊMICOS E CONSTRUÇÃO DE PLATAFORMA PARA APOIO A REDUÇÃO DA RETENÇÃO NO ENSINO SUPERIOR**

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Computação da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de Bacharel.

Orientador: Prof. Me. Francisco Pereira Junior  
Universidade Tecnológica Federal do Paraná

CORNÉLIO PROCÓPIO  
2021



4.0 Internacional

Esta licença permite remixe, adaptação e criação a partir do trabalho, para fins não comerciais, desde que sejam atribuídos créditos ao(s) autor(es) e que licenciem as novas criações sob termos idênticos. Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.



Ministério da Educação  
**Universidade Tecnológica Federal do Paraná**  
Câmpus Cornélio Procópio  
Departamento de Computação  
Engenharia de Computação



---

## **TERMO DE APROVAÇÃO**

### **ANÁLISE DE DADOS ACADÊMICOS E CONSTRUÇÃO DE PLATAFORMA PARA APOIO A REDUÇÃO DA RETENÇÃO NO ENSINO SUPERIOR**

**por**

**Lucas Ricardo de Lima Franco**

Este Trabalho de Conclusão de Curso de graduação foi julgado adequado para obtenção do Título de Bacharel em Engenharia de Computação e aprovado em sua forma final pelo Programa de Graduação em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Cornélio Procópio, 17/05/2021

---

Prof. Me. Francisco Pereira Junior

---

Prof. Dr. Alexandre Rômolo Moreira Feitosa

---

Prof. Dr. Claiton de Oliveira

“A Folha de Aprovação assinada encontra-se na Coordenação do Curso”

## **AGRADECIMENTOS**

Primeiramente agradeço a Deus por ter me dado a oportunidade e vigor para chegar até aqui.

Agradeço ao professor orientador Francisco Pereira Junior (Thesko), por todo o apoio e empenho dedicado para a realização deste trabalho e a todos os professores por todas as contribuições dadas, que reunidas fortaleceram muito mais do que o profissional, mas também na minha formação humana.

Agradeço também à toda a minha família, principalmente meus pais, meus irmãos e avós, por todo o incentivo, hoje percebo que sem eles nada disso seria possível.

E deixo aqui registrado a importância dos meus amigos em todos os momentos dessa jornada, pois percebo também o quanto cresci e aprendi com eles, e quanto o coletivo foi importante para a construção do indivíduo que cada um de nós somos.

Por fim agradeço a todos que direta ou indiretamente fizeram parte desta jornada.

*“Não vemos as coisas como elas são, mas como  
nós somos.” (Anais Nin)*

## RESUMO

FRANCO, Lucas R. L.. Análise de dados acadêmicos e construção de plataforma para apoio a redução da retenção no ensino superior. 2021. 79 f. Trabalho de Conclusão de Curso – Engenharia de Computação, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2021.

A retenção no ensino superior é um fenômeno complexo que causa grandes impactos na sociedade, devido a todos os fatores que estão intrinsecamente relacionados com sua ocorrência, como o financeiro, educacional e social. Este trabalho tem a retenção como objeto de estudo e apresenta uma análise de dados acadêmicos utilizando técnicas de mineração de dados e as etapas de construção de uma plataforma para tornar as análises mais acessíveis. Em conjunto foram implementados mecanismos automáticos para comunicação com alunos e artifícios para que docentes possam extrair *insights* e tomar ações a fim de contribuir para a redução da retenção acadêmica. Com este trabalho pôde-se verificar a influência parcial de algumas variáveis no desempenho dos alunos, como a frequência e a realização das atividades sugeridas pelos professores. Além disso, modelos de aprendizagem de máquina foram treinados considerando diferentes cenários, e dentre os algoritmos utilizados, para o conjunto de dados disponível o *Random Forest* apresentou os melhores resultados. Para os cenários mais comuns obteve-se acurácias entre 64,5% e 95,4%, dependendo do momento do período letivo no qual as previsões foram realizadas.

**Palavras-chave:** Mineração de dados. Ensino superior. Reconhecimento de padrões. Retenção acadêmica. Sistema acadêmico.

## ABSTRACT

FRANCO, Lucas R. L.. Analysis of academic data and implementation of a platform to support retention reduction in higher education. 2021. 79 f. Trabalho de Conclusão de Curso – Engenharia de Computação, Universidade Tecnológica Federal do Paraná. Cornélio Procópio, 2021.

Retention in higher education is a complex phenomenon that causes major impacts on society, due to all the factors that are intrinsically related to its occurrence, such as financial, educational and social. This monograph has retention as the object of study and presents an analysis of academic data using data mining techniques and the steps of building a platform to make the analysis more accessible. Moreover, automatic mechanisms for communication with students and devices are implemented so that teachers can extract *insights* and take actions in order to contribute to the reduction of academic retention. With this work it was possible to check the partial influence of some variables on the students' performance, such as the frequency and the performance of the activities suggested by the teachers. In addition, machine learning models were trained considering different scenarios, and among the algorithms used, for the available data set the *Random Forest* algorithm presented the best results. For the most common scenarios, accuracy between 64.5% and 95.4% was obtained, depending on the time of the school year in which the predictions were made.

**Keywords:** Data mining. Higher education. Pattern recognition. Academic platform.

## LISTA DE ABREVIATURAS E SIGLAS

|          |                                                                        |
|----------|------------------------------------------------------------------------|
| ACID     | Atomicidade, Consistência, Isolamento e Durabilidade                   |
| AVA      | Ambiente Virtual de Aprendizagem                                       |
| BD       | Banco de Dados                                                         |
| CSS      | <i>Cascading Style Sheets</i>                                          |
| CRISP-DM | <i>Cross-Industry Standard Process for Data Mining</i>                 |
| CRUB     | Conselho de Reitores das Universidades Brasileiras                     |
| CSV      | <i>Comma Separated Values</i>                                          |
| HTML     | <i>Hypertext Markup Language</i>                                       |
| IES      | Instituições de Ensino Superior                                        |
| IFES     | Instituições Federais de Ensino Superior                               |
| INEP     | Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira |
| IoT      | <i>Internet of Things</i>                                              |
| JSON     | <i>JavaScript Object Notation</i>                                      |
| KDD      | <i>Knowledge Discovery in Databases</i>                                |
| MD       | Mineração de Dados                                                     |
| ML       | <i>Machine Learning</i>                                                |
| Moodle   | <i>Modular Object-Oriented Dynamic Learning Environment</i>            |
| RA       | Registro Acadêmico                                                     |
| SPA      | <i>Single Page Application</i>                                         |
| UFRGS    | Universidade Federal do Rio Grande do Sul                              |
| URL      | <i>Uniform Resource Locator</i>                                        |
| USP      | Universidade de São Paulo                                              |
| UTFPR    | Universidade Tecnológica Federal do Paraná                             |
| WEKA     | <i>Waikato Environment for Knowledge Analysis</i>                      |
| XML      | <i>eXtensible Markup Language</i>                                      |



## LISTA DE FIGURAS

|                                                                                                                                                                      |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figura 1 – Fatores de destaque em relação à retenção e à evasão, segundo estudantes e professores do curso de Odontologia da UFRGS . . . . .                         | 16 |
| Figura 2 – Exemplo de regressão linear com a variável 'Salários Mínimos' em função da variável 'Anos de Experiências' . . . . .                                      | 18 |
| Figura 3 – Exemplo simplificado de árvore de decisão, com a classificação de faixa etária por meio da idade . . . . .                                                | 19 |
| Figura 4 – Exemplo simplificado de <i>Random Forest</i> utilizada para classificação . . . . .                                                                       | 20 |
| Figura 5 – Método CRISP-DM de modelo de processo para mineração de dados . . . . .                                                                                   | 21 |
| Figura 6 – <i>Trello</i> com as colunas e atividades planejadas para o trabalho . . . . .                                                                            | 30 |
| Figura 7 – Detalhes da atividade de implementação das <i>APIs</i> no <i>Trello</i> . . . . .                                                                         | 30 |
| Figura 8 – Subtarefas da atividade de implementação de <i>APIs</i> no <i>Trello</i> . . . . .                                                                        | 31 |
| Figura 9 – Modelo de planilha com notas exportada pelo Moodle . . . . .                                                                                              | 33 |
| Figura 10 – Modelo de planilha com faltas exportada do sistema acadêmico da UTFPR . . . . .                                                                          | 33 |
| Figura 11 – Modelo alternativo de planilha com faltas exportada pelo sistema acadêmico da UTFPR . . . . .                                                            | 34 |
| Figura 12 – Planilha de dados com notas e faltas separadas por abas . . . . .                                                                                        | 34 |
| Figura 13 – Modelo de planilha com dados exportada pela plataforma URI . . . . .                                                                                     | 35 |
| Figura 14 – Gráfico de dispersão com as frequências e notas dos alunos, considerando a realização dos trabalhos . . . . .                                            | 36 |
| Figura 15 – Histograma com as quantidades de alunos por faixa de notas . . . . .                                                                                     | 37 |
| Figura 16 – Histograma com as quantidades de alunos que realizaram 75% ou mais dos exercícios do URI, segmentando pela porcentagem realizada de trabalhos . . . . .  | 38 |
| Figura 17 – Histograma com as quantidades de alunos que realizaram menos de 75% dos exercícios do URI, segmentando pela porcentagem realizada de trabalhos . . . . . | 38 |
| Figura 18 – Fragmento do conjunto de dados de notas no formato pós preparação . . . . .                                                                              | 40 |
| Figura 19 – Fragmento do conjunto de dados de faltas no formato pós preparação . . . . .                                                                             | 41 |
| Figura 20 – Fragmento do conjunto de dados obtidos na plataforma URI no formato pós preparação . . . . .                                                             | 42 |
| Figura 21 – Fluxo de interações entre os componentes do sistema no envio de dados . . . . .                                                                          | 45 |
| Figura 22 – Fluxo de interações entre os componentes do sistema em requisições solicitando dados e previsões . . . . .                                               | 46 |
| Figura 23 – Fragmento superior do dashboard . . . . .                                                                                                                | 47 |
| Figura 24 – Fragmento inferior do dashboard . . . . .                                                                                                                | 48 |
| Figura 25 – Página de configurações . . . . .                                                                                                                        | 49 |
| Figura 26 – Página de configurações com campos expandidos . . . . .                                                                                                  | 50 |
| Figura 27 – Página de upload de planilhas . . . . .                                                                                                                  | 51 |

|                                                                                                                                                 |    |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figura 28 – Página de upload de planilhas com formatos para upload expandidos . . . . .                                                         | 51 |
| Figura 29 – Detecção de alunos por meio dos mecanismos de comunicação . . . . .                                                                 | 52 |
| Figura 30 – Exemplo de e-mail enviado à um aluno pelo sistema de predição . . . . .                                                             | 52 |
| Figura 31 – Página de envio de justificativa, com a descrição da ocorrência . . . . .                                                           | 53 |
| Figura 32 – Página de listagem de dados . . . . .                                                                                               | 54 |
| Figura 33 – Página de listagem de dados com as predições efetuadas . . . . .                                                                    | 55 |
| Figura 34 – Página de detalhes do estudante . . . . .                                                                                           | 56 |
| Figura 35 – Mensagem personalizada a ser enviada no e-mail por meio da página de<br>detalhes do estudante . . . . .                             | 57 |
| Figura 36 – E-mail enviado por meio da página de detalhes do estudante . . . . .                                                                | 58 |
| Figura 37 – Lista de feedbacks disponível na página de detalhes do estudante . . . . .                                                          | 58 |
| Figura 38 – Gráfico de dispersão com as frequências e notas dos alunos, considerando<br>somente a turma analisada com o MOSS . . . . .          | 67 |
| Figura 39 – Gráfico de barras com as notas médias para cada aluno, considerando<br>somente a turma analisada com o MOSS . . . . .               | 68 |
| Figura 40 – Gráfico de barras com as similaridades acumuladas para cada aluno, consi-<br>derando somente a turma analisada com o MOSS . . . . . | 68 |

## LISTA DE TABELAS

|                                                                                                                                                                                  |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Tabela 1 – Importância das <i>features</i> para o modelo treinado com <i>Gradient Boosting</i> .                                                                                 | 43 |
| Tabela 2 – Importância das <i>features</i> para o modelo treinado com <i>Random Forest</i> . . .                                                                                 | 44 |
| Tabela 3 – Cenários considerados para os modelos de aprendizado de máquina treinados.                                                                                            | 44 |
| Tabela 4 – Score e desvio padrão dos modelos construídos considerando os cenários selecionados . . . . .                                                                         | 60 |
| Tabela 5 – Importância das <i>features</i> considerando o cenário A para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . .  | 61 |
| Tabela 6 – Importância das <i>features</i> considerando o cenário B para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . .  | 62 |
| Tabela 7 – Importância das <i>features</i> considerando o cenário F para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . .  | 62 |
| Tabela 8 – Importância das <i>features</i> considerando o cenário H para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . .  | 62 |
| Tabela 9 – Importância das <i>features</i> considerando o cenário J para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . .  | 63 |
| Tabela 10 – Importância das <i>features</i> considerando o cenário A para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . . | 77 |
| Tabela 11 – Importância das <i>features</i> considerando o cenário B para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . . | 77 |
| Tabela 12 – Importância das <i>features</i> considerando o cenário C para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . . | 77 |
| Tabela 13 – Importância das <i>features</i> considerando o cenário D para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . . | 78 |
| Tabela 14 – Importância das <i>features</i> considerando o cenário E para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . . | 78 |
| Tabela 15 – Importância das <i>features</i> considerando o cenário F para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . . | 78 |
| Tabela 16 – Importância das <i>features</i> considerando o cenário G para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . . | 78 |
| Tabela 17 – Importância das <i>features</i> considerando o cenário H para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . . | 79 |
| Tabela 18 – Importância das <i>features</i> considerando o cenário I para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . . | 79 |
| Tabela 19 – Importância das <i>features</i> considerando o cenário J para <i>Gradient Boosting</i> e <i>Random Forest</i> e coeficientes lineares para Regressão Linear. . . . . | 79 |

## SUMÁRIO

|                                          |           |
|------------------------------------------|-----------|
| <b>1 – INTRODUÇÃO</b>                    | <b>12</b> |
| 1.1 PROBLEMA                             | 13        |
| 1.2 JUSTIFICATIVA                        | 13        |
| 1.3 OBJETIVOS                            | 13        |
| 1.3.1 Objetivos Específicos              | 13        |
| 1.4 ORGANIZAÇÃO DO TRABALHO              | 14        |
| <b>2 – FUNDAMENTAÇÃO TEÓRICA</b>         | <b>15</b> |
| 2.1 EVASÃO E RETENÇÃO NO ENSINO SUPERIOR | 15        |
| 2.2 MINERAÇÃO DE DADOS                   | 17        |
| 2.2.1 Ferramentas Estatísticas           | 17        |
| 2.2.2 Regressão Linear                   | 18        |
| 2.2.3 Árvore de Decisão                  | 18        |
| 2.2.4 <i>Random Forest</i>               | 19        |
| 2.2.5 <i>Gradient Boosting</i>           | 20        |
| 2.3 CRISP-DM                             | 20        |
| 2.3.1 Entendimento do Negócio            | 21        |
| 2.3.2 Entendimento dos Dados             | 22        |
| 2.3.3 Preparação dos Dados               | 22        |
| 2.3.4 Modelagem                          | 22        |
| 2.3.5 Validação                          | 23        |
| 2.3.6 Implementação                      | 23        |
| 2.4 VISUALIZAÇÃO DE DADOS                | 23        |
| 2.5 TRABALHOS RELACIONADOS               | 24        |
| <b>3 – DESENVOLVIMENTO</b>               | <b>27</b> |
| 3.1 TECNOLOGIAS E FERRAMENTAS            | 27        |
| 3.2 ENTENDIMENTO DO NEGÓCIO              | 31        |
| 3.3 ENTENDIMENTO DOS DADOS               | 32        |
| 3.4 PREPARAÇÃO DOS DADOS                 | 38        |
| 3.5 MODELAGEM                            | 42        |
| 3.6 IMPLEMENTAÇÃO                        | 44        |
| 3.6.1 Dashboard                          | 46        |
| 3.6.2 Configurações                      | 48        |
| 3.6.3 <i>Upload</i> de Dados             | 49        |
| 3.6.4 Feedback                           | 53        |

|          |                                                                                                                       |               |
|----------|-----------------------------------------------------------------------------------------------------------------------|---------------|
| 3.6.5    | Lista . . . . .                                                                                                       | 54            |
| 3.6.6    | Detalhes do Estudante . . . . .                                                                                       | 56            |
| <b>4</b> | <b>– ANÁLISE DOS RESULTADOS . . . . .</b>                                                                             | <b>59</b>     |
| 4.1      | MODELOS . . . . .                                                                                                     | 59            |
| 4.1.1    | Acurácia . . . . .                                                                                                    | 59            |
| 4.1.2    | Importância . . . . .                                                                                                 | 61            |
| 4.2      | CONSIDERAÇÕES GERAIS SOBRE A PLATAFORMA . . . . .                                                                     | 63            |
| 4.2.1    | Dashboard . . . . .                                                                                                   | 64            |
| 4.2.2    | Predição, Feedback e Mecanismos de Detecção . . . . .                                                                 | 64            |
| 4.3      | MOSS - Similaridades . . . . .                                                                                        | 65            |
| <b>5</b> | <b>– CONSIDERAÇÕES FINAIS . . . . .</b>                                                                               | <b>70</b>     |
| 5.1      | Trabalhos Futuros . . . . .                                                                                           | 71            |
|          | <b>Referências . . . . .</b>                                                                                          | <b>72</b>     |
|          | <br><b>Apêndices . . . . .</b>                                                                                        | <br><b>75</b> |
|          | <b>APÊNDICE A–Scores e Desvios Padrão dos modelos construídos para todos os cenários considerados . . . . .</b>       | <b>76</b>     |
|          | <b>APÊNDICE B–Importâncias e Coeficientes lineares para os modelos de aprendizagem de máquina treinados . . . . .</b> | <b>77</b>     |

## 1 INTRODUÇÃO

A evasão e a retenção acadêmica no ensino superior são fenômenos que possuem várias vertentes e trazem muitas discussões. Isto porque envolvem fatores que não estão somente relacionados ao ensino e à educação, mas à outros tópicos importantes para a sociedade, como o viés econômico. Segundo [Silva Filho et al. \(2007\)](#), fundadores do Instituto Lobo, que tem como objetivo contribuir com soluções para os problemas brasileiros relacionados às áreas de educação, ciência e tecnologia, o desperdício de recursos devido a evasão tanto em Instituições de Ensino Superior (IES) públicas quanto privadas somam milhões de reais anualmente, e mesmo que seja um tema discutido, não recebe a devida importância e ações necessárias.

[Silva Filho et al. \(2007\)](#) apresentam também, baseado em estudos nacionais e internacionais feitos pelo Instituto Lobo, uma lista de ações consideradas bem-sucedidas para o combate e redução da evasão nas IES brasileiras. A seguir são apresentadas algumas destas ações.

1. Análise estatística do fenômeno da evasão.
2. Criação de programas de aconselhamento e orientação de alunos.
3. Envolvimento da comunidade acadêmica no combate à evasão estimulando a visão centrada no aluno.
4. Criação de condições que satisfaçam os objetivos responsáveis por atrair os alunos às IES e apresentar casos positivos de satisfação.
5. Modernização da forma como são ministrados os cursos.
6. Introdução de atividades relacionadas à profissões desde o início do curso.
7. Estímulo para o desenvolvimento de atitudes inovadoras e empreendedoras pelos estudantes.

Além das ações consideradas convencionais apresentadas, existem também ações tecnológicas, que podem ser aplicadas a fim de torná-las mais eficientes, como por exemplo a implementação de Sistemas Inteligentes. Desenvolvidos respeitando as especificidades de cada instituição, estes sistemas podem auxiliar no acompanhamento do desempenho de alunos, professores e cursos, possibilitando a identificação de situações "críticas", como a de alunos que apresentam grande possibilidade de reprovação ou evasão, dificuldades pedagógicas por conta de professores, disciplinas com altos índices de reprovação e os possíveis motivos para estas ocorrências.

Das ações listadas anteriormente, a número 1 é a que mais se enquadra ao escopo deste trabalho, que visou, a partir da análise de dados acadêmicos, compreender a retenção no ensino superior, descrever estratégias que possibilitem prever a situação de alunos e dar suporte para que profissionais como: pedagogos, assistentes estudantis e professores, usufruam e atuem a partir delas.

## 1.1 PROBLEMA

Evidencia-se em números a problemática que se inserem as IES brasileiras quando trata-se da evasão e da retenção. Temas delicados devido a natureza subjetiva dos motivos que levam ao acontecimento, os índices de desistência e reprovação no Ensino Superior cresceu consideravelmente nos últimos anos, apresentando um aumento de mais de 37% entre os anos de 2010 e 2014, segundo o Censo da Educação Superior 2015, divulgado pelo [INEP \(2016\)](#).

Segundo [Santos et al. \(2018\)](#), o número de vagas ociosas nas instituições federais brasileiras é de mais de 25% do total de vagas ofertadas, devido à evasão e ao não preenchimento destas em processos seletivos. Nas instituições estaduais estes números aproximam-se de 17%.

[Silva Filho et al. \(2007\)](#) argumentam em seu trabalho que essa ociosidade de vagas na educação superior gera, além da perda da qualificação profissional dos estudantes, impactos econômicos negativos para a nação como um todo, uma vez que, os recursos públicos que foram aplicados para a criação destas vagas não surtiram um retorno correspondente.

Um outro problema, agora relacionado à dados, destacado por [Marques e Nicoletti \(2015\)](#), é a possibilidade do contexto de dados acadêmicos envolver um considerável volume de dados, uma vez que este pode ser escalável e evoluir para diferentes instituições de ensino. Além do volume, tem-se a substancial variedade dos dados, já que cada docente possui certa autonomia para estruturar as disciplinas vinculadas da forma que lhe convém, considerando ainda diferentes instituições de ensino, com diferentes padrões de avaliação.

## 1.2 JUSTIFICATIVA

Devido a atual situação das IES em relação as taxas de retenção, se faz necessário a criação de ferramentas que auxiliem as ações relacionadas à retenção em cursos de graduação, reduzindo estes números. Além de traçar perfis que permitam compreender o comportamento dos alunos e prever possíveis situações a partir da análise de dados acadêmicos prévios.

## 1.3 OBJETIVOS

O objetivo geral deste trabalho foi o desenvolvimento de uma plataforma web utilizando técnicas de mineração de dados, para realização de análises de dados acadêmicos, possibilitando aos usuários extrair *insights* e construir mecanismos de apoio à redução da retenção no ensino superior.

### 1.3.1 Objetivos Específicos

- Construção de gráficos para fácil apresentação e análise de dados;
- Implementação de algoritmos de aprendizado de máquina e reconhecimento de padrões que atuem a fim de identificar e prever a retenção em cursos de graduação;

- Desenvolvimento de um sistema web para que docentes tenham acesso aos dados acadêmicos, gráficos e mecanismos de comunicação com alunos;
- Desenvolvimento de mecanismos automatizados que possibilitem a comunicação com alunos em determinadas situações, como a situação de risco de reprovação por faltas em determinada disciplina.

#### 1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho é dividido em cinco capítulos. O capítulo 2 contém a fundamentação teórica, apresentando a contextualização de temas pertinentes ao trabalho e a revisão da literatura realizada. O capítulo 3 descreve o desenvolvimento realizado no trabalho, destacando as tecnologias e ferramentas utilizadas, as etapas executadas com o processo metodológico utilizado e o detalhamento da implementação efetuada, apresentando cada uma das partes do sistema desenvolvido. O capítulo 4 apresenta a análise dos resultados, levando em conta os modelos de predição e o sistema implementado. E o capítulo 5 contém as considerações finais apresentadas para este trabalho.



## 2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção são apresentados os conceitos básicos abordados neste trabalho. Inicialmente serão contextualizadas a evasão e a retenção acadêmica. Nas seções seguintes são abordados os conceitos de mineração de dados; o CRISP-DM – modelo de processo utilizado em projetos de mineração de dados; e, de visualização de dados.

### 2.1 EVASÃO E RETENÇÃO NO ENSINO SUPERIOR

A retenção no ensino superior é o objeto principal da aplicação deste trabalho, porém devido a relação intrínseca que esta possui com a evasão acadêmica, ambos serão melhor contextualizados.

Vale destacar também que o objeto de estudo deste trabalho envolveu questões relacionadas à reprovação, o que conceitualmente não é o mesmo que a retenção acadêmica, mas está diretamente relacionado à ela. Desta forma o termo retenção é utilizado em todo o trabalho podendo também se referir à reprovação.

A evasão e a retenção acadêmica são temas muito discutidos devido a importância e aos impactos sociais e econômicos que trazem consigo.

Em 1995 ocorreu o Seminário sobre evasão nas Universidades Brasileiras, evento promovido pelo Conselho de Reitores das Universidades Brasileiras (CRUB), que divulgou indicadores que apontavam uma taxa de evasão média consideravelmente elevada para as Instituições Federais de Ensino Superior (IFES) à época. A partir do evento constituiu-se uma comissão especial para estudos sobre a evasão nas IES públicas (PEREIRA, 2003; SILVA, 2012).

Esta comissão definiu os conceitos básicos que caracterizam a evasão, distinguindo-as da seguinte forma (PEREIRA, 2003; SILVA, 2012):

- **Evasão de curso:** ocorre quando o aluno desliga-se do curso superior, podendo caracterizar-se por abandono (deixar de efetuar a matrícula), desistência, transferência ou reopção de curso, trancamento ou exclusão por norma institucional.
- **Evasão da instituição:** ocorre quando o estudante desliga-se da instituição na qual está matriculado.
- **Evasão do sistema:** ocorre quando o estudante abandona temporariamente ou definitivamente o ensino superior.

E em relação aos fatores que podem levar um aluno a evasão acadêmica, estes são descritos como (PEREIRA, 2003; SILVA, 2012):

- **Fatores individuais do aluno:** relacionados à personalidade, ao interesse pelo curso e a expectativa do aluno quanto ao curso escolhido.
- **Fatores internos à instituição:** relacionados à questões acadêmicas, como a falta de formação pedagógica dos docentes, infraestrutura de pouca qualidade ou ausência de

um sistema de apoio psicopedagógico ao aluno.

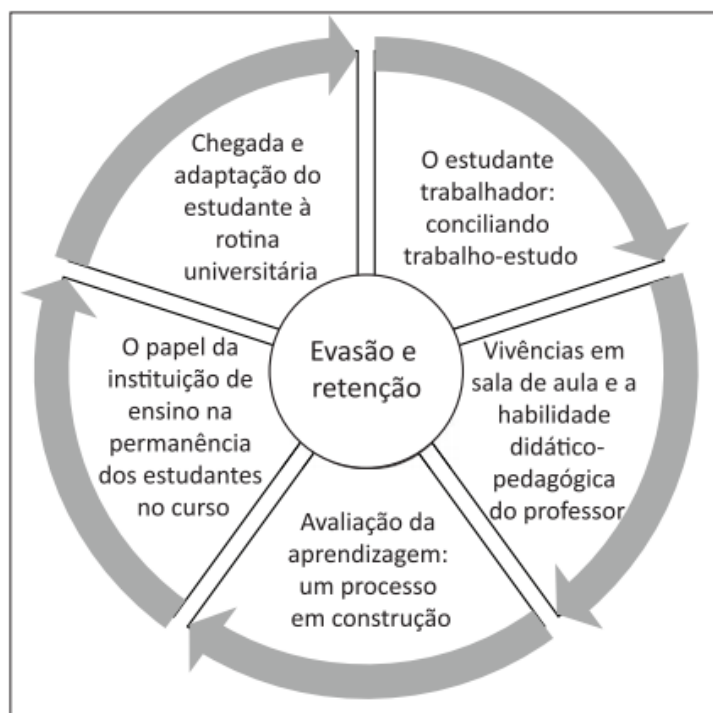
- **Fatores externos à instituição:** são fatores dos quais a instituição tem pouca predominância, como às condições familiar, social e financeira do estudante e o mercado de trabalho.

Já a retenção acadêmica é definida como a não seriação apropriada de um aluno devido à reprovação, ou seja, um atraso que faz com que seja necessário cursar disciplinas de períodos letivos anteriores ao esperado que o aluno esteja cursando. Pode-se distinguir a retenção em razão das suas características como (PEREIRA, 2003; SILVA, 2012):

- **Retenção parcial:** ocorre quando o aluno se encontra em atraso antes do tempo estipulado para término do curso, podendo ainda reverter a situação;
- **Retenção total:** ocorre quando o aluno extrapola o tempo estipulado para término do curso.

Lamers, Santos e Toassi (2017) apresentam em seu trabalho um estudo de caso sobre a evasão e retenção no curso de Odontologia da Universidade Federal do Rio Grande do Sul (UFRGS). Neste os autores efetuam pesquisas com discentes e docentes a fim de identificar os fatores de maior influência para a ocorrência de retenção ou evasão, estas são exibidas na Figura 1. Nota-se que grande parte dos fatores estão relacionados à comunicação entre docentes e discentes e à capacidade de adaptação dos discentes ao ambiente universitário.

Figura 1 – Fatores de destaque em relação à retenção e à evasão, segundo estudantes e professores do curso de Odontologia da UFRGS



Fonte: Lamers, Santos e Toassi (2017)

Assim, melhorar a comunicação entre os diversos atores (docentes, discentes, assistentes

psicopedagógicos) e a possibilidade de antever o desempenho dos estudantes em disciplinas do ensino superior podem ser muito relevante, uma vez que pode-se tomar ações acerca dessa situação e possivelmente revertê-la.

A ferramenta de comunicação desenvolvida neste trabalho será melhor apresentada em capítulos posteriores. Já no contexto de predizer o desempenho dos alunos, a mineração de dados pode ser empregada, com o uso de técnicas de aprendizado de máquina, a fim de realizar a detecção de padrões e definição de modelos que descrevam o comportamento da situação em discussão. Claramente, devido a natureza subjetiva do ambiente acadêmico, o número de variáveis que influenciam no desempenho dos alunos é muito grande, desta forma a acurácia associada à predição se torna muito dependente da qualidade, confiabilidade e volume dos dados disponíveis (CAETANO, 2016).

A seguir é contextualizada a mineração de dados, bem como algumas de suas técnicas.

## 2.2 MINERAÇÃO DE DADOS

Embora inicialmente direcionadas as áreas de negócios, as técnicas de mineração de dados se tornaram grandes ferramentas para a extração de conhecimento a partir de volumes de dados, e atualmente, diversos setores a utilizam para o reconhecimento de padrões e análise de dados (MARQUES; NICOLETTI, 2015).

Weis e Indurkha (1999) definem a mineração de dados como a busca e extração de informações valiosas de grandes volumes de dados, com a cooperação entre o homem - que projeta os bancos de dados, descreve os problemas e define os objetivos desejados, e o computador - que verifica os dados buscando padrões que harmonizem com as metas definidas anteriormente. Muitos autores classificam a mineração de dados como a mistura dos campos de estudo da estatística, da inteligência artificial e dos bancos de dados (CÔRTEZ; PORCARO; LIFSCHITZ, 2002).

Para este trabalho foram empregadas técnicas de mineração de dados e aprendizado de máquina como forma de extração de informação e predição dos volumes de dados acadêmicos.

### 2.2.1 Ferramentas Estatísticas

A primeira técnica consiste das ferramentas estatísticas utilizadas para efetuar análises simples do conjunto de dados a ser minerado. Muitas vezes com estes artefatos estatísticos é possível extrair informações relevantes sobre a distribuição dos conjuntos de dados. Alguns exemplos destas ferramentas são a média aritmética, valores mínimos e máximos, desvio padrão e a distribuição percentual dos dados. Com estas informações é possível construir gráficos e outras ferramentas visuais que possibilitam uma análise inicial dos dados (ADRIAANS; ZANTINGE, 1996).

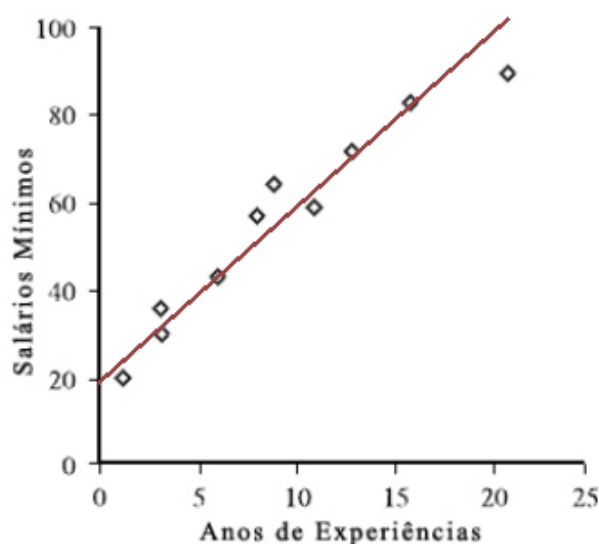
### 2.2.2 Regressão Linear

A regressão linear é uma técnica de predição numérica, e tem como objetivo estimar valores futuros para uma variável contínua, tentando encontrar uma relação com comportamento linear entre as variáveis preditoras. Ou seja, onde é possível criar um modelo no qual o valor de uma variável  $y$  pode ser descrita como uma função linear de uma variável  $x$ . Para esta técnica as variáveis preditoras são os atributos do conjunto de dados e usualmente é utilizada a abordagem dos mínimos quadrados para otimizar a função, minimizando a soma dos quadrados das diferenças entre o valor estimado e os dados reais (CAMILO; SILVA, 2009).

A Figura 2 apresenta um exemplo de regressão linear, no qual é possível verificar uma relação entre as variáveis envolvidas por meio da linearidade entre elas.

Segundo a Figura 2, com cerca de seis anos de experiência a remuneração anual é de aproximadamente 40 salários mínimos, já para 10 anos de experiência tem-se 60 salários mínimos e para 15 anos a remuneração é de 80 salários mínimos. Assim, verifica-se que o comportamento da variável 'Salários Mínimos' pode ser descrito aproximadamente como uma função linear da variável 'Anos de Experiências'. A linha contínua representa um exemplo de função linear que seria retornado por um modelo de regressão linear para este conjunto de dados.

Figura 2 – Exemplo de regressão linear com a variável 'Salários Mínimos' em função da variável 'Anos de Experiências'



Fonte: Camilo e Silva (2009) (Adaptado)

### 2.2.3 Árvore de Decisão

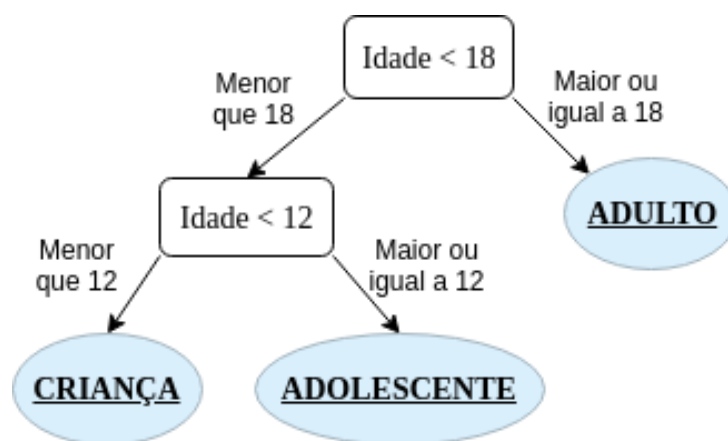
A árvore de decisão é um método simples de aprendizado de máquina construído por um processo chamado de indução, que visa dividir os dados em subconjuntos de forma a entender as segmentações criadas para cada atributo, ou *feature*, do modelo.

Ela funciona como um fluxograma no formato de árvore, onde cada nó não folha apresenta um teste realizado sobre um valor, as ligações entre os nós possuem os valores possíveis para o teste efetuado, e os nós folha indicam a classe ao qual o registro foi determinado como pertencente.

Apesar da simplicidade do método, a árvore de decisão apresenta acurácia semelhante a de métodos mais complexos, porém demanda de uma análise detalhada dos dados para garantir que bons resultados sejam obtidos (CAMILO; SILVA, 2009).

A Figura 3 apresenta um exemplo simplificado de árvore de decisão, onde é classificado a partir do atributo idade de uma determinada amostra a classe ao qual esta pertence. Os retângulos representam as regras ou testes, e as possíveis classes são representadas pelas elipses.

Figura 3 – Exemplo simplificado de árvore de decisão, com a classificação de faixa etária por meio da idade



Fonte: Autoria própria

#### 2.2.4 Random Forest

O *Random Forest* é uma técnica de aprendizado de máquina utilizada tanto para realizar regressões quanto classificações que utiliza múltiplas árvores de decisão para realizar suas predições.

Ela é considerada uma técnica do tipo *bootstrap aggregation*, comumente chamado de *bagging*. No *bagging* algoritmos mais simples são treinados diversas vezes de forma independente utilizando diferentes subconjuntos de dados, gerados por meio de re-amostragem com reposição em um conjunto de treino inicial.

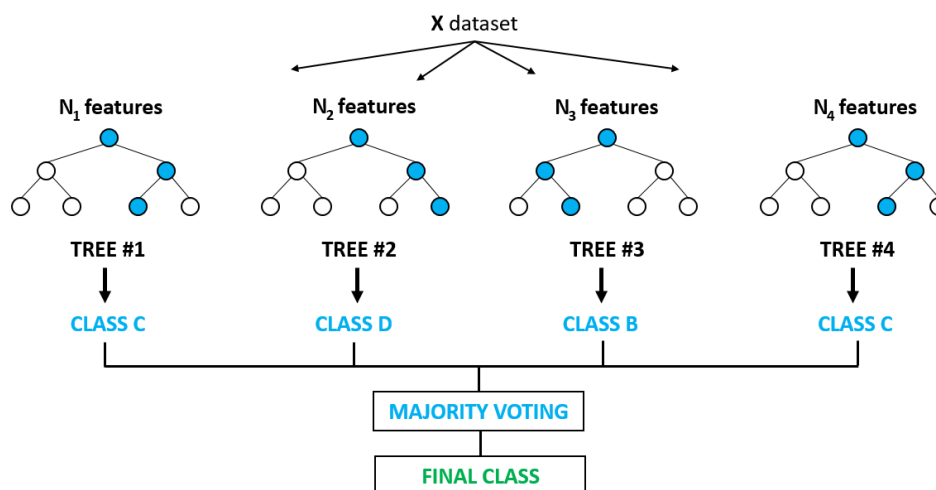
No caso do *Random Forest*, múltiplas árvores de decisão são treinadas independentemente, tomando como base subconjuntos de treinamento formados a partir dos dados de treinamento inicial. E o resultado final do algoritmo é definido com base na agregação dos resultados obtidos para cada uma das árvores de decisão.

Para classificação é comum o uso da maioria dos votos como estratégia para a definição da classe a ser retornada como resultado. Já para regressão são utilizadas ferramentas

estatísticas, como a média aritmética dos resultados retornados por cada árvore de decisão (LIAW; WIENER et al., 2002).

A Figura 4 apresenta um exemplo simplificado de uma *Random Forest* de classificação, nela são treinadas quatro árvores de decisão e o resultado da predição é determinado com base na classe que recebe a maioria dos votos, neste caso a classe C.

Figura 4 – Exemplo simplificado de *Random Forest* utilizada para classificação



Fonte: Abilash (2018)

### 2.2.5 Gradient Boosting

Paralelo ao *bagging*, apresentado anteriormente ao descrever o algoritmo *Random Forest*, tem-se os algoritmos do tipo *boosting*.

O princípio dos algoritmos de *boosting* consiste da combinação do resultado de diversos classificadores ou regressores não tão eficientes, a fim de obter um algoritmo final com melhor acurácia. Diferente das técnicas de *bagging*, que treinam modelos de forma independente, o *boosting* treina modelos com base no resultado e no erro de modelos anteriores, encadeando-os. Usualmente os algoritmos de *boosting* utilizam árvores de decisão como algoritmo interno, mas não se limitam a elas (GERON, 2017).

Tendo as técnicas de mineração de dados delimitadas, foi-se então escolhido um modelo metodológico a ser seguido durante o desenvolvimento do trabalho. A seguir será apresentado o CRISP-DM, modelo utilizado para demarcar as tarefas de um projeto de mineração de dados.

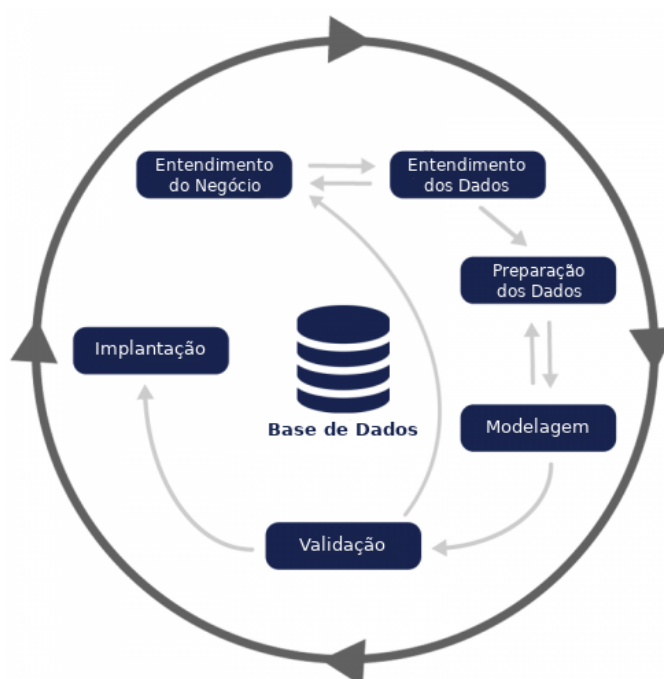
## 2.3 CRISP-DM

O CRISP-DM (*Cross-Industry Standard Process for Data Mining*) é um modelo de processo amplamente utilizado para descrever as etapas da mineração de dados. Um exemplo da

aplicação do método é apresentado por [Castro, Espitia e Montilla \(2018\)](#), no qual o CRISP-DM é aplicado à um cenário onde é realizado um processo de Busca de Conhecimento em Banco de Dados<sup>1</sup> visando a evasão no ensino superior.

O CRISP-DM, segundo [Chapman et al. \(2000\)](#) e [Shearer \(2000\)](#), é composto por seis etapas, que descrevem a sequência de fases para o desenvolvimento de um projeto de mineração de dados. Estas são apresentadas na [Figura 5](#) e são detalhadas a seguir.

Figura 5 – Método CRISP-DM de modelo de processo para mineração de dados



Fonte: [Chapman et al. \(2000\)](#)

### 2.3.1 Entendimento do Negócio

Permite compreender os objetivos do projeto considerando uma perspectiva do negócio, auxiliando na definição do problema de mineração de dados. É importante posteriormente para compreender os dados que serão analisados e quais os tipos de valores que se deseja extrair deles.

Pode ser resumido nas seguintes subetapas: definição dos objetivos do negócio, avaliação da situação, definição dos objetivos de mineração de dados e construção do plano de projeto ([CHAPMAN et al., 2000](#)).

<sup>1</sup>Conhecido também como KDD (*knowledge-discovery in databases*), é o processo de extração de informações valiosas implícitas em bases de dados.

### 2.3.2 Entendimento dos Dados

Esta fase consiste basicamente das seguintes subetapas: coleta inicial dos dados, descrição dos dados, análise dos dados e verificação da qualidade dos dados. Esta sequência de passos é efetuada de forma que, após a coleta inicial dos dados, objetiva-se desenvolver uma certa familiaridade com os dados, descrevendo-os, para assim identificar problemas de qualidade e subconjuntos de dados que possivelmente contenham informações implícitas (CHAPMAN et al., 2000).

### 2.3.3 Preparação dos Dados

Consiste do tratamento dos dados inicialmente coletados a fim de construir o conjunto final de dados que será utilizado para processamento e análise, e é subdividida em cinco etapas menores: seleção, limpeza, construção, integração e formatação dos dados.

Esta fase se dá inicialmente com a seleção dos dados que serão utilizados para a análise, verificando a sua relevância para os objetivos da mineração de dados, bem como a qualidade e as limitações técnicas quanto ao volume e tipos de dados.

Após, tem-se a limpeza dos dados, que é responsável por remover dados ruidosos e sem valor para a análise, evitando assim ambiguidades e anomalias nos resultados. Nela é feita a conversão de dados que possam estar em um formato não adequado e a remoção de caracteres inválidos, podendo também ser efetuada uma normalização dos dados para ajustá-los à determinadas situações. Remete a subfase do Entendimento dos Dados, que se refere a verificação da qualidade dos dados.

Em seguida, na construção dos dados, há a transformação de certos atributos coletados em atributos projetados para facilitar a etapa de processamento, reduzindo em certos casos a quantidade necessária de atributos, como por exemplo ao transformar os atributos *comprimento* e *largura* em somente um atributo *área*.

Também é importante a integração dos conjuntos de dados coletados relacionando-os para a criação de novos registros, por exemplo quando dois conjuntos de dados possuem valores que podem ser relacionados para gerar dados mais concisos, mas estão fragmentados.

Por fim tem-se a formatação dos dados, que redesenha o conjunto de dados a fim de permitir sua manipulação por algoritmos e ferramentas específicas (CHAPMAN et al., 2000).

### 2.3.4 Modelagem

Nesta etapa do processo são selecionadas técnicas de modelagem de acordo com o problema a ser resolvido. Em seguida são gerados testes empíricos para validar a aplicabilidade do modelo. Por exemplo para algoritmos de regressão e classificação são utilizadas as taxas de erro entre conjuntos de dados de treinamento e conjuntos de dados de teste para medir sua acurácia. Desta forma verifica-se se o modelo pode prever dados históricos com eficiência, para então utilizá-lo para prever dados futuros.



Então o modelo é gerado sobre o conjunto de dados preparados previamente e por fim é realizada a validação do modelo, avaliando a aplicação (CHAPMAN et al., 2000).

### 2.3.5 Validação

Nesta fase são avaliados os resultados apresentados pelo modelo, revisando a construção deste se necessário. Valida-se os resultados tomando em conta os objetivos do negócio definidos anteriormente, verificando se todos foram abordados de forma adequada ou se devem ser realizados ajustes nas etapas iniciais do processo.

Por fim deve-se decidir como utilizar os resultados obtidos com a mineração de dados, caso estes sejam considerados satisfatórios e respondam ao problema previamente definido (CHAPMAN et al., 2000).

### 2.3.6 Implementação

A implementação é a última etapa do CRISP-DM, mas isto não significa que o projeto é finalizado quando atinge esta etapa. Como apresentado na Figura 5, no qual o círculo com as setas mais externas representa a natureza cíclica da mineração de dados, onde se mantém em constante revisão e atualização.

Nesta fase as informações geradas a partir do processamento dos dados são organizadas e apresentadas ao usuário final (por exemplo: gerentes, analistas, clientes) por meio de gráficos e outras ferramentas de visualização de dados, a fim de auxiliar na análise e principalmente nas suas tomadas de decisões (CHAPMAN et al., 2000).

## 2.4 VISUALIZAÇÃO DE DADOS

A visualização de dados descreve técnicas que objetiva-se a auxiliar na compreensão e extração de significado de dados utilizando contextos visuais.

Em um conjunto de dados podem haver informações importantes ocultas que podem auxiliar nas tomadas de decisão e a promover negócios. Mas o desafio é que nem sempre é possível tirar conclusões apenas analisando números brutos. Quando se analisa os dados apresentados em um formato visual, surgem padrões, conexões e outras descobertas que, de outro modo, não seriam percebidos (Microsoft Power BI, 2018).

Segundo Ward, Grinstein e Keim (2015), a visualização de dados é extremamente útil como técnica de descobrimento de padrões, e embora possa parecer uma técnica não tão sofisticada, permite mensurar a qualidade inicial dos dados e identificar onde os padrões podem ser extraídos.

Outro ponto citado pelos autores é a capacidade de uma imagem ou gráfico conter uma grande quantidade de conteúdo, permitindo uma interpretação muito mais rápida quando utilizadas imagens quando comparado com páginas de texto que traduzem o mesmo conteúdo.

Isto ocorre porque a interpretação de imagens se dá de forma paralela pelo sistema de percepção humano, enquanto a velocidade de análise de texto é limitada ao processo sequencial de leitura.

Quando a visualização de dados é utilizada em processos mais avançados de mineração de dados é possível a criação de gráficos tridimensionais interativos e gráficos de segmentação de bancos de dados em forma hierárquica, como uma árvore (WARD; GRINSTEIN; KEIM, 2015).

A seguir são apresentados trabalhos relacionados que possuem propostas com características semelhantes à deste trabalho.

## 2.5 TRABALHOS RELACIONADOS

Em seu trabalho, Cornelius Junior (2015) apresenta um estudo sobre a evasão no ensino superior e a aplicação de técnicas de mineração de dados na tentativa de descobrir um perfil dos alunos com tendências evasivas e os padrões associados a essas tendências. O autor aplica as técnicas de classificação e associação utilizando a ferramenta de mineração de dados WEKA, com base em um conjunto de dados cedidos pela Universidade de Santa Cruz do Sul. Neste contexto foram constatados que os fatores que mais influenciaram para a evasão de cursos foram o total de disciplinas cursadas e o situação do aluno (Aprovado/Reprovado) nas disciplinas do primeiro semestre acadêmico.

Um dos objetivos do autor foi investigar se alunos que reprovam nas primeiras disciplinas do curso possuem maior tendência a evadir. Com os resultados verificou-se que em sua maioria os alunos que reprovaram nas primeiras disciplinas não continuavam no curso. Dos 302 alunos que reprovaram na disciplina de Cálculo, menos de 10% concluíram o curso, e dos 426 que reprovaram na disciplina de Algoritmos apenas 14 conseguiram se formar. O autor verificou também alguns padrões que levaram a maioria dos alunos a obterem melhores resultados, um destes foi a consistência no número de disciplinas cursadas, ou seja, cursar sempre a mesma quantidade de disciplinas nos semestres. Dos 82 alunos que fizeram cinco disciplinas durante os cinco primeiros semestres 69 se formaram e apenas 13 evadiram, sendo um dos poucos casos no qual o número de alunos formados superou o número de evadidos.

Ramos, Machado e Monteiro (2011) descrevem em seu trabalho uma pesquisa realizada com o objetivo de identificar determinantes para as questões de retenção e evasão no curso de Nutrição da UFRGS. Os autores utilizaram diversas abordagens para atingir o maior número de discentes, elaborando questionários eletrônicos para que fosse possível atingir tanto alunos em curso quanto alunos evadidos, e realizando entrevistas com alunos regulares e em situação de retenção. Com a conclusão do trabalho os autores explicitam a necessidade de criação de mecanismos de acompanhamento de alunos para a prevenção da retenção e evasão discente, bem como o desenvolvimento de competências relacionadas ao empreendedorismo, como forma de motivação para os alunos.

O trabalho de Marques e Nicoletti (2015) apresenta um projeto de pesquisa em andamento na área de mineração de dados educacionais, cujo objetivo é a predição do

desempenho de estudantes universitários em ambientes convencionais de ensino. No trabalho os autores afirmam, que devido ao número elevado e a natureza não linear das variáveis que determinam o desempenho dos estudantes, a previsão pode nem sempre ser satisfatória e os resultados não serem suficientemente bons para tomadas de decisões que ajudem a corrigir os problemas do tema. Mas ainda se faz necessária a produção de trabalhos deste cunho, devido a possibilidade de antever, mesmo com certa margem de erro e de tempo, o comportamento de um aluno que se enquadra em um determinado perfil de desempenho, e então motivar o docente responsável pela disciplina a utilizar recursos pedagógicos que busquem recuperar o aluno ainda no tempo regular das aulas da disciplina.

Em um trabalho posterior, [Caetano \(2016\)](#) apresenta uma aplicação de técnicas de aprendizado de máquina e mineração de dados com o foco na predição do desempenho acadêmico de alunos. Para isto a autora utilizou também o ambiente *WEKA*, empregando três diferentes algoritmos de aprendizado de máquina: i) o *Naive Bayes*; ii) o *Nearest Neighbor*; iii) e o *J48*, versão do *WEKA* do algoritmo de árvore de decisões *C4.5*. O resultado do seu trabalho mostrou que foi possível prever o desempenho de estudantes com precisão de aproximadamente 70% utilizando somente informações iniciais e 90% caso seja utilizado um conjunto de atributos completo ou somente com os atributos mais impactantes para o treinamento do modelo. Para este trabalho foram utilizados dados demográficos, como sexo, idade e estado civil, e dados associados ao desempenho parcial dos alunos, como notas dos trabalhos entregues durante o período letivo.

[Rigo, Cazella e Cambruzzi \(2012\)](#) apresentam em seu trabalho uma análise de melhorias possíveis na aplicação de mineração de dados educacionais, para que estes possam ser efetivamente utilizados para a detecção de padrões comportamentais relacionados à evasão escolar. Os autores salientam a necessidade de mapear os fatores associados ao fenômeno, bem como a necessidade de implantação de soluções interativas, que permitam o acesso dinâmico aos resultados e possibilite o diagnóstico precoce e a realização de ações pedagógicas relevantes para combater à evasão no ensino.

O modelo proposto por [Tinto \(1987\)](#) e posteriormente discutido por [Bandeira Andriola, Gomes Andriola e Moura \(2012\)](#) propõe que as características demográficas influenciam na integração do discente com o novo ambiente universitário. Dentre essas características pode-se citar o nível socio-econômico e os conhecimentos adquiridos previamente por meio de educação formal e informal. E já a condição social do discente, que envolvem a idade, o gênero, as habilidades pessoais e expectativas de desenvolvimento pessoal, está fortemente associada com a motivação para o desempenho acadêmico e o seu reconhecimento.

[Babosa, Carvalho e Andrade \(2017\)](#) exploram em seu trabalho aspectos da mineração de dados como recurso tecnológico para investigações em um Ambiente Virtual de Aprendizagem (AVA). Com os resultados obtidos os autores indicam que a mineração de dados permite acesso a informações sobre o processo educativo, tais como aproveitamento, desempenho e a predição dos resultados finais, identificando relações entre dados que podem produzir novos conhecimentos.

Isto permite refletir sobre os métodos de ensino empregados e então tomar decisões para a melhoria do desempenho acadêmico.

### 3 DESENVOLVIMENTO

Nesta seção são apresentadas as tecnologias e ferramentas utilizadas para a implementação do trabalho. Em seguida são descritas as etapas do processo metodológico CRISP-DM, detalhando as atividades realizadas e os artefatos gerados em cada uma das etapas.

#### 3.1 TECNOLOGIAS E FERRAMENTAS

Para a implementação do *front-end* do sistema web foi utilizado o framework *Javascript React.js*, devido a sua capacidade de construção de *Single Page Applications (SPAs)*, que permitem a interação do usuário com as páginas da aplicação sem demandar de recarregamento das páginas. Nas *SPAs* todos os *templates*, ou seja, códigos implementados utilizando *HyperText Markup Language (HTML)*, *Cascading Style Sheets (CSS)* e *Javascript* dos quais o navegador necessita interpretar para a renderização das páginas são carregados no início do uso da aplicação, restando somente o carregamento dos dados a serem feitos durante a navegação.

Tesarik, Dolezal e Kollmann (2008) mostram em seu trabalho as dificuldades na construção de *SPAs*, por conta do design mais sofisticado que esta deve ter para permitir a movimentação do usuário entre as páginas. Mas, em contrapartida, apresentam uma das principais vantagens deste tipo de construção, que é a melhor fluidez que pode ser obtida na experiência do usuário ao interagir com a aplicação.

Em conjunto ao *React.js* foram utilizadas as seguintes bibliotecas:

- **Bootstrap<sup>1</sup>**: um pacote de ferramentas para construção da aplicação do lado do cliente, ou seja, o *front-end*. Contém componentes que facilitam a construção e estilização de páginas responsivas, disponibilizando recursos em *HTML*, *CSS* e *Javascript*.
- **Charts.js<sup>2</sup>**: uma biblioteca rápida, responsiva e flexível, utilizada por *designers* e desenvolvedores para a construção de gráficos.
- **D3.js<sup>3</sup>**: abreviação para *Data-Driven Documents*, ela é uma biblioteca para criação dinâmica e interativa de gráficos, grafos e outras ferramentas de visualização de dados.

Para a implementação do *back-end* foram criados dois programas isolados. O primeiro consiste de uma *API* construída utilizando a plataforma *Node.js* com a linguagem de programação *Typescript*. Esta *API* é responsável por dispor todos os dados para as operações que não envolvem previsões, ou seja, dados para construção de gráficos e listagem no *front-end*, e também pelo envio de e-mails e preparação dos dados a partir das planilhas carregadas para persistência. As bibliotecas utilizadas no código em *Node.js* são:

---

<sup>1</sup>Disponível em <https://getbootstrap.com/>

<sup>2</sup>Disponível em <https://chartjs.org/>

<sup>3</sup>Disponível em <https://d3js.org>

- **Express**<sup>4</sup>: *framework* leve e flexível que possibilita a uma aplicação Node.js criar uma API robusta rapidamente.
- **Mongoose**<sup>5</sup>: responsável por mapear os objetos entre o Node.js e o banco de dados MongoDB, gerenciando relacionamentos de dados e provendo a construção de esquemas para validação de regras na inserção de dados, o que garante uma maior consistência dos dados.
- **SheetJS**<sup>6</sup>: plugin leve e robusto para manipulação de planilhas, permitindo a criação, leitura e conversão de planilhas para *Javascript Object Notation (JSON)*, entre outros.
- **Moment.js**<sup>7</sup>: biblioteca que possui um amplo ferramental para manipulação, validação e apresentação de datas em *Javascript*, contemplando diferentes fusos horários.
- **Nodemailer**<sup>8</sup>: módulo gratuito que permite aplicações Node.js realizar o envio de e-mails.
- **SimpleCrypto**<sup>9</sup>: possibilita o processo de encriptação e decriptação por meio de uma chave criptográfica de forma simples.

O segundo programa também consiste de uma *API*, desta vez construída utilizando a linguagem de programação Python. Esta é responsável pelo treinamento dos modelos de aprendizado de máquina e pelas requisições enviadas pelo *front-end* relacionadas às predições. Para isto são utilizadas algumas bibliotecas. São estas:

- **Flask**<sup>10</sup>: semelhante ao Express para Python, ele é um módulo leve que disponibiliza toda a estrutura para a criação de uma *API*.
- **PyMongo**<sup>11</sup>: biblioteca que dispõe um conjunto de ferramentas para se trabalhar com o MongoDB.
- **NumPy**<sup>12</sup>: pacote de ferramentas matemáticas para computação científica em Python;
- **Pandas**<sup>13</sup>: biblioteca *open-source* para Python que provém estruturas de dados e ferramentas de manipulação e análise de dados com alta performance.
- **Matplotlib**<sup>14</sup>: biblioteca para construção de gráficos 2D com poucas linhas de código.
- **Scikit-learn**<sup>15</sup>: biblioteca que disponibiliza diversas ferramentas para mineração e análise de dados, como algoritmos de aprendizado de máquina para classificação, regressão e *clusterização*, bem como ferramentas de pré-processamento e validação de modelos.

Para persistência foi utilizado o banco de dados não relacional MongoDB. A escolha do banco de dados (BD) se fez devido a algumas características do modelo não relacional,

---

<sup>4</sup>Disponível em <https://expressjs.com/>

<sup>5</sup>Disponível em <https://mongoosejs.com/>

<sup>6</sup>Disponível em <https://github.com/SheetJS/sheetjs>

<sup>7</sup>Disponível em <https://momentjs.com>

<sup>8</sup>Disponível em <https://nodemailer.com/about/>

<sup>9</sup>Disponível em <https://www.npmjs.com/package/simple-crypto-js>

<sup>10</sup>Disponível em <https://palletsprojects.com/p/flask/>

<sup>11</sup>Disponível em <https://api.mongodb.com/python/current/>

<sup>12</sup>Disponível em <https://numpy.org/>

<sup>13</sup>Disponível em <https://pandas.pydata.org/>

<sup>14</sup>Disponível em <https://matplotlib.org/>

<sup>15</sup>Disponível em <https://scikit-learn.org/stable/>

que permitem uma maior escalabilidade, e mesmo não satisfazendo as propriedades ACID (Atomicidade, Consistência, Isolamento e Durabilidade), que garantem a não duplicidade de dados, é mais flexível e permite estruturar os dados de forma a facilitar a utilização para a geração de gráficos.

O estudo apresentado por [Côrrea, Almeida e Graça Neto \(2017\)](#), que compara o MongoDB com o PostgreSQL, sistema de gerenciamento de banco de dados relacional (SGBD), mostra uma maior performance do primeiro BD para a maioria dos casos testados.

Considerando o cenário de dados acadêmicos, o volume de dados pode se tornar consideravelmente grande, bem como a variedade de dados, desta forma a performance é essencial para a escalabilidade do sistema. Assumindo o crescimento do volume de dados, pode ser desejável o emprego de ferramentas de *Big Data*, e observando que o modelo não-relacional é bastante utilizado e eficaz para o emprego dessas técnicas, o MongoDB se fez adequado.

A fim de tornar mais organizadas as atividades do trabalho utilizou-se o *Trello*<sup>16</sup>, uma ferramenta on-line que auxilia na gestão de projetos utilizando o *Kanban*<sup>17</sup>. O espaço de planejamento do *Trello* foi compartilhado com o professor orientador, desta forma este possuía visibilidade das atividades planejadas e que estavam sendo desenvolvidas.

O desenvolvimento deste trabalho passou por um período de pausa, devido à isso algumas das tarefas do *Trello* possuem datas referentes à anos anteriores ao da finalização do trabalho.

A [Figura 6](#) apresenta a tela do *Trello* com as colunas e parte das tarefas criadas. As colunas criadas para a divisão das tarefas foram as seguintes:

- **Backlog:** criada para manter todas as atividades planejadas para o trabalho.
- **To-Do:** contém as tarefas que estão na fila e serão executadas logo ao finalizar as tarefas em execução. Geralmente são movidas para *Doing* em até duas semanas.
- **Doing:** contém as tarefas que estão sendo executadas. Normalmente se mantém no máximo três atividades em simultâneo com esse status.
- **Testing:** contém as atividades que tiveram sua implementação finalizada, mas precisam passar por testes para verificar se cumprem com o resultado previsto.
- **Done:** contém as atividades executadas, testadas e então concluídas.

Das atividades planejadas, algumas possuem detalhes que foram preenchidos conforme o avanço e conclusão destas. A [Figura 7](#) traz como exemplo a atividade de implementação das *APIs*. Na atividade foram listadas algumas das rotas criadas para inserção e busca de dados e predições, trazendo também uma breve descrição de cada uma das rotas.

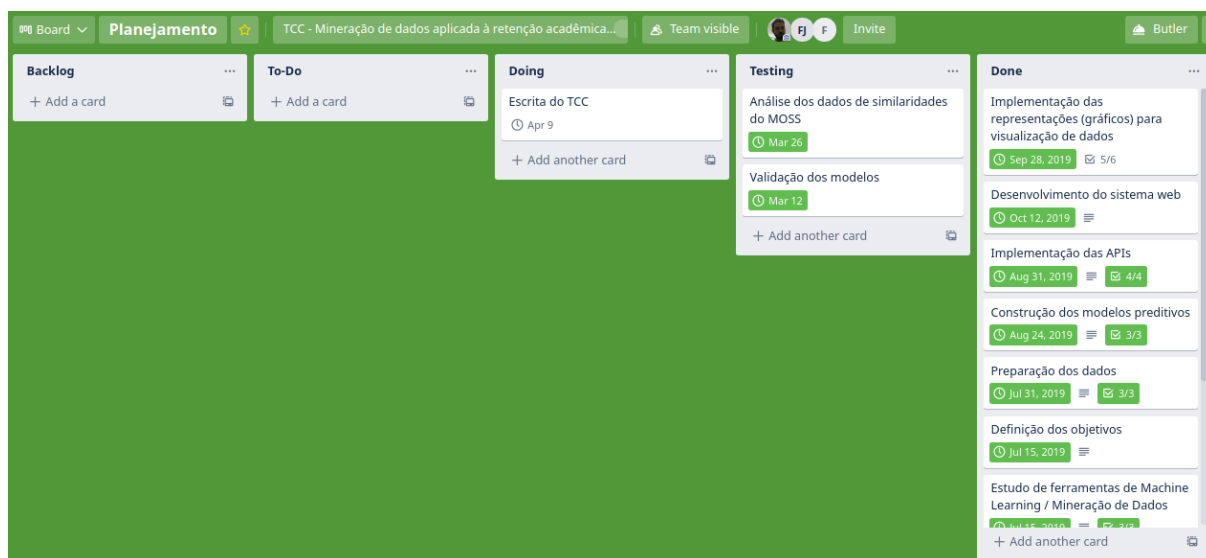
Ainda nos detalhes da atividade foram criadas subtarefas, e com a conclusão de cada uma delas, estas foram assinaladas como finalizadas, sinalizando o término dessa subtarefa. A [Figura 8](#) mostra as subtarefas criadas para a atividade de implementação das *APIs*.

Para o desenvolvimento deste trabalho foi utilizado o método CRISP-DM, apresentado

<sup>16</sup>Disponível em <https://trello.com/en>

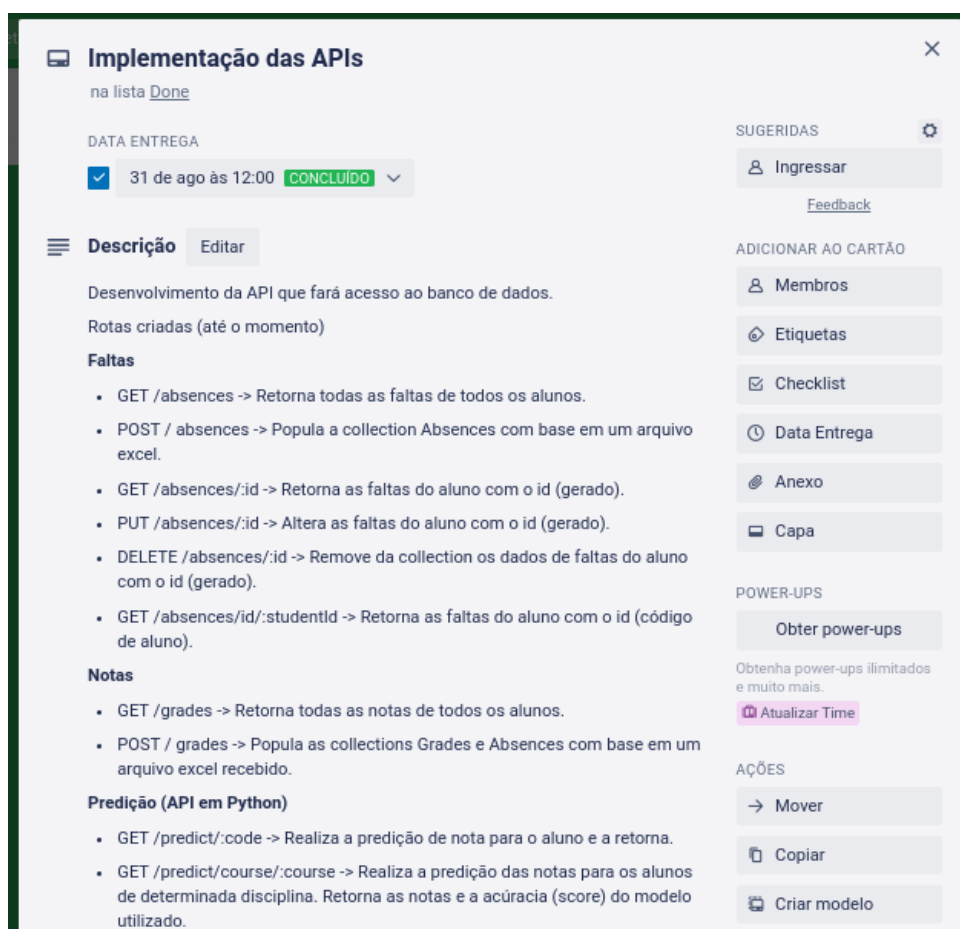
<sup>17</sup>Sistema de gestão visual para o controle de tarefas e fluxo de trabalho por meio de colunas e *post-its*.

Figura 6 – Trello com as colunas e atividades planejadas para o trabalho



Fonte: Autoria Própria

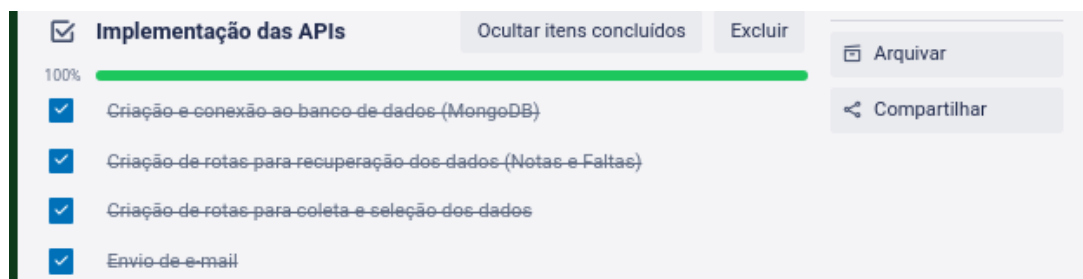
Figura 7 – Detalhes da atividade de implementação das APIs no Trello



Fonte: Autoria Própria



Figura 8 – Subtarefas da atividade de implementação de APIs no Trello



Fonte: Autoria Própria

na seção 2.3. Este processo metodológico é subdividido em seis fases principais, que foram seguidas de forma não linear durante o processo de implementação do projeto, desta forma em diversos momentos fez-se necessário revisitar etapas anteriores.

### 3.2 ENTENDIMENTO DO NEGÓCIO

No contexto apresentado, a principal intenção era analisar dados acadêmicos e com isso construir modelos para estimar o desempenho de alunos durante o período letivo e implementar mecanismos para interagir com alunos em situação prevista de retenção.

Com o objetivo de compreender fatores relacionados ao contexto acadêmico, foram definidos objetivos de negócio por meio das seguintes questões:

- Quais fatores sob o domínio de conhecimento da instituição de ensino influenciam para retenção de alunos?
- Como as notas e as frequências parciais de um aluno se relacionam?
- A partir das notas e frequências parciais de um aluno, é possível determinar o desempenho final dele?
- A partir das notas e frequências parciais de alunos, é possível determinar um comportamento comum que descreva uma tendência a retenção?

Dentre os fatores sob domínio de conhecimento da Instituição de Ensino tem-se os de cunho acadêmico, como: as notas das avaliações, as notas de todos os trabalhos e entregas e a frequência parcial do aluno (percentual de frequência durante o semestre letivo).

Foi considerado também o uso da plataforma *URI Online Judge* (URI, 2021), uma ferramenta on-line utilizada por alguns docentes que lecionam disciplinas de programação. A plataforma URI possibilita que os alunos testem seus conhecimentos resolvendo desafios computacionais e suporta diferentes linguagens de programação, como C/C++, Java, Python e PSQL.

Além disso, a plataforma possui uma versão acadêmica, que permite que um professor cadastre os alunos de sua turma, disponibilize listas de exercícios específicos e monitore quais alunos resolveram os exercícios, bem como a quantidade de exercícios que cada aluno

resolveu, tentou resolver, ou simplesmente não tentou. Complementar, a plataforma possibilita a comparação entre as soluções propostas e enviadas pelos alunos.

O *URI Academic* possui parceria com várias instituições nacionais, como a Universidade de São Paulo (USP) e a Universidade Tecnológica Federal do Paraná (UTFPR) e internacionais, como a *Baylor University*, do Texas, Estados Unidos, e a *Universidad Nacional Autónoma de México*, situada na Cidade do México ([URI ACADEMIC, 2021](#)).

O formato dos dados exportados pela plataforma *URI Online Judge* é apresentado na seção à seguir, na [Figura 13](#).

### 3.3 ENTENDIMENTO DOS DADOS

Nesta fase foi realizada a coleta dos dados iniciais para validação do sistema proposto. Para a implementação deste trabalho os dados foram disponibilizados pelo professor orientador e os dados acadêmicos foram suprimidos de identificação.

Neste trabalho foram utilizados dados de 10 turmas da UTFPR - campus Cornélio Procopio, contemplando dados de notas e faltas de 312 alunos. Complementar a isto foram utilizadas 20 planilhas de exercícios do URI, com aproximadamente oito exercícios cada, totalizando 164 exercícios. Vale ressaltar que todos os dados são referentes à disciplinas da área de computação, como Laboratório de Informática e Algoritmos.

Para a aquisição dos dados por parte do sistema de predição aqui desenvolvido, foi criada uma página para o envio de dados por meio de planilhas. Nesta página são aceitas planilhas nos formatos disponibilizados pelo sistema de gestão de aprendizagem Moodle, que contemplam as notas parciais de avaliações e atividades realizadas via Moodle, bem como o nome, sobrenome e e-mail de cada aluno. Vale ressaltar que os cabeçalhos das colunas de notas refletem os nomes designados para as atividades criadas no Moodle. Então, caso o docente se sinta interessado em utilizar o sistema desenvolvido neste trabalho, é importante realizar algumas adaptações na forma como são denominadas as atividades no Moodle, a fim de evitar trabalhos manuais posteriormente para carregar os dados.

O formato da planilha de notas disponibilizada atualmente pela plataforma Moodle é semelhante ao apresentado na [Figura 9](#).

Para a submissão das faltas é aceito o formato disponibilizado pelas planilhas exportadas pelo sistema acadêmico da UTFPR. Neste arquivo é disposto um código identificador ou Registro Acadêmico (R.A.), nome do aluno, total de faltas e colunas que representam cada um dos dias letivos do período, com o número de ausências de cada aluno em cada um desses dias.

O formato da planilha de faltas disponibilizada atualmente pelo sistema acadêmico da UTFPR é apresentado na [Figura 10](#), mas em alguns casos estas são exportadas com sua estrutura semelhante a apresentada na figura [Figura 11](#). A plataforma desenvolvida também está preparada para receber dados neste formato e ao carregar uma planilha neste padrão ele internamente realizará a re-estruturação desta para o formato apresentado na [Figura 10](#) e em seguida a persistência dos dados.

Figura 9 – Modelo de planilha com notas exportada pelo Moodle

|    | A       | B            | C                  | D                 | E                        |
|----|---------|--------------|--------------------|-------------------|--------------------------|
| 1  | Nome    | Sobrenome    | Endereço de e-mail | Micro-avaliação 1 | Avaliação 1 (03/10/2019) |
| 2  | Nome 1  | Sobrenome 1  | e-mail do aluno 1  | 4,00              | 70,00                    |
| 3  | Nome 2  | Sobrenome 2  | e-mail do aluno 2  | 9,00              | 81,00                    |
| 4  | Nome 3  | Sobrenome 3  | e-mail do aluno 3  | 5,50              | 30,00                    |
| 5  | Nome 4  | Sobrenome 4  | e-mail do aluno 4  | 9,00              | 85,00                    |
| 6  | Nome 5  | Sobrenome 5  | e-mail do aluno 5  | 7,50              | 37,00                    |
| 7  | Nome 6  | Sobrenome 6  | e-mail do aluno 6  | -                 | -                        |
| 8  | Nome 7  | Sobrenome 7  | e-mail do aluno 7  | 8,50              | 93,00                    |
| 9  | Nome 8  | Sobrenome 8  | e-mail do aluno 8  | 5,50              | 75,00                    |
| 10 | Nome 9  | Sobrenome 9  | e-mail do aluno 9  | -                 | 35,00                    |
| 11 | Nome 10 | Sobrenome 10 | e-mail do aluno 10 | 6,50              | 37,00                    |
| 12 | Nome 11 | Sobrenome 11 | e-mail do aluno 11 | 6,00              | 17,00                    |
| 13 | Nome 12 | Sobrenome 12 | e-mail do aluno 12 | 8,00              | 77,00                    |
| 14 | Nome 13 | Sobrenome 13 | e-mail do aluno 13 | 6,50              | 9,00                     |
| 15 | Nome 14 | Sobrenome 14 | e-mail do aluno 14 | 7,00              | 48,00                    |
| 16 | Nome 15 | Sobrenome 15 | e-mail do aluno 15 | 7,50              | 67,00                    |
| 17 | Nome 16 | Sobrenome 16 | e-mail do aluno 16 | -                 | 6,00                     |
| 18 | Nome 17 | Sobrenome 17 | e-mail do aluno 17 | -                 | 90,00                    |
| 19 | Nome 18 | Sobrenome 18 | e-mail do aluno 18 | 6,50              | 75,00                    |
| 20 | Nome 19 | Sobrenome 19 | e-mail do aluno 19 | 7,00              | 75,00                    |
| 21 | Nome 20 | Sobrenome 20 | e-mail do aluno 20 | 8,00              | 6,00                     |
| 22 | Nome 21 | Sobrenome 21 | e-mail do aluno 21 | 5,50              | 66,00                    |
| 23 | Nome 22 | Sobrenome 22 | e-mail do aluno 22 | 4,50              | 2,00                     |
| 24 | Nome 23 | Sobrenome 23 | e-mail do aluno 23 | 5,50              | 4,00                     |
| 25 | Nome 24 | Sobrenome 24 | e-mail do aluno 24 | 8,50              | 27,00                    |
| 26 | Nome 25 | Sobrenome 25 | e-mail do aluno 25 | -                 | 58,00                    |
| 27 | Nome 26 | Sobrenome 26 | e-mail do aluno 26 | 8,00              | 60,00                    |
| 28 | Nome 27 | Sobrenome 27 | e-mail do aluno 27 | -                 | 20,00                    |
| 29 | Nome 28 | Sobrenome 28 | e-mail do aluno 28 | 9,00              | 80,00                    |

Fonte: Autoria Própria

Figura 10 – Modelo de planilha com faltas exportada do sistema acadêmico da UTFPR

|    | A  | B       | C                | D        | E        | F        | G        | H        | I        | J        | K        | L        | M        | N        | O        | P        | Q        | R        | S        | T     |
|----|----|---------|------------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-------|
| 1  | ID | Código  | Nome             | 20/08/19 | 22/08/19 | 27/08/19 | 29/08/19 | 03/09/19 | 05/09/19 | 10/09/19 | 12/09/19 | 17/09/19 | 19/09/19 | 24/09/19 | 26/09/19 | 01/10/19 | 03/10/19 | 08/10/19 | 10/10/19 | Total |
| 2  | 1  | 1000001 | Nome do aluno 1  | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 2        | C        | 2     |
| 3  | 2  | 1000002 | Nome do aluno 2  | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 0     |
| 4  | 3  | 1000003 | Nome do aluno 3  | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 2        | C        | C        | 2     |
| 5  | 4  | 1000004 | Nome do aluno 4  | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 2        | C        | 2     |
| 6  | 5  | 1000005 | Nome do aluno 5  | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 2        | C        | C        | 2     |
| 7  | 6  | 1000006 | Nome do aluno 6  | 2        | C        | C        | C        | C        | C        | 2        | 2        | 2        | 3        | 2        | 3        | 2        | 3        | 2        | 3        | 26    |
| 8  | 7  | 1000007 | Nome do aluno 7  | C        | C        | C        | C        | C        | C        | C        | C        | 3        | C        | 3        | 2        | C        | 2        | C        | C        | 10    |
| 9  | 8  | 1000008 | Nome do aluno 8  | C        | C        | C        | 3        | C        | C        | 2        | C        | C        | C        | 3        | C        | C        | C        | C        | 8        |       |
| 10 | 9  | 1000009 | Nome do aluno 9  | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 2        | C        | 2        | C        | 2        | C        | 6     |
| 11 | 10 | 1000010 | Nome do aluno 10 | C        | C        | C        | C        | 2        | 3        | 2        | 2        | C        | C        | C        | C        | 2        | C        | C        | 3        | 14    |
| 12 | 11 | 1000011 | Nome do aluno 11 | C        | 3        | 2        | C        | 2        | C        | C        | 2        | 2        | C        | C        | C        | C        | C        | C        | C        | 11    |
| 13 | 12 | 1000012 | Nome do aluno 12 | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 0     |
| 14 | 13 | 1000013 | Nome do aluno 13 | C        | C        | C        | C        | C        | C        | 2        | C        | 2        | 3        | 2        | C        | C        | C        | C        | C        | 9     |
| 15 | 14 | 1000014 | Nome do aluno 14 | C        | C        | C        | 3        | C        | C        | C        | C        | C        | C        | C        | 3        | C        | C        | C        | C        | 6     |
| 16 | 15 | 1000015 | Nome do aluno 15 | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 0     |
| 17 | 16 | 1000016 | Nome do aluno 16 | C        | C        | C        | C        | C        | C        | C        | C        | C        | 3        | 2        | C        | C        | C        | 2        | 3        | 10    |
| 18 | 17 | 1000017 | Nome do aluno 17 | C        | C        | C        | C        | C        | C        | C        | 2        | C        | C        | C        | 2        | C        | C        | C        | C        | 4     |
| 19 | 18 | 1000018 | Nome do aluno 18 | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 0     |
| 20 | 19 | 1000019 | Nome do aluno 19 | C        | C        | C        | C        | 2        | C        | C        | C        | C        | C        | C        | C        | 2        | C        | C        | C        | 4     |
| 21 | 20 | 1000020 | Nome do aluno 20 | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 0     |
| 22 | 21 | 1000021 | Nome do aluno 21 | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 2        | C        | C        | C        | 2     |
| 23 | 22 | 1000022 | Nome do aluno 22 | --       | --       | 2        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 2        | C        | C        | C        | 4     |
| 24 | 23 | 1000023 | Nome do aluno 23 | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 0     |
| 25 | 24 | 1000024 | Nome do aluno 24 | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 2        | C        | 2     |
| 26 | 25 | 1000025 | Nome do aluno 25 | C        | C        | C        | C        | C        | C        | C        | C        | C        | C        | 2        | C        | 2        | C        | C        | C        | 4     |
| 27 | 26 | 1000026 | Nome do aluno 26 | C        | C        | C        | C        | C        | C        | 2        | C        | C        | C        | C        | 2        | C        | C        | C        | C        | 4     |
| 28 | 27 | 1000027 | Nome do aluno 27 | C        | 3        | C        | C        | C        | 3        | 2        | C        | C        | C        | C        | C        | 2        | C        | C        | C        | 10    |

Fonte: Autoria Própria

Figura 11 – Modelo alternativo de planilha com faltas exportada pelo sistema acadêmico da UTFPR

|    | A          | B          | C                       | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|----|------------|------------|-------------------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1  | 12/03/2019 |            |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 2  | [2]        | 14/03/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 3  | [2]        | 19/03/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 4  | [2]        | 21/03/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 5  | [2]        | 26/03/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 6  | [2]        | 28/03/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 7  | [2]        | 02/04/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 8  | [2]        | 04/04/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 9  | [2]        | 09/04/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 10 | [2]        | 11/04/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 11 | [2]        | 16/04/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 12 | [2]        | 18/04/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 13 | [2]        | 23/04/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 14 | [2]        | 25/04/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 15 | [2]        | 30/04/2019 |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 16 | [2]        |            |                         |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 17 |            | 1          | 1000001 Nome do aluno 1 | C | C | C | 2 | C | C | C | C | C | C | C | C | C | C | C |
| 18 |            | 2          | 1000002 Nome do aluno 2 | 2 | 2 | C | C | C | C | C | C | C | C | 2 | C | C | C | C |

Fonte: Autoria Própria

Figura 12 – Planilha de dados com notas e faltas separadas por abas

|    | A       | B            |           |
|----|---------|--------------|-----------|
| 1  | Nome    | Sobrenome    | Endereço  |
| 2  | Nome 1  | Sobrenome 1  | e-mail do |
| 3  | Nome 2  | Sobrenome 2  | e-mail do |
| 4  | Nome 3  | Sobrenome 3  | e-mail do |
| 5  | Nome 4  | Sobrenome 4  | e-mail do |
| 6  | Nome 5  | Sobrenome 5  | e-mail do |
| 7  | Nome 6  | Sobrenome 6  | e-mail do |
| 8  | Nome 7  | Sobrenome 7  | e-mail do |
| 9  | Nome 8  | Sobrenome 8  | e-mail do |
| 10 | Nome 9  | Sobrenome 9  | e-mail do |
| 11 | Nome 10 | Sobrenome 10 | e-mail do |
| 12 | Nome 11 | Sobrenome 11 | e-mail do |
| 13 | Nome 12 | Sobrenome 12 | e-mail do |
| 14 | Nome 13 | Sobrenome 13 | e-mail do |
| 15 | Nome 14 | Sobrenome 14 | e-mail do |
| 16 | Nome 15 | Sobrenome 15 | e-mail do |
| 17 | Nome 16 | Sobrenome 16 | e-mail do |

|    | A          | B          |                 |
|----|------------|------------|-----------------|
| 1  | 12/03/2019 |            |                 |
| 2  | [2]        | 14/03/2019 |                 |
| 3  | [3]        | 19/03/2019 |                 |
| 4  | [2]        | 21/03/2019 |                 |
| 5  | [3]        | 26/03/2019 |                 |
| 6  | [2]        | 28/03/2019 |                 |
| 7  | [3]        | 02/04/2019 |                 |
| 8  | [2]        | 04/04/2019 |                 |
| 9  | [3]        | 09/04/2019 |                 |
| 10 | [2]        | 11/04/2019 |                 |
| 11 | [3]        | 16/04/2019 |                 |
| 12 | [2]        | 18/04/2019 |                 |
| 13 | [3]        | 23/04/2019 |                 |
| 14 | [2]        | 25/04/2019 |                 |
| 15 | [3]        | 30/04/2019 |                 |
| 16 |            | 1          | 1000001 Nome do |
| 17 |            | 2          | 1000002 Nome do |

Para o envio de dados de notas e faltas, obtidos no Moodle e sistema acadêmico, respectivamente, é necessário que ambas as planilhas estejam consolidadas em somente um arquivo, como abas, como apresentado na Figura 12. Por mais que reflita em uma etapa de trabalho manual demandado do usuário, isto é usado como um recurso para evitar o envio de notas e de faltas de períodos divergentes ou planilhas que não estão sincronizadas quanto a data corrente do período letivo. Por exemplo, o envio de notas dos alunos do 1º mês de aula e envio das faltas referentes a uma data posterior.

Os nomes das abas não influenciam na importação dos dados, contanto que a ordem delas seja a mesma da apresentada na Figura 12, sendo a primeira aba destinada às notas e a segunda aba destinada às faltas dos alunos.

Além destes dois arquivos, foi considerado como fonte de dados a plataforma URI, citada anteriormente na seção 3.2, desta forma o sistema desenvolvido também possibilita o

envio de planilhas com dados do URI.

Figura 13 – Modelo de planilha com dados exportada pela plataforma URI

|    | A      | B       | C                | D         | E      | F                      | G          | H     | I           |
|----|--------|---------|------------------|-----------|--------|------------------------|------------|-------|-------------|
| 1  | uri    | codigo  | students         | exercises | solved | tried (but not solved) | percentage | score | total_score |
| 2  | 100001 | 1000001 | Nome do aluno 1  | 7         | 7      | 0                      | 100,0      | 700   | 700         |
| 3  | 100002 | 1000002 | Nome do aluno 2  | 7         | 7      | 0                      | 100,0      | 700   | 700         |
| 4  | 100003 | 1000003 | Nome do aluno 3  | 0         | 0      | 0                      | 0,0        | 0     | 700         |
| 5  | 100004 | 1000004 | Nome do aluno 4  | 7         | 7      | 0                      | 100,0      | 700   | 700         |
| 6  | 100005 | 1000005 | Nome do aluno 5  | 0         | 0      | 0                      | 0,0        | 0     | 700         |
| 7  | 100006 | 1000006 | Nome do aluno 6  | 0         | 0      | 0                      | 0,0        | 0     | 700         |
| 8  | 100007 | 1000007 | Nome do aluno 7  | 1         | 1      | 0                      | 14,3       | 100   | 700         |
| 9  | 100008 | 1000008 | Nome do aluno 8  | 0         | 0      | 0                      | 0,0        | 0     | 700         |
| 10 | 100009 | 1000009 | Nome do aluno 9  | 7         | 7      | 0                      | 100,0      | 700   | 700         |
| 11 | 100010 | 1000010 | Nome do aluno 10 | 0         | 0      | 0                      | 0,0        | 0     | 700         |
| 12 | 100011 | 1000011 | Nome do aluno 11 | 7         | 7      | 0                      | 100,0      | 700   | 700         |
| 13 | 100012 | 1000012 | Nome do aluno 12 | 7         | 7      | 0                      | 100,0      | 700   | 700         |
| 14 | 100013 | 1000013 | Nome do aluno 13 | 1         | 1      | 0                      | 14,3       | 100   | 700         |
| 15 | 100014 | 1000014 | Nome do aluno 14 | 0         | 0      | 0                      | 0,0        | 0     | 700         |
| 16 | 100015 | 1000015 | Nome do aluno 15 | 3         | 3      | 0                      | 42,9       | 300   | 700         |
| 17 | 100016 | 1000016 | Nome do aluno 16 | 5         | 4      | 1                      | 57,1       | 400   | 700         |
| 18 | 100017 | 1000017 | Nome do aluno 17 | 0         | 0      | 0                      | 0,0        | 0     | 700         |
| 19 | 100018 | 1000018 | Nome do aluno 18 | 7         | 7      | 0                      | 100,0      | 700   | 700         |
| 20 | 100019 | 1000019 | Nome do aluno 19 | 7         | 7      | 0                      | 100,0      | 700   | 700         |
| 21 | 100020 | 1000020 | Nome do aluno 20 | 0         | 0      | 0                      | 0,0        | 0     | 700         |
| 22 | 100021 | 1000021 | Nome do aluno 21 | 4         | 1      | 3                      | 14,3       | 100   | 700         |
| 23 | 100022 | 1000022 | Nome do aluno 22 | 1         | 0      | 1                      | 0,0        | 0     | 700         |
| 24 | 100023 | 1000023 | Nome do aluno 23 | 3         | 3      | 0                      | 42,9       | 300   | 700         |
| 25 | 100024 | 1000024 | Nome do aluno 24 | 7         | 7      | 0                      | 100,0      | 700   | 700         |
| 26 | 100025 | 1000025 | Nome do aluno 25 | 7         | 7      | 0                      | 100,0      | 700   | 700         |
| 27 | 100026 | 1000026 | Nome do aluno 26 | 7         | 7      | 0                      | 100,0      | 700   | 700         |
| 28 | 100027 | 1000027 | Nome do aluno 27 | 0         | 0      | 0                      | 0,0        | 0     | 700         |
| 29 | 100028 | 1000028 | Nome do aluno 28 | 7         | 6      | 1                      | 85,1       | 600   | 700         |

Fonte: Autoria Própria

O sistema foi configurado para permitir o envio de arquivos gerados com extensões:

- xls exlsx – formatos padrões utilizados pelo gerenciador de planilhas Excel, da Microsoft;
- ods – formato utilizado pelo software open-source LibreOffice Calc;
- csv – formato bastante utilizado por sistemas de exportação de texto formatado por vírgulas.

A plataforma também distingue os nomes das colunas, então é recomendado que estes sejam descritos conforme os apresentados nas Figuras 9, 10 e 13. Algumas alternativas de nomes de colunas também podem ser utilizadas, como Trabalho 1 ou Atividade 1 ao invés de Micro-avaliação 1, e Prova 1 ao invés de Avaliação 1. É importante a inserção das datas das avaliações em seu nome na atividade do Moodle, por exemplo Avaliação 1 - 03/10/2019 ou Avaliação 1 (03/10/19), para que seja possível associar as faltas até esta data. Não é necessário seguir um padrão específico de separador entre o nome da atividade e a data, uma vez que é utilizada uma expressão regular para capturar a data no título, caso exista.

Lembrando que todas as planilhas citadas anteriormente são disponibilizadas pelos seus respectivos sistemas, assim a coleta destes e inserção no sistema de predição ainda é feita de forma manual, mas o trabalho manual de manipulação dos dados das planilhas foi eliminado.

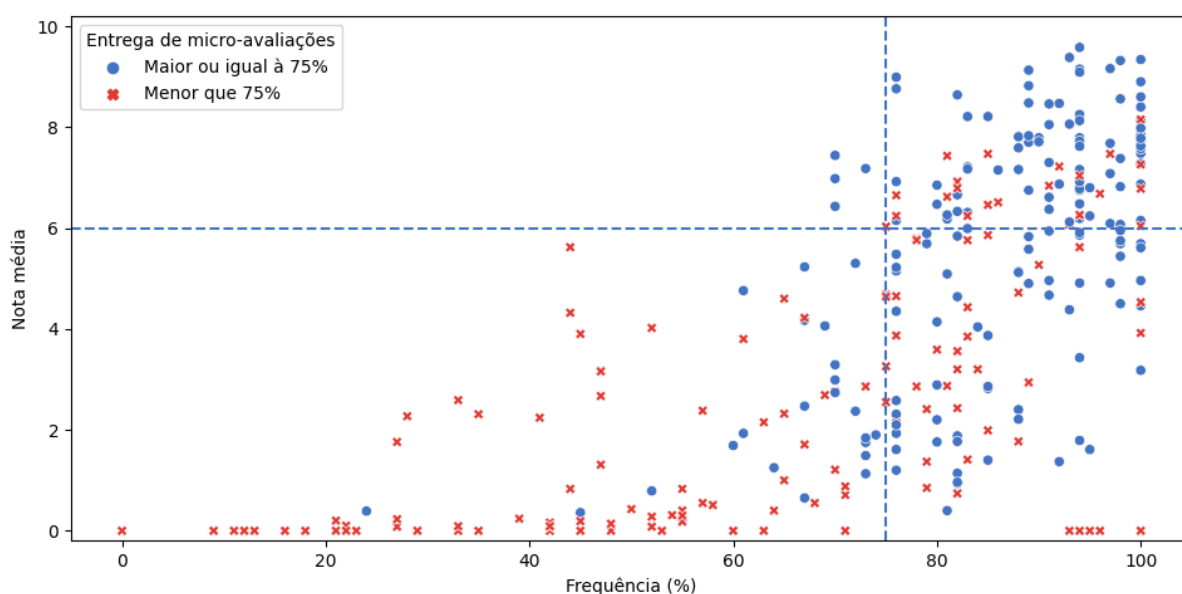
Após a aquisição dos dados iniciais, que foram utilizados para estruturar a aplicação, identificou-se a relação entre os tipos de planilhas inseridos por meio do código identificador do aluno. Foi possível verificar também que a planilha de notas poderia conter campos vazios ou

com caracteres especiais como o “-” (traço). Analisando, pode-se perceber que estes pertencem a alunos que não realizaram provas ou atividades, e estes foram substituídos por um valor padrão, neste caso o valor 0 (zero).

Foi-se então realizada uma análise dos dados, considerando a princípio as notas das avaliações, a realização de trabalho, a resolução de exercícios do URI e as frequências dos alunos, com isto gerou-se alguns gráficos.

A Figura 14 mostra um gráfico de dispersão contendo no eixo horizontal as frequências dos alunos e no eixo vertical a nota média destes mesmos alunos. Além disso, os pequenos círculos ilustram os alunos que realizaram 75% ou mais das atividades, enquanto que os símbolos em X ilustram os alunos que realizaram menos de 75% dos trabalhos. Pode-se perceber uma concentração de maiores notas quando os alunos possuem maiores frequências nas aulas e a realização dos trabalhos propostos torna ainda mais evidente esta relação. As linhas pontilhadas segmentam o gráfico em quatro quadrantes, tendo como critérios a nota igual à 6 e frequência igual à 75%, que são os critérios necessários para aprovação na UTFPR-CP. O quadrante superior direito é o considerado ideal, visto que os alunos que figuram ali obtiveram valores dentro dos critérios de aprovação.

Figura 14 – Gráfico de dispersão com as frequências e notas dos alunos, considerando a realização dos trabalhos

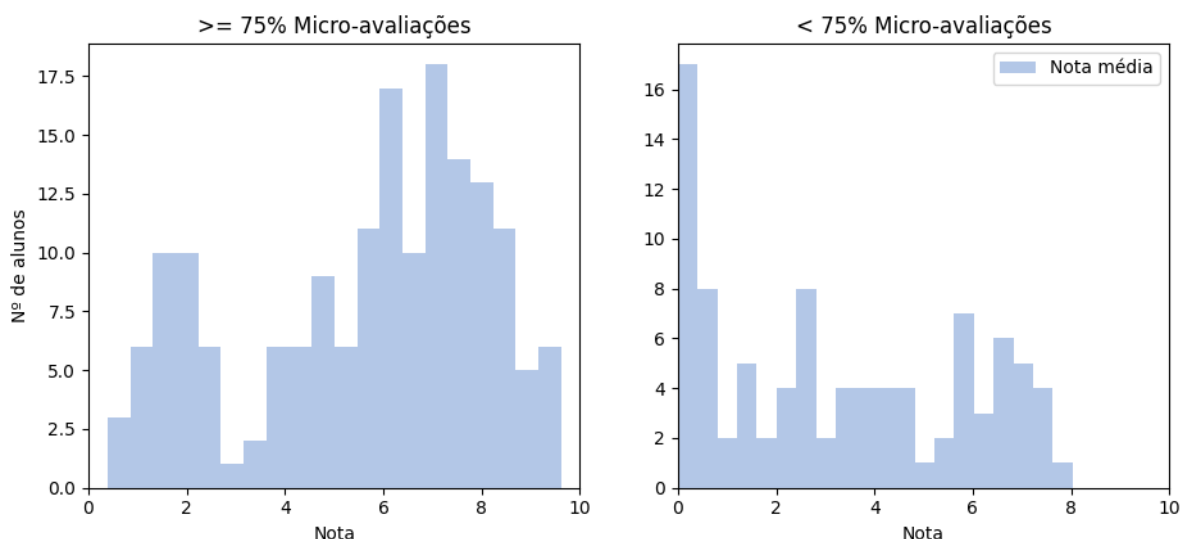


Fonte: Autoria Própria

A Figura 15 apresenta um gráfico com dois *subplots*, no eixo horizontal de ambos é apresentado a nota final obtida, já o eixo vertical mostra a quantidade de alunos com esta nota. O *plot* da esquerda é composto pelos alunos que realizaram 75% ou mais dos trabalhos/micro-avaliações e o da direita pelos alunos que realizaram menos de 75%. Pode-se observar uma maior concentração de alunos com boas notas na ilustração à esquerda, já na da direita a distribuição das notas é diferente, não havendo notas maiores do que oito.



Figura 15 – Histograma com as quantidades de alunos por faixa de notas



Fonte: Autoria Própria

As Figuras 16 e 17 também mostram as notas e o número de alunos, mas segmentando-os pela resolução dos exercícios do URI *Online Judge*. Para ambas as figuras, a ilustração da esquerda mostra os alunos que resolveram 75% ou mais das micro-avaliações propostas, enquanto que nos gráficos à direita são apresentados os alunos que resolveram menos de 75% das micro-avaliações.

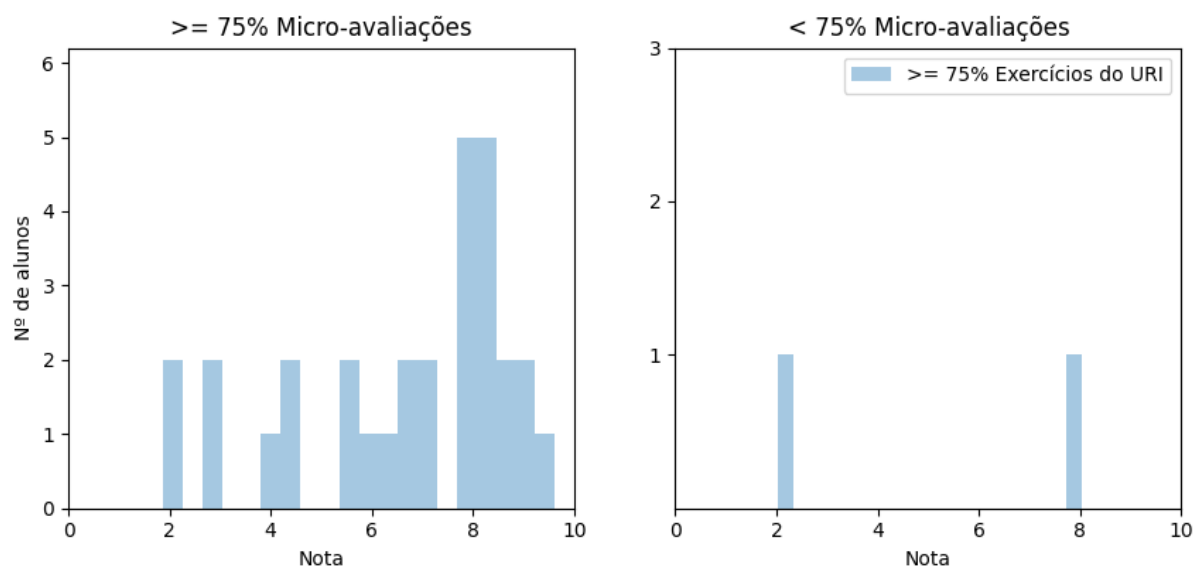
Ao averiguar as figuras é possível perceber que a realização dos trabalhos em conjunto da resolução dos exercícios do URI (Figura 16 - esq.) apresentou os melhores resultados, neste a distribuição das notas foi ampla, mas com uma maior concentração entre 5,5 e 9,5.

Ocorreram somente dois casos onde foram realizados menos de 75% parte dos trabalhos e maioria dos exercícios do URI (Figura 16 - dir.) e estes foram equilibrados, já que a média aritmética entre eles é cinco. Já para o caso contrário, alunos com mais de 75% dos trabalhos e poucos exercícios do URI (Figura 17 - esq.) a distribuição é esparsa e possui resultados relativamente bons, mas apresentando mais casos com notas finais entre 0,5 e 3. Esse gráfico também mostra que há uma tendência dos alunos que fazem os trabalhos também se engajarem a fazer os exercícios propostos do URI, mas o contrário não é válido.

Por fim tem-se os alunos que, considerando o critério de 75%, não realizaram os trabalhos nem os exercícios (Figura 17 - dir.). As notas refletem a falta de empenho dos alunos e evidentemente mostram a pior distribuição dentro os quatro gráficos. Houveram alguns casos positivos, mas em contrapartida a concentração se deu entre 0 e 3, desses, 15 alunos com notas iguais ou próximas à zero.

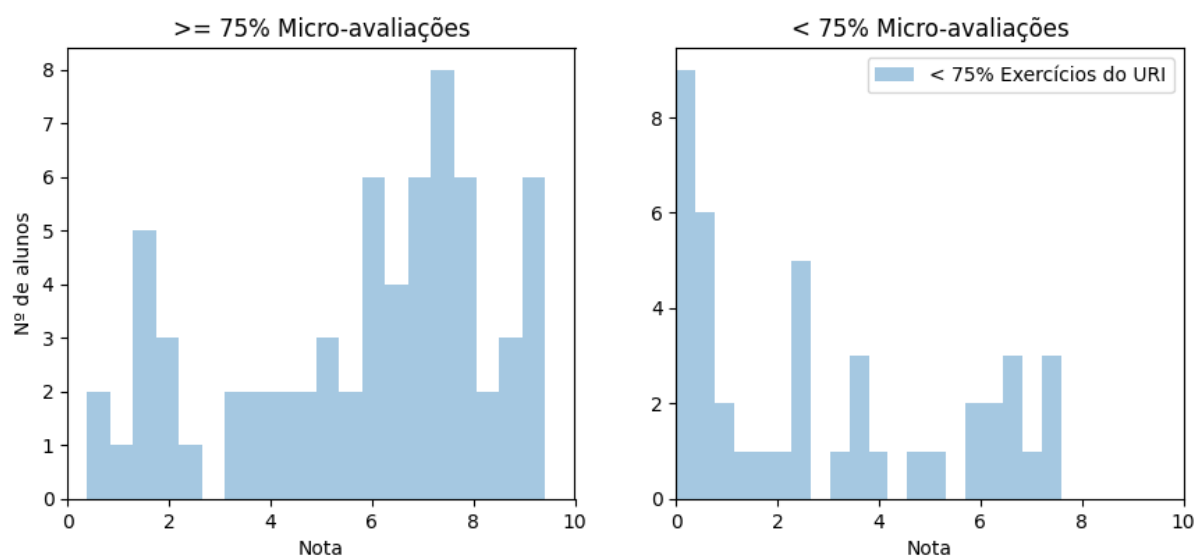
Para a concepção dos gráficos das figuras 15, 16 e 17 considerou-se um subconjunto de 263 alunos, que foram os alunos que obtiveram frequência acima de 25% nos períodos letivos, a fim de evitar viés nas análises devido a dados de alunos que nunca compareceram nas aulas ou em pouco tempo deixaram de frequentá-las.

Figura 16 – Histograma com as quantidades de alunos que realizaram 75% ou mais dos exercícios do URI, segmentando pela porcentagem realizada de trabalhos



Fonte: Autoria Própria

Figura 17 – Histograma com as quantidades de alunos que realizaram menos de 75% dos exercícios do URI, segmentando pela porcentagem realizada de trabalhos



Fonte: Autoria Própria

### 3.4 PREPARAÇÃO DOS DADOS

Como já descrito na seção 2.3.3, esta etapa consiste em tratar os dados a fim de dispor o conjunto de dados que será utilizado para o processamento e construção do modelo de mineração de dados, bem como para outros fins, como para apresentação e construção dos gráficos.

Nesta fase foram criados *scripts* para formatação dos dados, a fim de permitir o uso



destes para o processamento sem ruídos, incoerências ou redundâncias, reduzindo assim as chances de resultados anômalos.

Inicialmente implementou-se um *script* no *back-end* em Node.js, que efetua a formatação dos dados iniciais para um formato mais adequado para a construção de gráficos e uso nos modelos de predição. Neste *script* os dados das planilhas são convertidos para o formato JSON utilizando a biblioteca *SheetJS*, que torna mais simples a manipulação dos dados, já que o formato é nativo na linguagem de programação *Javascript/Typescript*, utilizada no projeto. Em seguida, as notas e frequências, inicialmente armazenadas em tipo texto são convertidas para valores numéricos, e as letras que eram utilizadas para representar o status de presença do aluno foram convertidas também para valores numéricos. Neste cenário a letra C representa os alunos presentes e a letra D representam que este havia sido dispensado com alguma justificativa, como a participação em algum evento ocorrente na universidade, e por isso não havia a contagem de faltas.

Como pode-se observar, as etapas do modelo CRISP-DM são interconectadas mas não lineares, visto que este último tratamento citado envolve também a etapa de entendimento dos dados.

A [Figura 18](#) apresenta a estrutura final dos dados de notas, que compõe além das notas alguns outros dados essenciais para o desenvolvimento do presente trabalho, que serão melhor detalhados posteriormente. A estrutura dos dados é composta pelos seguintes campos:

- **\_id:** ID de Objeto único gerado para este documento pelo MongoDB, ou seja, é a chave primária deste registro no banco de dados;
- **code:** código identificador do aluno;
- **course\_unit:** código da disciplina da turma que o aluno está vinculado;
- **period:** período letivo da turma;
- **name:** contém o nome completo do aluno;
- **email:** campo que contém o e-mail do aluno;
- **attendance:** contém a frequência, em percentual, do aluno para a disciplina associada a este registro, é calculada com base na quantidade de aulas dadas;
- **current\_grade:** contém a nota atual do aluno, é calculada com base no peso aplicado para avaliações e atividades;
- **feedbacks:** este campo é destinado à lista de notificações enviadas por e-mail ao aluno pela plataforma, juntamente com as justificativas ou *feedbacks* enviadas pelo aluno, para cada e-mail recebido.

Além desses, os campos a seguir são gerados dinamicamente conforme as colunas existentes na planilha origem do dado, podendo haver um ou mais campos para notas de avaliação, gerando assim campos *gradeN*, sendo N o número de cada avaliação. Bem como é possível não haver campos para datas de prova, caso esta não esteja presente no cabeçalho da coluna na planilha. A mesma lógica se aplica para as atividades, podendo assim existir zero, uma ou múltiplas atividades realizadas pelo aluno.

- **grade1**: contém a nota obtida pelo aluno para a 1ª avaliação da disciplina associada a este registro;
- **grade1\_date**: contém a data da 1ª avaliação;
- **micro\_grade1**: contém a nota obtida pelo aluno para a 1ª atividade realizada na disciplina.

Figura 18 – Fragmento do conjunto de dados de notas no formato pós preparação

```
{
  "_id": "5dc73fb7bcae0f812a61c3aa",
  "code": 1000001,
  "course_unit": "ES31C",
  "name": "Nome do aluno 1",
  "email": "aluno@email.com",
  "attendance": 95,
  "grade1": 70,
  "grade1_date": "03/10/2019",
  "micro_grade1": 40,
  "current_grade": 64,
  "feedbacks": []
}
```

Fonte: Autoria Própria

Então tem-se as faltas, que são apresentadas por cada aula dada e transformadas para um formato acumulativo, o que facilita também a construção dos gráficos. Todas estas transformações são realizadas sobre dados no formato apresentado anteriormente na [Figura 10](#), e o resultado é apresentado na [Figura 19](#), que exibe o conjunto de dados pós-preparação.

Por fim os dados são armazenados no banco de dados com o uso do *Mongoose*, e a [Figura 19](#) mostra a estrutura de dados JSON no formato final, que é composta pelos seguintes campos:

- **\_id**: ID de Objeto único gerado para este documento pelo MongoDB, ou seja, é a chave primária deste registro no banco de dados;
- **code**: código identificador do aluno;
- **course\_unit**: Código da disciplina ao qual o aluno está vinculado;
- **period**: período letivo da turma;
- **active**: campo booleano utilizado para definir o status do aluno na disciplina, caso a frequência seja verificada abaixo de 25% do total de aulas o aluno é considerado inativo, caso contrário é ativo;
- **total**: contém o número total de faltas do aluno para a disciplina relacionada no momento do envio dos dados;
- **absences**: é um campo do tipo *array*, que contém um objeto chave-valor para cada dia de aula dada até o momento. Os campos de cada objeto representam a data da aula e o número de faltas acumuladas até o momento.

- **id:** campo opcional, é populado caso a planilha carregada possua uma coluna nomeada como ID e é utilizado para ordenação.

Figura 19 – Fragmento do conjunto de dados de faltas no formato pós preparação

```
{
  "_id": "5dc73fb7bcae0f812ae1c3ca",
  "code": 1000001,
  "course_unit": "ES31C",
  "active": true,
  "total": 2,
  "absences": [
    { "date": "20/08/19", "absences": 0 },
    { "date": "22/08/19", "absences": 0 },
    { "date": "27/08/19", "absences": 0 },
    { "date": "29/08/19", "absences": 0 },
    { "date": "03/09/19", "absences": 0 },
    { "date": "05/09/19", "absences": 0 },
    { "date": "10/09/19", "absences": 0 },
    { "date": "12/09/19", "absences": 0 },
    { "date": "17/09/19", "absences": 0 },
    { "date": "19/09/19", "absences": 0 },
    { "date": "24/09/19", "absences": 0 },
    { "date": "26/09/19", "absences": 0 },
    { "date": "01/10/19", "absences": 0 },
    { "date": "03/10/19", "absences": 0 },
    { "date": "08/10/19", "absences": 2 },
    { "date": "10/10/19", "absences": 2 }
  ],
  "id": "1"
}
```

Fonte: Autoria Própria

A [Figura 20](#) mostra parcela dos dados após a preparação realizada ao receber a planilha de dados exportada pela plataforma URI. Estes são compostos pelos seguintes campos:

- **\_id:** ID de Objeto único gerado para este documento pelo MongoDB, ou seja, é a chave primária deste registro no banco de dados;
- **code:** código identificador do aluno;
- **course\_unit:** código da disciplina ao qual o aluno está vinculado e os exercícios do URI são referentes;
- **period:** período letivo da turma;
- **name:** contém o nome completo do aluno;
- **exercises:** contém a quantidade de exercícios aos quais o aluno solucionou completa ou parcialmente;
- **solved:** contém a quantidade de exercícios que o aluno finalizou corretamente;
- **tried:** contém a quantidade de exercícios que o aluno tentou solucionar, mas não os finalizou;
- **percentage:** contém a porcentagem de exercícios solucionados de forma integral;
- **total:** contém a quantidade total de exercícios para a lista em questão;

- **exam:** contém a referência da avaliação a qual este registro é associado.

Figura 20 – Fragmento do conjunto de dados obtidos na plataforma URI no formato pós preparação

```
{
  "_id": "5dc7150c0b0b3c545c715b55",
  "code": 1000001,
  "course_unit": "ES31C",
  "name": "Nome do aluno 1",
  "exercises": 7,
  "solved": 7,
  "tried": 0,
  "percentage": 100,
  "total": 7,
  "exam": 1
}
```

Fonte: Autoria Própria

Vale ressaltar que para um volume pequeno de dados, estas alterações poderiam ser efetuadas manualmente sem grande demanda de tempo, mas para um volume maior, como seria para o uso real da plataforma desenvolvida após implantação, se faz necessário a criação de rotinas de preparação de dados para a realização destas tarefas. Ressalta-se também que este trabalho tem como objetivo automatizar o maior número possível de tarefas, demandando o mínimo de atuação manual por conta do usuário após a aquisição dos dados.

### 3.5 MODELAGEM

Uma vez realizada a preparação dos dados, é possível seguir para a implementação dos modelos de aprendizado de máquina para mineração de dados.

Neste trabalho foram utilizados três algoritmos de aprendizado de máquina supervisionados, são estes: a regressão linear, o *Random Forest* e o *Gradient Boosting*, descritos na seção 2.2, que descreve a mineração de dados. Como um dos objetivos deste trabalho é prever o desempenho dos alunos (valores numéricos e contínuos) com base em suas notas e faltas, os algoritmos foram utilizados como técnicas de regressão, mesmo o *Random Forest* e o *Gradient Boosting* possuindo a capacidade de atuar como técnicas classificadoras de dados.

Desta forma foram implementados os algoritmos com o auxílio das bibliotecas de aprendizado de máquina *Scikit-Learn* e de manipulação de dados *Pandas*.

O trecho de código apresentado na listagem 3.1 é utilizado para treinar os modelos de aprendizado de máquina citados. O conjunto de dados foi dividido em subconjuntos de treinamento e teste, utilizando a razão 80%/20%. O modelo baseado em *Gradient Boosting* foi gerado com 400 estimadores, que internamente se traduzem em árvores de regressão em formato *boosting*, descrito na seção 2.2.5, com profundidade máxima igual a cinco e empregando a função erro *least squares*, ou mínimos quadrados. O modelo que utiliza *Random Forest* foi

gerado utilizando também 400 estimadores, que desta vez geram árvores de regressão em formato *bagging*, descrito na seção 2.2.4. Já o algoritmo de regressão linear não demanda nenhum argumento específico para a criação e treinamento de modelos.

Listagem 3.1 – Trecho de código para treinamento dos modelos de aprendizado de máquina

```
## Divisao do conjunto de dados em treinamento (80%) e teste (20%).
x_train, x_test, y_train, y_test = train_test_split(data.values,
    labels.values, test_size=0.2, random_state=42)

## Gradient Boosting Model
gradient_model = GradientBoostingRegressor(n_estimators=400, max_depth=5,
    loss='ls')
gradient_model.fit(x_train, y_train)

## Random Forest Model
random_forest = RandomForestRegressor(n_estimators=400)
random_forest.fit(x_train, y_train)

## Linear Regression Model
linear_regression = LinearRegression()
linear_regression.fit(x_train, y_train)
```

A biblioteca *Scikit-learn* possibilita extrair a importância, ou peso, de cada *feature* dos modelos treinados com os algoritmos *Gradient Boosting* e *Random Forest*, ou seja, é possível obter a porcentagem que cada variável utilizada no treinamento influencia nas estimativas retornadas pelos modelos. A Tabela 1 mostra a importância das *features* no modelo treinado com o algoritmo *Gradient Boosting*, considerando dados de turmas que realizaram duas avaliações e três atividades/trabalhos. A Tabela 2 também apresenta a importância das *features*, entretanto para o modelo treinado com o algoritmo *Random Forest*.

Tabela 1 – Importância das *features* para o modelo treinado com *Gradient Boosting*

| Feature      | Importância |
|--------------|-------------|
| grade2       | 67,4%       |
| grade1       | 26,5%       |
| attendance   | 3,0%        |
| micro_grade2 | 1,3%        |
| micro_grade3 | 1,1%        |
| micro_grade1 | 0,6%        |

As importâncias apresentadas nas Tabela 1 e Tabela 2 consideram um cenário onde

Tabela 2 – Importância das *features* para o modelo treinado com *Random Forest*.

| Feature      | Importância |
|--------------|-------------|
| grade2       | 71,5%       |
| grade1       | 24,0%       |
| attendance   | 1,9%        |
| micro_grade3 | 0,1 %       |
| micro_grade2 | 0,1 %       |
| micro_grade1 | 0,1 %       |

todos os dados dos alunos já estão disponíveis, o que só seria possível no final do período letivo. Como um dos objetivos deste trabalho é realizar a predição dos desempenhos durante o período letivo, foram treinados diversos modelos de aprendizado de máquina, considerando diferentes cenários nos quais parte dos dados ainda não estão disponíveis. No total foram gerados 10 cenários e para cada um deles treinou-se três modelos, um para cada um dos algoritmos apresentados. A [Tabela 3](#) apresenta a lista de cenários nomeados de **A** à **J**, com as respectivas *features* necessárias para o treinamento dos modelos. Com isso, já é possível ter uma predição preliminar do desempenho dos alunos com os dados parciais.

Tabela 3 – Cenários considerados para os modelos de aprendizado de máquina treinados.

| Feature       | Cenários |   |   |   |   |   |   |   |   |   |
|---------------|----------|---|---|---|---|---|---|---|---|---|
|               | A        | B | C | D | E | F | G | H | I | J |
| attendance    | x        | x | x | x | x | x | x | x | x | x |
| micro_grade1  | x        | x | x | x | x | x | x | x | x | x |
| micro_grade2  | x        | x | x | x | x | x | x | x | x | x |
| micro_grade3  |          | x | x |   | x | x | x | x | x | x |
| micro_grade4  |          |   | x |   |   |   | x | x |   | x |
| percentage    |          | x | x | x |   | x | x | x |   | x |
| grade1        |          |   |   | x | x | x | x | x | x | x |
| project_grade |          |   |   |   |   |   |   | x |   | x |
| grade2        |          |   |   |   |   |   |   |   | x | x |

Em contrapartida que cria-se esta oportunidade de identificar previamente o resultado final dos alunos, há uma queda proporcional na acurácia dos modelos, uma vez que conforme reduz-se a qualidade dos dados de entrada, a qualidade e acurácia dos dados de saída também é reduzida.

Tanto as acurácias quanto as importâncias das *features* para os modelos treinados são apresentadas em maiores detalhes na seção [4.1](#).

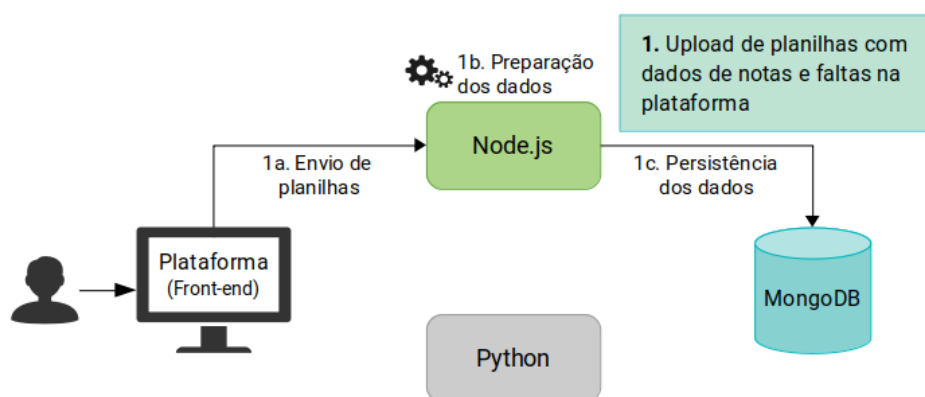
### 3.6 IMPLEMENTAÇÃO

Para esta etapa do trabalho foi construído um sistema web, que tem como finalidade permitir o acesso aos dados, gráficos e artefatos implementados. O sistema é composto por

seis páginas, sendo que quatro destas estão dispostas no menu lateral da plataforma, são elas: **Dashboard**, **Configurações**, **Upload de Dados**, e **Lista**. A quinta página é a de **Detalhes do Estudante** e pode ser acessada pelo dashboard ou pela Lista ao clicar em uma das linhas das listas de alunos. E, a última página é a de **Feedback**, acessada somente pelos alunos por meio de um link exclusivo. Este recurso será melhor explorado posteriormente na seção 3.6.4.

A Figura 21 apresenta as chamadas realizadas entre os componentes do sistema quando realizado o envio de uma planilha. Inicialmente o usuário carrega a planilha por meio da página de *upload* de dados do sistema, essa é enviada ao *back-end* escrito em Node.js, que por sua vez realiza a manipulação da planilha, transformando-a em formato JSON e executando o *script* de preparação de dados. Em seguida os dados formatados são inseridos na base de dados MongoDB com auxílio do *Mongoose*.

Figura 21 – Fluxo de interações entre os componentes do sistema no envio de dados

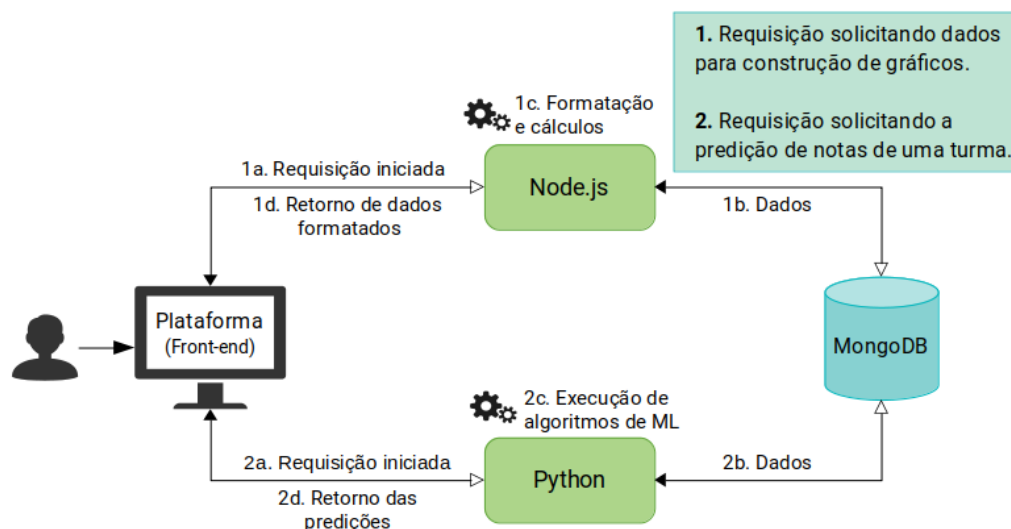


Fonte: Autoria Própria

Já para as chamadas que requisitam dados, a Figura 22 ilustra o processo, que pode percorrer dois caminhos distintos:

1. **Dados gerais:** estes são utilizados para apresentação, listagem ou construção de gráficos no *front-end*, e para este tipo de dados são realizadas chamadas ao *back-end* Node.js, que busca os dados no BD, transforma-os conforme a demanda específica, realiza os cálculos necessários para apresentação, como a média de faltas da turma para cada aula do período letivo e a média das notas dos alunos em cada disciplina, e então retorna-os ao *front-end*.
2. **Predições:** neste caso a chamada é enviada para o *back-end* em Python, que busca os dados necessários no BD, reformata-os para o uso com os modelos de aprendizado de máquina criados e citados anteriormente, busca o modelo correto a ser carregado com base nos dados atuais do aluno e então executa os algoritmos, retornando os resultados das predições.

Figura 22 – Fluxo de interações entre os componentes do sistema em requisições solicitando dados e previsões



Fonte: Autoria Própria

### 3.6.1 Dashboard

A página inicial da plataforma consiste de um dashboard, apresentado nas Figuras 23 e 24. Os indicadores apresentados nas cores verde, amarelo e vermelho no topo da página levam em conta a nota dos alunos e mostram o número de alunos em situações de tendência a aprovação, situações com notas medianas e situações com um risco maior de reprovação, respectivamente. Para esta separação foram criados dois *thresholds* e com base em um sistema de notas de 0 a 100 foram sugeridas as seguintes regras:

- **Adequado:** nota prevista maior ou igual a 70;
- **Atenção:** nota prevista entre 50 e 70;
- **Alto risco:** nota prevista abaixo de 50;

Estas regras foram criadas como uma forma de priorização de perfis de alunos na tomada de ações, ou seja, os usuários do sistema interessados em interagir com os alunos ou desenvolver mecanismos para prevenir a retenção podem iniciar pelos perfis de *Alto risco* como seu público alvo. Da mesma forma pode-se tomar diferentes ações com o perfil *Atenção*, uma vez que estes alunos, em tese, demandam menos esforço para garantir a aprovação nas disciplinas.

Localizados na direita superior da Figura 23 tem-se o número total de alunos cadastrados na plataforma e uma caixa de seleção flutuante que permite selecionar uma das disciplinas cadastradas e um período letivo. Ainda na Figura 23 tem-se o gráfico temporal de faltas dos alunos de uma turma, comparando-as com o limite parcial para reprovação por faltas, atualmente de 25% do total de aulas dadas no período letivo. O eixo horizontal é composto pelas datas das aulas até o momento em que os dados foram carregados na plataforma, e o

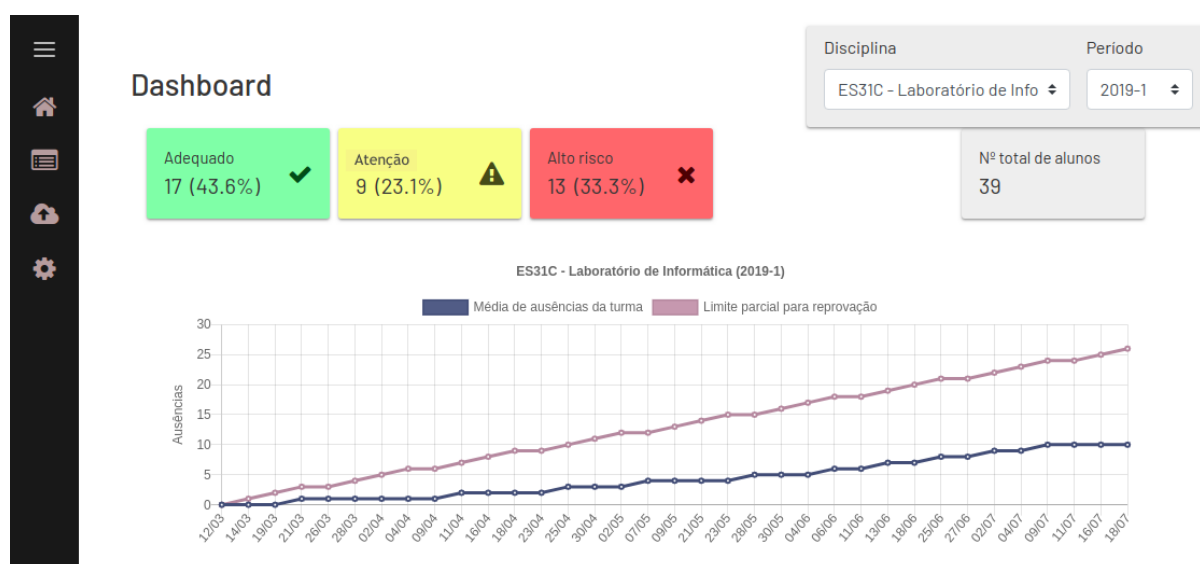


eixo vertical mostra a média das ausências em cada aula dos alunos de uma turma, neste caso (Figura 23) da disciplina ES31C - Laboratório de Informática para o primeiro semestre do ano de 2019.

Para obter números mais práticos e condizentes com a realidade da disciplina, no cálculo da média de faltas não são considerados os alunos inativos, ou seja, que possuem menos de 25% de frequência, pois a inserção de dados de alunos que cancelaram sua matrícula, nunca foram as aulas ou foram somente nas aulas iniciais, poderia influenciar negativamente nos números gerais. Estes alunos ainda podem ser analisados individualmente, porém, para efeitos de cálculos coletivos eles não são considerados.

A escolha do *threshold* para determinar se um aluno é ativo ou inativo foi estipulada de forma empírica, observando os dados disponíveis, mas caso seja observado ao longo do tempo um valor mais adequado para descrever esse comportamento, o *threshold* poderá ser alterado.

Figura 23 – Fragmento superior do dashboard



Fonte: Autoria Própria

A Figura 24 apresenta a segunda parte do dashboard. O primeiro gráfico apresentado nesta tela tem como objetivo mostrar a relação entre a frequência e notas dos alunos, caso exista, levando em conta se estes realizaram todas as atividades/trabalhos ou não. O eixo horizontal contém os percentuais de frequência dos alunos cadastrados, confrontando-os com as notas atuais dos mesmos, no eixo vertical. Os pontos de cor azul ilustram os alunos que entregaram todas as atividades, independente das notas obtidas. Já os pontos avermelhados ilustram alunos que deixaram de realizar pelo menos uma das atividades. Para a composição deste gráfico as atividades realizadas com nota igual a zero são consideradas como não realizadas.

A tabela à direita do gráfico citado anteriormente mostra os principais pontos de atenção, ou seja, a lista dos alunos com menores notas entre os cadastrados nesta disciplina,

descartando da listagem os alunos considerados inativos. Ao clicar em um dos alunos é apresentada a tela de detalhes do aluno.

Em seguida são apresentados os gráficos de barras e de rosca, que ilustram a proporção de alunos em situação de aprovação, média e reprovação. O primeiro segmenta os dados por disciplina, mostrando todas as disciplinas cadastradas na plataforma, enquanto o segundo procura mostrar uma visão das proporções para a disciplina selecionada.

Figura 24 – Fragmento inferior do dashboard



Fonte: Autoria Própria

### 3.6.2 Configurações

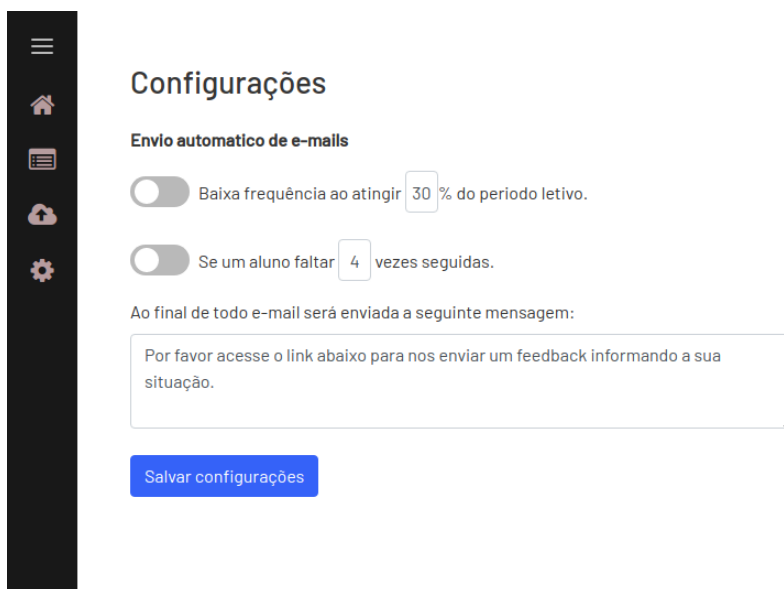
A primeira tela a ser utilizada ao acessar à plataforma pela primeira vez é a tela de configurações, esta é apresentada na [Figura 25](#).

Nesta tela o usuário pode selecionar dois mecanismos de detecção automática no carregamento de planilhas.

O primeiro permite que o sistema detecte, ao passar uma parte do período letivo (espécie de *checkpoints*), os alunos com frequência abaixo do limite de 75%. O segundo mecanismo permite detectar alunos que faltaram múltiplas vezes seguidas, o que pode refletir alguma situação incomum ocorrendo com o aluno, alguns exemplos são: i) o momento de desistência da disciplina; ii) algum impedimento médico; iii) a ocorrência de um evento rotineiro que o impossibilite de participar das aulas.

Com a detecção destes eventos, o sistema permite o envio de um e-mail para todos os alunos que se enquadram nessas situações. Desta forma, há um campo nesta página para que o usuário possa personalizar a mensagem a ser enviada aos alunos nestes e-mails. E, caso

Figura 25 – Página de configurações



Fonte: Autoria Própria

habilitado, os mecanismos também dispõem de um campo cada, que permite que a mensagem enviada por conta da detecção específica seja personalizada.

Como exemplificado na [Figura 26](#), as mensagens padrões para os mecanismos tentam ser brandas e trazer um tom de informalidade. As mensagens padrões são as seguintes:

- Baixa frequência ao atingir 30% do período letivo: Chegamos a #percentage do semestre e percebemos que você está com bastante falta. Você poderia nos dizer se há algum motivo específico para o ocorrido?
- Se um aluno faltar 4 vezes seguidas: Percebemos que você está faltando bastante, foram #times vezes seguidas. Existe algum motivo específico para essa sequência de faltas?
- Ao final do e-mail será enviada a seguinte mensagem: Você poderia acessar o link abaixo e por favor nos enviar uma justificativa para nos deixar a par do que está acontecendo?

Os termos precedidos pelo símbolo # representam variáveis que são alteradas automaticamente no envio do e-mail pelo valor gatilho selecionado para o evento. Ou seja, para este caso a expressão #percentage será substituída por 30% e a expressão #times será substituída por 4.

### 3.6.3 Upload de Dados

A próxima página do fluxo de uso da plataforma é a de *upload* de dados, esta é apresentada na [Figura 27](#). Nesta tela o usuário pode selecionar a disciplina e o período letivo associados aos dados que serão carregados.

A API em Node.js está preparada para a inserção de novas disciplinas e possui uma rota implementada para esse procedimento, porém o sistema não possui uma tela de cadastro

Figura 26 – Página de configurações com campos expandidos

**Configurações**

**Envio automatico de e-mails**

☒ Baixa frequência ao atingir  % do periodo letivo.

**Mensagem**

Atingimos #percentage do semestre e parece que você está com bastante faltas. Existe algum motivo específico para o ocorrido?

☒ Se um aluno faltar  vezes seguidas.

**Mensagem**

Parece que você está faltando bastante, foram #times vezes seguidas. Existe algum motivo específico para essa sequência de faltas?

Ao final de todo e-mail será enviada a seguinte mensagem:

Você poderia acessar o link abaixo e por favor nos enviar uma justificativa para nos deixar a par do que está acontecendo?

[Salvar configurações](#)

Fonte: Autoria Própria

de disciplinas. Desta forma, algumas disciplinas foram cadastradas a fim de exemplificação. Um dos campos esperados na inserção de uma nova disciplina no banco de dados é a data de término do período letivo, utilizado para o cálculo da frequência com base no número de aulas já dadas e faltantes.

Em seguida o usuário deve selecionar o tipo de planilha que será carregada, podendo esta ser de notas e faltas, consolidadas em um arquivo único, ou dados do *URI Online Judge*.

Caso selecionado a opção *Notas e Faltas* são exibidos campos de pesos, onde o docente deve inserir o peso das avaliações, o peso dos trabalhos e do projeto final para a composição da nota final dos alunos. Caso a disciplina não possua um projeto final, o docente pode deixar o campo com peso 0 e este será desconsiderado nas análises. Os valores são em porcentagem, desta forma a soma dos três pesos deve ser igual a 100. Ao modificar um dos campos de peso, automaticamente, o outro campo será ajustado, complementando-o a 100.

Por fim, seleciona-se a planilha a ser submetida e se tudo estiver correto, finaliza-se o envio.

Na parte inferior da página é disposta uma seção guia para os usuários, com os formatos recomendados para as planilhas a serem enviadas. Ao clicar nas setas ao lado de uma das opções esta é expandida e o *layout* é apresentado. A [Figura 28](#) mostra o resultado exibido na tela quando as opções estão expandidas.

Ao finalizar o envio da planilha, internamente o sistema realizará a detecção dos alunos

Figura 27 – Página de upload de planilhas

**Upload de Dados**

Disciplina: AS31A - Laboratório de Informática

Período: 2019-2

☒ Notas e faltas ☐ URI

Pesos (%)

Avaliações: 90      Micro-avaliações: 10      Projeto final: 0

Clique para selecionar o arquivo.

Enviar arquivo

**Formatos de planilhas para upload**

- ✓ Notas e faltas
- ✓ Dados do URI

Fonte: Autoria Própria

Figura 28 – Página de upload de planilhas com formatos para upload expandidos

**Formatos de planilhas para upload**

^ Notas e faltas

Notas

| Código  | Nome             | E-mail            | Micro-avaliação 1 | Micro-avaliação 2 | Avaliação 1 - 05/10/19 |
|---------|------------------|-------------------|-------------------|-------------------|------------------------|
| 1000001 | Mariana de Souza | mariana@email.com | 6                 | 6                 | 73                     |
| 1000002 | João da Silva    | joao@email.com    | 7                 | 9                 | 91                     |

Faltas

| Código  | 20/08/19 | 22/08/19 | 27/08/19 | 29/08/19 | ... | 10/10/19 | Total |
|---------|----------|----------|----------|----------|-----|----------|-------|
| 1000001 | C        | C        | C        | 2        | ... | C        | 2     |
| 1000002 | C        | 3        | C        | 3        | ... | C        | 6     |

^ Dados do URI

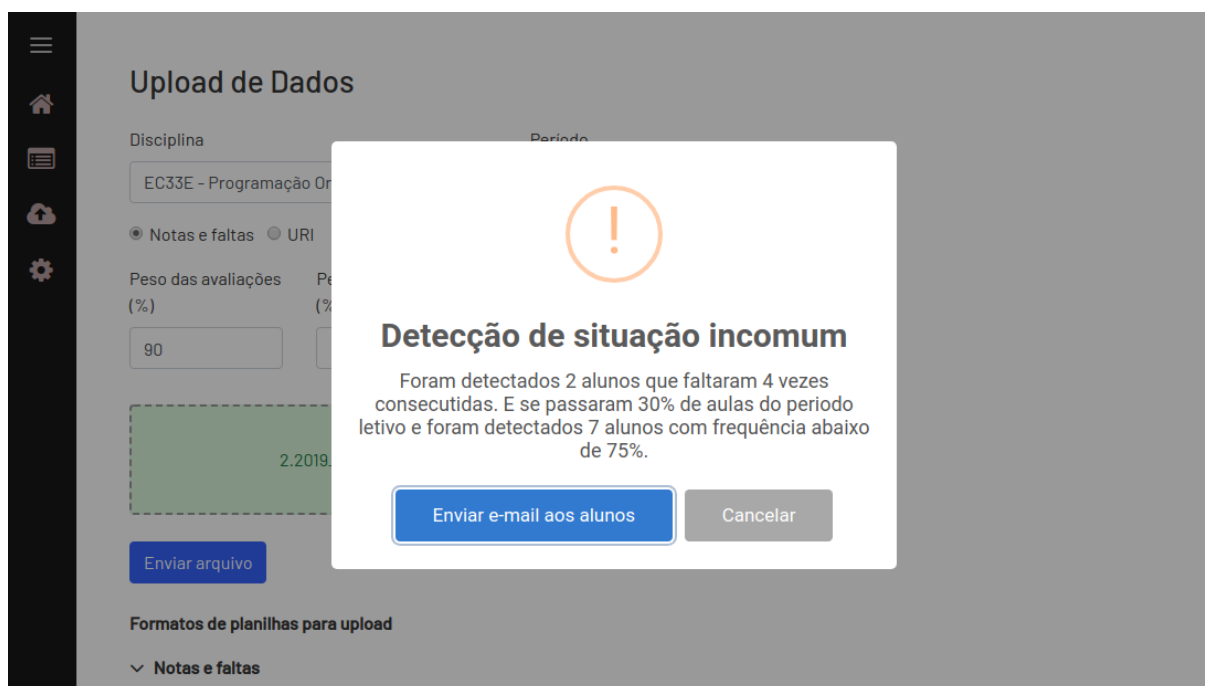
| URI    | Código  | Students         | Exercises | Solved | Tried (but not solved) | Percentage |
|--------|---------|------------------|-----------|--------|------------------------|------------|
| 380001 | 1000001 | Mariana de Souza | 3         | 3      | 0                      | 100        |
| 380002 | 1000002 | João da Silva    | 2         | 2      | 0                      | 66,67      |

Fonte: Autoria Própria

cujos dados coincidem com os mecanismos assinalados na tela de configurações e exibirá uma

página *modal* de alerta com um resumo da situação, solicitando ao usuário a confirmação para o envio dos e-mails a todos os alunos dentro dos perfis detectados. Caso confirmado, o e-mail é enviado pelo endereço de e-mail remetente *predicao.academica@gmail.com*, criado somente para a validação deste trabalho, e a mensagem enviada por e-mail aos alunos é construída com base nas mensagens especificadas na tela de configurações, apresentada anteriormente. A Figura 29 ilustra a situação de detecção descrita.

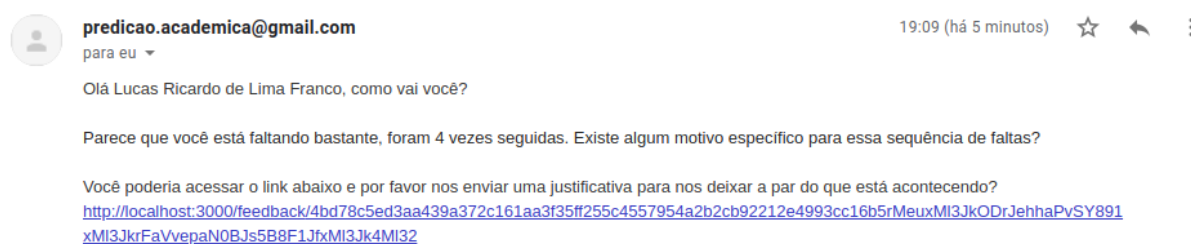
Figura 29 – Detecção de alunos por meio dos mecanismos de comunicação



Fonte: Autoria Própria

Para fins de validação todos os e-mails atualmente enviados pelo sistema são direcionados para *lucas.franco@alunos.utfpr.edu.br*, mas a plataforma já está preparada para enviá-los para os endereços de e-mail reais dos alunos, quando estes são carregados nas planilha de notas e faltas.

Figura 30 – Exemplo de e-mail enviado à um aluno pelo sistema de predição



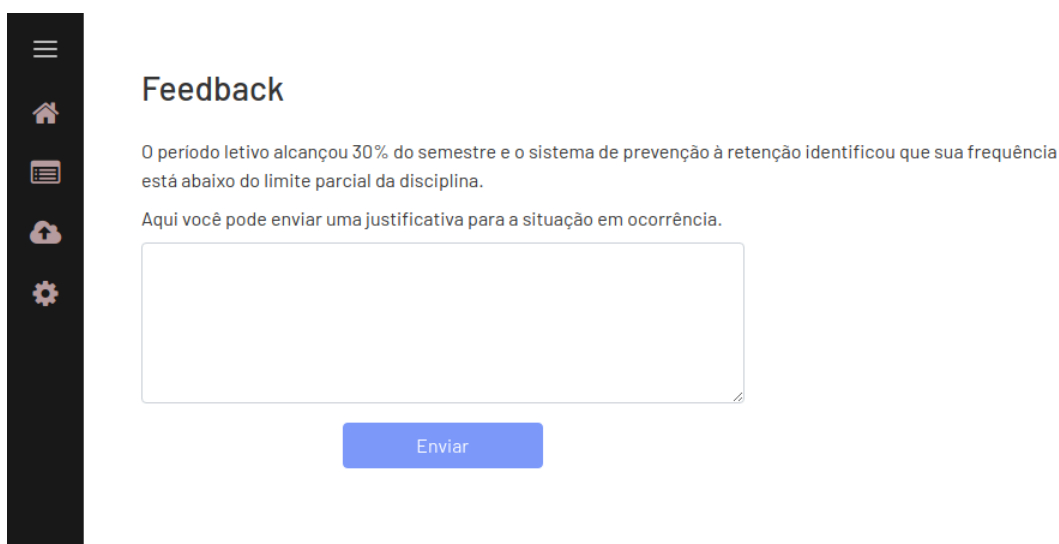
Fonte: Autoria Própria

A [Figura 30](#) exibe um exemplo dos e-mails enviados pelo sistema de forma automática por meio da detecção. Pode-se observar que é enviado junto à descrição da ocorrência um link para que o aluno acesse. Este link é referente a próxima página a ser descrita, a página de *feedback*.

#### 3.6.4 Feedback

A página de *feedback* é bastante simples e consiste de uma tela que só pode ser acessada pelo link enviado por e-mail. Ela contém um parágrafo que explica o motivo pelo qual o e-mail lhe foi enviado e o aluno precisa apresentar uma justificativa. Esse parágrafo também é construído com base nas mensagens definidas na tela de configurações. Além disso, a tela possui um campo para que o aluno envie a justificativa para o ocorrido. A [Figura 31](#) apresenta esta tela.

Figura 31 – Página de envio de justificativa, com a descrição da ocorrência



**Feedback**

O período letivo alcançou 30% do semestre e o sistema de prevenção à retenção identificou que sua frequência está abaixo do limite parcial da disciplina.

Aqui você pode enviar uma justificativa para a situação em ocorrência.

Enviar

Fonte: Autoria Própria

Com o auxílio da biblioteca *SimpleCrypto* é gerada a *URL* para acesso à tela de *feedback*. Nessa *URL* são encriptados alguns dados que possibilitam identificar o texto enviado pelo aluno. Os dados encriptados são o código identificador do aluno, o código da disciplina e um identificador para o e-mail enviado.

Não foi implementada uma tela de login nesta versão do sistema, mas ao enviar esses dados encriptados na *URL* elimina-se a necessidade de um controle de acesso para os alunos, pois caso não estejam autenticados no sistema estes só poderão acessar as páginas permitidas, que neste caso é somente a tela para o envio da sua justificativa.

Ao enviar a mensagem o campo para envio da justificativa é desabilitado, assim, se o estudante acessar novamente o mesmo link, este não conseguirá enviar outra mensagem ou modificar a mensagem já enviada.

Uma vez que o aluno enviar sua justificativa, o docente poderá monitorar os e-mails das ocorrências e as justificativas recebidas na tela de detalhes do estudante. O principal meio de acesso a tela de detalhes do estudante é pela página de lista. Esta é apresentada a seguir.

### 3.6.5 Lista

Esta página é responsável pela listagem geral dos dados dos alunos de cada turma. Ela é composta por duas caixas de seleção nas quais o usuário seleciona a disciplina e o período letivo a ser listado. A [Figura 32](#) mostra a página de listagem para a disciplina ES31C - Laboratório de Informática e período 2019-1.

Figura 32 – Página de listagem de dados

**Lista**

Disciplina: ES31C - Laboratório de Informática Período: 2019-1 [Buscar] [Executar Predição]

Situação Atual: Todos Situação prevista: Todos (Previstos)

39 estudantes - Laboratório de Informática

| # | Código  | Avaliação 1 | Avaliação 2 | Frequência | Nota Final | Status Final |
|---|---------|-------------|-------------|------------|------------|--------------|
| 3 | 1000001 | 88.0        | 79.0        | 83%        |            | [→]          |
| 4 | 1000002 | 57.0        | 84.0        | 83%        |            | [→]          |
| 5 | 1000003 | 0.0         | 0.0         | 24%        |            | [→]          |
| 6 | 1000004 | 21.0        | 28.0        | 70%        |            | [→]          |
| 7 | 1000005 | 2.0         | 0.0         | 55%        |            | [→]          |
| 8 | 1000006 | 16.0        | 0.0         | 71%        |            | [→]          |
| 9 | 1000007 | 57.0        | 65.0        | 91%        |            | [→]          |

Fonte: Autoria Própria

Em seguida estão dispostos dois botões, o primeiro é responsável por solicitar os dados armazenados no banco de dados ao *back-end* em Node.js, e o segundo é responsável por realizar a chamada ao *back-end* em Python, solicitando a execução dos algoritmos de aprendizado de máquina e o retorno das predições.

A sequência de interações realizadas ao pressionar ambos os botões é ilustrada na [Figura 22](#).

Como será detalhado na seção 4.1, que trata da análise dos modelos, o algoritmo *Random Forest* apresentou melhores resultados quando comparado aos algoritmos de regressão linear e *Gradient Boosting*. Devido a isso as predições apresentadas no sistema são provenientes dos modelos criados com o algoritmo *Random Forest*.

Logo abaixo da caixa de seleção e botões há dois filtros, um deles é habilitado ao pressionar o botão *Buscar* e permite filtrar a lista pelas regras definidas na seção 3.6.1, são



estes: Adequado, Atenção (ou média) e Alto risco. Este filtro considera a nota atual do aluno, calculado com base nas notas das avaliações e atividades realizadas até este determinado momento. O segundo filtro é habilitado ao pressionar o botão *Executar Predição*, e permite filtrar seguindo as mesmas regras citadas anteriormente, mas considerando a nota final prevista pelo modelo treinado com o *Random Forest*.

Após a execução das predições, os campos Nota Final e Status Final são preenchidos com os dados estimados pelos modelos. A Figura 33 apresenta a tela após a execução das predições. Os alunos que obtiveram uma nota estimada de 70 ou mais são listados com perfil *Adequado* e os campos preenchidos são coloridos em verde. Para alunos que obtêm notas entre 50 e 70 o status apresentado é o *Atenção* e os campos são coloridos em amarelo. Por fim, os alunos com notas inferiores a 50 são sinalizados com status *Alto risco* e os campos são coloridos em vermelho.

Figura 33 – Página de listagem de dados com as predições efetuadas

**Lista**

Disciplina: ES31C - Laboratório de Informática | Período: 2019-1 | Buscar | Executar Predição

Situação Atual: Todos | Situação prevista: Todos (Previstos)

39 estudantes - Laboratório de Informática

| # | Código  | Avaliação 1 | Avaliação 2 | Frequência | Nota Final | Status Final |
|---|---------|-------------|-------------|------------|------------|--------------|
| 3 | 1000001 | 88.0        | 79.0        | 83%        | 83.0       | Adequado     |
| 4 | 1000002 | 57.0        | 84.0        | 83%        | 72.7       | Adequado     |
| 5 | 1000003 | 0.0         | 0.0         | 24%        | 9.3        | Alto risco   |
| 6 | 1000004 | 21.0        | 28.0        | 70%        | 33.1       | Alto risco   |
| 7 | 1000005 | 2.0         | 0.0         | 55%        | 3.4        | Alto risco   |
| 8 | 1000006 | 16.0        | 0.0         | 71%        | 9.8        | Alto risco   |
| 9 | 1000007 | 57.0        | 65.0        | 91%        | 57.9       | Atenção      |

Fonte: Autoria Própria

Por mais que o status *Atenção* não traduza diretamente o status final do aluno, o uso desta faixa com 20% dos pontos (entre 50 e 70) permite à ferramenta uma margem maior entre os status *Adequado* e *Alto risco*, possibilitando que a determinação delicada do status do aluno com nota prevista nessa faixa seja feita de forma mais branda. Porém, isso requisita uma análise mais detalhada por conta do docente sobre os alunos que figuram neste perfil.

Ainda na tela de listagem, abaixo dos filtros é exibido o número de estudantes listados para a disciplina selecionada. Ao clicar em um dos botões exibidos na extrema direita de cada linha da tabela é possível acessar a tela de detalhes do estudante.

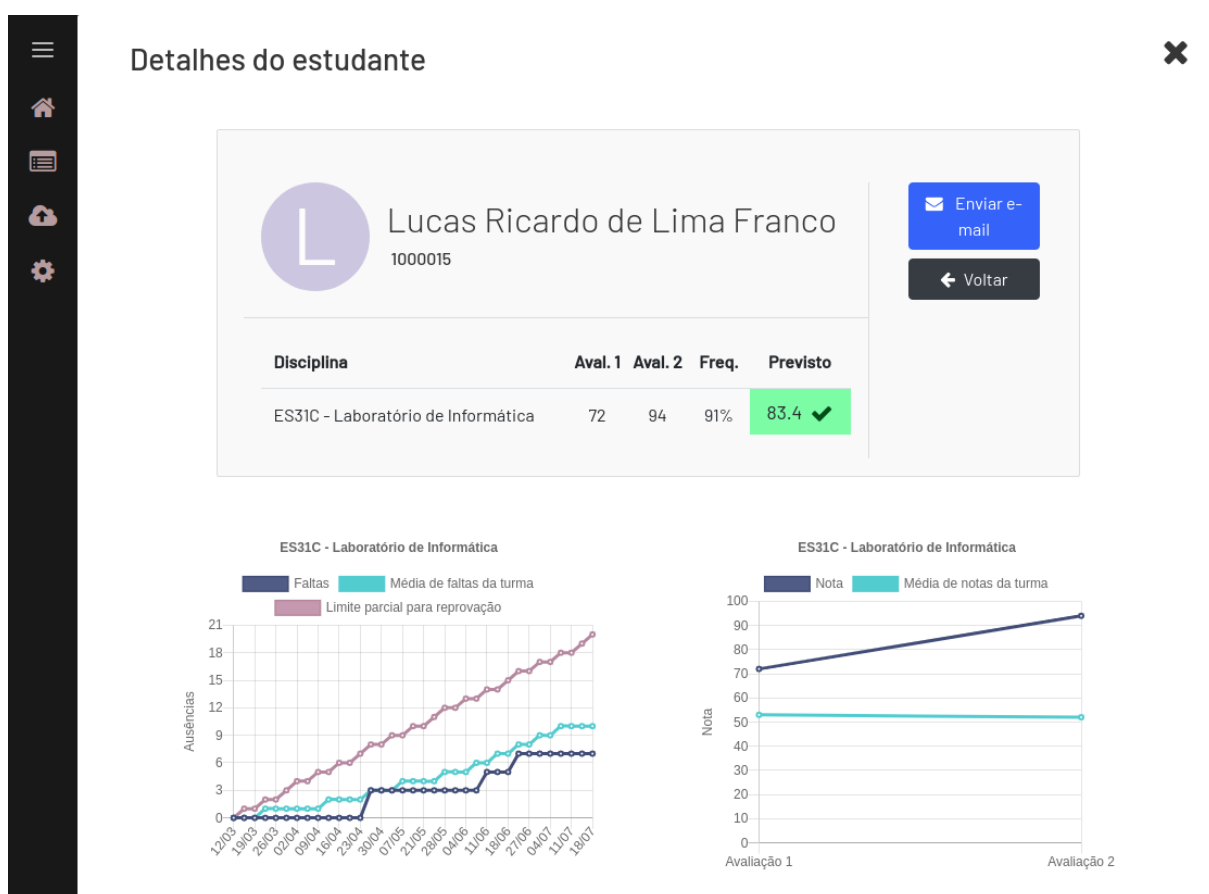
### 3.6.6 Detalhes do Estudante

A página de detalhes do estudante é responsável por apresentar os dados individuais do aluno selecionado.

Como é possível visualizar na [Figura 34](#), a página de detalhes possui um *card* na parte superior da tela que apresenta o nome e o código identificador do aluno, bem como a lista de disciplinas matriculadas com dados inseridos na ferramenta. Para cada disciplina são exibidas as notas obtidas pelo aluno até o momento, sua frequência e a nota final prevista pelo modelo de aprendizado de máquina.

Caso a nota prevista seja igual ou superior a 70 o campo é apresentado em verde e um ícone de *ok* é mostrado, caso a nota esteja entre 50 e 70 o campo é colorido de amarelo com um ícone de atenção. E para notas inferiores a 50 o campo é apresentado em vermelho com um ícone em forma de X. Semelhante aos indicadores do dashboard apresentados na [Figura 23](#).

Figura 34 – Página de detalhes do estudante



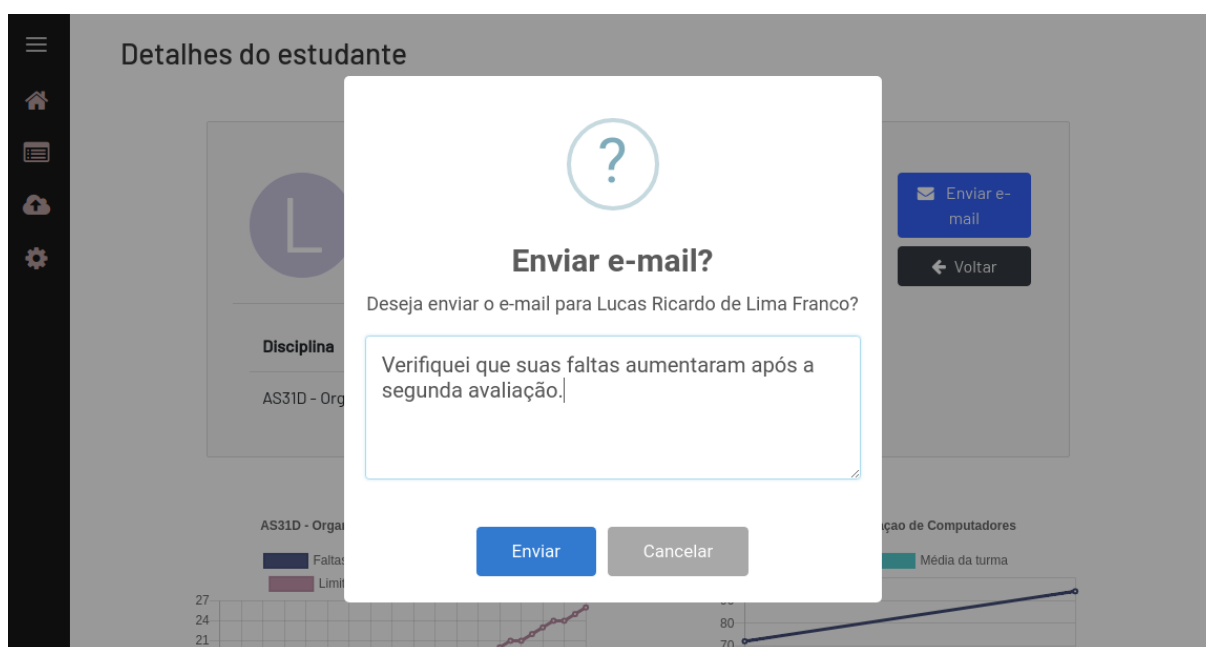
Fonte: Autoria Própria

Abaixo do *card* são gerados dinamicamente gráficos comparativos entre o desempenho do aluno e a média da turma. Para cada disciplina vinculada ao aluno é plotado um novo par de gráficos. Para cada par, o primeiro gráfico compara o número de ausências do aluno com a média de faltas da turma, trazendo também o limite de ausências que um aluno pode ter sem

que ocorra a reprovação por faltas. Já o segundo gráfico ilustra o comportamento das notas do aluno, possibilitando ver se estas estão em crescimento ou declínio, comparando-as com a média de notas da turma.

Retornando ao *card* na parte superior da tela é possível verificar dois botões na sua direita. Um deles é utilizado para voltar a tela anterior caso desejado e o outro é utilizado para realizar o envio de um e-mail personalizado ao aluno apresentado nesta tela. Ao clicar neste botão será apresentada uma página *modal* com um campo para que o usuário insira a mensagem a ser enviada junto de um link gerado automaticamente para que o aluno envie um *feedback*. A Figura 35 exemplifica a ação de envio de e-mail.

Figura 35 – Mensagem personalizada a ser enviada no e-mail por meio da página de detalhes do estudante



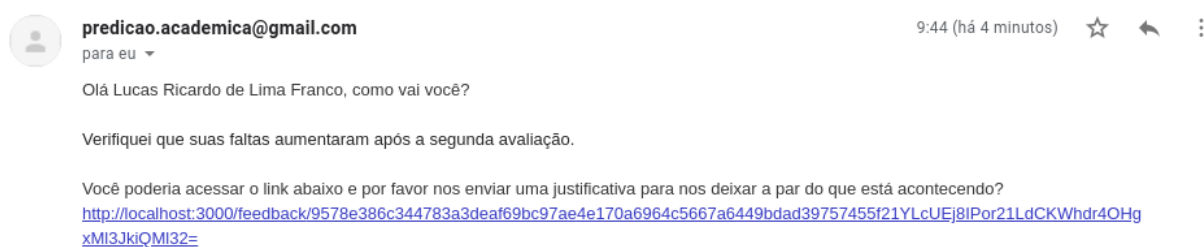
Fonte: Autoria Própria

A Figura 36 apresenta o e-mail recebido pela ação executada anteriormente, nesta é possível verificar que a construção do e-mail se deu com base na mensagem descrita na Figura 35. Ao final do e-mail encontra-se o link para que o aluno acesse e envie o *feedback* solicitado.

Por fim, a tela de detalhes do estudante permite visualizar a lista de e-mails enviados ao aluno, o status destes, ou seja, se o aluno já o retornou com o seu *feedback* ou se está pendente de resposta. Caso o aluno já o tenha respondido será possível ler a justificativa enviada. Também são exibidas a data e hora de envio dos e-mails e da resposta do aluno, caso existente.

A Figura 37 exemplifica a lista de *feedbacks* apresentando uma lista com dois envios de e-mail. Esta é ordenada por data de envio, iniciando pelos mais recentes.

Figura 36 – E-mail enviado por meio da página de detalhes do estudante



Fonte: Autoria Própria

O primeiro da lista é referente ao envio do e-mail da Figura 36, que ainda não recebeu uma resposta do aluno. Já o segundo é referente a uma notificação enviada ao aluno na detecção automática realizada no momento do envio das planilhas, por meio dos mecanismos estabelecidos anteriormente na tela de configurações. Este recebeu um retorno e pode ser visto no campo *Feedback* da lista.

Figura 37 – Lista de feedbacks disponível na página de detalhes do estudante

Notificações enviadas ao aluno

| Status      | Feedback                                                                                                                  | Data de envio    | Data de retorno  |
|-------------|---------------------------------------------------------------------------------------------------------------------------|------------------|------------------|
| ⚠ Pendente  |                                                                                                                           | 17/10/19 - 09:44 |                  |
| ✓ Retornado | Devido a uma cirurgia fiquei afastado por quinze dias. Já encaminhei o atestado médico para a secretaria da universidade. | 16/10/19 - 19:09 | 16/10/19 - 21:33 |

Fonte: Autoria Própria

## 4 ANÁLISE DOS RESULTADOS

Neste capítulo são apresentadas as análises dos resultados obtidos neste trabalho, considerando os modelos de aprendizado de máquina utilizados e a implementação da plataforma.

### 4.1 MODELOS

Esta seção é composta por duas subseções. A primeira possui a análise das acurácias obtidas com os modelos utilizando os algoritmos de regressão linear, *Gradient Boosting* e *Random Forest*. Já a segunda é composta pela análise da importância das *features* na composição de cada modelo.

#### 4.1.1 Acurácia

Como apresentado na seção 3.5, foram criados múltiplos cenários, com o treinamento de modelos para cada algoritmo de aprendizado de máquina testado para cada cenário. Os algoritmos utilizados foram o *Gradient Boosting*, o *Random Forest* e o de Regressão Linear, já descritos anteriormente e, ao realizar o treinamento dos modelos criados com o conjunto de dados disponível, foram obtidos seus *scores*, ou acurácias, com o auxílio da biblioteca *Scikit-learn*.

O cálculo do *score* dos modelos foi realizado utilizando o coeficiente de determinação  $R^2$ . O coeficiente de determinação  $R^2$  é a proporção de variância dos dados que pode ser explicada pelo modelo de regressão. Ela é utilizada para medir o sucesso da predição de uma variável com base em outras variáveis independentes e quanto mais próximo de 1, maior a acurácia do modelo (NAGELKERKE et al., 1991).

A função padrão de *score* do *Scikit-learn* retorna valores entre 0 e 1 para modelos onde a capacidade preditiva é melhor do que a média aritmética dos valores da amostra, ou seja, caso o modelo possua um *score* inferior à 0 o cálculo simples da média dos valores da amostra seria mais preciso que o modelo treinado.

Dentre os algoritmos, o *Gradient Boosting* e *Random Forest* são considerados mais adequados para um volume pequeno de dados de treinamento e mais robustos quanto a *overfitting*<sup>1</sup> quando comparados aos algoritmos de regressão linear. Desta forma, a fim de verificar a acurácia dos modelos para diferentes situações foi executada a validação cruzada dos modelos, que divide o conjunto de dados em subconjuntos, treina e testa o modelo várias vezes, retornando a acurácia média e o desvio padrão para os diferentes subconjuntos.

Após a obtenção dos *scores* para os modelos construídos, foram selecionados alguns dos cenários da Tabela 3 para melhor exemplificação. Foram selecionados os cenários A e J por

---

<sup>1</sup>Descreve um modelo estatístico que está muito ajustado para o conjunto de dados de treinamento, mas não é eficiente na predição de novos resultados

figurarem os extremos e os cenários B, F e H, que apresentaram mais comuns para o conjunto de dados utilizado.

Foram consideradas validações cruzadas de cinco subconjuntos para a obtenção dos *scores* e dos desvios padrão no *Scikit-learn*. Estes dados para os cenários selecionados são apresentados em porcentagens na [Tabela 4](#). A tabela completa com os *scores* para todos os cenários pode ser vista no [Apêndice A](#).

Tabela 4 – Score e desvio padrão dos modelos construídos considerando os cenários selecionados

| Cenário | Regressão Linear |           | Gradient Boosting |           | Random Forest |           |
|---------|------------------|-----------|-------------------|-----------|---------------|-----------|
|         | Score            | D. Padrão | Score             | D. Padrão | Score         | D. Padrão |
| A       | 44,6%            | ±6,11%    | 47,2%             | ±7,89%    | 53,3%         | ±6,63%    |
| B       | 58,2%            | ±6,78%    | 55,3%             | ±11,34%   | 64,5%         | ±6,92%    |
| F       | 88,1%            | ±2,43%    | 88,9%             | ±3,47%    | 91,4%         | ±2,35%    |
| H       | 95,7%            | ±0,73%    | 94,9%             | ±0,69%    | 95,4%         | ±0,88%    |
| J       | 97,5%            | ±1,22%    | 98,3%             | ±0,99%    | 97,3%         | ±0,67%    |

Na [Tabela 4](#) observa-se que para o cenário A o algoritmo de regressão linear obteve acurácia média de 44,6% e desvio padrão de 6,11%, ou seja, a dispersão média entre as cinco acurácias obtidas com a validação cruzada foi de 6,11%. O algoritmo *Gradient Boosting* obteve para o mesmo cenário uma acurácia média de 47,2% com um desvio levemente mais elevado, de 7,89%. Já o *Random Forest* obteve uma acurácia relativamente maior neste cenário, com 53,3% e desvio padrão de 6,63%.

Pode-se observar também um aumento proporcional na acurácia média dos algoritmos conforme avança-se pelos cenários seguintes ao A. Isso ocorre pois o cenário A possui um conjunto de dados de entrada muito simples para realizar o treinamento dos modelos para prever o desempenho dos alunos. Para os outros cenários vai-se proporcionalmente dispondo um conjunto de dados com maior qualidade para os algoritmos, possibilitando o treinamento de modelos mais robustos e acurados. Na [Tabela 3](#) é apresentada a lista de cenários com os respectivos níveis de dados providos aos modelos, na qual o cenário A é o mais simples e o cenário J é o que possui dados mais completos e consequentemente terá melhores modelos.

Com isso, pode-se realizar a predição preliminar do desempenho dos alunos logo no início do período letivo, considerando a penalidade esperada na acurácia destas predições, e ao decorrer do período letivo é possível acompanhar a evolução destas predições, que se tornarão mais precisas conforme as atividades, avaliações e dados de frequências dos alunos forem coletados e importados na ferramenta.

Dentre os cinco cenários apresentados na [Tabela 4](#), o cenário A considera dados muito limitados para a realização das predições e o cenário J leva em conta todos os dados, não havendo mais necessidade de uma predição, já que todos os dados necessários para calcular o desempenho dos alunos já estarão disponíveis. Os cenários B, F e H figuram situações mais realistas. O cenário B apresenta uma quantidade inicial de dados sem dados de avaliações, o cenário F considera a realização de uma avaliação e o cenário H considera quase todos os

dados, exceto pela nota da última avaliação.

Em relação aos resultados (acurácia média e desvio padrão), para a maioria dos casos os melhores números foram obtidos pelo *Random Forest*. Considerando o *Random Forest*, para predições em momentos do período letivo correspondentes aos três cenários mais comuns, pode-se dizer que o usuário do sistema obteria acurácias médias entre 64,5% e 95,4% (64,5% para o cenário B, 91,4% para o cenário F, e 95,4% para o cenário H).

Levando em conta todos os oito cenários entre os cenários B e I ([Apêndice A](#)), o resultado é semelhante, apresentando acurácias médias entre 62,2% e 95,4%.

Devido ao algoritmo *Random Forest* apresentar os melhores resultados para a maioria dos casos, esta foi a técnica empregada para retornar as predições na plataforma desenvolvida neste trabalho.

#### 4.1.2 Importância

Como mencionado na seção 3.5, após realizar o treinamento de modelos de aprendizado de máquina, o *Scikit-Learn* possibilita a obtenção das importâncias de cada uma das variáveis, ou *features*, para os resultados dos modelos.

O algoritmo de regressão linear não possui o artifício de identificação das importâncias como os outros, porém ele permite obter os coeficientes determinados pelo modelo para a composição da equação linear que descreve os dados. Diferente das importâncias para os outros modelos, os coeficientes lineares podem ser negativos.

As tabelas seguintes apresentam as importâncias das *features* para os algoritmos *Gradient Boosting* e *Random Forest* para os cenários B, F e H.

Apesar das importâncias e coeficientes lineares não possuírem exatamente o mesmo conceito, eles servem para mostrar o peso das variáveis nos modelos. Para melhor comparação ambos são apresentados lado a lado nas tabelas 5, 6, 7, 8 e 9, referentes aos cenários A, B, F, H e J, respectivamente. No qual as avaliações são representadas por *gradeN*, as micro-avaliações/trabalhos por *migro\_gradeN*, a frequência dos alunos por *attendance*, a nota do projeto final da disciplina por *project\_grade* e a porcentagem de exercícios da plataforma URI solucionados ou que houve tentativa de resolução pelo aluno por *percentage*.

Tabela 5 – Importância das *features* considerando o cenário A para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature      | Importância |           | Coeficiente  |
|--------------|-------------|-----------|--------------|
|              | Gradient B. | Random F. | Regressão L. |
| attendance   | 64,3%       | 61,7%     | 66,2%        |
| micro_grade2 | 22,1%       | 25,2%     | 19,3%        |
| micro_grade1 | 13,6%       | 13,1%     | 11,5%        |

A relação completa com as importâncias e coeficientes lineares para todos os cenários pode ser vista no [Apêndice B](#).

Tabela 6 – Importância das *features* considerando o cenário B para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature      | Importância |           | Coeficiente  |
|--------------|-------------|-----------|--------------|
|              | Gradient B. | Random F. | Regressão L. |
| attendance   | 52,5%       | 50,6%     | 50,7%        |
| micro_grade3 | 19,3%       | 23,5%     | 22,8%        |
| micro_grade2 | 14,9%       | 14,0%     | 10,8%        |
| micro_grade1 | 8,4%        | 7,9%      | 7,8%         |
| percentage   | 4,9%        | 4,0%      | 1,7%         |

Tabela 7 – Importância das *features* considerando o cenário F para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature      | Importância |           | Coeficiente  |
|--------------|-------------|-----------|--------------|
|              | Gradient B. | Random F. | Regressão L. |
| grade1       | 81,0%       | 82,4%     | 66,9%        |
| attendance   | 11,8%       | 11,4%     | 22,4%        |
| micro_grade3 | 3,5%        | 2,7%      | 10,6%        |
| micro_grade2 | 2,3%        | 1,8%      | 4,7%         |
| micro_grade1 | 0,8%        | 1,2%      | 1,0%         |
| percentage   | 0,6%        | 0,5%      | -3,1%        |

Tabela 8 – Importância das *features* considerando o cenário H para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature       | Importância |           | Coeficiente  |
|---------------|-------------|-----------|--------------|
|               | Gradient B. | Random F. | Regressão L. |
| grade1        | 66,8%       | 65,1%     | 57,4%        |
| project_grade | 27,7%       | 29,5%     | 28,6%        |
| attendance    | 3,0%        | 3,0%      | 7,8%         |
| micro_grade2  | 1,0%        | 0,8%      | 4,1%         |
| micro_grade3  | 0,8%        | 0,8%      | 1,8%         |
| micro_grade1  | 0,3%        | 0,5%      | 1,3%         |
| percentage    | 0,3%        | 0,2%      | -0,4%        |
| micro_grade4  | 0,2%        | 0,2%      | 2,9%         |

Os modelos do cenário B, que só possuía dados de frequência, micro-avaliações e exercícios do URI, apresentaram uma maior importância para a frequência, com valores entre 50,6% e 52,5% da importância total do modelo.

Após a frequência, tem-se as notas das micro-avaliações, que somadas figuram uma importância de cerca de 45% para os algoritmos não lineares e 48% para a regressão linear, com uma relevância um pouco maior para os trabalhos mais posteriores. Ou seja, a micro-avaliação 3 retornou uma importância sutilmente maior que a da micro-avaliação 2, que por sua vez retornou uma importância sutilmente maior que a da micro-avaliação 1. Esta relação entre as



Tabela 9 – Importância das *features* considerando o cenário J para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature       | Importância |           | Coeficiente  |
|---------------|-------------|-----------|--------------|
|               | Gradient B. | Random F. | Regressão L. |
| grade2        | 62,2%       | 67,7%     | 23,5%        |
| grade1        | 29,9%       | 27,1%     | 47,5%        |
| project_grade | 5,2%        | 2,4%      | 17,5%        |
| attendance    | 1,1%        | 1,1%      | 5,8%         |
| micro_grade3  | 0,7%        | 0,8%      | 2,3%         |
| micro_grade2  | 0,6%        | 0,5%      | 4,7%         |
| micro_grade1  | 0,2%        | 0,3%      | 2,4%         |
| micro_grade4  | 0,0%        | 0,1%      | 1,1%         |
| percentage    | 0,0%        | 0,1%      | -1,9%        |

micro-avaliações também se mostrou verdade para o cenário F.

Para todos os cenários que tinham disponíveis as avaliações ou a nota do projeto final da disciplina (cenários F, H e J), estas foram as variáveis que obtiveram a maior importância. O cenário J apresentou para o *Gradient Boosting* e o *Random Forest* uma importância muito mais significativa para a nota da segunda avaliação do que a primeira, com valores entre 62,2% e 67,7% para a segunda contra valores entre 27,1% e 29,9% para a primeira, mesmo que elas possuam o mesmo peso no cálculo da nota final. Isto demonstra a importância da realização da segunda avaliação e a influência que esta tem no desempenho final da disciplina.

No geral, os exercícios do URI tiveram uma participação muita baixa nas estimativas dos modelos, isto provavelmente se deu por este dado não estar diretamente relacionado ao cálculo da nota final dos alunos, além da entrega dos exercícios ser opcional (porém recomendada) para as disciplinas do conjunto de dados utilizado. Uma análise adicional quanto aos dados do URI é apresentada na seção 4.3.

Com os gráficos gerados na seção 3.3 para entendimento dos dados, foi possível verificar visualmente certo grau de correlação entre as notas dos alunos com as presenças e os exercícios realizados. Porém, percebeu-se que os algoritmos não identificaram a importância destas variáveis para a composição da nota final dos alunos.

## 4.2 CONSIDERAÇÕES GERAIS SOBRE A PLATAFORMA

Os resultados obtidos acerca da plataforma desenvolvida foram mais subjetivos, mas é possível visualizar os potenciais que a ferramenta têm para o auxílio dos usuários na análise dos dados acadêmicos.

#### 4.2.1 Dashboard

Com o dashboard implementado é possível extrair alguns *insights* sobre os dados. O gráfico temporal apresentado na [Figura 23](#) possibilita verificar a proporção de faltas da turma como um todo, e caso a linha de faltas da turma ultrapasse a linha que representa o limite permitido de faltas, o profissional pode tomar ações para compreender em detalhes quais os motivos que levam a esse acontecimento. Já com os indicadores da parte superior da mesma página, pode-se ter uma visão geral das turmas cadastradas ou de uma turma específica selecionado-a na caixa de seleção flutuante presente à direita.

A partir do gráfico de dispersão do dashboard, [Figura 24](#), pode-se verificar visualmente o nível de correlação presente entre a frequência e as notas dos alunos, e caso a influência dos trabalhos na nota final seja explícita, o profissional pode tomar ações ou apresentar este gráfico aos alunos, argumentando e influenciando-os a seguir o perfil que tende a receber melhores notas.

No mesmo gráfico pode-se verificar que há alunos com frequências significativamente altas mas com notas abaixo da média para aprovação, desta forma é possível acessar estes alunos a fim de averiguar quais os motivos que os levam a não obterem desempenho tão bons na disciplina. Com isso pode-se descobrir lacunas no método pedagógico utilizado pelo docente e oportunidades de uso de recursos para a melhoria do aprendizado, como ao destinar maior foco dos alunos monitores para apoiar estes alunos com baixo desempenho.

Ainda no dashboard, o gráfico de barras permite uma análise rápida das notas do alunos nas disciplinas e a comparação entre elas, assim pode-se identificar a proporção de alunos com um bom, médio e baixo desempenho, apurando também as disciplinas nas quais os alunos obtiveram menor aproveitamento.

#### 4.2.2 Predição, Feedback e Mecanismos de Detecção

A predição das notas dos alunos em conjunto com a detecção de situações incomuns complementam o sistema proposto e desenvolvido neste trabalho, que tem como objetivo fornecer ferramentas para que os profissionais possam analisar dados acadêmicos e tomar algumas ações simples sem a necessidade de muitos esforços manuais.

A página de detalhes do estudante proporciona ao docente uma visão da situação atual do aluno, permitindo nela comparar o aluno à média, tanto em faltas quanto em notas. Esta é uma forma de investigar um aluno a fim de identificar padrões no seu comportamento, como a diminuição das notas coincidentemente na época do período letivo que o aluno mais obteve faltas.

Ainda na mesma página o profissional poderá submeter uma notificação ao aluno, solicitando um *feedback*. Este recurso foi idealizado para o uso em situações que demandam uma "advertência", mas ele não é restrito a isso. Caso o docente observe alguma variação no perfil dos alunos, este pode pedir um *feedback* também em uma situação ainda considerada

adequada para o aluno. Um exemplo desse comportamento que o professor pode identificar é a ausência de determinados alunos sempre nos mesmos dias da semana. Deduz-se então que algum evento em especial pode estar ocorrendo sistematicamente nestes dias e o docente pode enviar um e-mail aos alunos a fim de comprovar seu *insight*.

Um dos objetivos almejados com a possibilidade de os alunos enviarem *feedbacks* é verificar se a comunicação com os alunos pode ser uma forma de conscientização e consequentemente melhores resultados no desempenho dos alunos.

Com a base de dados utilizada neste trabalho observou-se que existem diferentes perfis de estudantes. E esta plataforma é destinada a atingir a parcela de alunos que possui dedicação aos estudos mas por alguma razão, sob a influência do professor e da universidade, não conseguem obter um bom desempenho.

Como descrito anteriormente, o algoritmo selecionado para a predição das notas dos alunos foi o *Random Forest*. Com os modelos treinados com ele, utilizando uma base histórica, tornou-se possível estimar o desempenho dos alunos, a fim de tomar ações e tentar inibir a retenção acadêmica antes do seu acontecimento.

O sistema construído segmenta os alunos em três perfis, o primeiro com mais chances de aprovação na disciplina, o segundo com estimativas medianas e o terceiro com maiores riscos de reprovação. Com isso o profissional pode selecionar um perfil específico e atuar sobre os alunos contidos nele, com o intuito de melhorar suas situações e colaborar com a transição destes para perfis mais positivos.

#### 4.3 MOSS - Similaridades

Após as análises realizadas neste trabalho, foi possível ver claras relações entre as notas, faltas e atividades para o conjunto de dados utilizado. Entretanto, em alguns casos pode-se identificar alguns *outliers*<sup>2</sup> que não puderam ser explicados com a proposta deste trabalho (relação direta entre as notas e a frequência dos alunos). Nestes casos, os *outliers* são os alunos que obtiveram notas excelentes mesmo com frequências abaixo do considerado adequado para o ensino presencial, sendo o contrário também verdadeiro, alunos com notas baixas e altas taxas de frequências.

Desta forma, como uma tentativa de compreender estes *outliers*, analisou-se também um conjunto de dados complementar, disponibilizado pela plataforma URI em conjunto com a ferramenta MOSS. A seção a seguir apresenta o MOSS e a análise realizada com o dados obtidos a partir dele.

O MOSS é uma ferramenta disponibilizada pelo portal URI que identifica a similaridade e detecta plágios nos códigos-fonte (exercícios) enviados pelos alunos (BEZ NEILOR A. TONIN, 2015). Assim, tem-se um recurso a mais para tentar explicar principalmente a situação de alunos com boas taxas de frequência e que realizam as atividades/trabalhos solicitados, mas

---

<sup>2</sup>Um outlier é uma observação que se diferencia tanto das demais observações que levanta suspeitas de que aquela observação foi gerada por um mecanismo distinto (Hawkins, 1980)

continuam obtendo notas baixas. Se as taxas de similaridade forem altas para estes casos, então uma possível interpretação seria que estes alunos realizam a cópia das atividades de outros alunos e não estejam engajados em aprender e praticar o conteúdo, por isso não obtém bom desempenho nas avaliações.

Dados de similaridade foram extraídos pelo *MOSS* sobre exercícios entregues de uma das turmas do conjunto de dados utilizado para as análises já apresentadas. Com estes dados realizou-se o cruzamento com as notas e as frequências, observando principalmente os *outliers*. O *MOSS* realiza uma análise e disponibiliza dados para cada exercício que os alunos realizam. Esta turma realizou três listas que totalizaram 22 exercícios, sendo assim, 22 arquivos do *MOSS* foram extraídos.

A [Figura 38](#) traz os dados da turma citada. Como já apresentado na seção [3.3](#), esse gráfico traz a distribuição das notas dos alunos com base em suas frequências. O formato de cada item leva em conta se este entregou a maior parte dos trabalhos/micro-avaliações ou não.

Pode-se perceber que não há casos nos quais o aluno com frequência muito baixa, não realizou as atividades e obteve notas positivas (região extrema esquerda do quadrante superior esquerdo). Quando se trata de frequências altas com realização das atividades e notas muito baixas (região extremo inferior do quadrante inferior direito), tem-se alguns casos, destacado pelo círculo vermelho. Os três pontos mais à esquerda dentro do círculo foram considerados mesmo que possuam menos de 75% de presença, pois obtiveram frequências muito próximas à 75% e realizaram grande parte dos trabalhos.

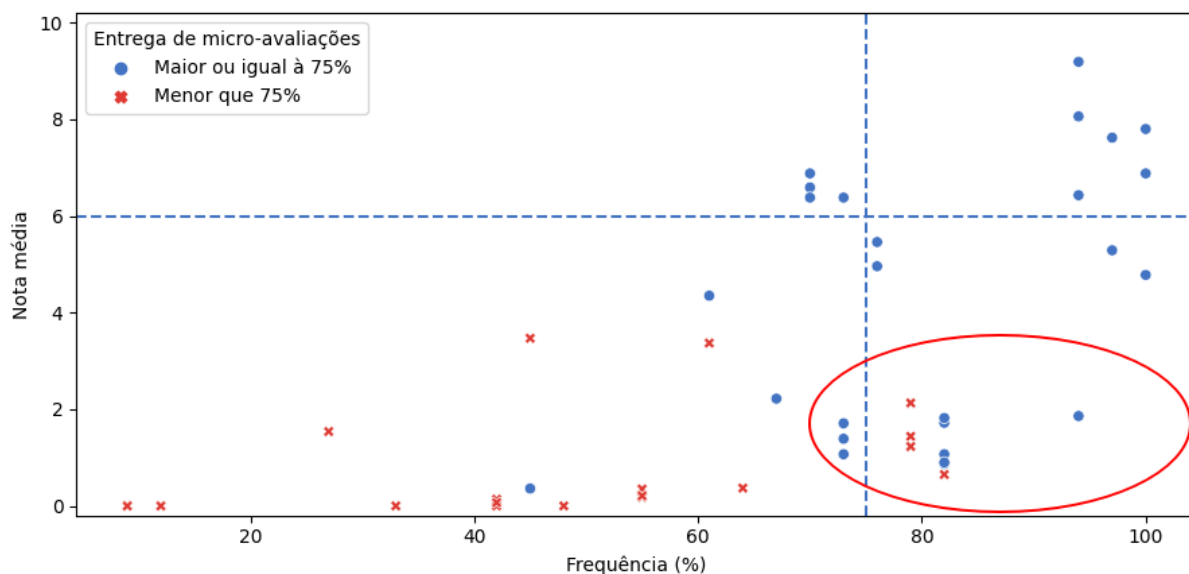
O critério inicial utilizado para a seleção dos *outliers* foi a obtenção de uma frequência acima de 75% (exceto pelos três pontos mencionados anteriormente) e nota inferior a 4, que é a nota utilizada por padrão na UTFPR como nota mínima para quebra de pré-requisitos, ou seja, é a nota mínima que os alunos devem obter em uma disciplina para considerá-los aptos a cursar outras disciplinas que a possuam como um pré-requisito.

Como abordado neste texto, a execução e entrega dos trabalhos podem ser variáveis que influenciam, mesmo que leve e indiretamente, no desempenho final dos alunos nas disciplinas. Desta forma, para esta análise inicial foram observados os nove pontos dentro do círculo em destaque, que foram os casos considerados *outliers* - entregaram mais de 75% dos trabalhos, tem alta frequência mas obtiveram notas finais baixas.

A [Figura 39](#) apresenta a nota média final dos alunos desta turma em ordem decrescente. Como a identificação do aluno não é importante para essa análise, o eixo X contém somente um índice crescente. Já a [Figura 40](#) mostra os dados de similaridade dos mesmos alunos e na mesma ordem apresentada na [Figura 39](#). A similaridade apresentada é a média considerando os exercícios entregues por cada aluno, desta forma tem-se uma proporção apropriada para alunos que entregaram quantidades diferente de exercícios. Vale lembrar que um aluno pode ser reportado múltiplas vezes para o mesmo exercício se for encontrado similaridade em seu código com mais de um aluno, e por isso a similaridade média pode ultrapassar 100.

Em ambas as imagens foram destacados os nove casos considerados *outliers*, na

Figura 38 – Gráfico de dispersão com as frequências e notas dos alunos, considerando somente a turma analisada com o MOSS



Fonte: Autoria Própria

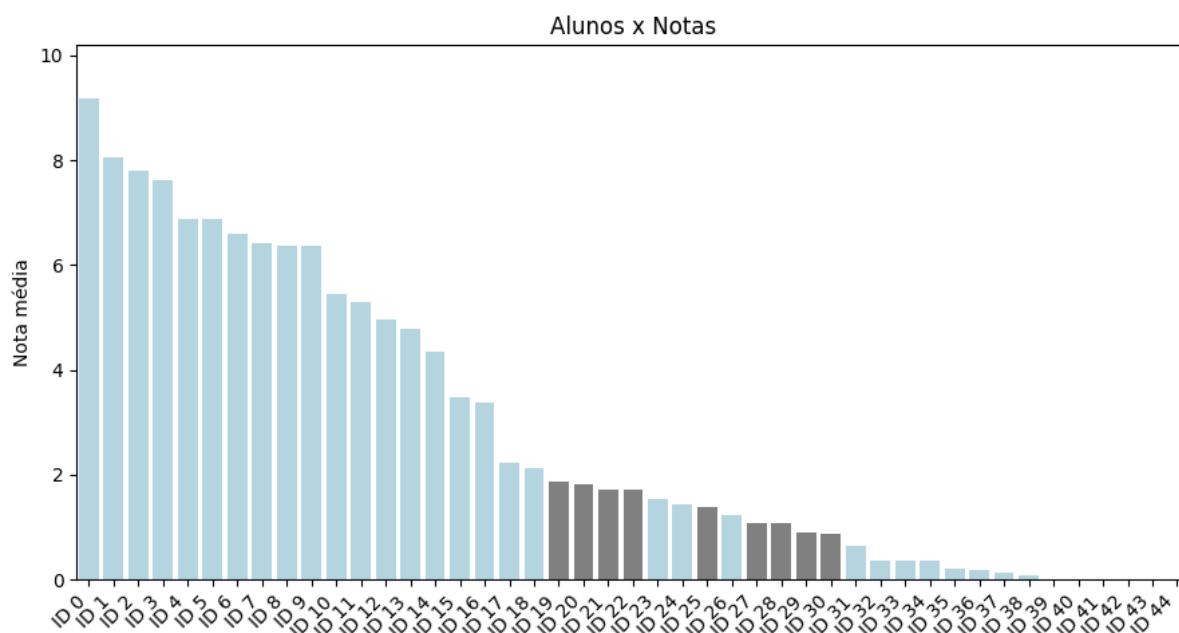
primeira imagem observa-se a nota obtidas por eles entre 1 e 2. Já na segunda imagem aparecem somente cinco casos, isto acontece pois quatro dos *outliers* não entregaram nenhum exercícios do URI, logo não podem obter nenhuma taxa de similaridade. Para os outros cinco casos, tem-se valores distintos de similaridade, com somente dois alunos obtendo um valor evidentemente alto de similaridade média.

Concluindo, dos nove casos considerados *outliers*, para quatro deles o *MOSS* não apresentou nenhuma taxa de similaridade, para três deles foi reportado um valor relativamente baixo e somente para dois dos casos foi-se reportada uma alta taxa de similaridade. Além disso, considerando todos os 44 alunos com dados do *MOSS*, não é possível identificar um padrão claro entre as similaridades e as notas dos alunos.

Para o pequeno conjunto de dados utilizado para a análise do *MOSS* não constatou-se uma relação direta entre as similaridades e o desempenho final dos alunos. Mas vale notar que dos *outliers* que entregaram exercícios, todos obtiveram taxa média de similaridade por exercício acima de 30%.

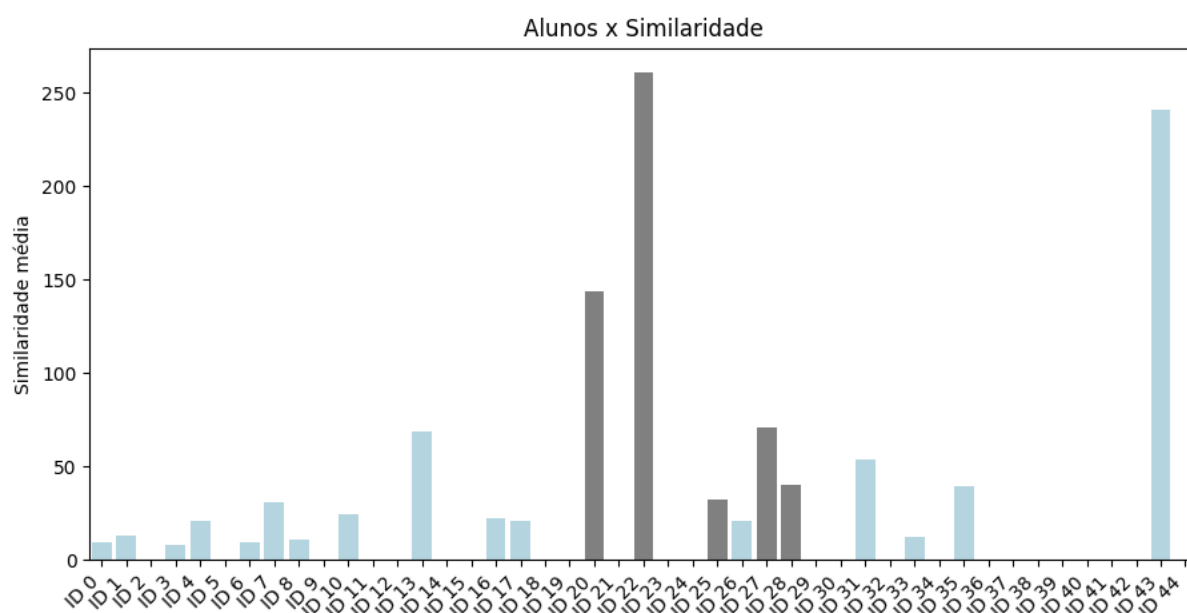
Uma análise adicional foi realizada em relação ao três pontos abaixo de 75% de presença com uma alta entrega de exercícios para tentar compreender a razão pela qual essa situação ocorreu. Verificou-se que dois dos três alunos iniciaram uma sequência grande de faltas na disciplina e desistiram das aulas logo após realizarem e obterem notas muito baixas na última avaliação da disciplina. Estes alunos não realizaram nem mesmo a avaliação substitutiva para recuperação da nota. O terceiro aluno não realizou nem a última avaliação da disciplina, com uma frequência inicial adequada na disciplina este simplesmente parou de ir às aulas sem uma razão aparente.

Figura 39 – Gráfico de barras com as notas médias para cada aluno, considerando somente a turma analisada com o MOSS



Fonte: Autoria Própria

Figura 40 – Gráfico de barras com as similaridades acumuladas para cada aluno, considerando somente a turma analisada com o MOSS



Fonte: Autoria Própria

O sistema proposto e desenvolvido neste trabalho poderia auxiliar no acompanhamento de casos como estes, e a partir disso compreender melhor a motivação pelos quais estes e outros alunos desistem das disciplinas que cursam. Com esta base de conhecimento, torna-se

possível identificar padrões e assim antecipar casos de desistência em turmas seguintes.

## 5 CONSIDERAÇÕES FINAIS

O desenvolvimento deste trabalho possibilitou compreender o impacto que a retenção acadêmica possui em uma universidade brasileira, e refletir sobre quais são as formas de contribuir para a redução deste problema. A retenção acadêmica é uma das variáveis relacionadas à evasão acadêmica e uma das formas de se iniciar a trabalhar sobre a evasão é observar anteriormente o fenômeno da retenção acadêmica.

Então, como objetivo deste trabalho, implementou-se uma plataforma de análise de dados acadêmicos para o fornecimento de mecanismos que permitam atuar sobre a retenção acadêmica e estimar a situação final dos alunos nas disciplinas. Este pode ser a tendência à aprovação, atenção ou tendência à reprovação.

A ferramenta desenvolvida neste trabalho possibilita que docentes tenham um maior proveito sobre os dados que possuem, permitindo que estes carreguem dados das suas disciplinas e, a partir disso, utilize mecanismos para comunicar-se com alunos enquadrados em um perfil que tende à retenção, a fim de compreender o ocorrido e tentar reverter a situação.

Com o estudo realizado sobre os dados de 10 turmas de disciplinas da área de computação da UTFPR - Campus Cornélio Procopio, foi possível implementar modelos preditivos que atuam a fim de antecipar a retenção acadêmica ocorrida principalmente devido a ausências. Gerando um recurso para que docentes engajados possam realizar tomadas de decisão para auxiliar os alunos na regularização das suas situações.

Alguns algoritmos de aprendizado de máquina foram empregados para a criação dos modelos de predição, foram estes a Regressão Linear, o *Random Forest* e o *Gradient Boosting*. Após as análises realizadas, verificou-se que dentre eles o *Random Forest* se mostrou mais adequado para o conjunto de dados utilizado, com uma acurácia obtida entre 64,5% e 95,4%, considerando três dos cenários mais comuns dentre os 10 cenários avaliados. A acurácia de 64,5% é obtida em um cenário inicial do período letivo a fim de prever preliminarmente o desempenho dos alunos, conforme o período letivo decorre e mais dados são incorporados aos modelo, melhores acurácias são retornadas, atingindo até 95,4% em estágios mais avançados do período letivo.

A análise realizada com os dados do *MOSS* evidencia alguns casos complexos de se analisar e corrobora com a ideia de se ter uma forma apropriada de se acompanhar os alunos. Como já mencionado, há inúmeras variáveis que influenciam no desempenho acadêmico dos alunos, e com uma ferramenta que facilite a comunicação entre os docentes e alunos, possivelmente a compreensão de parte destas variáveis poderia ser feita de forma ágil e eficaz.

A seguir são apresentadas algumas possibilidades de trabalhos futuros considerando o escopo deste trabalho.



## 5.1 Trabalhos Futuros

O tema abordado e a ferramenta proposta e desenvolvida neste trabalho apresentam grande potencial de expansão e aperfeiçoamentos, assim, como oportunidades de trabalhos futuros tem-se a exploração em mais detalhes de dados dos alunos, dentre esses os dados sociais, como a idade, gênero, histórico familiar e educacional, e os de cunho financeiro como a renda familiar e o recebimento de bolsas de auxílio. Com estes dados é possível ampliar o objeto de estudo deste trabalho para além da retenção, alcançando também o âmbito da evasão acadêmica, que possui uma natureza muito mais subjetiva.

Outra possibilidade de trabalho é a análise de diferentes tipos de cursos, comparando por exemplo cursos presenciais e de educação à distância, no qual a frequência não é um dos fatores chave para a aprovação.

O acesso de assistentes psicopedagógicos na ferramenta também é uma possibilidade de expansão, desta forma profissionais com habilidades mais específicas na interação com os estudantes poderiam realizar um trabalho direcionado aos alunos que foram identificados com risco de retenção. Além disso estes e outros profissionais poderiam acompanhar os alunos e assim identificar a motivação por trás das desistências, coletando mais insumos para a criação de um perfil completo do aluno não somente em uma disciplina mas para toda a graduação. Com isso seria possível tomar ações coordenadas baseadas em dados.

Quanto aos dados do *MOSS*, também observa-se oportunidades futuras de análise. Mesmo aqui não apresentando resultados expressivos para compreender o comportamento dos *outliers*, os dados de similaridades dos exercícios podem mostrar outras informações para os docentes, como por exemplo a interação entre os alunos em sala de aula e o reflexo disso nos dados oferecidos pelo *MOSS*. Um volume maior de dados pode ser também importante para uma análise mais rica e detalhada.

Análises mais aprofundadas de como são geradas as similaridades retornadas pelo *MOSS* podem também fazer parte de um trabalho futuro, pois a análise apresentada neste trabalho pode não abranger todos os aspectos destes dados. Por exemplo, parte dos exercícios solicitados aos alunos podem ter uma complexidade muito baixa e poucas opções de solução, neste caso o *MOSS* poderia apresentar uma alta taxa de similaridade para muitos alunos que simplesmente resolveram os exercícios da forma mais simples possível, mesmo que não tenha havido plágio.

## Referências

- ABILASH, R. **Applying Random Forest (Classification) — Machine Learning Algorithm From Scratch With Real Datasets**. 2018. Disponível em: <<https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57>>. Citado na página 20.
- ADRIAANS, P.; ZANTINGE, D. Data mining. 1996. Citado na página 17.
- BABOSA, A. A. S.; CARVALHO, R. N. de; ANDRADE, F. S. Mineração de dados em ambientes virtuais de aprendizagem: Aportes para a pesquisa em educação a distância. Rio de Janeiro, v. 14, n. 52, p. 365–382, 2017. Citado na página 25.
- Bandeira Andriola, W.; Gomes Andriola, C.; MOURA, C. P. Opiniões de docentes e de coordenadores acerca do fenômeno da evasão discente dos cursos de graduação da universidade federal do ceará (ufc). Rio de Janeiro, v. 14, n. 52, p. 365–382, 2012. Citado na página 25.
- BEZ NEILOR A. TONIN, M. S. J. L. **URI Online Judge Academic: Integração e Consolidação da Ferramenta no Processo de Ensino/Aprendizagem**. 2015. Disponível em: <[https://www.researchgate.net/publication/279516635\\_URI\\_Online\\_Judge\\_Academic\\_Integracao\\_e\\_Consolidacao\\_da\\_Ferramenta\\_no\\_Processo\\_de\\_EnsinoAprendizagem](https://www.researchgate.net/publication/279516635_URI_Online_Judge_Academic_Integracao_e_Consolidacao_da_Ferramenta_no_Processo_de_EnsinoAprendizagem)>. Acesso em: 23 de abril de 2021. Citado na página 65.
- CAETANO, M. M. **O Uso de Técnicas de Aprendizado de Máquina na Predição de Desempenho Acadêmico de Alunos em Cursos Superiores**. Outubro 2016. Tese (Mestrado em Ciência da Computação) — Faculdade Campo Limpo Paulista, 2016. Citado 2 vezes nas páginas 17 e 25.
- CAMILO, C. O.; SILVA, J. C. da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. 2009. Citado 2 vezes nas páginas 18 e 19.
- CASTRO, L. F.; ESPITIA, E.; MONTILLA, A. F. Applying crisp-dm in a kdd process for the analysis of student attrition. **Colombian Conference on Computing**, v. 13, p. 386–401, 2018. Citado na página 21.
- CHAPMAN, P. et al. **CRISP-DM 1.0: Step-by-step data mining guide**. [S.l.], 2000. Citado 3 vezes nas páginas 21, 22 e 23.
- Cornelius Junior, R. Uso da mineração de dados na identificação de alunos com perfil de evasão do ensino superior. 2015. Citado na página 24.
- CÔRREA, T. da S.; ALMEIDA, D. E. C. de; Graça Neto, A. F. Comparação entre banco de dados relacional e não relacional em arquitetura distribuída. **III Seminário de Desenvolvimento em SOA com Cloud Computing e Conectividade**, 2017. Citado na página 29.
- CÔRTEZ, S. da C.; PORCARO, R. M.; LIFSCHITZ, S. Mineração de dados: Funcionalidades, técnicas e abordagens. 2002. Citado na página 17.
- GERON, A. **Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems**. [S.l.]: O'Reilly Media, Inc., 2017. Citado na página 20.

- INEP. **Censo da Educação Superior 2015**. 2016. Disponível em: <<https://www.abmes.org.br/documentos/detalhe/491/principais-documentos-censo-da-educacao-superior-2015>>. Acesso em: 23 de abril de 2021. Citado na página 13.
- LAMERS, J. M. de S.; SANTOS, B. S. dos; TOASSI, R. F. C. Retenção e evasão no ensino superior público: Estudo de caso em um curso noturno de odontologia. **Educação em Revista**, n. 33, 2017. Citado na página 16.
- LIAW, A.; WIENER, M. et al. Classification and regression by randomforest. **R news**, v. 2, n. 3, p. 18–22, 2002. Citado na página 20.
- MARQUES, M.; NICOLETTI, M. do C. Mineração de dados educacionais para a predição de desempenho acadêmico em cursos universitários. 2015. Citado 3 vezes nas páginas 13, 17 e 24.
- Microsoft Power BI. **O que é visualização de dados?** 2018. Disponível em: <<https://powerbi.microsoft.com/pt-br/data-visualization/>>. Acesso em: 23 de abril de 2021. Citado na página 23.
- NAGELKERKE, N. J. et al. A note on a general definition of the coefficient of determination. **Biometrika**, Oxford University Press, v. 78, n. 3, p. 691–692, 1991. Citado na página 59.
- PEREIRA, F. C. B. **Determinantes da Evasão de Alunos e os Custos Ocultos para as Instituições de Ensino Superior**: Uma aplicação na universidade do extremo sul catarinense. Outubro 2003. 46-49 p. Tese (Doutorado em Engenharia de Produção) — Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina, Florianópolis, 2003. Citado 2 vezes nas páginas 15 e 16.
- RAMOS, M.; MACHADO, A.; MONTEIRO, M. Ações para o combate a retenção e evasão acadêmica e qualificação do ensino da graduação em nutrição da universidade federal do rs. **Salão de Ensino**, n. 7, 2011. Citado na página 24.
- RIGO, S. J.; CAZELLA, S. C.; CAMBRUZZI, W. Minerando dados educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. 2012. Citado na página 25.
- SANTOS, R. B. et al. Evasão na educação superior: Um estudo do impacto da evasão sobre os custos das universidades federais brasileiras. **SIMPCONT**, v. 3, n. 1, 2018. Citado na página 13.
- SHEARER, C. The crisp-dm model: The new blueprint for data mining. **Journal of Data Warehousing**, v. 5, n. 4, p. 14–19, 2000. Citado na página 21.
- Silva Filho, R. L. L. e et al. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, v. 37, n. 132, p. 641–659, 2007. Citado 2 vezes nas páginas 12 e 13.
- SILVA, G. P. da. Análise de evasão no ensino superior: Uma proposta de diagnóstico de seus determinantes. 2012. Citado 2 vezes nas páginas 15 e 16.
- Tesarik, J.; Dolezal, L.; Kollmann, C. User interface design practices in simple single page web applications. In: **2008 First International Conference on the Applications of Digital Information and Web Technologies (ICADIWT)**. [S.l.: s.n.], 2008. p. 223–228. Citado na página 27.
- TINTO, V. **Leaving College**: Rethinking the causes and cures of student attrition. 2. ed. Chicago: University of Chicago Press, 1987. Citado na página 25.

URI ACADEMIC, O. J. **URI Online Judge Academic**. 2021. Disponível em: <<https://www.urionlinejudge.com.br/academic/>>. Acesso em: 23 de abril de 2021. Citado na página 32.

URI, O. J. **URI Online Judge**. 2021. Disponível em: <<https://www.urionlinejudge.com.br/>>. Acesso em: 23 de abril de 2021. Citado na página 31.

WARD, M. O.; GRINSTEIN, G.; KEIM, D. **Interactive data visualization: foundations, techniques, and applications**. [S.l.]: AK Peters/CRC Press, 2015. Citado 2 vezes nas páginas 23 e 24.

WEIS, S. M.; INDURKHYA, N. **Predict Data Mining**. [S.l.]: Morgan Kaufmann Publishers, Inc., 1999. Citado na página 17.

## Apêndices

**APÊNDICE A – Scores e Desvios Padrão dos modelos construídos para todos os cenários considerados**

| <b>Cenário</b> | <b>Regressão Linear</b> |                  | <b>Gradient Boosting</b> |                  | <b>Random Forest</b> |                  |
|----------------|-------------------------|------------------|--------------------------|------------------|----------------------|------------------|
|                | <b>Score</b>            | <b>D. Padrão</b> | <b>Score</b>             | <b>D. Padrão</b> | <b>Score</b>         | <b>D. Padrão</b> |
| A              | 44,6%                   | ±6,11%           | 47,2%                    | ±7,89%           | 53,3%                | ±6,63%           |
| B              | 58,2%                   | ±6,78%           | 55,3%                    | ±11,34%          | 64,5%                | ±6,92%           |
| C              | 56,4%                   | ±5,67%           | 56,0%                    | ±6,32%           | 62,2%                | ±5,39%           |
| D              | 87,8%                   | ±2,01%           | 85,7%                    | ±1,75%           | 88,7%                | ±2,18%           |
| E              | 89,1%                   | ±3,55%           | 90,1%                    | ±1,91%           | 91,2%                | ±2,91%           |
| F              | 88,1%                   | ±2,43%           | 88,9%                    | ±3,47%           | 91,4%                | ±2,35%           |
| G              | 87,5%                   | ±2,55%           | 89,8%                    | ±1,94%           | 90,9%                | ±1,65%           |
| H              | 95,7%                   | ±0,73%           | 94,9%                    | ±0,69%           | 95,4%                | ±0,88%           |
| I              | 95,8%                   | ±0,85%           | 95,3%                    | ±1,23%           | 95,3%                | ±0,82%           |
| J              | 97,5%                   | ±1,22%           | 98,3%                    | ±0,99%           | 97,3%                | ±0,67%           |

## APÊNDICE B – Importâncias e Coeficientes lineares para os modelos de aprendizagem de máquina treinados

Tabela 10 – Importância das *features* considerando o cenário A para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature      | Importância |           | Coeficiente  |
|--------------|-------------|-----------|--------------|
|              | Gradient B. | Random F. | Regressão L. |
| attendance   | 64,3%       | 61,7%     | 66,2%        |
| micro_grade2 | 22,1%       | 25,2%     | 19,3%        |
| micro_grade1 | 13,6%       | 13,1%     | 11,5%        |

Tabela 11 – Importância das *features* considerando o cenário B para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature      | Importância |           | Coeficiente  |
|--------------|-------------|-----------|--------------|
|              | Gradient B. | Random F. | Regressão L. |
| attendance   | 52,5%       | 50,6%     | 50,7%        |
| micro_grade3 | 19,3%       | 23,5%     | 22,8%        |
| micro_grade2 | 14,9%       | 14,0%     | 10,8%        |
| micro_grade1 | 8,4%        | 7,9%      | 7,8%         |
| percentage   | 4,9%        | 4,0%      | 1,7%         |

Tabela 12 – Importância das *features* considerando o cenário C para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature      | Importância |           | Coeficiente  |
|--------------|-------------|-----------|--------------|
|              | Gradient B. | Random F. | Regressão L. |
| attendance   | 38,3%       | 38,9%     | 46,4%        |
| micro_grade3 | 29,7%       | 32,5%     | 23,8%        |
| micro_grade2 | 15,8%       | 13,3%     | 11,5%        |
| micro_grade4 | 7,4%        | 5,7%      | 7,5%         |
| micro_grade1 | 5,8%        | 6,8%      | 5,6%         |
| percentage   | 3,0%        | 2,8%      | -4,6%        |

Tabela 13 – Importância das *features* considerando o cenário D para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature      | Importância |           | Coeficiente  |
|--------------|-------------|-----------|--------------|
|              | Gradient B. | Random F. | Regressão L. |
| grade1       | 82,5%       | 83,4%     | 71,0%        |
| attendance   | 12,7%       | 11,4%     | 26,9%        |
| micro_grade2 | 2,7%        | 2,7%      | 10,2%        |
| micro_grade1 | 1,3%        | 1,6%      | 0,5%         |
| percentage   | 0,8%        | 0,9%      | 1,5%         |

Tabela 14 – Importância das *features* considerando o cenário E para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature      | Importância |           | Coeficiente  |
|--------------|-------------|-----------|--------------|
|              | Gradient B. | Random F. | Regressão L. |
| grade1       | 81,9%       | 82,9%     | 69,4%        |
| attendance   | 11,9%       | 11,7%     | 20,1%        |
| micro_grade2 | 2,9%        | 2,2%      | 6,8%         |
| micro_grade3 | 2,6%        | 2,0%      | 10,1%        |
| micro_grade1 | 0,6%        | 1,1%      | 0,3%         |

Tabela 15 – Importância das *features* considerando o cenário F para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature      | Importância |           | Coeficiente  |
|--------------|-------------|-----------|--------------|
|              | Gradient B. | Random F. | Regressão L. |
| grade1       | 81,0%       | 82,4%     | 66,9%        |
| attendance   | 11,8%       | 11,4%     | 22,4%        |
| micro_grade3 | 3,5%        | 2,7%      | 10,6%        |
| micro_grade2 | 2,3%        | 1,8%      | 4,7%         |
| micro_grade1 | 0,8%        | 1,2%      | 1,0%         |
| percentage   | 0,6%        | 0,5%      | -3,1%        |

Tabela 16 – Importância das *features* considerando o cenário G para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature      | Importância |           | Coeficiente  |
|--------------|-------------|-----------|--------------|
|              | Gradient B. | Random F. | Regressão L. |
| grade1       | 80,9%       | 82,4%     | 66,7%        |
| attendance   | 10,3%       | 10,6%     | 21,7%        |
| micro_grade2 | 2,5%        | 1,8%      | 5,6%         |
| micro_grade3 | 2,5%        | 2,1%      | 9,7%         |
| micro_grade4 | 2,4%        | 1,3%      | 7,7%         |
| micro_grade1 | 1,0%        | 1,3%      | -0,1%        |
| percentage   | 0,4%        | 0,5%      | -3,1%        |



Tabela 17 – Importância das *features* considerando o cenário H para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature       | Importância |           | Coeficiente  |
|---------------|-------------|-----------|--------------|
|               | Gradient B. | Random F. | Regressão L. |
| grade1        | 66,8%       | 65,1%     | 57,4%        |
| project_grade | 27,7%       | 29,5%     | 28,6%        |
| attendance    | 3,0%        | 3,0%      | 7,8%         |
| micro_grade2  | 1,0%        | 0,8%      | 4,1%         |
| micro_grade3  | 0,8%        | 0,8%      | 1,8%         |
| micro_grade1  | 0,3%        | 0,5%      | 1,3%         |
| percentage    | 0,3%        | 0,2%      | -0,4%        |
| micro_grade4  | 0,2%        | 0,2%      | 2,9%         |

Tabela 18 – Importância das *features* considerando o cenário I para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature      | Importância |           | Coeficiente  |
|--------------|-------------|-----------|--------------|
|              | Gradient B. | Random F. | Regressão L. |
| grade2       | 67,4%       | 71,5%     | 37,1%        |
| grade1       | 26,5%       | 24,0%     | 47,0%        |
| attendance   | 3,0%        | 1,9%      | 9,6%         |
| micro_grade2 | 1,3%        | 1,0%      | 5,3%         |
| micro_grade3 | 1,1%        | 1,1%      | 4,3%         |
| micro_grade1 | 0,6%        | 0,5%      | 2,0%         |

Tabela 19 – Importância das *features* considerando o cenário J para *Gradient Boosting* e *Random Forest* e coeficientes lineares para Regressão Linear.

| Feature       | Importância |           | Coeficiente  |
|---------------|-------------|-----------|--------------|
|               | Gradient B. | Random F. | Regressão L. |
| grade2        | 62,2%       | 67,7%     | 23,5%        |
| grade1        | 29,9%       | 27,1%     | 47,5%        |
| project_grade | 5,2%        | 2,4%      | 17,5%        |
| attendance    | 1,1%        | 1,1%      | 5,8%         |
| micro_grade3  | 0,7%        | 0,8%      | 2,3%         |
| micro_grade2  | 0,6%        | 0,5%      | 4,7%         |
| micro_grade1  | 0,2%        | 0,3%      | 2,4%         |
| micro_grade4  | 0,0%        | 0,1%      | 1,1%         |
| percentage    | 0,0%        | 0,1%      | -1,9%        |