Projeto Orientado em Computação II

Vitor Rodarte Ricoy

Ferramenta de Gestão de Turmas para o Curso de Introdução à Computação para Bioinformática

2023/1

Relatório final do Projeto Orientado em Computação II

Vitor Rodarte Ricoy

Ferramenta de Gestão de Turmas para o Curso de Introdução à Computação para Bioinformática

Apresentado como requisito da disciplina de Projeto Orientado em Computação II do Curso de Bacharelado em Ciência da Computação da UFMG

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Orientadora: Raquel Cardoso de Melo Minardi

2023/1

Resumo

O principal objetivo do presente trabalho é apresentar as atividades realizadas durante o Projeto Orientado em Computação II. O projeto é de cunho tecnológico e tem como objetivo a construção de uma ferramenta de gestão de turmas para o curso de extensão: "Introdução à Computação para Bioinformática", ministrado pela professora Raquel Minardi. A ferramenta será importante para executar a analítica de aprendizagem por meio dos dados gerados pelos alunos, permitindo o acompanhamento mais detalhado e menos trabalhoso destes, visto que o curso é ministrado em um ambiente remoto e possui turmas com um número grande de estudantes.

Para demonstrar os resultados do projeto, primeiro é apresentada a contextualização e os trabalhos relacionados com as atividades desenvolvidas. Após isso, é apresentada a metodologia e o processo de desenvolvimento das atividades realizadas durante o projeto, sendo as principais: a análise exploratória dos dados, a definição das visualizações e do modelo preditivo, a implementação do modelo preditivo e a implementação da ferramenta. Em seguida, são apresentados os resultados alcançados no Projeto Orientado em Computação II, em conjunto com discussões e observações sobre tais resultados. Por fim, são apresentadas as conclusões e também os trabalhos futuros.

Palavras-chave: Ciência de Dados Educacional. Learning Analytics. Analítica de Aprendizagem. Moodle. Introdução à Computação para Bioinformática. Ferramenta Educacional. Gestão de turmas. Python.

Abstract

The main goal of this work is to present the activities carried out during the Oriented Computing Project II. The project is of a technological nature and aims to build a class management tool for the extension course: "Introduction to Computing for Bioinformatics", taught by Professor Raquel Minardi. The tool will be important to execute learning analytics using data generated by students, allowing for more detailed and less laborious monitoring of these, since the course is taught in a remote environment and has classes with a large number of students.

To demonstrate the results of the project, the context and works related to the activities developed are first presented. After that, the methodology and development process of the activities carried out during the project are presented, the main ones being: exploratory data analysis, definition of the visualizations and the predictive model, implementation of the predictive model and the tool implementation. Then, the results achieved in the Computer Oriented Project II are presented, together with discussions and observations on these results. Finally, the conclusions and future works are presented.

Keywords: Educational Data Science. Learning Analytics. Moodle. Introduction to Computing for Bioinformatics. Educational Tool. Class management. Python.

Sumário

1 INTRODUÇÃO 6

2 CONTEXTUALIZAÇÃO E TRABALHOS RELACIONADOS 8
2.1 Contextualização 8
2.1.1 Características e Estrutura do Curso 8
2.1.2 Referencial Teórico 9
2.2 Trabalhos Relacionados 11

3 DESENVOLVIMENTO DO TRABALHO 12
3.1 Metodologia 12
3.2 Execução 13
3.2.1 Análise exploratória 13
3.2.2 Definição das visualizações e modelo preditivo 13
3.2.3 Implementação do modelo preditivo 15
3.2.3.1 Desempenho do modelo preditivo 16
3.2.4 Definição das tecnologias 18
3.2.5 Implementação da ferramenta 18
3.2.5.1 Definição do banco de dados 19
3.2.5.2 Implementação do Login e Cadastro 19
3.2.5.3 Implementação da Importação de Dados 20
3.2.5.4 Implementação do Dashboard 20

4 RESULTADOS E DISCUSSÃO 21
4.1 Login e Cadastro 21
4.2 Importação de Dados 24
4.3 Dashboard 27
4.3.1 Entrega de atividades 27
4.3.2 Nota de atividades 30
4.3.3 Finalização de cada módulo 33
4.3.4 Materiais assistidos 34
4.3.5 Visualização da página do curso 37
4.3.6 Probabilidade de completar o curso 40

5 CONCLUSÕES 41

6 TRABALHOS FUTUROS 42

REFERÊNCIAS 43

1 Introdução

O Ensino à Distância (EAD) já vinha ganhando popularidade nos últimos anos, mas a partir de 2020, a sua importância aumentou consideravelmente, pois ele se tornou parte do dia a dia de praticamente todos os estudantes e professores, em consequência do distanciamento social decorrente da pandemia de COVID-19.

Segundo dados do Censo da Educação Superior de 2021[1], desde 2015 já se tem uma tendência de crescimento do número de vagas para ensino superior à distância. De 2015 a 2016 se teve aproximadamente 61% de crescimento, de 2016 a 2017 se teve aproximadamente 5%, de 2017 a 2018 aproximadamente 52%, de 2018 a 2019 aproximadamente 44%, de 2019 a 2020 aproximadamente 45% e, por fim, de 2020 a 2021 aproximadamente 24%. Com isso, se teve um crescimento total de por volta de 602% nesse período. Também vale observar que de 2018 a 2021 se teve uma redução total de aproximadamente 8% do número de vagas na modalidade presencial. Ou seja, o ensino à distância possui um crescimento expressivo, enquanto o ensino presencial apresenta uma lenta retração.

Esses dados são importantes para percebermos como, consequentemente, as metodologias e ferramentas aplicáveis ao ensino a distância estão se tornando cada vez mais importantes, já que elas possibilitam um andamento e aprendizado melhor dos alunos de cursos EAD, além de tornar o trabalho dos professores mais eficiente e menos maçante.

Uma das grandes vantagens dos cursos à distância é a possibilidade de aperfeiçoá-los por meio da aplicação da Analítica de Aprendizagem, ou *Learning Analytics* (LA). A autora Andrea Filatro define a LA como "a medição, coleta, análise e o reporte de dados sobre os alunos e seus contextos, com o propósito de entender a aprendizagem e os ambientes em que ela ocorre"(FILATRO, 2021, p. 82)[2].

A grande quantidade de dados gerados pelos alunos em ambientes de aprendizado online possibilitam, por meio das técnicas da Ciência de Dados Educacionais, uma aplicação relevante da LA, devido à disponibilidade abundante de informações. Essa aplicação se torna ainda mais importante quando se trata de um MOOC (*Massive Online Open Course*), em que se tem um grande número de alunos, geralmente de origens diversas. O uso de LA nesse contexto pode trazer diversos benefícios aos estudantes e aos professores, ao possibilitar um melhor entendimento do andamento dos alunos no curso.

Portanto, o problema que foi tratado no meu Projeto Orientado em Computação é a análise dos dados, por meio de *Learning Analytics*, dos alunos de uma turma do curso de extensão à distância de Introdução à Computação para Bioinformática, com o objetivo de compreender melhor o relacionamento dos alunos com o curso.

Assim, o objetivo geral do projeto, a ser atingido no final do seu desenvolvimento, é a construção de uma ferramenta de gestão de turmas, utilizando *Learning Analytics*, com o propósito de fornecer meios para o entendimento dos dados de alunos gerados por uma

turma de um curso online, lecionado pelo Moodle.

Vale notar que, por ser uma ferramenta de gestão de turma, o objetivo principal do projeto é fornecer meios para a análise do desempenho dos alunos de uma turma e para a tomada de ações no escopo dessa turma. Dessa forma, o problema engloba tratamentos que envolvam uma turma ativa, objetivando auxiliar os alunos no seu processo de aprendizado durante o curso, e auxiliar o professor nas tomadas de decisões relativas à turma atual. Ou seja, o projeto não envolve análises temporais do curso e suas turmas passadas, nem objetiva auxiliar em alterações permanentes de materiais ou metodologias.

Exemplos de decisões que poderiam ser tomadas a partir da utilização da ferramenta são: entrar em contato com um aluno que está tendo dificuldades, ministrar uma aula de dúvidas para um determinado assunto que a turma teve dificuldade, propor novos exercícios caso a turma tenha tido dificuldade ou uma grande facilidade no exercício anterior, entre outros. Vale notar que a ferramenta também auxilia o docente a ter uma visão geral do comportamento da turma.

Por fim, nesse trabalho são apresentados os resultados alcançados após o desenvolvimento do Projeto Orientado em Computação II. Portanto, são apresentadas as atividades realizadas durante o processo de construção da ferramenta, atingindo o objetivo geral do projeto. Além disso, são discutidos os resultados alcançados e propostos possíveis trabalhos futuros.

2 Contextualização e Trabalhos Relacionados

2.1 Contextualização

O principal contexto do projeto é o ambiente de ensino à distância. Tal ambiente permite uma coleta ativa ou passiva de dados dos alunos, o que aumenta consideravelmente o volume e detalhamento dos dados que são coletados pelos professores ou pela própria plataforma em que o curso é lecionado.

Como cada curso possui suas peculiaridades, a estrutura e características do curso de extensão de Introdução à Computação para Bioinformática serão apresentadas. Isso é feito para que se tenha uma ideia do formato do curso, assim como do contexto e natureza dos dados coletados e analisados ao longo do trabalho.

Também será apresentado o referencial teórico utilizado para o desenvolvimento do trabalho. Tal referencial consiste basicamente em conceitos importantes para a Ciência de Dados Educacional, assim como referenciais de ferramentas já existentes que possuem alguma similaridade com a ferramenta sendo desenvolvida no projeto e, portanto, trouxeram alguma inspiração de funcionalidades.

2.1.1 Características e Estrutura do Curso

Para contextualizar o problema abordado, é importante destacar as características do curso que será analisado pela ferramenta. Idealmente, tais características poderão ser abstraídas como configurações da ferramenta em seu estado final. Entretanto, nas atividades realizadas nessa primeira parte do projeto, é importante entender essas particularidades.

O primeiro ponto importante é a inexistência de avaliações tradicionais no curso. O aluno possui dois momentos em que ele pode testar o que ele aprendeu. O primeiro deles é nos exercícios de revisão, que são corrigidos automaticamente pelo Moodle. Portanto, os alunos recebem uma nota por esses exercícios, mas essa nota não é utilizada ao se considerar a aprovação ou não do aluno.

Outro momento de avaliação são os exercícios práticos, em que os alunos entregam uma resposta pelo Moodle, mas essa resposta não é avaliada pelos docentes, e sim pelos próprios alunos. Após o prazo de entrega, a resposta do exercício é divulgada, em conjunto com um vídeo explicativo. Isso é feito para que o aluno possa se autoavaliar e entender o que ele errou.

Essa abordagem de avaliação é característica de cursos com muitos alunos, pois se torna inviável para o professor corrigir manualmente as atividades de todos os estudantes. Outro fator decisivo para se adotar essa abordagem é o curso não ter um caráter avaliativo para

emitir certificados. O aluno recebe grande autonomia em seu processo de aprendizagem, possuindo o papel de solicitar o certificado de conclusão ao finalizar o conteúdo do curso.

Também vale citar qual a estrutura do curso abordado. Ele é dividido em quatro módulos principais, além de possuir um módulo de preparação e um módulo bônus. Os quatro módulos principais são divididos em: introdução, programação, análise de complexidade de algoritmos e algoritmos para bioinformática. Vale notar que o segundo módulo concentra a maior parte das aulas e atividades avaliativas.

Todas as aulas do curso possuem dois materiais: um conjunto de slides e uma video-aula. Alguns conteúdos específicos também são acompanhados de um desafio, contido no próprio slide da aula. Nesse tipo de conteúdo, também se tem um link para um questionário, a ser respondido pelos alunos, perguntando sobre a dificuldade do desafio e sobre comentários acerca dele.

Por fim, algumas aulas são seguidas de exercícios práticos. Já os exercícios de revisão se referem a um conjunto de aulas, ou até um módulo inteiro, como é o caso do primeiro exercício de revisão.

2.1.2 Referencial Teórico

O principal referencial teórico são os conceitos de *Learning Analytics* (LA) apresentados no livro Data Science na Educação: Presencial, a Distância e Corporativa da autora Andrea Filatro. O conceito mais importante é própria definição de LA, apresentada pela autora (2021, p.82)[2] como sendo a medição, coleta, análise e o reporte de dados de alunos e seus contextos, com o objetivo de entender melhor o processo de aprendizagem.

Outro conceito apresentado nesse mesmo livro, que é importante para o trabalho realizado, é a diferenciação entre modelos de dados educacionais descritivos, preditivos e prescritivos. Segundo a autora (2021, p. 20)[2], o primeiro tipo de modelo retrata e analisa o que está acontecendo agora ou o que aconteceu no passado. O segundo tipo possibilita prever resultados futuros, com algum grau de precisão, geralmente fazendo uso dos dados de modelagens descritivas. Por fim, o terceiro tipo de modelo ajudam os agentes das instituições de ensino a tomarem decisões informadas. Geralmente os modelos prescritivos são construídos com base em *insights* obtidos a partir dos modelos descritivos e preditivos. Esses conceitos estão relacionados com a caracterização das funcionalidades que foram implementadas. Podemos caracterizar as visualizações prototipadas como modelos de dados educacionais descritivos, e o modelo de regressão implementado como um modelo de dados educacional preditivo.

O conceito de MOOCs também é importante para o entendimento do contexto do problema. Os Cursos On-line Abertos e Massivos, conhecidos como MOOCs (*Massive Open Online Courses*), são definidos por Agonács e Matos (2020, p. 19)[3] como uma forma inovadora de aprendizagem à distância, que tende a ser livre e aberta, tendo como o maior diferencial o fato dela ser desenhada para receber participantes em massa, o que

implica que cada um dos estudantes deve gerar a sua própria aprendizagem. Essa definição mostra a importância do *Learning Analytics* para esse tipo de curso, já que as ferramentas de LA também possibilitam que os alunos acessam as análises geradas, permitindo que se tenha um auxílio automatizado aos alunos, facilitando com que seu aprendizado seja mais autônomo.

No mesmo artigo, Agonács e Matos (2020, p. 17)[3] também destacam que os MOOCs tendem a possuir alunos com uma grande diversidade de origens, competências e experiências. Como o curso sendo tratado no projeto é um curso de extensão que possui um grande número de alunos com experiências relativamente heterogêneas, apesar da maioria dos alunos serem da área de ciências biológicas, ele é semelhante a um MOOC nesse aspecto. Outros aspectos em que o curso abordado é semelhante a um MOOC são: a independência do aluno em seu processo de aprendizagem e o número massivo de alunos que podem fazer o curso simultaneamente. Devido a essas semelhanças, os estudos acerca de MOOCs foram uma fonte importante de referências para metodologias que puderam ser aplicadas nesse projeto.

Por fim, existem algumas ferramentas que resolvem um problema similar ao apresentado, compatíveis com o Moodle. A primeira delas é o IntelliBoard[4], uma plataforma de análise de sistemas educacionais. Ela fornece visões variadas de dados educacionais, oferecendo uma experiência bem completa que exige pouco esforço por parte do usuário. Além disso, a plataforma funciona para diversos outros Sistemas de Gestão da Aprendizagem além do Moodle. Entretanto, seu grande ponto negativo é o fato do sistema ser pago e, portanto, menos acessível.

Outras ferramentas que resolvem um problema similar ao apresentado são dois plugins do Moodle. O primeiro se chama SmartKlass[5], e fornece um *dashboard* dentro do Moodle. Apesar de ser pouco flexível e ser a ferramenta mais simples, possuindo poucas análises, ela é fácil de usar e é gratuita. O segundo plugin se chama Google Analytics[6] e fornece uma integração dos dados gerados pelo Moodle com o Google Analytics. Tal integração proporciona ao usuário todo o poder da ferramenta do Google Analytics, entretanto, ela também trás o ponto negativo de ser necessário conhecer a ferramenta para usar o plugin adequadamente. Por fim, ambos plugins possuem o ponto negativo de ser necessário instalá-los no Moodle, o que faz com que eles sejam completamente dependentes da plataforma.

Assim, podemos concluir com a análise dessas ferramentas que o projeto sendo desenvolvido será benéfico, pois será uma ferramenta fácil de usar, gratuita e ao menos tão completa quanto as ferramentas existentes, além de permitir o acesso de alunos às análises geradas. Ademais, a única dependência que a ferramenta terá com o Moodle será a importação dos dados, podendo ser adaptada futuramente a outras plataformas.

2.2 Trabalhos Relacionados

Um trabalho relacionado importante é o artigo *Leveraging the Teaching of Programming Skills to Life Science Students Through Learning Analytics* (DE MELO-MINARDI; DE MELO; BASTOS, 2022)[7], que aborda o mesmo curso de extensão que foi tratado nesse projeto. Dessa forma, esse trabalho é fortemente relacionado por aplicar o *Learning Analytics* no mesmo curso, apesar do artigo possuir um escopo menor. Nele o LA é usado para propor um modelo de previsão da desistência de alunos, de acordo com dados sobre sua participação no curso. Tal modelo preditivo foi uma boa inspiração para a funcionalidade de previsão da ferramenta sendo desenvolvida.

Outro trabalho relacionado, já citado anteriormente, é *Os Cursos On-line Abertos e Massivos (MOOC) como ambientes heutagógicos* (AGONÁCS; MATOS, 2020) [3]. Esse trabalho traz algumas definições e entendimentos dos MOOCs que foram utilizados como conceitos guiadores na definição das funcionalidades da ferramenta.

Por fim, outros dois trabalhos relacionados, que serviram de inspiração para o entendimento de como se aplicar a análise de dados nos MOOCs, mesmo que as metodologias apresentadas nesses artigos não tenham sido utilizadas diretamente, são *Learning analytics and MOOCs.* (INAN; EBNER, 2020) [8], e *Expectativas e desafios: Learning analytics em MOOC* (SILVA; CARVALHO; TEIXEIRA, 2018) [9].

3 Desenvolvimento do Trabalho

Para apresentar o desenvolvimento do trabalho primeiramente será explicada a metodologia adotada para executar as atividades e em seguida serão explicadas a execução das tarefas em si.

3.1 Metodologia

Para atingir o objetivo de construir a ferramenta de gestão de turmas, foram realizadas, além de estudos sobre a Ciência de Dados Educacional, diversas atividades para compreender os dados disponíveis, de nir as funcionalidades que a ferramenta conterá e implementar tais funcionalidades.

O estudo dos dados disponíveis se iniciou com a leitura da documentação do Moodle [10], para de nir o formato preferível de obtenção dos dados. Após isso, foi realizada uma análise exploratória dos dados da turma 12 do curso, em um Jupyter Notebook [11], utilizando a biblioteca Pandas [12].

Após a análise exploratória dos dados, foram decididas quais seriam as visualizações e o modelo de predição presentes na ferramenta. Os dados e os *insights* gerados pela análise exploratória foram a base para esse processo de decisão.

Após a de nição das funcionalidades, foram prototipadas as visualizações e implementado o modelo preditivo a ser implementado na ferramenta. As visualizações foram prototipadas também em um Jupyter Notebook [11], utilizando a biblioteca Plotly [13]. O modelo preditivo foi implementado no mesmo Jupyter Notebook [11], utilizando a biblioteca Scikit Learn [14]. Em ambas as atividades foram utilizadas as bibliotecas Pandas [12] e Numpy [15] como apoio.

Após a prototipação das visualizações e a implementação do modelos preditivos, foram de nidas as tecnologias a serem utilizadas para implementar a ferramenta. Como a ferramenta consiste em um aplicativo web, uma aplicação de servidor e um banco de dados, foram de nidas as tecnologias para essas peças de infraestrutura.

Após a de nição das tecnologias, a ferramenta foi implementada com as tecnologias de nidas. A ferramenta possui três funcionalidades chave: o sistema de login e cadastro, o sistema de importação de dados das turmas e o *dashboard* que contém todas as visualizações de uma turma.

Por m, a ferramenta foi publicada na Azure, para que ela que acessível publicamente. Foi utilizada a conta de estudante da UFMG, portanto os serviços usados são gratuitos. Como o serviço gratuito para o servidor Python apresentou um desempenho bem abaixo do tolerável, o servidor foi hospedado em um servidor do DCC na UFMG. O aplicativo web e o banco de dados, seguem hospedados na Azure gratuitamente.

3.2 Execução

3.2.1 Análise exploratória

A análise exploratória foi executada em um Jupyter Notebook [11], com o objetivo de analisar os dados da turma 12, iniciada em 1 de julho de 2022 e terminada em 31 de outubro de 2022, com 72 alunos.

Primeiramente, foram baixados os três conjuntos de dados da turma 12: olog detalhado, a entrega de atividades e abertura de aulas, e as notas das atividades avaliativas.

Esses dados foram analisados, tratados e algumas informações foram coletadas. O tratamento foi feito removendo dolog detalhado membros que não eram alunos, que consistem basicamente em professores, monitores e administradores. Após a remoção dos membros, foi feito um tratamento para colocar os alunos como atores de ações automáticas, em casos que tais ações são feitas pelo Moodle e um administrador consta como autor. Por m, também foram feitas algumas padronizações nos nomes de atividades, o que mostrou tornar mais fácil integração com dados de outras turmas, como foi necessário para o treinamento do modelo preditivo.

Com a análise, foi descoberto que existem 38.032 entradas no olog detalhado, além de ter sido feito um diagrama de pareto para os diferentes tipo de entradas nesse log.

Também foi feito um diagrama de pareto para visualizar a participação de alunos nas entradas dolog detalhado, para identi car se olog cobria bem a base de estudantes. Por m, também foram gerados grá cos da nota média dos alunos que entregaram os exercícios de revisão, do número de alunos que entregaram os exercícios de revisão e do número de alunos que entregaram o exercício prático.

3.2.2 De nição das visualizações e modelo preditivo

Primeiramente foi feita a de nição das visualizações de dados que seriam implementadas. Foram de nidos seis grandes temas para as visualizações, tendo cada um deles um número variado de visualizações:

- Entrega de atividades: visualizações que dizem respeito à entrega de atividades por alunos. Dentro desse tema foram feitos:
  - um heatmapde entregas por aluno;
  - um grá co de barras da média de entregas por aluno;
  - um histograma da média de entrega das atividades por aluno;
  - um grá co do número de entregas por atividade;
- Nota de atividades: visualizações que dizem respeito às notas das atividades. Dentro desse tema foram feitos:

um heatmapde notas nas atividades por aluno;

um grá co de barras da nota percentual por aluno;

um histograma da nota percentual dos alunos;

um grá co de barras da nota dos alunos, feito para cada exercício;

- Finalização de cada módulo: visualizações que dizem respeito à nalização dos módulos do curso. Dentro desse tema foram feitos:

  um heatmapda porcentagem completa de cada módulo por aluno;

  um grá co de barras da conclusão percentual por aluno de cada módulo do curso;

- Materiais assistidos: visualizações que dizem respeito à visualização de materiais por alunos. Dentro desse tema foram feitos:

  um grá co de barras do número médio de vezes que um material é repetido por aluno;

  uma tabela da frequência de repetição por material e por aluno, ordenada de forma decrescente pela diferença entre a repetição do material pelo aluno e a média de repetições do aluno;

  um histograma da repetição média de visualização de materiais por aluno;

  uma tabela da repetição média de cada material por material, ordenada pelo material com a maior média;

  um histograma da repetição média de visualização por material;

- Visualização da página do curso: visualizações que dizem respeito a entrada dos alunos na página do curso. Dentro desse tema foram feitos:

  um heatmapdo número de vezes que cada aluno entrou na página do curso, por semana;

  um grá co de barras com a média de acessos ao curso por aluno;

  um histograma da média de entrada no curso por aluno;

  um grá co de barras da média de entradas no curso por semana;

  um histograma do número médio de entradas por semana do curso;

- Probabilidade de completar o curso: Visualizações que dizem respeito à probabilidade dos alunos de completar o curso, usando como base os dados gerados pelo modelo preditivo implementado. Dentro desse tema foram feitos:

  um grá co de barras da probabilidade de completar o curso por aluno;

um histograma das probabilidades de completar o curso por aluno;

Assim, também podemos notar que foram de nidas 22 visualizações a serem implementadas.

Já sobre o modelo preditivo, foi de nido que o modelo trataria da probabilidade de um aluno ser reprovado. Tal de nição foi alcançada após analisar a diferença entre realizar apenas uma predição de se o aluno seria aprovado ou não no curso, e atribuir uma probabilidade à aprovação do aluno. A predição com uma probabilidade foi preferida para dar uma ideia quantitativa do seu desempenho ao aluno e também para dar ideia de progresso, caso o aluno esteja se esforçando para aumentar a sua probabilidade de ser aprovado.

O critério adotado para aprovação é o aluno ter visto 75% das aulas e entregue 60% das atividades. Esse critério foi escolhido por já ser usado na prática em diversos cursos e, após uma análise dos dados, foi veri cado que ao utilizar ele na turma 12, 23 alunos seriam aprovados, ou seja, 32% dos alunos. Esse número nos pareceu razoável, visto que é um curso à distância e que não possui critérios de reprovação.

Outro ponto importante é relembrar que o curso não possui o conceito de reprovação, portanto o modelo preditivo tem o seu próprio critério de aprovação que representa um desempenho satisfatório do aluno. Assim, mesmo que o modelo possa ser usado como referência pelos alunos e professores, o próprio modelo e os seus critérios não implicam que um aluno será reprovado ou não.

Finalmente, a metodologia utilizada para testar o modelo foi calcular as métricas R^2, Mean Absolute Error, Mean Squared Error e Root Mean Squared Error, usando como treinamento os dados das turmas de 8 a 11 e como teste os dados da turma 12.

Por m, para realizar a avaliação, os dados foram divididos cronologicamente e o treinamento e teste foi feito para o conjunto de dados de forma incremental. Ou seja, primeiro se treinou o modelo com uma atividade e testou o resultado da predição utilizando uma atividade. O mesmo é feito considerando duas atividades, três atividades e assim por diante, sempre em ordem cronológica. Isso foi feito para simular diferentes momentos do curso em que o modelo poderá ser aplicado, considerando a ferramenta nal.

3.2.3 Implementação do modelo preditivo

A implementação do modelo preditivo foi feita por meio da biblioteca Scikit Learn [14], com apoio do Pandas [12] e do Numpy [15]. A importação e tratamento dos dados foram realizados de acordo com o processo feito na análise exploratória.

Após o treinamento do modelo, ele foi avaliado utilizando os dados da turma 12. A avaliação foi feita utilizando quatro métricas: R^2, Mean Absolute Error, Mean Squared Error e Root Mean Squared Error

Por m, o modelo treinado foi salvo, utilizando a extensão .joblib [16]. Com isso foi gerado um arquivo .joblib que é carregado no aplicativo de servidor da ferramenta para executar o modelo em dados de turmas importados, salvando os resultados no banco de dados.

3.2.3.1 Desempenho do modelo preditivo

Figura 1 Score R2 para os dois modelos implementados

Figura 2 Erro quadrado médio para os dois modelos implementados

Figura 3 Erro absoluto médio para os dois modelos implementados

Figura 4 Raiz do erro quadrado médio para os dois modelos implementados

Os scores do modelo implementado podem ser vistos nas Figuras 1, 2, 3 e 4. Nelas podemos perceber que os resultados alcançados são razoáveis e as taxas de erro do modelo são aceitáveis, principalmente depois de, aproximadamente, 30 atividades realizadas.

3.2.4 De nição das tecnologias

A de nição das tecnologias foi feita com base na praticidade de implementação dos recursos necessários. Para o banco de dados foi escolhido o SGDB Postgres.

Para o aplicativo web foi escolhida a framework React, em conjunto com a biblioteca Plotly [17], para auxiliar na implementação das visualizações.

Já para o servidor foi escolhida a linguagem Python, utilizando a framework Flask, para implementar a API que será consumida pelo aplicativo web.

3.2.5 Implementação da ferramenta

A implementação da ferramenta foi dividida na implementação de três subsistemas, além da estrutura do banco de dados.

3.2.5.1 De nição do banco de dados

Figura 5 Diagrama do banco de dados

Antes da implementação de qualquer parte do sistema, foi de nida a estrutura do banco de dados da ferramenta. A estrutura de nida segue o diagrama da Figura 5, e atende as necessidades de dados de todas as visualizações de dados, além do sistema de login e cadastro e do sistema de importação de dados.

3.2.5.2 Implementação do Login e Cadastro

Para implementar o sistema de login e cadastro, primeiramente, foi de nido que todas as senhas passariam por um algoritmo de hash antes de serem salvas no banco, com isso foi usado um algoritmo SHA256 para essa nalidade. Além disso, a sessão do usuário é mantida utilizando um token JWT, utilizando um algoritmo HMAC com SHA256.

Esse token possui apenas informações não sensíveis (o código do usuário no sistema, o seu nome e seu e-mail), sendo mais importante para chamadas futuras a sua assinatura, atestando a autenticidade dele, do que o seu conteúdo em si.

A implementação do login e do cadastro, fora essas questões de segurança foi bem simples, seguindo a lógica comum de criar uma entrada na tabela do usuário no banco de dados no cadastro, e consultar essa tabela no login.

3.2.5.3 Implementação da Importação de Dados

Já a importação de dados foi implementada com uma tela simples, que permite enviar os três arquivos necessários para o funcionamento da ferramenta: o log detalhado dos alunos, as notas das atividades e as entregas de atividades. O funcionamento desse sistema é mais complexo no lado do servidor, que é responsável por tratar os dados enviados pela tela, unir os três conjuntos de dados, aplicar o modelo de predição e salvar o resultado de todo esse processamento no banco de dados.

O tratamento dos dados foi feito de uma forma que funciona para qualquer curso, recebendo como parâmetro a associação dos módulos e materiais, a lista de materiais ignorados e a lista de alunos ignorados. Entretanto, a tela não suporta a configuração desses parâmetros, enviando sempre um conjunto de parâmetros que funciona para o curso de Introdução à Computação para Bioinformática.

3.2.5.4 Implementação do Dashboard

As visualizações foram implementadas no React, com o apoio da biblioteca Plotly [17]. As visualizações foram agrupadas nos seis temas de nidos anteriormente, tendo uma aba para cada tema. Dentro de cada tema, temos uma aba para cada visualização.

A implementação das visualizações foi feita seguindo a lógica de implementar um componente React, que se comunica com o servidor para obter os dados, buscados por meio de uma query SQL pelo servidor no banco de dados. Após o componente obter os dados, ele renderiza a visualização utilizando esses dados.

4 Resultados e Discussão

Serão apresentados os resultados do projeto. Tais resultados estão estritamente relacionados com a ferramenta nal desenvolvida.

A ferramenta nal pode ser acessada pelo link. Como a ferramenta possui um sistema de login o seu link de acesso não revela nenhum dado privado de alunos, e os leitores podem se cadastrar e importar dados, caso desejado.

Vale notar que caso o leitor não possua acesso a dados de turmas do curso tratado, a ferramenta não terá utilidade, já que ela não se adapta a dados de qualquer turma.

4.1 Login e Cadastro

Figura 6 Tela de login

Figura 7 Tela de cadastro

Figura 8 Tela de escolha de turma

O ponto de entrada do sistema são as telas de login e cadastro, vistas nas Figuras 6 e 7, respectivamente. Após o login ou cadastro é exibida a tela de selecionar turma, vista na Figura 8.

4.2 Importação de Dados

Figura 9 Tela de importação de dados de uma nova turma

Figura 10 Tela de atualização dos dados de uma turma