INSTITUTO FEDERAL DE PERNAMBUCO CAMPUS BELO JARDIM
BACHARELADO EM ENGENHARIA DE SOFTWARE

FELIPE MONTEIRO DA SILVA
ALYSSON PEREIRA ASSUNÇÃO

ANÁLISE E PREDIÇÃO DO DESEMPENHO DOS ESTUDANTES DO ENSINO MÉDIO DE BELO JARDIM-PE

Belo Jardim, Pernambuco, 28/12/2023

FELIPE MONTEIRO DA SILVA
ALYSSON PEREIRA ASSUNÇÃO

ANÁLISE E PREDIÇÃO DO DESEMPENHO DOS ESTUDANTES DO ENSINO MÉDIO DE BELO JARDIM-PE

Trabalho de conclusão de Curso (TCC) apresentado como requisito parcial para obtenção do grau de Bacharel em Engenharia de Software. TCC aprovado no curso de Engenharia de Software do IFPE - Campus Belo Jardim

Banca de Qualificação:

Francisco Ariaildo Da Costa Sá Lucena - IFPE - Campus Belo Jardim
João Almeida e Silva - IFPE - Campus Belo Jardim
Sionise Rocha Gomes - IFAM - Campus Presidente Figueiredo

Belo Jardim, Pernambuco, 28/12/2023.

Dados Internacionais de Catalogação - CIP

S586a Silva, Felipe Monteiro da

Análise e predição do desempenho dos estudantes do ensino médio de Belo Jardim - PE / Felipe Monteiro da Silva, Alysson Pereira Assunção. – Belo Jardim-PE, 2023.
50f.: il. ; 29 cm.

Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Software) – Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco, Campus Belo Jardim- PE, 2023.

Orientador: Prof.º Francisco Ariaildo da Costa Sá Lucena.

Inclui referências.

1. Tecnologia da Informação - Educação. 2. Ferramenta de TI. 3. Desempenho de estudantes - análise. I. Título. II. Assunção, Alysson Pereira. III. Lucena, Francisco Ariaildo da Costa Sá. III. Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco.

CDD 004

Ficha catalográfica elaborada pelo Sistema Integrado de Bibliotecas - SIBI/IFPE.
Bibliotecária: Louise Machado Freire Dias CRB – 4 /2267.

FELIPE MONTEIRO DA SILVA
ALYSSON PEREIRA ASSUNÇÃO

ANÁLISE E PREDIÇÃO DO DESEMPENHO DOS ESTUDANTES DO ENSINO MÉDIO DE BELO JARDIM-PE

Francisco Ariaildo Da Costa Sá Lucena

Professor Orientador

João Almeida e Silva

Convidado 1

Sionise Rocha Gomes

Convidado 2

AGRADECIMENTOS

A elaboração deste trabalho não teria sido possível sem o apoio e a colaboração de diversas pessoas, amigos, familiares e instituições, às quais gostaríamos de agradecer e reconhecer. Agradeço ao meu orientador Francisco que me conduziu na iniciativa deste trabalho importante para educação.

Por fim e mais importante que todos, agradeço a Deus por ter me dado forças e paciência para conseguir terminar esse projeto, só Ele sabe o tamanho da luta!! .

“Não fui eu que ordenei a você? Seja forte e corajoso! Não se apavore nem desanime, pois o Senhor, o seu Deus, estará com você por onde você andar". Josué 1:9”

RESUMO

Este trabalho de conclusão de curso (TCC) apresenta um estudo sobre o desempenho escolar dos estudantes que realizaram o Exame Nacional do Ensino Médio(ENEM), na cidade de Belo Jardim-PE. Ao decorrer desta pesquisa foi realizada a análise e predição dos dados desses participantes de acordo com a média das notas nas áreas de conhecimento, sendo Matemática, Linguagens, Humanas e Ciências da Natureza e suas tecnologias. Deste modo, podemos analisar fatores que possam impactar no desempenho dos alunos com base no questionário socioeconômico disponibilizado pelo INEP, onde podemos analisar e verificar os resultados do algoritmo de Machine learning.

Palavras-chave: Educação; Aprendizagem de máquina; Análise; Predição; ENEM; Desempenho.

ABSTRACT

This course completion work (TCC) presents a study on the academic performance of students who took the National High School Examination (ENEM), in the city of Belo Jardim-PE. During this research, analysis and prediction of these participants was carried out according to the average grades in the areas of knowledge, being Mathematics, Languages, Humanities and Natural Sciences and their technologies. Based on this, we can analyze factors that may impact students' performance based on the socioeconomic questionnaire provided by INEP, where we can analyze and verify the results of the Machine learning algorithm.

Keywords: Education; Machine Learning; Analysis; Prediction; ENEM; Performance.

LISTA DE ABREVIATURAS

ENEM Exame Nacional do Ensino Médio

INEP Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

KDD Knowledge Discovery in Database(Descoberta de Conhecimento em Banco de Dados)

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 Introdução

O processo de Revolução Industrial foi iniciado no século XVIII. Desde então, o mesmo vem passando por eras e contextos, onde cada um deles traz uma mudança significativa nas relações humanas. Mas é a sua terceira e mais recente onda (conhecida como Revolução Digital) que traz sentido completo a conceitos como Globalização e Inclusão Digital.

Agora, em uma realidade de acesso e uso massivo de Tecnologias da Informação, é notório que o comportamento do coletivo social sofre alterações influenciadas pela quantidade exorbitante de dados que temos acesso da palma da nossa mão. Para Schwab (2018), "As mudanças são tão profundas que, na perspectiva da história humana, nunca houve um momento tão potencialmente promissor ou perigoso." Sobre a Revolução Digital, o autor ainda acrescenta que "ela envolve a transformação de sistemas inteiros entre países e dentro deles, em empresas, indústrias e em toda sociedade" (Klaus 2018).

Diante disto, podemos perceber a relevância da informação em uma comunidade global que sofre constantes transformações. Entretanto, nosso papel não precisa e não deve ser necessariamente de espectador destas mudanças, mas sim, de agentes ativos e de promotores de perspectivas positivas para o futuro a partir do tratamento e da análise dos dados que podemos obter.

E não há contexto onde esta análise se faz mais necessária do que na educação. Porém, pelo menos quando falamos no contexto nacional, surpreendentemente não temos uma cultura de análise de dados.

> Atualmente, na área da pesquisa educacional, excluindo análises de dados de avaliações de rendimento escolar realizadas em alguns sistemas educacionais no Brasil, poucos estudos empregam metodologias quantitativas. (...) No entanto, há problemas educacionais que para sua contextualização e compreensão necessitam ser qualificados através de dados quantitativos. Por exemplo, 1 1.0 2 como compreender a questão do analfabetismo no Brasil, e discutir políticas em relação a esse problema, sem ter dados sobre seu volume e a sua distribuição segundo algumas variáveis, como gênero, idade, condição socioeconômica, região geográfica, cidade-meio rural, etc (BERNARDETTE, 2004).

O Brasil é um país de dimensões continentais e de estado federalista. Sua diversidade cultural, política e étnica não permite formular um diagnóstico geral a respeito do seu cenário

educacional. É importante, portanto, fomentar a pesquisa acadêmica neste âmbito, valendo-se de dados quantitativos.

Motivados por esta lacuna na pesquisa brasileira e local, nos propomos a realizar uma investigação quantitativa do cenário educacional da cidade de Belo Jardim - Agreste de PE.

Atualmente, temos o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) como principal e mais completa fonte de dados socioeconômicos dos estudantes que procuram ingressar no ensino superior através do Exame Nacional do Ensino Médio (ENEM). Com este recurso em mãos, podemos nos valer das ferramentas da Revolução Digital e, por meio de técnicas e algoritmos de Aprendizagem de Máquina, desenvolver uma predição de dados que auxiliem a tracionar e reconhecer padrões no contexto educacional local.

Para prever a nota de um aluno utilizando somente as informações do questionário econômico, escolhemos um modelo de regressão capaz de trabalhar com um volume muito grande de dados: a Árvore de Decisão.

Esta é uma técnica de modelagem amplamente utilizada em ciência da computação e aprendizado de máquina para a tomada de decisões baseadas em dados. Ela é uma representação gráfica de um processo de decisão que consiste em uma estrutura hierárquica de nós, onde cada nó representa uma escolha entre diferentes alternativas, e cada ramo que se origina desses nós representa uma possível consequência dessa escolha. O objetivo principal de uma árvore de decisão é segmentar o espaço de entrada em regiões distintas, permitindo a classificação de dados ou a previsão de resultados com base nas características dos dados de entrada (BREIMAN, 1984).

As árvores de decisão são uma ferramenta valiosa em áreas como mineração de dados, reconhecimento de padrões e tomada de decisões automatizadas, e sua eficácia é amplamente documentada na literatura de aprendizado de máquina e mineração de dados (QUINLAN, 1986).

Acreditamos que este estudo pode nos ajudar a olhar para a educação local sob um novo prisma, além de ser um exemplo de como as ferramentas de Machine Learning podem ser aliadas da reflexão e do progresso da sociedade.

1.1 Objetivo Geral

Analisar o cenário educacional da cidade de Belo Jardim - PE, por meio do desenvolvimento de um modelo de predição de dados que utilize informações socioeconômicas e culturais dos estudantes que realizaram o Exame Nacional do Ensino Médio (ENEM) na região, e assim, identificar correlações entre essas variáveis e o desempenho dos alunos no ENEM, bem como fornecer insights relevantes para a compreensão do ambiente educacional local e, por fim, oferecer recomendações que possam contribuir para o aprimoramento da equidade educacional na cidade.

1.2 Objetivo Específicos

Realizar a coleta de dados socioeconômicos e culturais dos estudantes que participaram do Exame Nacional do Ensino Médio (ENEM) na cidade de Belo Jardim.

Realizar uma análise descritiva e gráfica dos dados coletados, de forma que se ilustram as principais características dos estudantes, como renda familiar, nível de escolaridade dos pais e acesso a recursos educacionais.

Investigar se existem relações significativas entre o ambiente educacional e o desempenho dos alunos.

Desenvolver um modelo de predição estatística que utilize as informações socioeconômicas e culturais dos estudantes como variáveis independentes para prever seu desempenho no Enem.

Interpretar os resultados do modelo de predição, destacando as variáveis mais influentes no desempenho dos estudantes no ENEM.

Discutir as implicações dos resultados para a compreensão do cenário educacional da cidade de Belo Jardim.

2 Fundamentação teórica

Foi a partir dos anos 2000 que vimos crescer o número de universidades no Brasil. Crescimento este que contemplou uma diversidade demográfica e, consequentemente, populacional.

> De 2003 a 2010, houve um salto de 45 para 59 universidades federais, o que representa uma ampliação de 31 por cento; e de 148 campi para 274 campi / unidades, com um crescimento de 85 por cento. A interiorização das universidades e dos campi também proporcionou uma elevação no número de municípios atendidos: de 114 para 272, com um crescimento de 138 por cento (SOARES, 2013).

A mudança dos últimos vinte anos no cenário do ensino superior no Brasil, por si só, já poderia ser tema de estudo. Poderíamos observar o novo cenário do ensino acadêmico e identificar seus impactos no mercado de trabalho e na produção científica do Brasil.

Ainda assim, é notório que a educação básica não evoluiu no mesmo ritmo que o ensino superior. Isto configura um problema grave, uma vez que como consequência, a educação superior (ainda que pública) continua majoritariamente elitista.

> Se é verdade que a expansão recente, especialmente de vagas públicas, abre oportunidades para que maior número de jovens ingressem num curso de graduação, é verdade também que o sistema brasileiro de educação superior (...) continua sendo basicamente de acesso de elite (SOARES, 2013).

Partindo deste ponto, é necessária uma análise qualitativa e quantitativa do ensino médio brasileiro, a fim de identificar os problemas que impedem a acessibilidade do ensino superior para todos, e assim, tracionar medidas que possam mitigar estas questões.

Sabemos que em um país com dimensões continentais e estado federalista como o Brasil, é praticamente impossível prover um diagnóstico geral para a educação básica nacional. Cada contexto deve ser analisado considerando as suas nuances e variáveis próprias.

Recolhendo os dados obtidos a partir do Exame Nacional do Ensino Médio dos últimos quatro anos, podemos usar ferramentas de análise e predição de dados a fim de identificar o cenário socioeconômico e educacional da cidade de Belo Jardim (Agreste de PE).

Com o decorrer das últimas décadas, o Exame Nacional do Ensino Médio vem aumentando sua relevância social. Por meio dele, podemos aprimorar a matriz curricular tanto do ensino superior quanto do ensino médio e básico como um todo.

> O ENEM teve reestruturações metodológicas e teóricas em 2009 com o objetivo de aproximar a sua matriz das proposições das Diretrizes Curriculares do Ensino Médio; viabilizar o uso dos resultados nos processos de seleção nas universidades; chamar a responsabilidade delas tanto para a formação básica como para a docente, além de induzir as mudanças curriculares no ensino médio (Maceno 2011).

Este fator intrínseco ao exame nos permite realizar uma análise apurada da realidade dos inscritos na avaliação, o que torna muito mais fácil analisar o cenário do ensino médio brasileiro numa escala nacional e local, bem como suas variações no decorrer dos anos e contextos políticos e históricos.

Esta rica fonte de informação pode ser muito bem trabalhada quando aliada à técnicas de predição de dados. Tema que abordaremos com mais afinco no decorrer deste capítulo.

2.1 Engenharia de Software

É a partir dos anos 1970 que a literatura menciona a chamada ‘crise do software’. Este termo, à primeira vista, não parece condizer com a realidade, uma vez que as soluções de software se fazem mais necessárias e presentes do que nunca. O que seria então esta crise já temida há tantas décadas?

Waslawick afirma que “a crise do software continuará enquanto os desenvolvedores continuarem a utilizar processos artesanais e a não capitalizar erros e acertos aplicando as modernas técnicas da Engenharia de Software” (WASLAWICK, 2013).

A Engenharia de Software é a área do saber responsável pela gerência e produção de processos e produtos de software, garantindo que estes sejam implantados seguindo todos os parâmetros de qualidade, desde a fase do design e projeção até a fase de teste e implantação.

Nas palavras de Pressman (1995), a Engenharia de Software é o estabelecimento e uso de sólidos princípios de engenharia para que se possa obter economicamente um software que seja confiável e que funcione eficientemente em máquinas reais (PRESSMAN, 1995).

Um elemento muito comum e necessário a qualquer software que atenda aos requisitos mínimos de eficácia e qualidade são a tratativa dos seus dados. E é aqui que focaremos nossa atenção e esforços para a produção de um algoritmo assertivo.

Segundo Figueira (1998), “a tecnologia tornou relativamente fácil o acúmulo de dados. A consequência é a ampliação do uso dos Data Warehouses. Ao mesmo tempo, a informação é valorizada como nunca antes na história, e os dados armazenados nos Data Warehouses são vasculhados por profissionais especializados, à procura de tendências e padrões.” (FIGUEIRA, 1998)

Mais do que nunca, os dados podem ser considerados tanto uma mina de ouro nas relações comerciais quanto a chave para a construção de uma sociedade que evolua e mitigue paradigmas até então, prejudiciais para a ideia de uma comunidade global justa e próspera.

Esta última utilização dos dados ainda é subaproveitada no Brasil, o que não nos permite criar estratégias baseadas em dados concretos e atuais, principalmente quando falamos do uso de dados dados na educação brasileira:

> O uso de dados quantitativos na pesquisa educacional no Brasil nunca teve, pois, uma tradição sólida, ou uma utilização mais ampla. Isto dificultou, e dificulta, o uso desses instrumentais analíticos de modo mais consistente, bem como dificulta a construção de uma perspectiva mais fundamentada e crítica sobre o que eles podem ou não podem nos oferecer (...) (BERNARDETTE, 2004).

É papel da ciência buscar padrões nos fenômenos sociais e culturais, a fim de examinálos, entendê-los e alterá-los. Nas Tecnologias da Informação, isto não seria diferente. A engenharia de software permite que analisemos dados a fim de reconhecer padrões e trabalhar baseado neles.

Temos como exemplo, um e-commerce. Quando bem implementado, a aplicação pode fazer mais do que o serviço de compra e venda. O proprietário / cliente do projeto, pode salvar e obter dados referentes às preferências dos usuários finais do sistema. Adicionar múltiplas variáveis neste cenário permite que se tenha controle também nas vendas por época do ano, região, tipo de produto, faixa etária, gênero, faixa de preço, etc.

Ainda podemos tirar mais informações deste cenário. Podemos também trabalhar com logs de usuários, nos permitindo entender qual o caminho que o usuário tomou para chegar a determinada situação. Em que ponto o usuário desistiu ou deixou de usar o sistema? Quais as pesquisas que o usuário faz que o levam a comprar?

E se pudermos prever estes cenários, evitando os maus e enfatizando os bons? E se pudermos fazer isto, agora não para propósitos comerciais, mas para atender demandas educacionais e culturais? E se pudermos evitar a evasão de discentes em instituições de ensino? Isto é possível quando estudamos o conceito de predição de dados.

A predição de dados, também conhecida como modelagem preditiva, é uma área da ciência de dados que busca prever resultados futuros com base em padrões e tendências identificados em dados históricos.

Trata-se de uma área interdisciplinar, mobilizando principalmente conhecimentos de análise estatística de dados, aprendizagem de máquina, reconhecimento de padrões e visualização de dados.

2.2 Análise e Predição de dados na Educação

A análise e predição de dados pode se tornar uma ferramenta vital no setor educacional, permitindo que instituições de ensino tomem decisões informadas e personalizadas para melhorar o desempenho dos alunos e otimizar o processo de aprendizado. Discutiremos agora os princípios, métodos e aplicações da análise e predição de dados na educação, destacando suas possibilidades de significativas contribuições na melhoria da qualidade do ensino.

Personalização do ensino: A análise de dados permite que os educadores adaptem o conteúdo do curso com base nas necessidades individuais dos alunos, oferecendo uma abordagem mais personalizada.

Previsão de desistência: Identificar alunos em risco de desistência permite que as instituições de ensino tomem medidas preventivas para apoiar esses alunos.

Avaliação da eficácia do ensino: A análise de dados pode ajudar a identificar quais métodos de ensino são mais eficazes, possibilitando melhorias contínuas no currículo.

Aprimoramento da retenção de alunos: Através da identificação de problemas comuns que levam à desistência, as instituições podem implementar estratégias de retenção mais eficazes.

Podemos perceber como a análise e a predição de dados na educação têm o potencial de revolucionar a forma como ensinamos e aprendemos. Essas técnicas permitem que as instituições educacionais tomem decisões mais informadas e personalizadas, melhorando a eficácia do ensino e a retenção de alunos.

No entanto, é essencial garantir que a coleta e análise de dados sejam realizadas com responsabilidade e consideração pela privacidade dos alunos, garantindo que os benefícios sejam alcançados sem comprometer a ética educacional.

Por isso, para fazer este estudo, usaremos aquela que tem sido a maior fonte de dados da educação no Brasil: O próprio Exame Nacional do Ensino Médio (ENEM).

O Enem coleta uma vasta quantidade de informações sobre o desempenho dos estudantes em várias disciplinas, suas características socioeconômicas e outras variáveis relevantes. Essa riqueza de dados transformou o Enem em uma fonte valiosa para análises educacionais.

A análise dos dados do Enem tem múltiplas finalidades:

Avaliação do Sistema Educacional: Os resultados do Enem permitem avaliar a qualidade do ensino médio no país e identificar áreas onde melhorias são necessárias.

Seleção Universitária: O Enem é usado como critério de seleção para ingresso em universidades públicas e privadas, tornando-o uma ferramenta fundamental na alocação de vagas.

Identificação de Desigualdades: A análise dos dados do Enem pode destacar disparidades educacionais entre grupos demográficos, contribuindo para políticas de equidade.

Predição de Desempenho: Com o uso de algoritmos de aprendizado de máquina, é possível prever o desempenho futuro dos alunos com base em seu histórico e características pessoais.

Além disso, a análise dos resultados do Enem tem implicações práticas para a formulação de políticas educacionais. As informações obtidas podem orientar a alocação de recursos para áreas com maiores necessidades, o desenvolvimento de programas de apoio aos alunos e a adaptação do currículo escolar para atender às demandas dos estudantes.

3 Metodologia, Técnicas Ferramentas utilizadas

O objetivo deste capítulo é explanar como foi estruturada a exploração, análise e interpretação dos dados. Nesta seção, apresentaremos a metodologia adotada para a condução deste estudo, destacando o processo de Descoberta de Conhecimento em Banco de Dados, ou Knowledge Discovery in Database (KDD) como base de nossa abordagem.

Além disso, exploraremos aspectos relacionados à mineração de dados, a técnicas de classificação e a utilização dos algoritmos de aprendizagem de máquina, como por exemplo, a Árvore de Decisão. Abordaremos ainda, a aplicação da regra de classificação e as ferramentas utilizadas na etapa de análise de dados, como as bibliotecas do Python e o Power BI que desempenham um papel fundamental ao longo desse processo.

3.1 Processo de KDD (Knowledge Discovery in Databases

A Descoberta de Conhecimento em Banco de Dados(KDD) é um processo essencial na era da informação, sendo utilizado para extrair informações valiosas e significativas a partir de conjunto de dados grandes e complexos. Segundo Steiner, "o processo de KDD é um

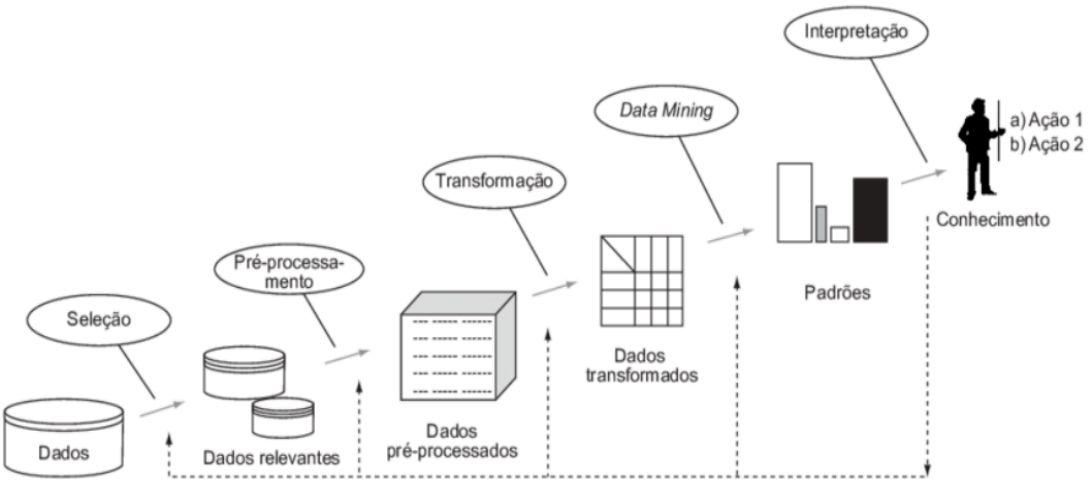
conjunto de atividades contínuas, que compartilham conhecimento descoberto a partir de dados".

Este processo oferece uma estrutura sólida para transformar dados em conhecimento, auxiliando não apenas na análise mas como também na tomada de decisão. Segundo Fayyad, Piatetsky-Shapiro e Smyth(1996), o processo de KDD é um "processo não trivial de identificação de padrões, a partir de dados, que sejam válidos, novos, potencialmente úteis e compreensíveis." Isso quer dizer que esse processo remete a necessidade de seguir várias etapas, para alcançar um objetivo de identificar padrões através da análise de dados.

Desse modo, o processo de KDD é uma abordagem interdisciplinar que envolve a extração, limpeza, transformação, mineração e interpretação dos dados, com o intuito de descobrir padrões e tendências importantes. Ele é utilizado amplamente em ciências de dados, na aprendizagem de máquina e na inteligência artificial para a tomada de decisões, realização de previsões e resolução de problemas complexos. Suas etapas incluem a seleção de dados, pré-processamento, transformação, mineração de dados e interpretação dos resultados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Podemos perceber na figura 3.1, as etapas que compõem o processo de KDD. É importante ter em mente que este processo pode ser dinâmico e contínuo, adaptando-se às mudanças nas condições em que novos dados são disponíveis.

Figura 3.1: Etapas do processo KDD



Fonte: Fayyad et al.(1996)

Nesse contexto, destacamos as seguintes etapas do processo KDD:

1. Seleção dos Dados: A primeira etapa envolve a escolha dos dados relevantes para o problema ou tarefa de descoberta de conhecimento. Isso pode incluir a definição de critérios de seleção e a identificação de fonte de dados. Nesta fase, buscamos fazer o levantamento das principais informações que dificultam no aprendizado dos estudantes do ensino médio, utilizando a base de dados disponibilizado no site do INEP, a fim de selecionar do questionário socioeconômico os dados mais importantes e que impactam na educação, como acesso a recursos educacionais por meio de computador e internet, renda dos pais, etc. A qualidade e representatividade dos dados selecionados são essenciais para o sucesso do processo.
2. Pré-Processamento dos Dados: Na segunda etapa, os dados coletados e selecionados da base são preparados para a análise. Isso inclui a limpeza dos dados a fim de remover valores ausentes ou inconsistentes. Com base nisso, foi feita a mesclagem dos dados dos últimos quatro anos do ENEM, com o intuito de ter uma base mais consistente. Nessa etapa, houve o entendimento da base como algumas informações importantes como a contagem, média dos estudantes, como também a remoção de valores ausentes ou inconsistentes presentes nos dados, mas com todo cuidado para não eliminar dados necessários que podem impactar nos resultados. Além disso, a fim de identificar algumas informações, foi realizado algumas filtragens simples dos dados selecionados tendo como objetivo a obtenção de detalhes. Isso normalizou os valores e garantiu que todas as variáveis tivessem a mesma escala. O pré-processamento é vital para garantir a qualidade dos dados.
3. Transformação dos dados: A terceira etapa se concentra em realçar os padrões e estruturas dos dados. Isso pode envolver a aplicação de técnicas de engenharia de recursos ou redução de dimensionalidade para simplificar a análise. Nesta fase, os dados coletados e pré-processados são submetidos a uma série de operações destinadas a realçar padrões, estruturas e informações relevantes, tornando-os mais adequados para análise. Com isso, os dados foram transformados em um formato mais adequado para a entrada no algoritmo de aprendizagem de máquina. No nosso contexto, as variáveis categóricas foram convertidas em valores numéricos para facilitar no aprendizado do algoritmo. Por exemplo: quando a base de dados do questionário socioeconômico apresenta as opções [A, B, C, D, E] para acesso à internet por parte dos alunos, o algoritmo entenderá as opções como [0, 1, 2, 3, 4], o que torna os dados mais adequados para a mineração.

4. Mineração dos dados: A mineração de dados é o cerne do processo KDD, onde algoritmos de aprendizado de máquina e técnicas de mineração são aplicados aos dados transformados. Nesta fase, os dados preparados e transformados são analisados por meio de técnicas de aprendizagem de máquina para descobrir padrões, tendências, relações e informações valiosas. Dessa forma, foram desenvolvidos gráficos que permitem fazer análise dessas informações importantes, como também o desenvolvimento dos algoritmos de aprendizagem de máquina, e permitindo extrair informações importantes para identificar as principais dificuldades no desempenho dos estudantes de Belo Jardim. Com isso, os algoritmos são treinados usando dados de treinamento. Isso envolve alimentá-lo com exemplos rotulados e permitindo que ele faça previsões do desempenho dos alunos, com o objetivo de descobrir padrões e tendências na educação.
5. Interpretação: É nesta última etapa que ocorre a análise e representação dos resultados obtidos durante as outras etapas. Com isso, podemos identificar os padrões, e informações importantes e analisar os principais influenciadores no desempenho dos estudantes que realizam o Exame Nacional, o que pode auxiliar na tomada de decisão na educação.

3.2 Árvore de Decisão

Uma Árvore de Decisão é uma estrutura hierárquica que se assemelha a uma árvore, composta por nós e arestas. Cada nó na árvore representa uma decisão ou um teste sobre uma variável específica, enquanto as arestas conectam os nós e indicam o fluxo das decisões. O nó superior da árvore é chamado de nó raiz, e os nós folha são as conclusões ou resultados das decisões(LAURETTO, 2010). O processo de construção de uma Árvore de Decisão começa com o nó raiz, que é associado à variável mais importante ou aquela que melhor separa os dados. Em seguida, a árvore se ramifica em nós filhos, representando testes sobre outras variáveis, com base em critérios de divisão. À medida que avançamos na árvore, os ramos se ramificam ainda mais, representando decisões adicionais. Esse processo continua até que se atinja um ponto em que os nós folha contenham informações suficientes para fazer uma previsão ou decisão. Os nós folha representam as classes ou valores previstos(CARVALHO, 2005).

Dessa forma, a árvore de decisão permite uma análise minuciosa dos padrões socioeconômicos dos estudantes em que foi realizado o estudo com base nos dados do ENEM em Belo Jardim. Ao explorar os nós e ramos da árvore, é possível, por meio do algoritmo, identificar quais as variáveis socioeconômicas que têm maiores impactos no desempenho acadêmico, proporcionando uma compreensão mais profunda do contexto local. Os resultados da Árvore de Decisão servem como guia valioso para a tomada de decisão no âmbito educacional. Administradores, educadores e formuladores de políticas podem utilizar essas informações para direcionar recursos de maneira mais eficiente, implementar programas de intervenção direcionados e promover a equidade no sistema educacional local.

3.3 Regra de classificação

A regra de classificação no algoritmo é essencial no processo preditivo, representando a lógica que fundamenta a atribuição de categorias ou valores aos diferentes ramos da árvore. O algoritmo de classificação, por regra, tem a finalidade de encontrar relacionamentos entre os atributos, buscando a predição com base em um novo registro de dados na base.(FILHO, 2017). No contexto do nosso estudo, se aplica na predição do rendimento dos estudantes no ENEM em Belo Jardim, com base nas notas de cada área de conhecimento e o questionário socioeconômico.

Desse modo, a estrutura hierárquica da Árvore de Decisão implica em cada nó, ao longo dos vários níveis, que é associado a uma regra de classificação específica. Com isso, essas regras são derivadas dos dados de treinamento e representam as condições que, quando satisfeitas, guiam a previsão do desempenho de um estudante. A compreensão dessa regra proporciona uma visão clara das condições que levam a diferentes destinos na árvore(FILHO, 2017).

Com isso, a regra é representada pelo tipo SE-ENTÃO, que é uma estrutura lógica que encapsula as condições sob as quais uma determinada predição é realizada. Essa representação é fundamental para interpretar e aplicar efetivamente os resultados do algoritmo. Na árvore de decisão, cada nó é associado a uma condição específica pela parte "SE", com base em uma variável específica, que levam à ramificação da árvore. Já a parte "ENTÃO"da regra indica a conclusão ou a classificação que é atribuída se a condição especificada no "SE", for satisfeita. Então a conclusão pode ser a previsão de uma classe,

um valor numérico ou qualquer resultado específico ao qual a árvore está sendo treinada para realizar a predição dos dados(FILHO, 2017).

No nosso estudo, fizemos a regra de classificação com o intuito de identificar os perfis dos estudantes e relacionar o desempenho com as condições sociais dos participantes.

Objetivando isto, foi realizada uma regra com base em 5 classificações: Satisfatório (Condição Social Alta com nota maior do que 500 possuindo todos o recursos para uma preparação melhor), Desempenho Satisfatório (Condição Social Média com nota também maior do 500 mas com uma renda familiar boa apresentando vários dos benefícios sociais), Desempenho Satisfatório(Condição social baixa com nota média maior do que 500), Desempenho Insatisfatório(condição social alta mas com nota inferior a 500 e com falta de recursos para preparação no exame), Desempenho Insatisfatório(condição social média com nota menor do que 500 com alguns dos benefícios), Desempenho Insatisfatório(condição social baixa com nota média menor do que 500 e quase nenhum recurso para preparação no ENEM) e participantes com outro perfil.

3.4 Ferramentas utilizadas

No nosso estudo utilizamos alguns recursos bastantes importantes para analisar e predizer o desempenho dos estudantes que realizaram o ENEM. Dentre essas ferramentas, destacamos o Python contendo suas bibliotecas especializadas e o Power BI para a criação de gráficos e visualizações importantes. Com isso, utilizamos algumas bibliotecas do python para a visualização dos dados como:

- Pandas: O Pandas foi fundamental para a manipulação eficiente dos conjuntos de dados. Seu poderoso conjunto de estruturas de dados e funções facilitou a limpeza, filtragem e organização dos dados, preparando-os para a etapa de visualização.
- Seaborn: Construído sobre o Matplotlib, trouxe uma camada de abstração adicional para a criação de gráficos estatísticos atraentes. Sua sintaxe simples e recursos avançados permitiram a geração rápida de gráficos informativos, adicionando uma dimensão visual valiosa à nossa análise.

- Matplotlib: Ferramenta robusta e versátil, foi empregada para criar gráficos estáticos detalhados. Sua flexibilidade permitiu a personalização minuciosa dos gráficos, adaptando-os às necessidades específicas de nossa análise.
- NumPy: Contribuiu para manipulações numéricas eficientes, sendo essencial para realizar operações matemáticas sobre os dados. Sua integração perfeita com o Pandas e outras bibliotecas de análise de dados foi crucial para a coerência e eficácia do processo.
- Plotly Express: A inclusão do Plotly Express trouxe uma dimensão interativa à nossa visualização. Sua capacidade de gerar gráficos interativos e painéis dinâmicos enriqueceu na ilustração, permitindo uma exploração mais aprofundada dos dados por meio de interfaces dinâmicas.

Utilizamos também o Power BI que é uma ferramenta que possui vários gráficos integrados que permitem também uma visualização dinâmica, com filtros de informações importantes para análise.

Na implementação do algoritmo de Árvore de Decisão para a predição do desempenho estudantil no ENEM, recorremos a bibliotecas especializadas em aprendizado de máquina fornecidas pelo pacote Scikit-learn. Cada biblioteca desempenhou um papel crucial na construção, avaliação e interpretação do modelo preditivo.

- DecisionTreeClassifier (Scikit-learn): O DecisionTreeClassifier do Scikit-learn foi a espinha dorsal da nossa implementação. Essa classe permite a construção de uma Árvore de Decisão para tarefas de classificação. Configuramos parâmetros como critério de divisão e profundidade da árvore para otimizar o desempenho do modelo.
- Train-test-split (Scikit-learn): A função train-test-split do Scikit-learn foi essencial para dividir nosso conjunto de dados em conjuntos de treinamento e teste. Isso permitiu avaliar o desempenho do modelo em dados não utilizados durante o treinamento, garantindo uma avaliação mais realista.
- Confusion-matrix, accuracy-score, precision-score, recall-score (Scikit-learn): As funções para avaliação de desempenho, como confusion-matrix, accuracy-score, precision score e recall-score, foram utilizadas para analisar o quão bem o modelo de Árvore de Decisão estava se saindo. Essas métricas ofereceram insights sobre a precisão, recall e acurácia do modelo, além de fornecer uma matriz de confusão para uma compreensão mais detalhada dos resultados.

Essas bibliotecas do Scikit-learn proporcionaram uma base sólida para o desenvolvimento e avaliação do modelo preditivo. A combinação dessas ferramentas permitiu uma implementação eficaz da Árvore de Decisão, contribuindo para a compreensão do desempenho dos estudantes com base em dados socioeconômicos e notas do ENEM.

4 Resultados

Nesta seção, apresentamos os resultados da nossa análise que visa a compreensão do desempenho dos estudantes de Belo Jardim com base em dados do Exame Nacional do Ensino Médio(ENEM), realizado por estudantes do ensino médio com o intuito de ingressarem nas universidades, bem como a criação de um dashboard interativo no Power BI e gráficos desenvolvidos com a linguagem de programação Python, além de um algoritmo de machine learning para prever esse desempenho. O objetivo central deste estudo é explorar as relações entre as notas obtidas no ENEM e o sucesso acadêmico dos estudantes, considerando também fatores socioeconômicos importantes para o desempenho estudantil.

Ao longo deste capítulo, forneceremos uma visão detalhada dos resultados obtidos, destacando as principais conclusões e insights extraídos de nossa análise. Inicialmente apresentando os dados utilizados em nossa pesquisa, incluindo suas fontes e a preparação que foi realizada a fim de torná-la adequada para análise (HAN; KAMBER; PEI, 2012). Em seguida, discutiremos as características do dashboard desenvolvido no Power Bi e dos gráficos gerados em Python, demonstrando como essas ferramentas contribuíram para a visualização clara e acessível dos dados. Além disso, apresentaremos os resultados da predição do desempenho dos estudantes de Belo Jardim.

4.1 Base de Dados

Os dados para realizar o estudo sobre o desenvolvimento escolar dos estudantes foi utilizado com os Microdados do ENEM, disponibilizados no portal do INEP. A base de dados conta com variáveis relativas dos estudantes que realizaram a prova, sendo algumas delas relacionadas às notas obtidas em cada área de conhecimento, sendo elas, Linguagens, Códigos, Matemática, Ciências Humanas e Ciências da Natureza e suas tecnologias. Esta

base de dados é um recurso valioso para pesquisas relacionadas ao desempenho dos estudantes e aos fatores que o influenciam.

Os Microdados do ENEM são disponibilizados em formato de arquivo CSV e trazem informações sobre notas obtidas pelos alunos, como também, dados socioeconômicos relacionados aos mesmos. Com isso, foi realizada a limpeza e preparação com o objetivo de ter eficiência na análise e predição dos resultados. Para a realização dessa etapa, foi utilizado o Python, que foi responsável por organizar os dados em formato em planilha. Os arquivos estão divididos em 6 pastas, sendo elas: Dados, Dicionário, Inputs, Leia-me, Planilhas e Provas/Gabaritos.

Dessa forma, foi utilizado várias informações dessa base de dados dos últimos quatro anos de realização do exame com o intuito de identificar e analisar os diferentes aspectos que podem influenciar o desempenho acadêmico dos estudantes de Belo Jardim. Sendo assim, algumas variáveis foram usadas na análise:

1. Análise de Notas por Tipo de Escola: Uma das variáveis-chave desta pesquisa é a análise das notas dos estudantes em relação ao tipo de escola em que estudam. Essa variável nos permite investigar se existem diferenças significativas no desempenho entre estudantes de escolas públicas e privadas. Essa análise ajuda a compreender a influência do ambiente escolar na formação educacional dos estudantes.
2. Dependência Administrativa da Escola: A variável que descreve a dependência administrativa das escolas, categorizada em federal, estadual, municipal e privada, é essencial para avaliar como a gestão escolar pode afetar o desempenho acadêmico dos estudantes. Ela nos permite explorar as diferenças de recursos e abordagens pedagógicas entre esses tipos de instituições.
3. Nível de Educação dos Pais: A educação dos pais desempenha um papel crucial na formação dos estudantes. Essa variável nos permitirá analisar como o nível de escolaridade dos pais pode impactar o desempenho acadêmico dos estudantes.
4. Acesso a Tecnologia: As variáveis "Na sua residência tem telefone celular"e "Na sua residência tem computador"fornecem insights sobre o acesso dos estudantes a tecnologias que podem influenciar seu processo de aprendizagem. A disponibilidade desses recursos pode ser um fator importante na qualidade da educação.
5. Renda Familiar: A renda familiar é um indicador significativo do contexto socioeconômico dos estudantes. Esta variável nos permitirá explorar como a renda da família está associada ao desempenho acadêmico dos estudantes.

6. Grupo de Trabalho dos Pais: A variável relacionada ao grupo de trabalho dos pais nos permitirá entender o ambiente familiar dos estudantes e como a situação de trabalho dos pais pode influenciar seu desempenho na escola.
7. Localização da Escola: A variável que descreve a localização da escola (zona rural ou urbana) nos ajudará a entender como o ambiente escolar impacta o na aprendizagem dos alunos.

4.2 Análise Exploratória dos Dados

Nesta etapa fundamental da nossa pesquisa, apresentamos nossas análises detalhadas, conduzidas com base nos microdados do ENEM, enriquecidas por ferramentas poderosas e intuitivas de visualização de dados. A análise de dados desempenha um papel central na compreensão das informações permitindo extrair insights significativos que envolvem o desempenho dos estudantes na educação.

Nosso objetivo ao realizar esta análise é desvendar os padrões, as relações e as informações cruciais que residem nos dados, e assim, proporcionar uma visão mais detalhada de fatores que dificultam a aprendizagem dos alunos de Belo Jardim. Os gráficos e visualizações neste capítulo não são apenas representações visuais, mas sim, janelas para o entendimento profundo do nosso objeto de estudo.

4.2.1 Exploração visual dos dados com o Power BI

Nessa seção da nossa pesquisa, iremos aprofundar nos dados preparados, explorando sua riqueza, por meio de análises detalhadas e visualizações dos relatórios com base nos dados do ENEM. Permitindo, assim, a compreensão dos desafios que envolvem o desempenho dos participantes.

O Power BI é uma ferramenta importante e interativa na visualização de dados e nosso aliado na criação de um panorama interativo das informações coletadas. Gartner, 2020, em seu estudo sobre visualização de dados, enfatiza que ferramentas como o Power BI proporcionam a capacidade de mergulhar profundamente nos dados, destacar tendências e identificar correlações. Por meio de gráficos dinâmicos e painéis intuitivos, exploraremos as principais variáveis que impactam no desempenho dos alunos. Ele oferece a capacidade de mergulhar profundamente nos dados, destacar tendências, comparar variáveis e

identificar correlações. Além disso, em seu estudo destaca como um dos software com melhor custo benefício. Ainda Segundo Gartner (2020), "além do preço, as capacidades analíticas do Power BI com a facilidade de criação de relatórios e outras tecnologias embutidas, como Machine Learning e serviços de Inteligência Artificial, expandiu ainda mais as suas capacidades após a integração nativa com serviços Azure". Essa abordagem interativa facilita na identificação de conclusões valiosas que podem ajudar a melhorar na análise dos dados do exame realizado no ensino médio.

Dessa forma, desenvolvemos um dashboard com o propósito de oferecer insights valiosos com base nos dados do ENEM. Nossa intenção é contribuir para a melhoria da educação, identificando tendências e desafios significativos no cenário educacional de Belo Jardim. Ele proporciona uma visão abrangente de informações, podendo auxiliar professores e gestores nas tomadas de decisões e na implementação de práticas pedagógicas mais eficazes, direcionadas para o sucesso dos estudantes.

Figura 4.1: Tela inicial do relatório educacional no Power BI

Fonte: Felipe Monteiro, Alysson Assunção, 2023

Na Figura 4.1, apresentamos a tela inicial do nosso relatório de análise do cenário educacional com o total de 1.397 estudantes que realizaram o exame em Belo Jardim PE nos últimos anos do exame, que está dividido em duas partes distintas. A primeira parte exibe um relatório abrangente que compila informações dos estudantes com base em seus dados socioeconômicos. A segunda parte se concentra nas informações relacionadas à situação da pandemia causada pelo COVID-19 relacionado ao ano de 2021.

Desse modo, na exibição apresentado na figura 4.2, para podermos acompanhar as variáveis no gráfico, o INEP disponibiliza em um arquivo, sendo um dicionário informando sobre cada questão no formulário coletado, em que podemos ver o que significa cada uma das variáveis. Com isso, podemos perceber nos gráficos relacionados aos estudantes a relação ao grupo pertencente aos pais, se em sua residência tem algum computador e celular, como também acesso a internet e sobre a renda familiar representado no gráfico por profissão. Esses gráficos ajudam a identificar em que grupo cada aluno está e qual dificuldade pode apresentar ao realizar o exame necessitando de recursos para um melhor desempenho no exame. Além disso, podemos filtrar as infomações de acordo com os anos (2019, 2020, 2021 e 2022), sexo ou por tipo de escola, permitindo analisar por escola pública ou privada. Também é possível analisar a faixa etária dos candidatos.

Com base nisso, podemos analisar no primeiro gráfico abaixo, qual profissão é exercida pelas mães dos participantes sendo o grupo B, formado por diaristas, empregadas domésticas, cuidadoras de idosos, babás, cozinheiras (em casas particulares), jardineiras, faxineiras de empresas e prédios, vigilantes, vendedoras, atendentes de loja, auxiliares administrativos, recepcionistas e repositoras de mercadoria. Juntas, estas compõem um total de 567. Já o grupo D, composto por 267 pessoas, é composto por professoras, técnicas, policiais militares, corretoras de imóveis, supervisoras, gerentes, etc.

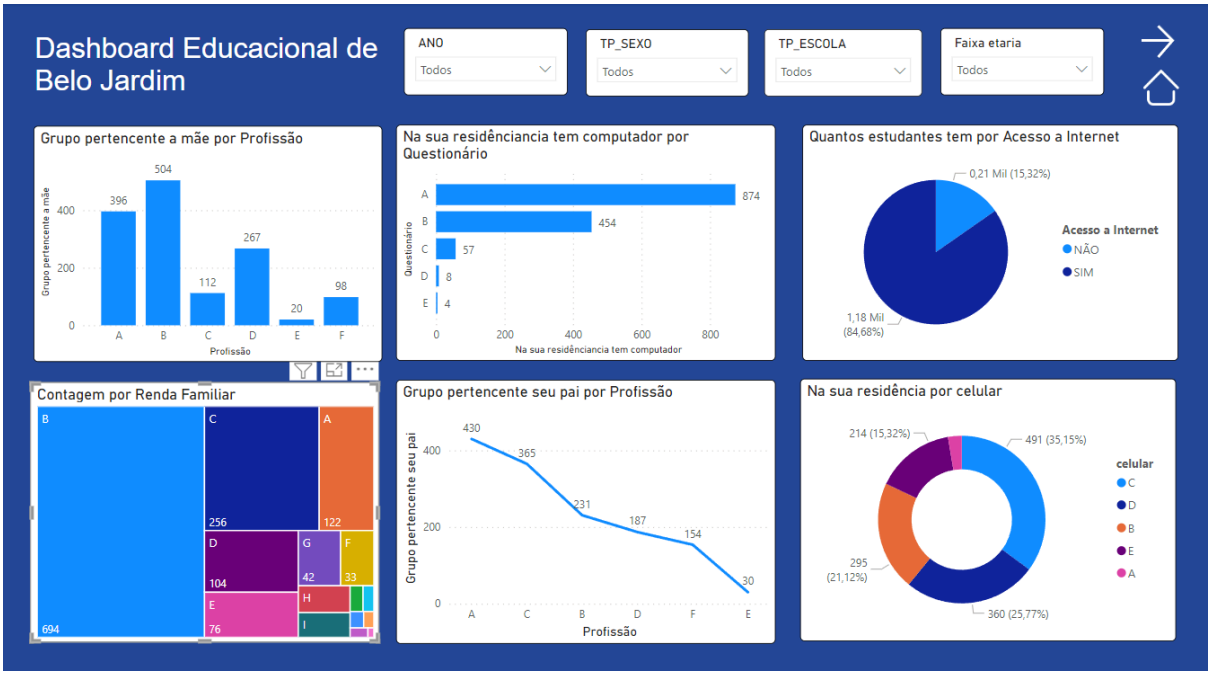
Há também um grande número de mulheres identificadas pelo grupo A, com 396 sendo composto por agricultoras e junto com atividades da pecuária. Já nos demais, conforme mostra o gráfico temos um número mais reduzido, pois apresentam mais estudo e consequentemente têm profissões como grupo C para trabalhos em empresas, e no grupo D, sendo professoras com 267 e grupo E com 20 médicas, engenheiras, além dos alunos que não souberam informar com total de 98.

Já em relação à profissão exercida pelos pais, temos o gráfico de linha, em que o grupo A é o maior grupo e contém 430 compostos por agricultores e atividades relacionadas à pecuária. O grupo C com 365 conta com caminhoneiros, taxistas e operários de fábrica. No grupo D podemos observar uma predominância de quem teve maior grau de estudos,

com 187 sendo professores, policiais e microempresários. O grupo F conta 154 alunos, sendo aqueles que não informaram nenhum dado relacionado. Por fim, temos o grupo E com 30 sendo médicos e engenheiros.

Podemos observar no outro gráfico, a renda familiar classificada em grupo A como nenhuma renda, grupo B até R$1.212, grupo C de R$ 1.212,01 até R$ 1.818,00, grupo D de R$ 1.818,01 até R$ 2.424,00, grupo E de R$ 2.424,01 até R$ 3.030,00, grupo F de R$ 3.030,01 até R$ 3.636,00, grupo G de R$ 3.636,01 até R$ 4.848,00, entre outros. Com isso, podemos perceber que o grupo B tem em sua maioria com 694 apresentado uma renda menor ou igual a um salário mínimo, tendo também com o grupo A 122, que não tem nenhuma renda que são pessoas que trabalham na agricultura, e tem grupo C com 256, grupo D com 104, e os outros apresentam baixo valor, por serem já considerados com boa condição de vida.

Figura 4.2: Relatório Educacional de Belo Jardim



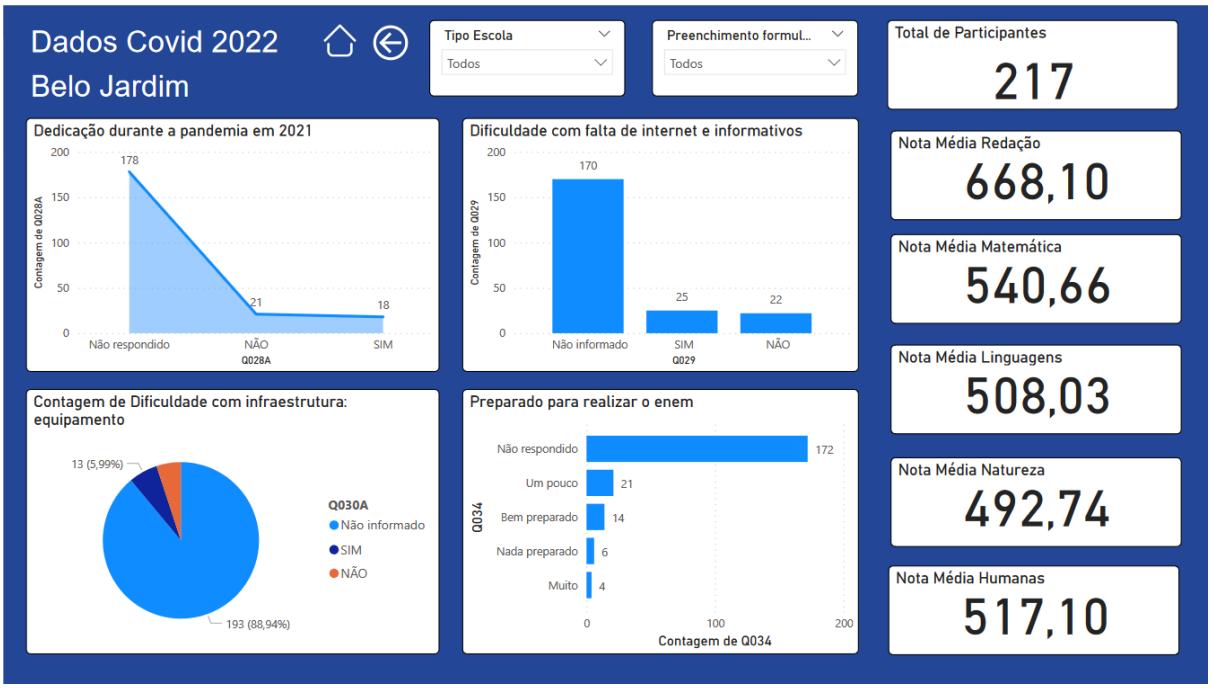
Fonte: Felipe Monteiro, Alysson Assunção, 2023

Nesse contexto, abordamos o acesso a computadores, celulares e internet como fatores determinantes no apoio ao aprendizado dos estudantes. No que diz respeito ao acesso a computadores, o gráfico revela que a maioria dos participantes pertencem ao grupo A, com 874 (62,75%) relatando a ausência de computador em suas residências, enquanto o grupo B, representando aqueles com pelo menos um computador para fins de estudo,

totaliza 454 (32,61%). Em relação ao acesso a dispositivos móveis, notamos que a grande maioria dos estudantes possui um aparelho celular que pode contribuir para a preparação para o exame. O grupo A, composto por aqueles que não possuem aparelho celular, é representado por apenas 37 (2,65%) dos participantes. No grupo B, temos 295 (21,12%) que possuem pelo menos um aparelho celular, e no grupo C, 491 (35,15%) relataram ter mais de um celular. Quanto ao acesso à internet, 1.183 (84,68%) dos participantes responderam afirmativamente, indicando que tinham acesso à internet. No entanto, 214 (15,32%) relataram não ter acesso. Conforme o gráfico, aqueles com acesso à internet e dedicação, desfrutaram de vantagens na utilização de recursos contemporâneos para estudos, proporcionando maior facilidade na preparação para o ENEM.

Na segunda página do relatório, conforme ilustrado na Figura 3, apresentamos informações relacionadas à pandemia do COVID-19, que teve início em 2020. Neste contexto, na base foi o total de 217 estudantes que participaram do ENEM em 2021. Através desses dados, elaboramos diversos gráficos que abordam a dedicação e preparação dos estudantes para o exame nacional, bem como as dificuldades enfrentadas devido à falta de acesso à internet, disponibilidade de informações, infraestrutura e equipamentos para estudo. Além disso, analisamos a média das notas obtidas pelos estudantes em cada área de conhecimento, incluindo Redação, Matemática, Linguagens, Ciências da Natureza e Ciências Humanas. Essa análise nos permite identificar claramente o impacto da pandemia no desempenho dos alunos. Ela revela as adversidades enfrentadas pelos estudantes e como esses desafios influenciaram suas notas no ENEM. Essas informações são essenciais para entender o contexto educacional em um período de turbulência global e para ajudar a informar estratégias futuras na educação.

Figura 4.3: Relatório Educacional de Belo Jardim Covid



Fonte: Felipe Monteiro, Alysson Assunção, 2023

Porém, como podemos perceber, vários estudantes optaram por não responder algumas das perguntas, que poderiam ajudar melhor no entendimento em relação ao contexto da pandemia no cenário educacional. Por isso, vamos analisar somente uma parte desses dados que são essenciais para análise. Em relação às notas se mantiveram ideal, obtendo no ano de 2021 uma média maior em redação. E com base nisso, podemos analisar sobre a preparação e dedicação durante o ano de pandemia, em que ficou na média de 50% dos que responderam que tiveram algumas dificuldades. Já em questão a falta de informações também ficou nessa faixa entre os participantes, e com relação a infraestrutura podemos perceber que poucos sentiram dificuldades.

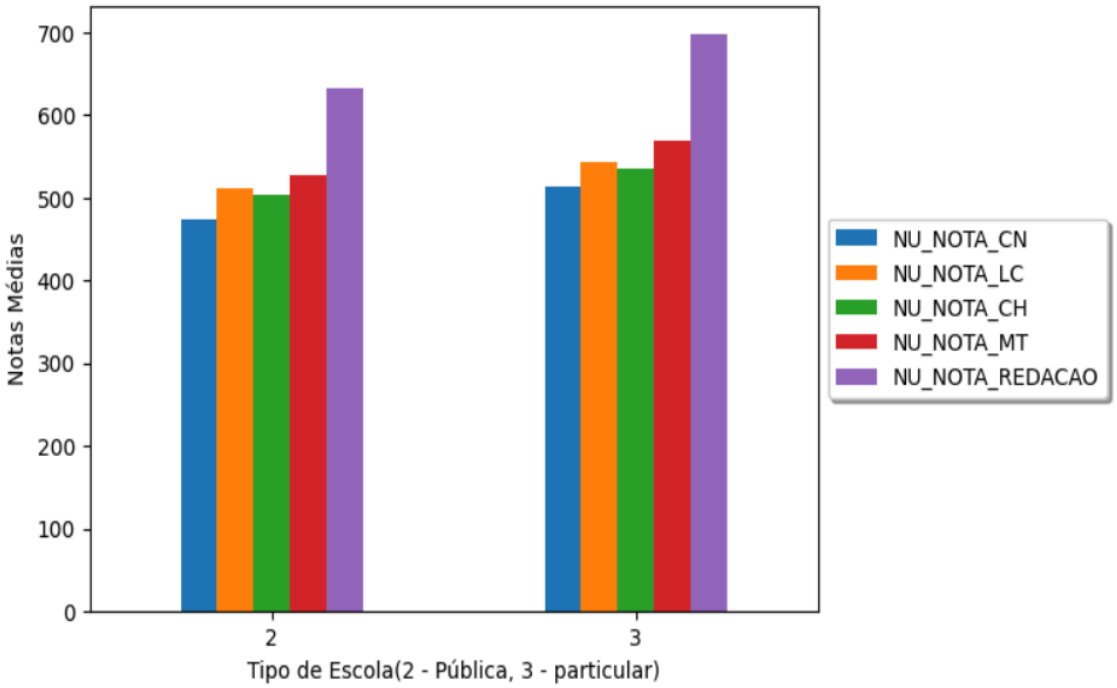
4.2.2 Análise Visual com Python

Os gráficos e visualizações que vamos apresentar são ferramentas que comunicam, esclarecem e inspiram, permitindo assim, em cada visualização insight valioso nos resultados, com objetivo de descobrir os padrões, as relações e as conclusões que emergem dos dados.

O primeiro gráfico de barras na figura 4.4, apresenta uma análise comparativa na evolução dos estudantes em escolas públicas e particulares com relação à média nas

diferentes áreas de conhecimento avaliadas no ENEM, sendo elas Ciências Humanas, Linguagens, Ciências da Natureza, Matemática e suas tecnologias e Redação.

Figura 4.4: Médias das notas por tipo escola



Fonte: Felipe Monteiro, Alysson Assunção, 2023

Observamos que, em todas as áreas de conhecimento, os estudantes de escola particular tendem a ter um desempenho superior em relação aos de escolas públicas, como podemos visualizar no gráfico. Além disso, podemos perceber que na prova de redação essa diferença é bastante superior entre essas duas escolas, sendo que essa competência é a de maior peso para os estudantes que querem ingressar nas universidades.

Isso está relacionado a vários fatores que podem impactar no desempenho dos estudantes de escola pública, apesar de nos últimos anos ter melhorado bastante na educação pública, porém ainda podemos perceber discrepâncias, incluindo recursos educacionais, qualificação de professores, ambiente de aprendizado e apoio familiar.

No gráfico de linha da figura 4.5, representa a média das notas nas diversas áreas de conhecimento, categorizadas por dependência administrativa das escolas identificados como: 1 - Federal, 2 - Estadual, 3 - Municipal e 4 - Privada, como podemos perceber na imagem representando esta análise. Com base nisso, torna-se evidente que, as notas dos estudantes de escolas particulares mantêm-se consistentemente superiores em comparação com as notas das escolas estaduais e federais.

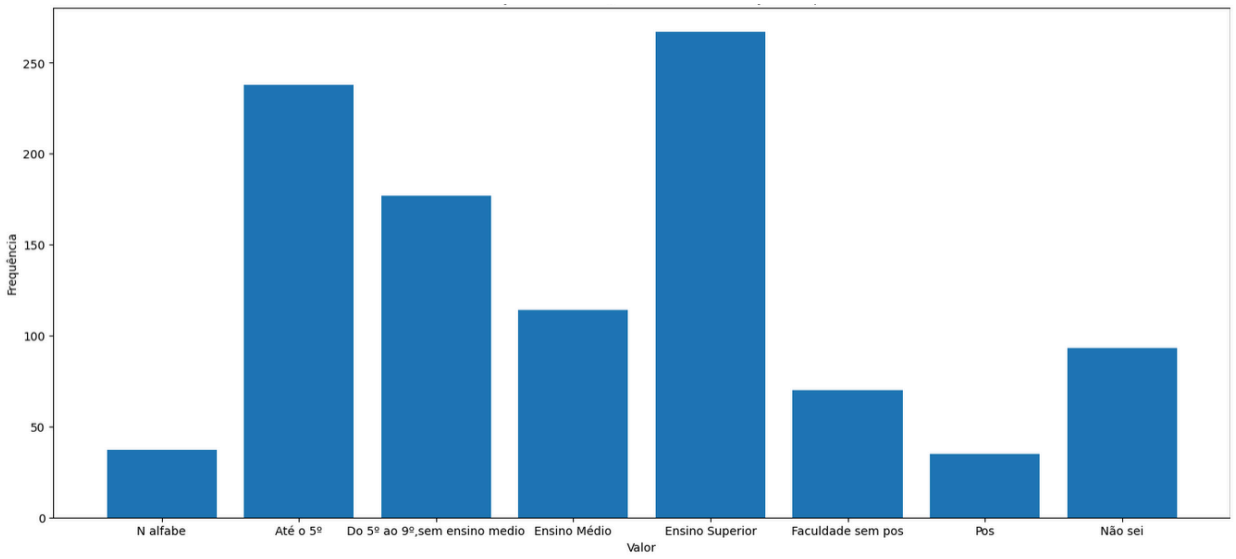
Figura 4.5: Médias das notas por competência e dependência administrativa

Fonte: Felipe Monteiro, Alysson Assunção, 2023

No entanto, é interessante observar que, em algumas áreas de conhecimento, as escolas estaduais e federais apresentam desempenho similar, em algum momento a escola estadual tem um valor um pouco superior a escola estadual, como também acontece com a escola federal, sendo superior em relação a escola estadual. Isso sugere que, embora haja uma clara diferença de desempenho em favor das escolas particulares, as outras duas acabam se equiparando em termos de notas obtidas pelos estudantes que realizaram o ENEM.

Dessa forma, podemos observar no gráfico da Figura 4.6, uma análise da educação dos pais dos estudantes, com categorias que incluem: Não alfabetizado, estudaram até o 5º ano, do 5º ao 9º ano do ensino fundamental, Ensino Médio completo, Ensino Superior, Pós-Graduação e aqueles que não souberam informar a escolaridade dos pais. Com isso, permite um entendimento acerca da educação dos pais que fornece uma visão abrangente das influências e desafios educacionais que podem impactar também na performance dos estudantes que têm pouca condição de vida.

Figura 4.6: Gráfico representando quantidade relacionado a educação dos pais



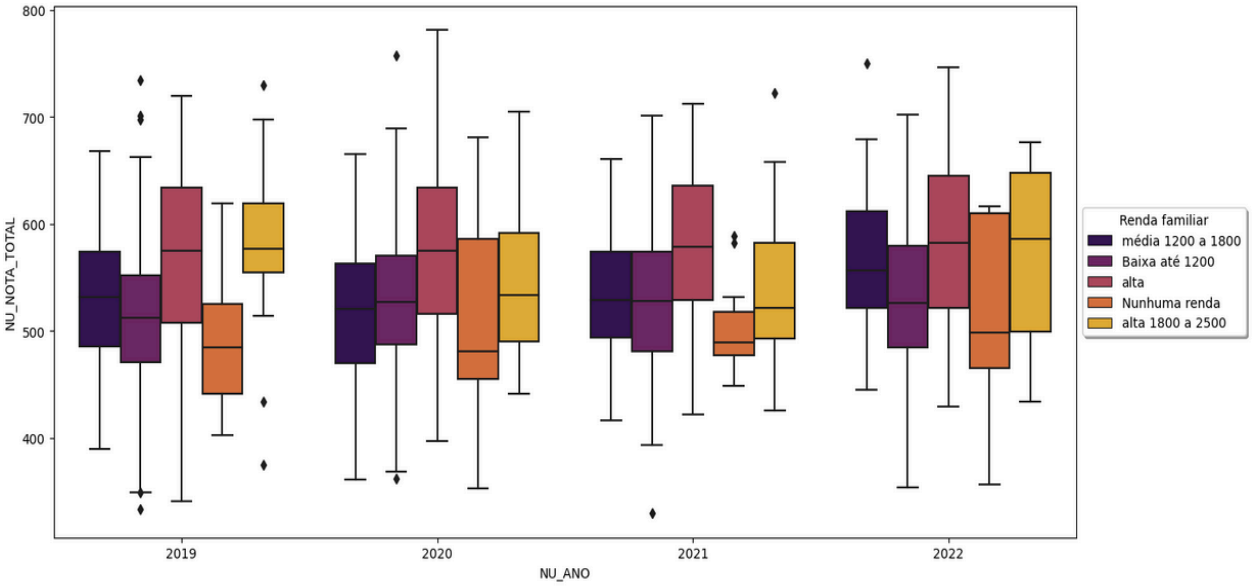
Fonte: Felipe Monteiro, Alysson Assunção, 2023

Nesse contexto, o gráfico de barras revela a distribuição dos níveis de escolaridade dos pais. Nota-se que a maioria deles possuem formação no ensino superior, conforme evidenciado no gráfico. No entanto, é igualmente evidente a presença de vários estudantes cujos pais têm apenas escolaridade até o 5º ano. Essa disparidade ressalta a coexistência de estudantes com antecedentes educacionais variados. Também podemos observar que os pais que têm ensino superior ou nível mais avançado em estudo conseguem proporcionar melhores condições e consequentemente adquirir melhores recursos para aprendizagem dos seus filhos.

Essa diversidade na escolaridade dos pais pode ter implicações significativas nas condições de vida dos estudantes, bem como em seu desempenho no exame. Alguns podem enfrentar desafios na aprendizagem devido à falta de recursos para a educação. Outros podem se deparar com a necessidade de ajudar a sustentar a renda familiar, levando até na desistência dos estudos. Portanto, a análise dos níveis de escolaridade dos pais revela uma realidade complexa que exige considerações cuidadosas para abordar as necessidades educacionais e as condições de vida dos estudantes de forma mais abrangente e equitativa.

No gráfico da figura 4.7, podemos visualizar a representação dos dados de forma mais nítida sobre a relação entre a nota obtida no exame e renda da família nos anos de 2019, 2020, 2021 e 2022. Esse gráfico mostra o quanto o desempenho dos estudantes pode variar, dependendo da situação financeira da sua família, e que influencia bastante na nota obtida no exame, e o que enfatiza, a relevância dessas análises, quando se busca entender e comparar, de fato, qual a situação da renda familiar dos candidatos.

Figura 4.7: Médias das notas pela renda familiar dos estudantes



Fonte: Felipe Monteiro, Alysson Assunção, 2023

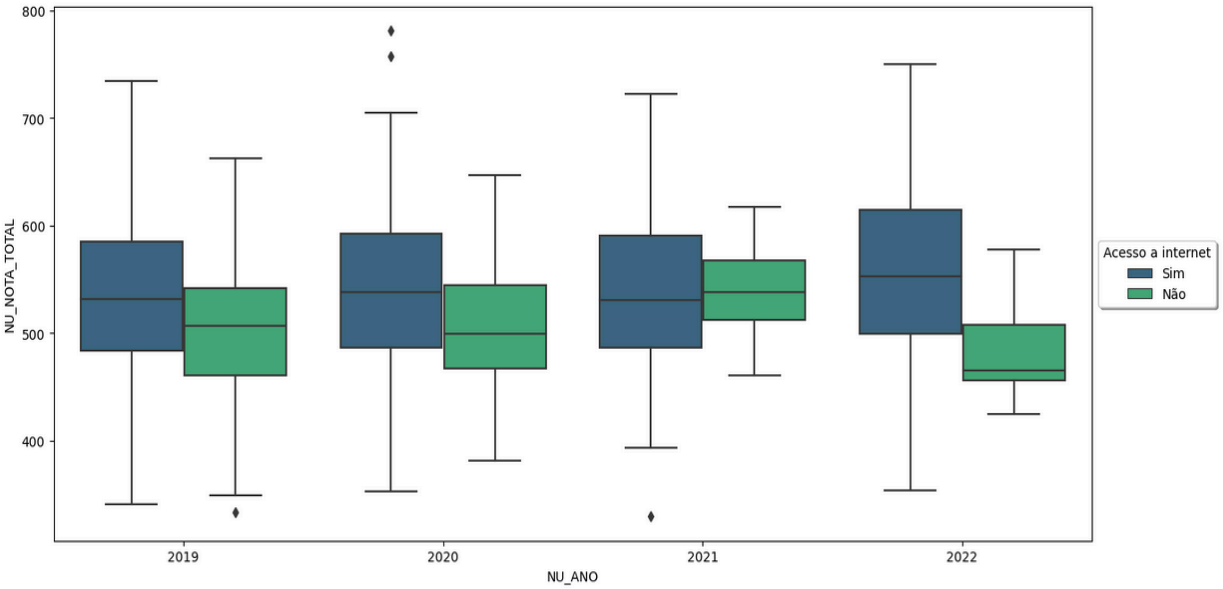
No gráfico apresentado, analisamos a classificação da renda familiar dos estudantes com base na média geral obtida por eles. A classificação considera diferentes faixas de renda: ’Nenhuma Renda’, ’Baixa Renda’(com valor até 1200), ’Média Renda’(de 1200 a 1800) e ’Alta Renda’(com valores superiores).

A análise do gráfico revela que em 2019, os estudantes classificados como ’Nenhuma Renda’ obtiveram notas mais baixas em comparação com os demais grupos. No entanto, em 2020, observamos um equilíbrio nas notas entre as diferentes categorias. Nos anos subsequentes, as notas tendem a superar em relação aos outros, com exceção de 2021, quando há uma redução em comparação às outras categorias, vale destacar que foi nesse tempo que ocorreu a pandemia e causando consequências para o ensino na cidade de belo jardim e principalmente para aqueles alunos com dificuldades financeiras.

No gráfico da Figura 4.8, também é apresentado um gráfico boxplot que facilita a comparação e a visualização das diversas pontuações obtidas pelos estudantes nos últimos anos do ENEM, levando em conta se tiveram ou não acesso à internet. Esse gráfico possibilita a análise dos valores medianos, que revelam pontuações médias entre 500 e 600 em todas as áreas de conhecimento. Além disso, é evidente que os participantes sem acesso à

internet tendem a obter notas inferiores em comparação com aqueles que têm acesso. Isso se deve ao fato de que enfrentam mais dificuldades no acesso à informação.

Figura 4.8: Médias das notas pelo acesso a internet

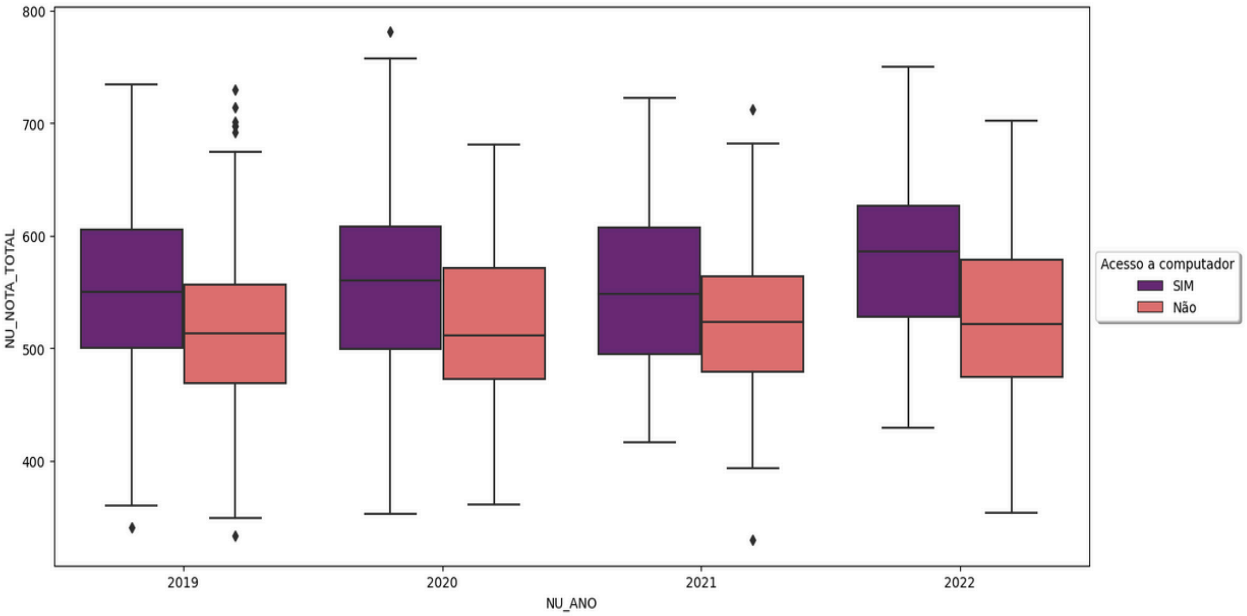


Fonte: Felipe Monteiro, Alysson Assunção, 2023

Além das diferenças nas notas médias entre esse grupo de estudantes, o gráfico também nos permite observar a variação das pontuações. Com isso, os alunos que têm acesso a internet tem vantagem de poderem acessar uma gama de recursos educacionais online, como materiais de estudo, videoaulas e tutoriais, facilitando assim na aprendizagem. Isso pode contribuir bastante para um melhor desempenho nas provas, refletido nas notas mais altas. Por isso, esta análise demonstra a importância de considerar o impacto do acesso à internet, apesar de ter melhorado no acesso à conectividade nos últimos anos, mas ainda diversos estudantes enfrentam desafios significativos na obtenção de informações e recursos de aprendizado.

Neste segundo gráfico, representado na Figura 4.9, é realizada uma análise sobre o acesso a computador em casa sobre as notas médias, sendo um fator de grande importância. Pois, esse acesso desempenha um papel crucial no aprendizado dos alunos, pois facilita a inclusão digital e realização de trabalhos escolares.

Figura 4.9: Médias das notas dos estudantes que têm acesso a computador em casa



Fonte: Felipe Monteiro, Alysson Assunção, 2023

Com isso, podemos observar que os estudantes que também possuem acesso a computador apresentam um desempenho superior comparado aos demais que não possuem computador em casa, permitindo maior facilidade no aprendizado, permitindo assim, aos alunos explorar recursos educacionais online, realizar pesquisas, acessar materiais de estudo e a aprimorar suas habilidades de computação. Realizar simulados de provas anteriores do ENEM, permitindo um treinamento antes de fazer o exame. Com isso, esse diferencial possibilita, além de um desenvolvimento de competências digitais, habilidades cada vez mais cruciais no cenário educacional atual.

4.3 Resultados da Árvore de Decisão

Nessa etapa, iremos mergulhar nos resultados obtidos por meio da implementação do algoritmo de aprendizagem de máquina chamado Árvore de Decisão, sobre o desempenho estudantil de Belo Jardim, com base nos dados do ENEM. A combinação refinada desses dados é composta pelo questionário socioeconômico e notas obtidas pelos estudantes, permitindo explorar padrões e identificar também fatores que podem influenciar no desempenho.

tempo em atividades diárias das pessoas com TEA. Esses elementos complementam os resultados quantitativos, proporcionando uma compreensão mais abrangente do contexto em que os requisitos para o aplicativo de mapa de rotina são fundamentados.

Dessa forma, com regra definindo a predição de acordo com as informações relacionadas a média das notas obtidas em cada área de conhecimento, e de acordo com dados sobre a educação dos pais, a qual grupo pertencem e educação. Além disso, utilizamos informações sobre recursos para a preparação para o exame, como acesso a internet e computador em sua residência. Com base nisso, foi criado uma nova coluna chamada "Desempenho Satisfatório", sendo que cada regra irá definir com base na condição social, baixa, média e alta, em relação a nota média maior que 500, considerado desempenho satisfatório e valor menor um desempenho insatisfatório. Dessa forma, foi feito com um total de 1397 estudantes que realizaram o ENEM nos últimos quatro anos de ENEM, sendo 2019, 2020, 2021 e 2022.

Na tabela 4.1 podemos observar as classificações proporcionadas pelo algoritmo, oferecendo insights importantes sobre quais aspectos das notas do ENEM e o questionário mais influenciam no desempenho.

Tabela 4.1: Classificação de Desempenho da Árvore de Decisão

Classificação(Desempenho)	**Quantidade(Estudantes)**
Satisfatório(condição social alta)	37
Satisfatório(condição social média)	1212
Satisfatório(condição social baixa)	3
Insatisfatório(condição social alta)	8
Insatisfatório(condição social média)	7
Insatisfatório(condição social baixa)	3
Outro perfil	128
Total	**1397**

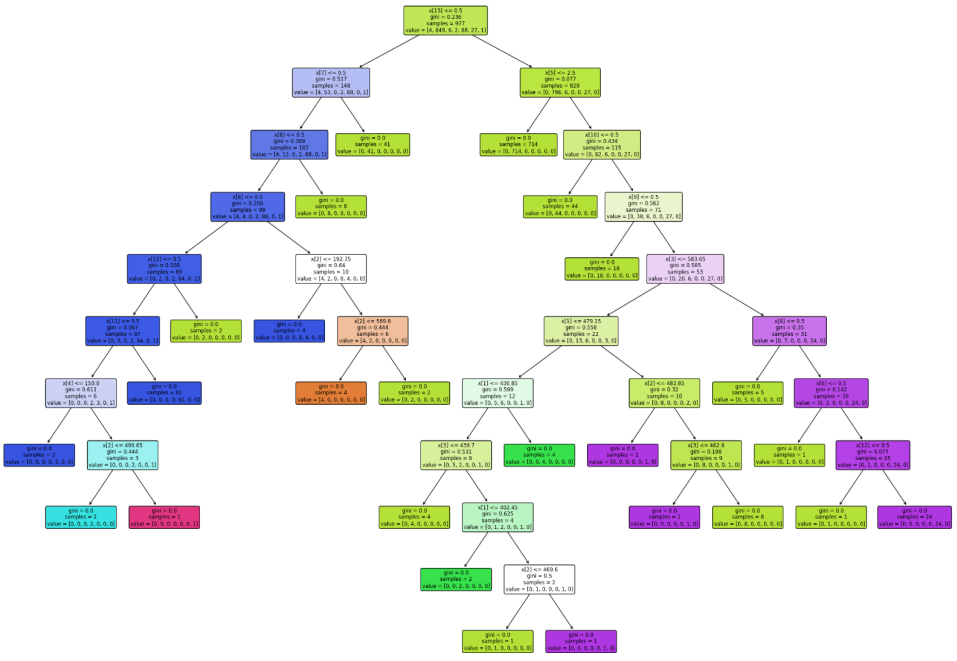
Fonte: Felipe Monteiro, Alysson Assunção, 2023

Dessa maneira, podemos observar em que de acordo com a regra de classificação em que prevalece mais uma quantidade maior de estudantes com desempenho satisfatório com condição social média com 1212, em que conforme a regra de classificação pertencem

a rede privada de ensino, os pais têm ensino médio completo, grupo pertence a professores, gerente e etc, a renda familiar considerada boa, já com relação às outras classificações temos mais poucos em que podemos observar com satisfatório com condição alta com 37, satisfatório condição baixa 3, e com resultados insatisfatório temos com condição alta 8, média 7 e baixa 3.

Com base nisso, desenvolvemos o algoritmo com base nessa regra, em que adicionamos para treinamento e aprendizagem 30% dos dados e assim podemos observar a ilustração da árvore na figura 4.10, cada nó representado a nota média das áreas de conhecimento associado ao questionário socioeconômico dos participantes.

Figura 4.10: Árvore de Decisão do desempenho dos estudantes



Fonte: Felipe Monteiro, Alysson Assunção, 2023

Desse modo, a imagem 4.10 da árvore não é apenas uma representação gráfica, é um mapa interpretativo que nos guia pelos ramos decisórios, revelando variáveis específicas influenciam as predições do modelo.(FILHO, 2017) Cada nó é uma encruzilhada de informações, onde as decisões algorítmicas são tomadas com base em padrões discerníveis nas notas do ENEM e nas características socioeconômicas dos estudantes. Sendo definido

na regra de classificação quando o estudante apresentar um desempenho satisfatório ou não com a realidade social.

Desse modo, podemos adicionar em um conjunto de valores informações sobre nota, e informações sociais, com isso podemos obter uma predição do desempenho do estudante, como podemos observar na figura 4.11.

Figura 4.11: Conjunto de valores de um estudante

```
# criando um novo conjunto de dados para um aluno fictício
novo_aluno = pd.DataFrame({
    'NU_NOTA_CN': [550],
    'NU_NOTA_CH': [580],
    'NU_NOTA_LC': [650],
    'NU_NOTA_MT': [650],
    'NU_NOTA_REDACAO': [650],
    'TP_ESCOLA': [3],
    'Q001' :[1],
    'Q002' :[1],
    'Q003' :[1],
    'Q004' :[1],
    'Q006' :[1],
    'Q022' :[1],
    'Q024' :[1],
    'Q025' :[1]

})
```

Fonte: Felipe Monteiro, Alysson Assunção, 2023

Figura 4.12: Resultado da predição com base nas informações do estudante

```
['Satisfatório(condição social alta)']
```

Fonte: Felipe Monteiro, Alysson Assunção, 2023

Essas informações estão relacionadas a nota em Matemática(NU-NOTA-MT), Ciências da Natureza(NU-NOTA-CN), Ciências Humanas(NU-NOTA-CH) e redação(NU-NOTA REDAÇÃO), incluindo tipo escolar como 3 sendo particular e em questão a informações de renda familiar, escolaridade, acesso a internet e recursos para preparação para o enem como valor 1(sim). Com isso, podemos analisar na figura 4.12 de acordo com essas informações a predição de desempenho satisfatório com condição alta, sendo o resultado esperado com base nessas informações.

4.4 Avaliação e métricas de eficiência da Árvore de decisão

Nesta etapa, iremos abordar sobre a avaliação, desvendando as camadas de eficiência da Árvore de Decisão, de extrema importância no entendimento do desempenho estudantil no ENEM. Este momento em que verificamos e avaliamos o quão é a eficiência deste algoritmo na predição dos dados. As métricas que apresentaremos ilustrada na figura 4.13, serão a matriz de confusão, precisão e recall, que são fundamentais permitindo analisar a eficiência do modelo em diferentes perspectivas, compreendendo a capacidade do algoritmo de aprendizagem de máquina em discernir entre categorias distintas de desempenho. A matriz de confusão, precisão, acurácia e recall, emergem como indicadores críticos, lançando luz sobre a habilidade da Árvore de Decisão em predizer corretamente as classificações e identificar possíveis áreas de melhoria.

Figura 4.13: Métricas de eficiência do algoritmo

```
Matriz de confusão [[  3   0   0   0   0   0   0]
 [  2 355   1   0   0   5   0]
 [  0   1   1   0   0   0   0]
 [  0   0   0   1   0   0   0]
 [  0   4   0   0  36   0   0]
 [  0   2   1   0   0   7   0]
 [  0   0   0   0   0   0   1]]
Acurácia:  0.9619047619047619
Precisão: 0.9673349118652986
Recall: 0.9619047619047619
```

Fonte: Felipe Monteiro, Alysson Assunção, 2023

Desse modo, podemos observar na figura 4.13 essas métricas essenciais em que temos a matriz de confusão, a acurácia com 0.9617(96,17%), precisão em 0.9673(96%) e Recall com 0.96(96%) também, permitindo perceber a efetividade do modelo de aprendizagem no cenário educacional dos estudantes de Belo Jardim. Para isso, utilizamos a biblioteca do python do pacote sklearn, metrics composto por confusion-matrix, accuracy-score, precision-score, recall-score permitindo analisar essas métricas de desempenho da Árvore de Decisão.

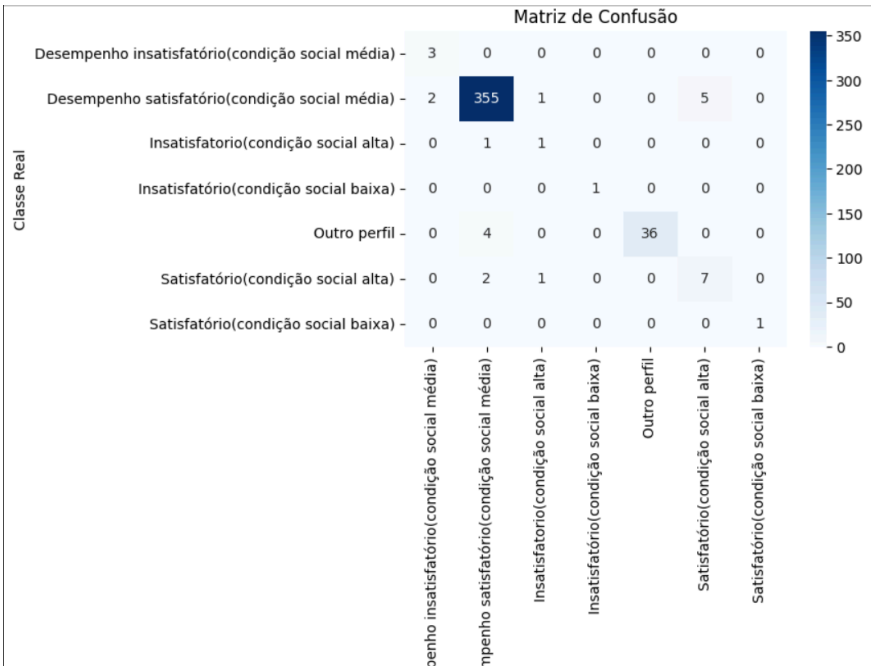
A matriz de confusão representada na figura 4.13, é uma poderosa ferramenta que nos permite avaliar a eficácia do modelo em categorizar corretamente as diferentes classes. Neste contexto específico, a matriz de confusão apresenta uma visão detalhada das classificações

feitas pela Árvore de Decisão, fornecendo uma compreensão mais profunda de seu desempenho no estudo da evolução estudantil no ENEM.

Vamos explorar cada componente dessa matriz:

- Verdadeiros Positivos (TP): Na célula (2,2), temos 355 casos em que o modelo corretamente classificou estudantes, como "Desempenho Satisfatório (Condição Social Alta)". Este é um indicativo positivo da capacidade do modelo em acertar nessa categoria
- Falsos Positivos (FP): Na célula (2,6), encontramos 5 casos em que o modelo erroneamente classificou estudantes como "Desempenho Satisfatório (Condição Social Alta)", quando, na verdade, pertenciam a outras categorias. Este é um aspecto a ser considerado para refinamento do modelo
- Verdadeiros Negativos (TN): As células diagonais fora da principal (1,1), (3,3), (5,5) indicam os casos corretamente classificados como não pertencentes a determinada categoria. Por exemplo, em (1,1), temos 3 casos corretamente identificados como "Satisfatório (Condição Social Alta)".
- Falsos Negativos (FN): Na célula (6,2), temos 2 casos em que o modelo falhou em identificar estudantes que realmente pertenciam à categoria "Desempenho Satisfatório (Condição Social Média)". Esses são casos que merecem atenção para melhoria.

Figura 4.14: Matriz de confusão

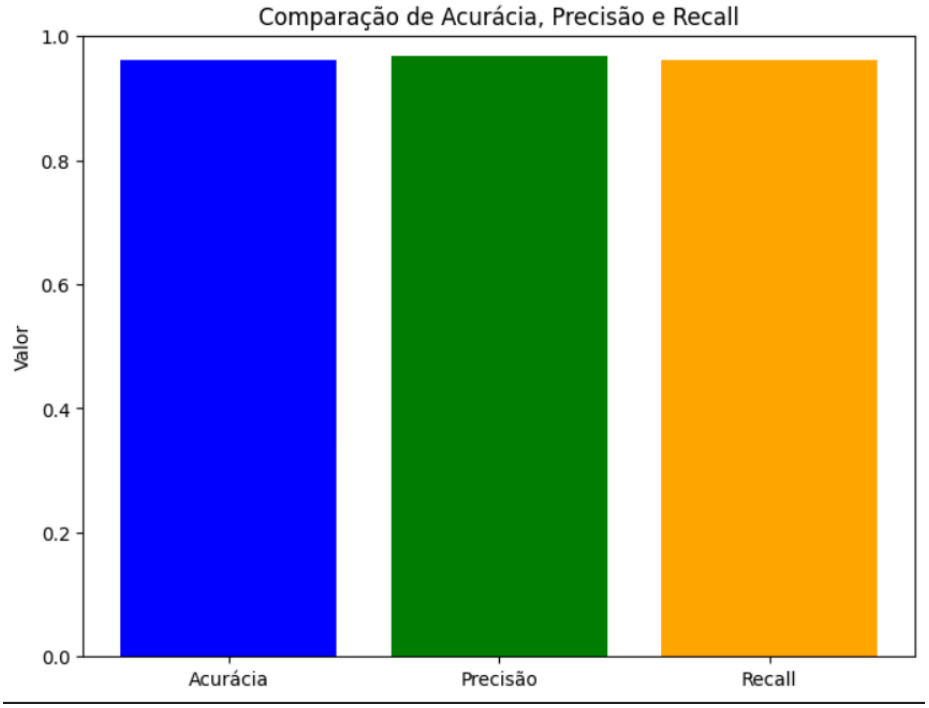


Fonte: Felipe Monteiro, Alysson Assunção, 2023

Essa interpretação da matriz de confusão nos guia além dos números brutos, permitindo-nos entender onde o modelo acerta e onde pode ser otimizado. Este é um componente vital na avaliação do desempenho da Árvore de Decisão, fornecendo insights críticos para aprimorar sua precisão e capacidade de previsão.

O gráfico da figura 4.15, representa a comparação entre precisão, acurácia e recall, proporciona uma visão abrangente do desempenho da Árvore de Decisão, destacando sua habilidade em realizar classificações precisas e identificar corretamente as instâncias de interesse. Cada métrica revela aspectos específicos da performance do modelo, oferecendo uma compreensão aprofundada do quão bem ele se sai em diferentes aspectos.

Figura 4.15: Comparação entre Precisão, Acurácia e Recall



Fonte: Felipe Monteiro, Alysson Assunção, 2023

Com isso, este conjunto de métricas trabalham em conjunto para fornecer uma visão holística da eficiência do modelo. A acurácia nos dá uma visão geral, enquanto a precisão e o recall aprofundam nosso entendimento sobre como o modelo lida com classificações específicas. Em última análise, este gráfico não apenas quantifica o desempenho, mas esboça a confiabilidade e a robustez da Árvore de Decisão nas complexas tarefas de classificação em nosso estudo.

5 Conclusões e Trabalhos Futuros

Ao longo dessa pesquisa, podemos fazer uma análise e uma predição do desempenho dos estudantes de Belo Jardim com base nos dados do ENEM, o que nos proporciona um olhar detalhado das dificuldades educacionais deste micro universo brasileiro.

Exploramos, por meio das notas e questionários socioeconômicos, as dificuldades que permeiam a realidade deste contexto. Esta compreensão pode ser um pontapé inicial na tomada de decisão em prol da melhoria educacional.

Concluímos que estes resultados demonstram que o algoritmo criado junto com os gráficos e práticas adotadas podem trazer benefícios para o desenvolvimento educacional a fim de permitir a tomada de decisão ao alocar recursos para os estudantes que precisam, com o intuito de ajudar no desempenho estudantil permitindo um futuro melhor na educação.

5.1 Trabalhos Futuros

Trabalhos Futuros:

- Aumentar as variáveis no formulário disponibilizado pelo INEP, como quanto tempo de preparação para realização da prova, conhecimento sobre a área que deseja seguir para propor um monitoramento nos primeiros anos do ensino médio e analisar o desempenho para que no último ano se obtenha um resultado de sucesso.
- Evoluir o algoritmo de machine learning com base nessas novas variáveis permitam acompanhar o desempenho nos primeiros anos do Ensino Médio. Nessa melhoria o algoritmo será capaz de indicar em qual área o estudante terá que se dedicar mais a fim de alcançar o ingresso na universidade desejada.
- Implementação de um algoritmo mais aprofundado utilizando redes neurais que usam nós ou neurônios interconectados em uma estrutura em camadas, semelhante ao cérebro humano. Nesse sentido, podemos evoluir na regra de classificação, sendo que podemos classificar com valores maiores com o intuito que aprenda e melhore continuamente.
- Integração deste algoritmo com um sistema web a fim de permitir por meio de uma interface ao usuário uma predição do desempenho no ENEM de forma mais detalhada e proporcionando melhor acompanhamento dos resultados. Além disso, disponibilização dos relatórios e gráficos para o monitoramento dos estudantes.

- Implementação do projeto em uma infraestrutura em nuvem(Cloud). Com essa melhoria o projeto e os dados estarão disponíveis em tempo real e em cada nova atualização dos estudantes o sistema estará atualizado.

Através dessas iniciativas, pretende -se melhorar a eficiência, a escalabilidade e a integração de um algoritmo de machine learning capaz de ajudar no desempenho educacional dos estudantes que realizam o exame em Belo jardim e demais localidades do país, aproveitando as tecnologias atuais e garantindo uma base mais estruturada para o conhecimento e sucesso contínuo do projeto desenvolvido.

Referências

BERNARDETTE, Gatti A. **Estudos quantitativos em educação**., jan. 2004. Disponível em: .http://educa.fcc.org.br/pdf/ep/v30n01/v30n01a02.pdf. Acesso em: 01/03/2022.

BREIMAN, Friedman L. **Classification and regression trees.** 1. ed. [S.l.: s.n.], 1984. Acesso em: 20/06/2022.

CARVALHO, Deborah Ribeiro. **Árvore de decisão algoritmo:** genérico para tratar o problema de Pequenos disjuntos em classificação de dados, dez. 2005. Disponível em: https://www.ipardes.pr.gov.br/sites/ipardes/arquivos_restritos/files/documento/2019-09/deborah_carvalho_tese_2005.pdf. Acesso em: 20/06/2022.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery in Databases**, mar. 1996. Disponível em: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230. Acesso em: 20/02/2022.

FIGUEIRA, R. M. **A. Miner:** um software de inferência de dependências funcionais, jan. 1998.

FILHO, ROGÉRIO LUIZ CARDOSO SILVA. **Modelo de análise e predição do desempenho dos alunos dos Institutos Federais de Educação usando o ENEM como indicador de qualidade escolar**, p. 1–94, ago. 2017. Disponível em: https://repositorio.ufpe.br/handle/123456789/28008. Acesso em: 20/07/2022.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining:** Concepts and Techniques Third Edition. 3RD EDITION. [S.l.: s.n.], 2012. P. 1–560. Disponível em: https://d1wqtxts1xzle7.cloudfront.net/43034828/Data_Mining_Concepts_And_Techniques_3rd_Edition.pdf?1456374619=&response-contentdisposition=inline%3B+filename%3DDATA_MINING_CONCEPTS_AND_TECHNIQUES_3RD.pdf&Expires=1699970323&Signature=e18gfymAu5Dplix19QJNdn6hUKJOE9sEV08eQXIDmG9wMz~ipxoKEYhu3xzsBI9~8odjQ0S0MK79feb-xMwLKZjAh-t~6pu73UQMHbg12r2-xsB8heQfJ8X74cxjTWebU0avmkiINPoC85CK4xiwM2XsuqWO3oISrIYSbucA9xTetvcZWuGoQ4lazRw-zYnxp

URXvbZNgP7k7iDHYKFeFovZNYbyu5OmHPoEwvqxKEe9tn9SyXwxwJGj3uTJGkVeubI395.1 40CP0KDhDP~vzLFLFcWeiJwTJ2OPmvD6BN6NF6dOWXpJC0YRhNHRuBuf6WKyyfw7RVIFeVLGPovtEgX3A__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA. Acesso em: 20/08/2022.

LAURETTO, Marcelo S. **Árvores de Decisão**, nov. 2010. Disponível em: https://edisciplinas.usp.br/pluginfile.php/4469825/mod_resource/content/1/ArvoresDecisao_normalsize.pdf. Acesso em: 20/08/2022. Acesso em: 10/09/2022.

PRESSMAN, Roger S. **Engenharia de Software**. 3. ed. [S.l.: s.n.], 1995.

QUINLAN, J. R. **Induction of decision trees**. 1. ed. [S.l.: s.n.], 1986.

SOARES, Laura Tavares. **Vinte e um anos de Educação Superior:** Expansão e Democratização, jan. 2013. Disponível em: https://biblioteca.flacso.org.br/files/2015/03/Caderno_GEA_N3.pdf. Acesso em: 20/09/2022.

WASLAWICK, Raul Sidnei. **Engenharia de Software:** Conceitos e Práticas. 1. ed. [S.l.: s.n.], 2013.