PROJETO DE GRADUAÇÃO

ANÁLISE DE DADOS ESTRUTURADOS DO ENEM PARA APRIMORAR POLÍTICAS EDUCACIONAIS PÚBLICAS

Por,

Edison Miranda Júnior

Brasília, 07 de julho de 2023

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

UNIVERSIDADE DE BRASÍLIA

Faculdade de Tecnologia

Departamento de Engenharia de Produção

PROJETO DE GRADUAÇÃO

ANÁLISE DE DADOS ESTRUTURADOS DO ENEM PARA APRIMORAR POLÍTICAS EDUCACIONAIS PÚBLICAS

Por, Edison Miranda Júnior.

Relatório submetido como requisito parcial para obtenção
do grau de Engenheiro de Produção

Banca Examinadora

Prof. Dr. Clóvis Neumann, UnB/EPR (Orientador)

Prof. Dr André Luiz Marques Serrano (Avaliador)

FICHA CATALOGRÁFICA

MIRANDA, Edison Júnior

ANÁLISE DE DADOS ESTRUTURADOS DO ENEM PARA APRIMORAR POLÍTICAS EDUCACIONAIS PÚBLICAS

. [Distrito Federal] 2023.

(EPR/FT/UnB, Engenheiro, Engenheiro de Produção, 2023).

Trabalho de conclusão de curso – Universidade de Brasília, Faculdade de Tecnologia.

Departamento de Engenharia de Produção.

REFERÊNCIA BIBLIOGRÁFICA

MIRANDA, EDISON JUNIOR. ANÁLISE DE DADOS ESTRUTURADOS DO ENEM PARA APRIMORAR POLÍTICAS EDUCACIONAIS PÚBLICAS

Trabalho de conclusão de curso, Departamento de Engenharia de Produção, Universidade de Brasília, Brasília, DF.

CESSÃO DE DIREITOS

AUTOR: EDISON MIRANDA JUNIOR

TÍTULO: ANÁLISE DE DADOS ESTRUTURADOS DO ENEM PARA APRIMORAR POLÍTICAS EDUCACIONAIS PÚBLICAS

GRAU: Engenheiro em Engenharia de Produção ANO: 2023

RESUMO

O presente trabalho de conclusão de curso consistiu em desenvolver uma análise de dados a partir dos dados do ENEM 2021 e tem como objetivo analisar como as variáveis socioeconômicas influenciam na nota final dos candidatos. Por meio de estudos estatísticos, buscou-se compreender as relações existentes entre os fatores socioeconômicos e o desempenho dos estudantes em cada uma das provas do exame. A coleta de dados foi realizada a partir das informações disponibilizadas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), órgão responsável pela aplicação do ENEM. A metodologia adotada consiste em uma análise estatística dos dados e métodos de aprendizado de máquina, empregando técnicas como regressão linear e árvore de decisão. Estas técnicas permitem estabelecer relações estatísticas entre as variáveis socioeconômicas e as notas finais dos candidatos, identificando quais fatores exercem maior influência no desempenho dos estudantes. Neste trabalho, buscou-se apresentar uma metodologia de análise de dados para a produção de conhecimentos a partir dos dados oficiais do ENEM e com isso oferecer informações úteis que possam guiar novos estudos na área de educação uma vez que compreender as relações existentes entre as variáveis socioeconômicas e o rendimento dos estudantes ajuda na tomada de decisão em processos de promoção de políticas públicas voltadas à melhoria da qualidade da educação no país.

Palavras chaves: Dados, Estatística, Aprendizado de máquina.

ABSTRACT

The present undergraduate thesis consisted of developing a data analysis based on the ENEM 2021 data and aims to analyze how socioeconomic variables influence the final scores of the candidates.. Through statistical studies, we sought to understand the relationships between socioeconomic factors and students' performance in each of the exam's tests. Data collection was carried out using information made available by the National Institute for Educational Studies and Research Anísio Teixeira (INEP), the organization responsible for administering the ENEM. The adopted methodology consists of statistical analysis of the data and machine learning methods, employing techniques such as linear regression and decision trees. Through these techniques, it is possible to establish statistical relationships between socioeconomic variables and candidates' final grades, identifying which factors have the greatest influence on students' performance. In this work, we aimed to present a data analysis methodology for generating knowledge from official ENEM data and thereby provide useful information that can guide new studies in the field of education. Understanding the relationships between socioeconomic variables and students' performance helps in making decisions regarding the promotion of public policies aimed at improving the quality of education in the country.

Keywords: Data, Statistics, Machine Learning.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE FÓRMULAS

LISTA DE EXEMPLOS

LISTA DE SÍGLAS

CNN	Convolutional Neural Networks
CPF	Cadastro de Pessoa Física
ENEM	Exame Nacional do Ensino Médio
GBM	*Gradient Boosting Machine*
IA	Inteligência Artificial
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IP	*Internet Protocol*
LBPH	*Local Binary Patterns Histograms*
MSE	*Mean Squared Error*
RG	Registro Geral
RNC	Rede Neural Convolucional
SQL	*Structured Query Language*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

O Exame Nacional do Ensino Médio (ENEM) é um exame aplicado em todo o país e tem como principal objetivo medir o domínio que seus participantes possuem dos conhecimentos e competências esperados daqueles que finalizaram o ensino médio.

O resultado deste exame também disponibiliza parâmetros indicadores sobre a educação brasileira de forma a contribuir no aperfeiçoamento dos currículos do ensino médio e na criação de políticas públicas voltadas para a educação.

O Enem é aplicado em dois dias distintos e conta com 180 questões objetivas de múltipla escolha e uma proposta de redação. De acordo com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), o exame nacional do ensino médio aborda quatro áreas de conhecimento conforme mostra o Quadro 01.

Quadro 01 – Áreas de Conhecimento abordadas no ENEM

Área do Conhecimento	Componente Curricular
Linguagens, Códigos e suas tecnologias	Língua Portuguesa, Literatura, Língua Estrangeira (Inglês ou Espanhol), Artes, Educação Física e Tecnologias da Informação e Comunicação.
Ciências Humanas e suas tecnologias	História, Geografia, Filosofia e Sociologia.
Ciências da Natureza e suas tecnologias	Química, Física e Biologia.
Matemática e suas tecnologias	Matemática.

Fonte: Microdados do ENEM (2022)

Além das provas objetivas e da redação, todo participante é solicitado no ato da inscrição a responder um questionário socioeconômico. Com isso, ao final do exame, é compilado um conjunto de dados contendo não só o rendimento nas provas objetivas e redação como também vários aspectos socioeconômicos dos participantes. Estes dados são tratados para que não haja informações pessoais de estudantes e assim disponibilizados no portal do INEP para que qualquer interessado possa utilizá-los a fim de gerar conhecimentos educacionais.

Inicialmente, o ENEM consistia em um exame que oferecia aos estudantes parâmetros para que os mesmos pudessem executar uma autoavaliação de sua formação escolar. Com o tempo, muitas instituições de ensino superior passaram a utilizar o resultado do exame como parte do processo seletivo para o ingresso ao

nível superior, tornando o exame muito mais cobiçado pelos estudantes. Hoje, instituições privadas de ensino oferecem bolsas parciais e integrais em seus cursos a alunos que obtiverem boas notas no ENEM. Além disso, instituições públicas de ensino superior passaram a adotar o ENEM como processo avaliativo de ingresso à universidade oferecendo uma forma a mais de entrar no ensino superior além do vestibular tradicional.

1.2 Definição do problema

Os dados do ENEM 2021 tem o potencial de gerar uma grande quantidade de informações capazes de direcionar novos estudos que busquem compreender o panorama educacional no Brasil. A definição do problema nesse contexto consiste em propor uma metodologia de analise de dados a partir dos dados oficiais do ENEM que possibilite identificar e propor questões relevantes a serem investigadas a cerca da realidade dos estudantes brasileiros destacando-se o desempenho dos estudantes em diferentes áreas do conhecimento em função da disparidade regional, do acesso à educação e das desigualdades socioeconômicas.

1.3 Justificativa

O Exame Nacional do Ensino Médio fornece uma vasta quantidade de informações que pode ser utilizada para identificar tendências, padrões e correlações e com isso, evidenciar possíveis disparidades a cerca da realidade da educação brasileira.

As informações obtidas pela análise dos dados do ENEM 2021 podem ser usadas em estudos que busquem promover políticas públicas voltadas à melhoria da qualidade da educação.

1.4 Objetivo Geral

Desenvolver uma metodologia de análise de dados estruturados do ENEM 2021 para aprimorar políticas educacionais públicas.

1.5 Objetivos Específicos

- Identificar quais variáveis socioeconômicas estão correlacionadas com as notas finais do ENEM.
- Obter parâmetros capazes de mensurar a correlação entre as variáveis socioeconômicas e as notas finais do ENEM.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo engloba uma revisão sobre os principais conceitos que englobam o estudo de mineração de dados e aprendizado de máquina.

2.1 Ciência de Dados

No final do século XX houve um grande avanço das tecnologias da informação. Nos primórdios da tecnologia da informação, os computadores eram máquinas gigantescas, muito caras e com baixo poder de processamento e armazenamento de dados. Seu uso era restrito em universidades e algumas corporações além do fato de que seu manuseio exigia um elevado nível de conhecimentos de comandos capazes de operar a máquina.

Até a década de 1970 era muito comum o armazenamento e a manipulação de informações através de diários, cadernetas e outros tipos de arquivos baseados em papel. Com o avanço e a popularização dos computadores, as informações foram sendo armazenadas e acessadas a partir de dispositivos eletrônicos numa tendência crescente e sem volta. Caldas (2016) explica que o aumento do volume de dados e da velocidade com que as informações passaram a ser processadas tornou o uso da informática indispensável.

Com o advento da microeletrônica, em especial a tecnologia dos semicondutores, foi possível construir pequenas pastilhas de silício com milhares e posteriormente milhões de transistores num espaço muito pequeno. Com isso os computadores passaram a ser cada vez menores ao mesmo tempo em que sua capacidade de processamento e armazenamento de dados cresceram. Assim, surgiu o microcomputador pessoal, um computador pequeno e com preço acessível às pessoas em geral.

Esse avanço das tecnologias da informação ampliou a utilização do computador fazendo com que este dispositivo fosse utilizado nos mais diversos setores das atividades humana. A partir daí, houve um grande aumento na geração de dados e com isso acumulou-se uma quantidade enorme de dados tornando impossível que se extraia qualquer informação relevante a partir de métodos tradicionais. No entanto, o avanço das tecnologias da informação também possibilitou o surgimento de novas técnicas de analises baseadas em algoritmos

computacionais, muito mais rápidas e capazes de trabalhar com quantidades gigantescas de dados. O processo de analisar e extrair informações úteis a partir de grande quantidade de dados é chamado de Mineração de Dados ou *Data Mining*.

Castro (2016) afirma que a produção de dados cresceu exponencialmente chegando ao ponto de gerar mais dados em um curto período de tempo do que muitos séculos de história da humanidade.

A mineração de dados surgiu como uma área de pesquisa independente na década de 1990. Antes disso, já se efetuava a mineração de dados, embora ainda não existisse esse termo, e tal tarefa era desempenhada por profissionais da matemática, estatística e da computação.

O termo “mineração de dados” foi assim denominado por se basear na atividade de mineração de pedras preciosas. Castro (2016) afirma que o termo mineração de dados foi cunhado como alusão ao processo de mineração de ouro e pedras preciosas a partir de uma mina, comparando uma base de dados com a mina, os algoritmos com as ferramentas e os conhecimentos obtidos com os minerais preciosos.

Em paralelo à crescente aplicação da Ciência de Dados, se desenvolveu a Engenharia de dados, que consiste em um ramo da engenharia voltado para desenvolver e manter sistemas de dados capazes de processar e armazenar grandes volumes de dados. Entende-se por sistema de dados o conjunto de equipamentos (hardware) e os algoritmos (softwares) voltados para o armazenamento e a disponibilização de dados. São exemplos de *software* voltados para sistema de dados: Oracle Database, SQL Server e MySQL.

> Data science é a ciência da computação que extrai informações significativas a partir de dados brutos e comunica-as, com eficiência, em atividades práticas. A engenharia de dados, por outro lado, é um domínio da engenharia dedicado a criar e manter sistemas que superam as obstruções do processamento de dados e os problemas do tratamento de dados para os aplicativos que consomem, processam e armazenam grandes volumes, variedades e velocidades de dados (PIERSON, 2019, p. 33).

Pierson (2019) explica que tanto na ciência de dados como na engenharia de dados, os dados são classificados em três tipos: Dados estruturados, semiestruturados e não estruturados. Os dados estruturados são classificados e organizados de tal forma que se encaixam em banco de dados relacional e podem

ser facilmente acessados, processados e gerenciados. Os dados semiestruturados, embora não sejam classificados de forma a se encaixarem em banco de dados relacionais, são organizados a partir de etiquetas apresentando uma ordem hierárquica. Os dados não estruturados não são classificados, ou seja, não se encaixam em nenhum tipo de banco de dados e também não possuem nenhum tipo de organização que facilite definir qualquer hierarquia entre suas partes.

A importância da mineração de dados se justifica pelos potenciais benefícios que ela pode trazer. Uma empresa ou indústria que empregue a mineração de dados pode obter informações que auxiliem na tomada de decisão fazendo com que as decisões sejam executadas de forma mais precisa. Com isso, é possível, por exemplo: otimizar a produção, reduzir desperdícios, potencializar vendas e melhorar a estratégia da empresa perante o mercado.

Devido à crescente tendência de competição mercadológica, obter melhores informações para a tomada de decisão pode ser crucial para que uma empresa seja capaz de oferecer preços compatíveis ao mercado quando comparada à concorrência e com isso manter-se no mercado. A mineração de dados está deixando de ser apenas uma ferramenta que traz melhorias, se tornando algo essencial e indispensável.

> Nos dias de hoje, o mundo empresarial está cada vez mais competitivo e, consequentemente, concerne às organizações inovar e manterem-se prontas para as alterações no mercado de modo a ajustarem-se e permanecerem competitivas e com sucesso, independentemente da sua área de negócio (METELO *et al*., 2021, p.588).

Embora os dados sejam essenciais para o processo de tomada de decisão, apenas recentemente a área de análise de dados, conhecida como Ciência de Dados, tem se formalizado gerando um novo tipo de profissional (cientista de dados), sendo este muito requisitado no mercado de trabalho. Castro (2016) afirma que após surgir os termos “analista de dados” e “cientista de dados”, estes profissionais estão sendo cada vez mais requisitados e bem pagos devido à crescente produção de dados.

De acordo com um relatório do Linkedin (2020 *apud* GUIMARÃES, 2022) a profissão de cientista de dados ocupa a terceira posição entre as profissões mais emergentes na atualidade. Guimarães (2022) acrescenta que há uma dificuldade por

parte do mercado em encontrar profissionais com pensamento computacional ainda mais quando se trata de profissionais experientes.

> A demanda por profissionais com capacidades analíticas e técnicas para lidar com grandes e heterogêneos volumes de dados, tem resultado em uma expansão na oferta de cursos para a formação deste perfil profissional (CURTY; SERAFIN, 2016, p. 309.).
>
> Com papéis divididos e diversificados, predominantemente orientados a dados, esse profissional impacta áreas relacionadas a pesquisa, inovação, economia e na sociedade em geral (CAO, 2019 apud GUIMARÃES *et. al*. 2022, p.56).

Por ser uma área nova, a ciência de dados é geralmente ocupada por profissionais de outras áreas.

> Por ainda ser uma área em desenvolvimento, muitos cientistas de dados possuem formação em cursos universitários já estabelecidos, como física, economia, engenharia, mas principalmente, estatística e ciência da computação (BAŠKARADA; KORONIOS, 2017 apud GUIMARÃES *et. al*. 2022, p.57).

Pode-se dizer que os conhecimentos mais necessários para um profissional que queira atuar com ciência de dados são a estatística e a computação.

> A literatura é unânime em enfatizar que os cientistas de dados devem apresentar domínio estatístico e computacional para a programação e uso de sistemas capazes de processar grandes volumes de dados (CHATIFELD et al., 2014; GRANVILLE, 2014 apud GUIMARÃES et. al. 2022, p.311)

A mineração de dados faz uso de diversos métodos. Estes métodos são executados por algoritmos computacionais. Cada método é capaz de efetuar um determinado tipo de análise. Logo, o método a ser utilizado depende do tipo de análise que se pretende efetuar.

Exemplos de algoritmos de mineração de dados são: Algoritmo de regressão, Algoritmo baseado em instância, Algoritmo de regularização, Método Naïve Bayes, Árvore de decisão, Algoritmo de agrupamento, Método de redução da dimensão, Rede neural, Método de aprendizado profundo e Algoritmo de grupo.

2.2 Inteligência Artificial e Aprendizado de Máquina

Inteligência Artificial é a tecnologia que permite que máquinas simulem uma inteligência similar à humana, ou seja, que torne a máquina capaz de tomar decisões de forma autônoma, diferentemente de algoritmos convencionais que fazem a máquina executar ordens específicas anteriormente definidas.

Aprendizado de Máquina (*Machine Learning*) é um ramo da Inteligência Artificial voltado para algoritmos capazes de aprender determinadas tarefas automaticamente a partir da análise de um conjunto de dados. Modelos gerados por Aprendizado de Máquina são capazes de reconhecer padrões nos dados que lhe são dispostos para efetuar o aprendizado e a partir deles, a máquina é capaz de melhorar seu desempenho.

O fato de não ser possível programar um computador para executar determinadas tarefas é que diferencia os algoritmos de inteligência artificial. Ao invés disso, aplicam-se técnicas de aprendizado de máquina que consistem em fazer o computador aprender a partir de um conjunto de dados. A aplicação de inteligência artificial é muito eficiente para problemas em que não se sabe como desenvolver programas que os resolvam, mas que se têm muitos dados relacionados a tais problemas.

> Os humanos têm muito conhecimento intuitivo, que não conseguem expressar verbalmente com facilidade. Não se tem acesso consciente a esse conhecimento intuitivo. Sem uma compreensão formal desse conhecimento intuitivo não é possível escrever programas para representá-lo. Então qual é a solução? A solução é a máquina aprender esse conhecimento por si mesma, de maneira similar a como os seres humanos aprendem (LUDERMIR, 2021, p.87).

O uso de inteligência artificial e aprendizado de máquina ampliaram a capacidade de resolver problemas. Antes, o computador só executava o que um ser humano programava, ou seja, o computador era limitado a desempenhar tarefas que um ser humano consegue descrever como fazer de forma detalhada, apresentando apenas a vantagem de executar tal tarefa muito mais rapidamente.

> É por meio do Aprendizado de Máquina que o computador está adquirindo novas habilidades. As técnicas de Aprendizado de Máquina permitem que o computador aprenda por exemplos, ou seja, aprenda por meio dos dados. O Aprendizado de Máquina tornou-se chave para colocar conhecimento nos computadores (LUDERMIR, 2021, p.86).

Pierson (2019) afirma que os métodos que utilizam aprendizado de máquina consistem em três etapas principais: configuração, aprendizado e aplicação.

A configuração consiste na aquisição dos dados assim como seu processamento inicial e na divisão dos dados em duas partes, sendo uma para treinamento e outra para testes.

O aprendizado consiste em utilizar a parte dos dados separada para treinamento para construir um modelo de classificação dos dados. A máquina irá buscar relações que permitam identificar e classificar os dados.

A aplicação será a execução do modelo gerado utilizando a parte de dados reservada para testes. Ao aplicar o modelo gerado nos dados de testes, é possível estimar o aprendizado da máquina a partir da quantidade de erros e acertos efetuados com estes dados.

> Para se efetuar uma boa avaliação do modelo gerado, é importante que a divisão dos dados para treinamento e testes seja feita de forma aleatória. “Uma boa regra prática para dividir os dados em conjuntos de testes e treinamento é aplicar uma amostragem aleatória” (PIERSON, 2019, p. 52).

Os métodos de Aprendizado de Máquina podem ser divididos em três tipos: Supervisionado, Não-Supervisionado e o Aprendizado por reforço. O tipo supervisionado consiste em criar um modelo capaz de predizer rótulos a partir de um conjunto de dados previamente rotulado. O tipo não supervisionado consiste em, a partir de um conjunto de dados sem nenhum tipo de rótulo, descobrir relações e similaridades entre os objetos. O tipo de aprendizado por reforço baseia-se na tentativa e erro. A máquina é inserida num cenário dinâmico e aprende a se comportar a partir da interação e recepção de *feedback* externo.

> Existem três estilos principais na aprendizagem de máquina: supervisionado, não supervisionado e semissupervisionado. Os métodos supervisionado e não supervisionado estão por trás de quase todas as aplicações da aprendizagem de máquina e o aprendizado semissupervisionado é promissor (PIERSON, 2019, p. 53).

2.3 Mineração de texto

A mineração de texto consiste na descoberta de conhecimentos contidos em grandes volumes de dados textuais.

> Mineração de textos consiste em extrair regularidades, padrões ou tendências de grandes volumes de textos em linguagem natural, normalmente, para objetivos específicos (ARANHA; PASSOS, 2006, p.1).

Dados textuais diferem de outros tipos de dados devido à forma como são organizados. Geralmente, dados são organizados em tabelas, ou seja, os dados são classificados. Dizemos então que tais dados são estruturados. No entanto, dados textuais não possuem uma pré-classificação, necessitando de algoritmos diferenciados para sua análise.

> Inspirado pelo data mining ou mineração de dados, que procura descobrir padrões emergentes de banco de dados estruturados, a mineração de textos pretende extrair conhecimentos úteis de dados não estruturados ou semi-estruturados (ARANHA; PASSOS, 2006, p.1).

Segundo Barcelos (2005 *apud* ARANHA; PASSOS, 2006), como as técnicas desenvolvidas para Mineração de Dados foram desenvolvidas para dados estruturados, técnicas específicas para Mineração de Textos têm sido desenvolvidas para processar uma parte importante da informação disponível, que pode ser encontrada na forma de dados não-estruturados.

A mineração de texto, por se tratar de um tipo de mineração de dados, faz uso de conhecimentos e ferramentas da Informática e Estatística. No entanto, a mineração de dados textuais também faz uso de conhecimentos em Linguística e Ciência Cognitiva uma vez que a informação contida em texto está diretamente relacionada com as características da língua na qual o texto foi escrito e na relação sintática entre as palavras diferente de dados estruturados em que a informação existe na relação matemática entre os elementos. “Um único padrão essencial para entender todas as línguas ainda é utópico. Sendo assim, qualquer sistema de processamento de textos tem parâmetros específicos para cada língua” (ARANHA; PASSOS, 2006, p.3).

A criação de métodos voltados para a análise de dados textuais foi bastante estimulada devido ao grande número de dados textuais gerados em mídias sociais, *e-mails, blogs* e outras mídias digitais. Estes dados contêm informações muito úteis, por exemplo, para empresas que vendem produtos e serviços *online*, uma vez que a partir de tais textos é possível identificar e selecionar o perfil dos clientes em potencial tornando o processo de *marketing* mais barato e eficiente.

2.4 Análise de Imagens

A área de visão computacional apresentou grande crescimento acompanhando o crescimento de ferramentas computacionais. A utilização de máquinas na identificação de imagens trouxe inúmeras possibilidades de aplicações nas mais diversas áreas.

Pode-se citar como aplicações do uso de algoritmos de detecção de imagens a detecção de faces para autenticação em sistemas de segurança, a detecção de placas de veículos em estradas e a identificação de insetos potencialmente perigosos.

> Os estudos na área de visão computacional vêm proporcionando inúmeros benefícios para sociedade. Entre esses benefícios, está a melhoria em sistemas de segurança (tecnologias de detecção facial estão sendo utilizadas em aplicações de reconhecimento de face para automatizar o processo de autenticação, buscando substituir o uso de senhas e impressões digitais) (COSTA, 2021, p.2).

A construção de algoritmos capazes de efetuar analise de imagem sempre foi um grande desafio. Embora seja muito simples para um ser humano analisar e identificar objetos a partir de imagens, para uma máquina, esta tarefa sempre se apresentou como algo de grande complexidade. No entanto, com o surgimento de ferramentas de inteligência artificial e aprendizado de máquina, a análise automática de imagens passou a ser uma realidade e está sendo amplamente usada em diversas aplicações.

> O que anteriormente poderia ser classificado como uma tarefa extremamente desafiadora e complexa, a detecção de faces e de suas características como olhos, nariz, boca e, até mesmo, derivar emoções de suas formas, tornou-se uma tarefa possível de ser resolvida com o uso de técnicas de Deep Learning, e bibliotecas de processamento digital de imagens e visão computacional, como a biblioteca OpenCV (PRADO, 2017 apud COSTA, 2021, p.2).

Uma aplicação muito interessante foi apresenta por Sousa (2019) em um estudo sobre o uso de aprendizado de máquina para análise de imagens voltadas para a identificação de insetos vetores de doenças tornando sua vigilância mais eficaz. O estudo aborda a doença de Chagas, uma infecção crônica potencialmente fatal por causar problemas cardíacos graves. Como essa doença é transmitida pelo inseto da subfamília *Triatominae*, conhecidos como barbeiros, é preciso efetuar uma boa vigilância desse vetor, o que torna necessário a sua identificação.

A partir de 2.331 imagens de 86 espécies do inseto existente no Brasil e no México, conseguiu-se aplicar um aprendizado de máquina de tal forma que tal inseto pode ser identificado com sucesso a partir de fotografias em um processo totalmente automatizado.

Da mesma forma que a máquina aprendeu a identificar o inseto *Triatominae*, é possível fazer com que a máquina aprenda a identificar outras coisas, em outro contexto completamente diferente. Para isso, é necessário aplicar técnicas de aprendizado de máquina a partir de um grande conjunto de imagens classificadas. Se o conjunto de imagens usadas no processo de aprendizagem for suficientemente grande e as imagens forem devidamente classificadas, após o aprendizado, a máquina será capaz de identificar padrões que possibilitem a identificação e classificação de novas imagens de acordo com a classificação aprendida.

> Deep Learning é uma técnica de aprendizado de máquina na qual o programa computacional – que constitui uma Rede Neural Convolucional (RNC) – aprende a distinguir entre imagens diferentes tal como humanos fazem.” [...] “Para uma RNC aprender a diferenciar objetos, é necessário fornecer imagens classificadas. Na prática, arquivos contendo várias imagens de cada tipo são fornecidos ao programa.” [...] “O aprendizado também é mais bem-sucedido quando se oferece um número maior de fotos e que apresentem boa qualidade (SOUZA, 2019, p.148).

Algoritmos capazes de identificar faces possibilitam uma série de aplicações. Atualmente, os telefones celulares oferecem a função de tirar foto, ou seja, funcionam como máquinas fotográficas. Alguns destes dispositivos possuem um algoritmo capaz de detectar o sorriso do usuário e com isso efetuar a fotografia automaticamente.

Costa (2021) apresenta um estudo cujo objetivo é realizar uma análise comparativa entre alguns dos principais métodos de detecção e reconhecimento facial a partir do tempo de execução e do desempenho de cada um deles.

Foram testados os algoritmos *Eigenface, Fisherface e LBPH* (*Local Binary Patterns Histograms*) e a rede neural CNN (*Convolutional Neural Networks*). Foram usados três conjuntos de imagens de faces públicos, tendo 445 imagens no primeiro, 750 imagens no segundo e 750 imagens no terceiro. Os conjuntos de imagens foram divididos em duas partes. A primeira, contendo, em média, 86% das imagens foi usada para treinar o algoritmo enquanto a segunda foi usada para testar seu desempenho. O resultado obtido com a rede neural CNN apresentou a melhor

taxa de acerto, entretanto, não apresentou um bom tempo de execução se comparado com os outros métodos. Sendo assim, (DA COSTA, 2021) concluí que a escolha de um método deve estar relacionada com os requisitos da aplicação que se deseja efetuar.

Embora o surgimento de técnicas baseadas em aprendizado de máquina tenham se mostrado muito eficiente para fazer o computador aprender e desempenhar a tarefa de detectar imagens, o auxilio do ser humano ainda se mostra crucial para o correto funcionamento destes algoritmos. As amostras de imagens utilizadas precisam representar corretamente a realidade daquilo que se deseja que a máquina aprenda a identificar, caso contrário, pode-se induzir a máquina a efetuar um comportamento enviesado.

> Os conjuntos de treinamento fornecidos aos algoritmos podem não ser uma boa representação do mundo real, e as amostras podem estar enviesadas. Por exemplo, um sistema pode ser desenvolvido para distinguir gatos de cachorros, mas no conjunto de treinamento apresentado ao algoritmo, todos os cachorros são pretos e todos os gatos são brancos, então o sistema pode aprender a distinguir preto de branco e não gato de cachorro. Nesses casos, é necessária a experiência dos humanos para descobrir e solucionar o problema (LUDERMIR, 2021, p.92).

No exemplo citado, percebe-se que o algoritmo apresentaria um grave erro caso tivesse que classificar uma foto como uma imagem de um gato ou de um cachorro se em tal foto fosse apresentado um gato preto ou um cachorro branco. Como todos os cachorros apresentados nas fotos, durante o processo de aprendizagem eram pretos ao mesmo tempo em que todos os gatos eram brancos, é presumível que a cor do animal tenha sido um dos padrões identificados, talvez o principal deles, e utilizados na identificação do animal.

2.5 Riscos e Benefícios no Uso da Inteligência Artificial

São inegáveis os benefícios obtidos nos últimos anos a partir do uso de Inteligência Artificial nas mais diversas áreas das atividades humanas. Esse crescimento está relacionado com o desenvolvimento de novas tecnologias da informação que trouxe novas formas de extração, armazenamento, transmissão e processamento de dados.

É importante considerar que toda tecnologia pode ser usada de forma incorreta ou mesmo perigosa. No caso da inteligência artificial, uma questão muito relevante é a relação entre as máquinas e os seres humanos. A capacidade de aprender somada com a velocidade com que as máquinas estão sendo melhoradas trás um alerta sobre até que ponto os seres humanos estão de fato no controle, lembrando que a sociedade atual está cada vez mais dependente de dispositivos autônomos. No entanto, é notório que toda nova tecnologia desperta esperanças e receios.

> Uma das principais incertezas sobre nossa relação com máquinas inteligentes é como lidar com conflitos entre máquinas e seres humanos. Havendo uma disputa entre máquinas e seres humanos, é importante observar que enquanto a estrutura e o comportamento dos seres humanos são guiados por lentos processos bioquímicos, as máquinas, com os avanços na ciência de materiais e na capacidade de processamento de dispositivos computacionais, são cada vez mais resistentes e eficientes (CARVALHO, 2021, p.27).

Uma preocupação real e muito debatida diz respeito à questão de como a Inteligência Artificial afeta o mercado de trabalho. Muitas atividades exercidas por humanos estão sendo substituídas por máquinas causando a extinção de várias atividades profissionais.

> Um dos receios associados a isso é como a IA afetará o mercado de trabalho. Vários estudos mostram que atividades profissionais desaparecerão, sendo substituídas por atividades que até o momento são desconhecidas ou inimagináveis. Com frequência, são divulgadas listas com as profissões com maior probabilidade de desaparecer no futuro (CARVALHO, 2021, p.28).

Em contrapartida, não se pode negar os benefícios obtidos com a substituição de seres humanos por máquinas em trabalhos cuja atividade apresenta riscos à vida humana. A questão não está simplesmente no fato de as máquinas substituirem os seres humanos, mas sim, em como a organização social humana está lidando com toda essa mudança.

Outra questão muito relevante é o fato de que a Inteligência Artificial está sendo amplamente usada no processo de tomada de decisão. Para confiar e aceitar decisões tomadas por máquinas, Carvalho (2021) afirma que as pessoas devem sentir que estas decisões são justas. O autor trás a tona a ocorrência de noticias sobre decisões preconceituosas geradas por modelos de algoritmo de inteligência

artificial voltados para a identificação de criminosos em que o modelo gerado fazia uso de dados raciais e apresentava maior probabilidade de ter alguém como criminoso se tal pessoa fosse de uma dada raça.

Este fato mostra que modelos gerados por Inteligência Artificial não estão isentos de ideologia e preconceito.

Duarte (2022) relata um caso de uma estudante negra que após tentar desenvolver um algoritmo de identificação de faces a partir de imagens de um banco de dados percebeu que o algoritmo não funcionava bem quando as imagens analisadas eram de pessoas da sua etnia.

Com isso, é levantada a seguinte questão: “como é possível construir um modelo de observação, identificação e busca de soluções capaz de lidar com o problema da discriminação perpetrada por algoritmos de aprendizado de máquina?” (DUARTE, 2022, p.4).

A construção de programas de treinamento e aprendizado de máquina pode ser direcionada, propositalmente ou não, para gerar um processo de decisão com vieses discriminatórios. Isso acontece devido à forma como a máquina aprende. Se um desenvolvedor rotula os dados a serem usados pela máquina no processo de aprendizagem, os dados irão conter o viés do desenvolvedor e consequentemente serão reproduzidos pela máquina. Além do mais, os próprios dados utilizados podem ter sido produzidos dentro de uma situação nas quais crenças e vieses humanos interferem nesse processo. Com isso, dados tendenciosos irão gerar modelos de algoritmos também tendenciosos.

De acordo com Carvalho (2021), o uso da inteligência artificial é algo presente e sem volta e por isso entende-se que o que deve ser discutido não é o uso dessa tecnologia, mas sim a forma como essa tecnologia será usada. O uso de novas tecnologias deve ser usado de forma que promova a inclusão social e que o uso de inteligência artificial traga benefícios para todos.

2.6 Mineração de Dados Educacionais

O avanço tecnológico ocorrido nas últimas décadas trouxe consigo modificações no processo de ensino e aprendizagem. Muitos *softwares* voltados para o ensino foram desenvolvidos além da vasta quantidade de informações largamente acessíveis devido a popularização da *internet*. Com tudo isso, houve um

enorme crescimento no volume de dados educacionais de tal forma que se torna imprescindível o uso de recursos computacionais para se efetuar análises nestes dados.

Segundo a Sociedade Internacional de *Educational Data Mining* (EDM, 2020 *apud* Souza, 2021, p.185),

> A Mineração de Dados Educacionais pode ser definida da seguinte forma: É uma disciplina emergente, preocupada com o desenvolvimento de métodos para explorar dados únicos e cada vez mais em larga escala, provenientes de contextos educacionais e usa esses métodos para entender melhor os alunos e as configurações em que aprendem.

A utilização de ferramentas de mineração de dados e inteligência artificial é muito utilizada em aplicações relacionadas com a área de educação. A seguir, são apresentadas algumas aplicações de mineração de dados e inteligência artificial em diferentes contextos educacionais obtidas em pesquisas bibliográficas. No entanto, as possibilidades de aplicações são imensamente maiores que os exemplos a seguir apresentados.

Souza (2021) apresenta a realização de previsões de desempenho de alunos nas disciplinas de português e matemática a partir da análise de um conjunto de dados públicos utilizando técnicas de mineração de dados educacionais.

A mineração de dados se apresenta como uma boa ferramenta na descoberta de relações entre a realidade social e econômica dos alunos com as notas dos respectivos alunos. Com isso, é possível identificar possíveis causas de dificuldade dos alunos e consequentemente, Fórmular hipóteses que auxiliem os gestores educacionais em decisões importantes acerca do processo educativo nas escolas.

A evasão escolar se apresenta como um grande problema recorrente quando se trata de Educação. O abandono escolar, denominado também de evasão escolar, provoca graves consequências sociais, acadêmicas e econômicas (BAGGI; LOPES, 2011 *apud* BITENCOUT *et. al*. 2022, p.669).

O abandono escolar é um problema bastante frequente e pode ter como causa muitos fatores, sendo alguns internos e outros externos, como por exemplo, atributos demográficos, acadêmicos, pessoais e familiares.

Sonnenstrah *et al*. (2021) apresenta um estudo cujo objetivo foi analisar possíveis evasões em cursos de modalidade à distância através da Mineração de Dados Educacionais a partir de dados gerados pela interação dos alunos no

Ambiente Virtual de Aprendizagem e com isso disponibilizar dados estratégicos que auxiliem a tomada de decisão dos gestores educacionais.

A educação voltada para estudantes com necessidades especiais se apresenta como uma parte da área da educação que possui características muito singulares. Embora existam leis que favoreçam o ingresso de pessoas com necessidades especiais em universidades, a forma como se dá o processo de ensino e aprendizagem e as estruturas funcionais nos locais de ensino não apresentam condições suficientes para inserir de forma inclusiva este tipo de aluno.

Segundo a entrevista realizada na Universidade Federal de Mato Grosso do Sul, uma professora afirma que:

> A acessibilidade já avançou muito na instituição, mas é preciso ir além, avançar ainda mais com a transformação da divisão em um núcleo, com o aumento da estrutura para atendimento ao aluno, melhoria na acessibilidade da instituição e maior capacitação dos profissionais. É preciso evoluir (VIEGAS, 2016).

Oliveira *et al*. (2020) apresenta um trabalho de aplicação de técnica de mineração de dados e aprendizado de máquina nos registros do ENEM do ano de 2018 com o intuito de predizer o desempenho final dos candidatos com deficiência a partir de suas características socioeconômicas, demográficas, étnicas, faixa-etária e formação escolar.

A partir da análise feita, os autores conseguiram resultados interessantes: alunos com Déficit de Atenção são propensos a apresentarem excelentes resultados enquanto Deficientes Mentais e Auditivos são propensos a apresentarem resultados ruins. Além disso, percebeu-se também que alunos que apresentaram boa competência em redação, tendem a apresentar bom resultado nas áreas de Matemática e melhores desempenho na média final.

A partir de tais observações, os autores concluem que o acesso à escola particular e uma renda mais alta contribuem muito significativamente para o acesso ao ensino superior. Poucos deficientes tentam ingressar em uma universidade e dos que tentam, poucos conseguem ingressar no ensino superior.

2.7 Teste de Aderência

Teste de aderência consiste em um teste estatístico cujo objetivo é verificar se um conjunto de dados de uma amostra apresenta uma distribuição suficientemente próxima de uma distribuição teórica.

2.7.1 Teste qui-quadrado de aderência

Existem diferentes testes de aderência. O teste qui-quadrado de aderência pode ser utilizado quando se deseja testar dados distribuidos em diferentes categorias.

> O teste qui-quadrado de aderência pode ser aplicado quando estamos estudando dados distribuídos em categorias e há interesse em verificar se as frequências observadas nas K diferentes categorias(BARBETTA et. al, 2010, p.275).

Caso o teste seja positivo, significa que existe aderência, ou seja, há pouca diferença entre os dados testados e o valor esperado sendo tal diferença casual. Caso o teste seja negativo, siginifica que a diferença entre os dados testados e o valor esperado é grande o suficiente a ponto de ser considerada improvável que existe devido a uma causalidade.

2.7.2 Teste de Kolmogorov-Smirnov

Quando é necessário testar a aderência de um conjunto de dados numéricos, o teste qui-quadrado de aderência não se apresenta como uma boa alternativa. O teste de aderência de *Kolmogorov-Smirnov* se apresenta como melhor opção para estes casos, pois se trata de um teste estatístico utilizado para determinar se uma amostra de dados segue uma distribuição específica. Este teste compara a função de distribuição acumulada dos dados a serem testados com a função de distribuição acumulada da distribuição teórica a qual se deseja comparar.

> Considere uma situação em que desejamos verificar a aderência de um conjunto de valores em relação a uma distribuição de probabilidades especificada (discreta ou contínua). Embora seja possível aplicar o teste qui-quadrado de aderência, geralmente é melhor aplicar o chamado teste de aderência de Kolmogorov-Smirnov, que é uma alternativa mais poderosa do que o teste qui-quadrado, nestas situações (BARBETTA et. al, 2010, p.277).

Ao aplicar o teste de *Kolmogorov-Smirnov*, busca-se determinar se há evidências estatísticas suficientes para rejeitar a hipótese nula de que os dados apresentam a distribuição equivalente à distribuição teórica testada. O resultado do teste é obtido a partir de um valor *p*, que indica a probabilidade de se obter uma estatística igual ou maior do que a observada, assumindo que a hipótese nula seja verdadeira. Se o valor *p* for menor que um nível de significância pré-definido, rejeita-se a hipótese nula e conclui-se que os dados não seguem a distribuição teórica.

O teste de *Kolmogorov-Smirnov* possibilita verificar a adequação de um conjunto de dados a uma distribuição específica, permitindo uma avaliação objetiva da similaridade entre os dados observados e a distribuição teórica escolhida. Além disso, este teste pode ser facilmente aplicado em um conjunto de dados utilizando a linguagem *python* e a biblioteca *scipy.stats*.

2.8 Regressão Linear Simples

A regressão linear simples é uma técnica estatística capaz de estabelecer uma relação entre uma variável independente (x) e uma variável dependente (y). Tal regressão é denominada linear por assumir que a relação entre as variáveis segue o comportamento de uma função afim ao mesmo tempo em que é denominada simples por possuir apenas uma variável independente.

O modelo de regressão linear simples supõe que a relação entre as variáveis independente (x) e dependente (y) possa ser descrito pela equação:

$$y = \alpha + \beta x \quad \text{(Fórmula 01)}$$

sendo α e β os parâmetros do modelo.

A partir do conjunto de observações representadas por dados do tipo (x_1, y_1), (x_2,y_2), ..., (x_n,y_n), ou seja, por n pares ordenados, defini-se a seguinte relação:

$$yi = \alpha + \beta xi + \varepsilon i \quad \text{(Fórmula 02)}$$

O ε_i é o erro da i-ésima observação, ou seja, é um valor relacionado a uma infinidade de fatores que influenciam a variável dependente tornando a distribuição dos dados não linear.

Embora as variáveis dependente e independente possam não estar relacionadas a ponto de estabelecer uma função afim, o modelo de regressão linear simples é capaz de encontrar os parâmetros α e β que apresente o menor erro (ε) acumulado além de disponibilizar um parâmetro que indique o quanto essa relação se aproxima de uma função linear.

2.8.1 Método dos mínimos quadrados

Para se obter a melhor equação linear que represente a relação entre as variáveis dependentes e independentes é possível fazer uso do método dos mínimos quadrados.

“Há vários métodos para estimar os parâmetros do modelo. O mais usual é o método de mínimos quadrados, que consiste em fazer com que a soma dos erros quadráticos seja a menor possível” (BARBETTA *et. al*, 2010, p.326).

O valor de ε_i pode ser definido pela seguinte equação:

$$\varepsilon i = yi - \alpha - \beta xi \quad \text{(Fórmula 03)}$$

A ideia é encontrar os valores de α e β para os quais os erros acumulados sejam mínimos. No entanto, se somarmos todos os erros, diferenças negativas serão anuladas por diferenças positivas. Sendo assim, minimizamos o quadrado do erro, pois tanto as diferenças positivas quanto as negativas serão sempre somadas.

O método dos mínios quadrados consiste em minimizar calcular os valores de α e β que minimizem o valor de S da seguinte equação:

$$S = \sum_{i=1}^{n} \varepsilon i^2 = \sum_{i=1}^{n} (yi - \alpha - \beta xi)^2 \quad \text{(Fórmula 04)}$$

Para minimizar o valor de S, basta igualar as derivadas parciais de S em função das variáveis α e β a zero.

$$\frac{\partial S}{\partial \alpha} = 0 \; \frac{\partial S}{\partial \beta} = 0 \quad \text{(Fórmula 05)}$$

Relacionando as duas Fórmulas é possível obter um sistema linear que após ser resolvido, disponibiliza os valores de α e β:

$$\beta = \frac{n \cdot \sum_{i=1}^{n} (xi \cdot yi) - (\sum_{i=1}^{n} xi) \cdot (\sum_{i=1}^{n} xi)}{n \cdot \sum_{i=1}^{n} xi^2 - (\sum_{i=1}^{n} xi)^2} \quad \text{(Fórmula 06)}$$

$$\alpha = \frac{\sum_{i=1}^{n} yi - \beta \sum_{i=1}^{n} xi}{n} \quad \text{(Fórmula 07)}$$

2.8.2 Coeficiente de determinação

Se os valores da variável independente (x) não influenciar o valor da variável dependente (y) então a melhor forma de se estimar o valor de y é a partir da média aritmética das observações dos y_i, ou seja:

$$Y = \frac{\sum_{i=1}^{n} yi}{n} \quad \text{(Fórmula 08)}$$

sendo Y o valor esperado de y.

No entanto, se a variável independente (x) causar alguma influência na variável dependente (y) então é esperado que a equação obtida a partir do método de regressão linear simples apresente uma melhor forma de se estimar o valor de y. Sendo assim, é preciso obter um estimador que apresente o quanto o modelo de regressão linear é capaz de descrever a relação entre as variáveis x e y.

Este estimador é chamado de R^2 e pode ser obtido a partir da razão entre o erro acumulado no modelo de regressão linear e a soma dos erros acumulados no modelo de regressão linear e na média aritmética:

$$R^2 = \frac{\sum_{i=1}^{n} (Yi - \overline{Y})^2}{\sum_{i=1}^{n} (Yi - \overline{Y})^2 + \sum_{i=1}^{n} (Yi - \widehat{Y})^2} \quad \text{(Fórmula 09)}$$

sendo $\overline{Y}$ o valor estimado de y pela regressão e $\widehat{Y}$ a média das amostras de y.

O valor de R^2 pode variar entre 0 e 1. Se o valor de R^2 for próximo de 1 entende-se que a variável dependente (y) tem seu comportamento fortemente relacionado com a variável independente (x), ou seja, que o modelo de regressão se apresenta como uma boa estimativa para se obter o valor de y a partir dos valores de x. No entanto, se o valor de R^2 for próximo de 0 entende-se que a variável dependente (y) não tem seu comportamento relacionado com a variável

independente (x), ou seja, que o modelo de regressão não se apresenta como um modelo capaz de estimar o valor de y a partir dos valores de x.

2.9 Regressão Linear Múltipla

A Regressão Linear Múltipla é uma técnica estatística capaz de estabelecer uma relação entre uma variável dependente (y) e duas ou mais variáveis independentes (x_1, x_2, x_3, ...). Esse modelo é uma extensão da regressão linear simples, onde é possível incluir mais de uma variável independente no modelo de análise.

Na regressão múltipla, a equação de regressão inclui o coeficiente α e vários coeficientes (β_1, β_2, β_3, ...), sendo um para cada variável independente. Assim como na Regressão linear simples, busca-se encontrar a combinação linear dessas variáveis que melhor explica a variação na variável dependente.

2.10 Redes Neurais Artificiais

Redes neurais artificiais são modelos computacionais compostos por unidades interconectadas chamadas de neurônios artificiais. Tais modelos tiveram seu desenvolvimento inspirado no funcionamento do cérebro de animais. As redes neurais são muito úteis em tarefas complexas que exige o uso de aprendizado de máquina. Assim como um cérebro biológico, redes neurais artificiais aprendem a partir de exemplos sem serem previamente programadas.

> No reconhecimento de imagem, as redes neurais podem aprender a identificar imagens que contenham gatos analisando exemplos que tenham sido rotulados manualmente como "gato" ou "não gato" e, em seguida, usar os resultados para identificar gatos em outras imagens. As redes neurais fazem isso sem qualquer conhecimento prévio sobre gatos (FALCÃO et. al. 2019, p.2).

2.10.1 Neurônio

O neurônio artificial é a unidade básica de processamento em uma rede neural artificial. Cada neurônio recebe um conjunto de entradas numéricas e produz uma saída. Essa saída alimenta a entrada de outros neurônios formando uma grande rede.

O processamento do dado de entrada de cada neurônio é realizado por meio de uma função de ativação seguida de uma função não-linear. Embora a capacidade de processamento de um neurônio seja simples, a junção de muitos neurônios torna a rede neural capaz de realizar operações matemáticas altamente complexas.

2.10.2 Funcionamento da Rede

As conexões entre os neurônios, assim como o comportamento dos mesmos, têm seu comportamento vinculadas a variáveis de ajuste (pesos). Para efetuar o aprendizado, utiliza-se um conjunto de dados que contêm os valores de *input* (entrada) e *output* (saída). Assim, a rede neural processa os dados de *input* e compara o dado obtido na saída da rede com o valor de *output*. Nesse processo, as variáveis de ajuste são modificadas de forma a minimizar a diferença da comparação entre o dado calculado e o esperado.

Os valores de ajuste podem ser iniciados com valores aleatórios e ajustados durante o processo de aprendizagem. O processo de aprendizado pode fazer uso de algoritmos de otimização. Tal processo varia de acordo com o modelo de rede neural adotado.

2.10.3 TensorFlow

TensorFlow é uma biblioteca de *software* voltada para efetuar aprendizado de máquina e inteligência artificial. Desenvolvida pela equipe do *Google Brain*, essa biblioteca permite a criação e treinamento de redes neurais em diferentes linguagens como por exemplo a linguagem *python*.

Segundo *Tensorflow* (2019 *apud* FALCÃO et. al. 2019) a biblioteca *TensorFlow* é uma das principais do mercado para a criação de redes neurais e aprendizado de máquina.

A biblioteca *TensorFlow* oferece uma ampla gama de recursos, desde o pré-processamento de dados até o treinamento de modelos complexos. Além disso, ela também possui uma comunidade ativa de desenvolvedores e usuários que fornecem suporte e contribuem com recursos adicionais.

> O maior benefício que o TensorFlow oferece para o desenvolvimento de sistemas inteligentes é a abstração. Em vez de lidar com os detalhes básicos da implementação de algoritmos ou de descobrir formas adequadas de ligar a saída de uma função à entrada de outra, o desenvolvedor pode se concentrar na lógica geral da aplicação ou no problema que se deseja resolver (TENSORFLOW, 2019 apud FALCÃO et. al. 2019, p.5).

2.11 Árvore de Decisão

O método de árvore de decisão é um algoritmo de aprendizado de máquina. Seu algoritmo funciona a partir de uma estrutura em forma de árvore. Este método é muito útil quando se deseja prever o valor de uma variável dependente em função de varias variáveis independentes.

> Árvore de decisão: Uma estrutura de árvore é útil como ferramenta de suporte de decisão. Você pode usá-la para criar modelos que prevejam possíveis consequências associadas a uma determinada decisão (PIERSON, 2019, p. 56).

A partir de um conjunto de dados estruturados, o algoritmo da árvore de decisão gera uma sequência de decisões em função dos dados das variáveis independentes a fim de obter um valor final da variável dependente. A ordem dos testes é definida de forma a minimizar o erro de previsão, ou seja, de forma a diminuir a diferença entre o valor previsto e o valor real da variável que se deseja prever.

> O algoritmo da árvore de decisão desenvolve um conjunto de regras do tipo sim ou não, que você pode aplicar aos novos dados e ver exatamente como serão caracterizados pelo modelo (PIERSON, 2019, p. 92).

2.12 Expressões Regulares

Expressão Regular (Regex – Regular Expression) é um método formal de se especificar um padrão de texto. Utilizando expressões regulares é possível definir de forma concisa e flexível um conjunto de sequências de caracteres formadas por determinado padrão.

> Uma expressão regular é um método formal de se especificar um padrão de texto. [...], é uma composição de símbolos, caracteres com funções especiais, que, agrupados entre si e com caracteres literais, formam uma sequência, uma expressão. Essa expressão é interpretada como uma regra que indicará sucesso se uma entrada de dados qualquer casar com essa

> regra, ou seja, obedecer exatamente a todas as suas condições (JARGAS, 2016, p.12).

As Expressões Regulares tiveram seu início a partir do trabalho do matemático norte-americano Stephen Cole Kleene, que desenvolveu uma notação capaz de definir um conjunto de texto que apresentasse um certo padrão desejado.

As Expressões Regulares são muito úteis na busca e substituição de texto em editores de texto que disponibilizam o uso de Regex uma vez que facilita a busca de textos que apresentem certos padrões como por exemplo CPFs, CEPs, Telefones, Datas ou qualquer tipo de texto que siga uma notação especifica.

Segundo Jargas (2016), as expressões regulares podem ser usadas em uma infinidade de tarefas, sendo difícil citar todas as possibilidades de uso, pois elas são úteis sempre que se precisar buscar ou validar um padrão de texto, que pode ser variável, como por exemplo: data, horário, número de IP, endereço de *e-mail*, endereço de *Internet*, número de telefone, RG, CPF etc.

As expressões regulares baseiam-se em metacaracteres que são caracteres que possuem significados especiais dependendo da posição em que aparece nas expressões regulares. Estes caracteres, quando usados em uma expressão regular, não são interpretados de maneira literal. Os metacaracteres utilizados em Regex são apresentados no Quadro 02.

Quadro 02 – Caracteres utilizados em Expressão regular

Caractere	**Descrição**
.	Ponto
+	Sinal de mais
\	Barra invertida
()	Parênteses
[]	Colchetes
{}	Chaves
^	Acento agudo
?	Ponto de interrogação
$	Cifrão
!	Ponto de exclamação
|	Barra vertical
=	Sinal de igualdade

Fonte: Elaborado pelo autor

2.12.1 Ponto

O primeiro metacaractere é o ponto. O ponto é um curinga, ou seja, uma expressão regular entende o ponto como qualquer caractere, exceto quebra de

linha. Sendo assim, a expressão regular “ca.a”, por exemplo, pode ser interpretada como qualquer sequência de caracteres que comece com as letras “ca”, seguida de um caractere qualquer, e por fim, que termine com o caractere “a”.

O Exemplo 01 apresenta um pequeno texto em que as sequencias de caracteres que satisfazem a expressão regular “ca.a” estão destacadas com fundo amarelo:

Exemplo 01 – Caracteres utilizados em Expressão regular

Encantado, casa, cada, acatada, cocada, calado, encarado.

Fonte: Elaborado pelo autor

É possível perceber que tal cadeia de caractere pode estar localizada dentro de palavras. Para definir que a sequência de caracteres que se deseja buscar esteja no início ou no fim de uma palavra, deve-se fazer uso do um metacaractere “\b”. Da mesma forma, para definir o inicio e o fim de ma linha deve-se fazer uso de metacaracteres específicos, que são o acento circunflexo (^) e o cifrão ($).

O acento circunflexo indica o início de uma linha assim como o cifrão indica o fim de uma linha. Estes metacaracteres podem ser utilizados em conjunto conforme necessário.

2.12.2 Colchetes

Os metacacteres colchetes são usados para definir uma lista de caracteres. Às vezes, deseja-se buscar por sequências de caracteres que só apresentem caracteres pertencentes a uma lista, como por exemplo: somente números, somente letras ou qualquer conjunto especificado. Para isso, é possível definir uma lista, conforme apresentado no Exemplo 02.

Exemplo 02 – Lista de caracteres em Expressão Regular

[0123456789ABCDEF]

Fonte: Elaborado pelo autor

Esta lista especificada significa que os caracteres aceitos são os numéricos e as letras maiúsculas “ABCDEF”. Sendo assim, ao escrever a expressão regular 0x[0123456789ABCDEF] busca-se o caractere ‘0’ seguido do caractere ‘x’ seguido de um caractere pertencente à lista.

O Exemplo 03 apresenta um pequeno texto em que as sequências de caracteres que satisfazem a expressão regular “0x[0123456789ABCDEF]” estão destacadas com fundo amarelo:

Exemplo 03 – Lista de caracteres em Expressão Regular

A0xz940xF25 9AFx030x7 0x1234 91x7325

Fonte: Elaborado pelo autor

Dentro de uma lista, todos os caracteres são tidos como literais, com exceção dos caracteres traço (-) e acento circunflexo (^). O traço, quando posicionado entre dois caracteres, informa que a lista é formada por todos os caracteres compreendidos entre o primeiro e segundo caractere de acordo com a tabela ASC. A lista “[0123456789ABCDEF]”, por exemplo, poderia ser mais facilmente representada conforme mostrado no Exemplo 04:

Exemplo 04 – Lista de caracteres sequenciados em Expressão Regular

```
[0-9A-F]
```

Fonte: Elaborado pelo autor

Para acrescentar o traço em uma lista, basta posicioná-lo no final da lista. Por exemplo: define-se uma lista que contenham apenas os caracteres numéricos, o caractere vírgula e o caractere traço, conforme mostra o Exemplo 05.

Exemplo 05 – Lista de caracteres sequenciados e traço

```
[0-9,-]
```

Fonte: Elaborado pelo autor

O primeiro traço indica que todos os caracteres compreendidos entre o caractere 0 e o 9 (incluindo estes caracteres) fazem parte da lista. O segundo traço, como está no final da lista, é interpretado de forma literal. Sendo assim, as expressões “[0123456789,-]” e “[0-9,-]” são equivalentes.

O acento circunflexo, quando posicionado no início da lista, indica que todos os caracteres definidos posteriormente dentro da lista serão negados, ou seja, será formada uma lista com todos os caracteres exceto os caracteres definidos na lista. Isso é muito útil, por exemplo, quando se deseja definir um caractere que pode ser qualquer coisa exceto uma determinada lista de caracteres. Se usado em outra posição da lista, este acento é interpretado de forma literal.

2.12.3 Chaves

As chaves são usadas como quantificadores. Quantificadores são úteis para definir cadeias de caracteres em que vários caracteres seguidos satisfazem uma mesma condição, ou seja, quando se deseja buscar uma cadeia de caracteres que satisfazem uma determinada lista. Por exemplo: Suponha-se que seja necessário identificar em um texto todos os números de CPF sabendo que tais números, no texto em questão, são escritos da seguinte forma: XXX.XXX.XXX-XX sendo cada X um caractere numérico. Pode-se definir esse padrão de texto com a expressão regular apresentada no Exemplo 06.

Exemplo 06 – Expressão Regular de um CPF

```
[0-9][0-9][0-9][.] [0-9][0-9][0-9][.] [0-9][0-9][0-9]- [0-9][0-9]
```

Fonte: Elaborado pelo autor

Embora essa expressão regular esteja correta e funcione, é possível fazer uso dos quantificadores para tornar a expressão mais simples e legível. Veja que o caractere ponto foi colocado dentro de uma lista, pois fora dela, este caractere seria interpretado como qualquer caractere já que ele é o curinga das expressões regulares. É possível escrever a mesma expressão regular conforme mostra o Exemplo 07.

Exemplo 07 – Expressão Regular de um CPF com quantificadores

```
[0-9]{3}[.][0-9]{3}[.][0-9]{3}-[0-9]{2}
```

Fonte: Elaborado pelo autor

Quando é necessário definir uma sequência de caracteres cuja quantidade de caracteres não é exata, ou seja, que pode variar dentro de certa quantidade, deve-se utilizar os quantificados. Suponha-se, por exemplo, que seja necessário identificar em um texto todas as sequencias de caracteres alfabéticos cuja quantidade de caracteres esteja compreendida entre três e cinco. Para isso, pode-se utilizar a expressão regular apresentada no Exemplo 08.

Exemplo 08 – Expressão Regular com quantificadores

```
[A-Za-z]{3,5}
```

Fonte: Elaborado pelo autor

O Exemplo 09 apresenta um pequeno texto em que as sequências de caracteres que satisfazem a expressão regular “[A-Za-z]{3,5}” estão destacadas com fundo amarelo e verde alternadamente.

Exemplo 09 – Expressão Regular com quantificadores

```
Teste de funcionamento da expressao regular anterior.
```

Fonte: Elaborado pelo autor

Veja que a expressão regular tenta sempre buscar o maior conjunto possível. Como a palavra “regular” possui sete letras, as primeiras cinco letras são identificadas como uma sequência de caracteres que satisfaz a expressão proposta e, no entanto, os dois caracteres restantes não satisfazem a expressão. Pode-se imaginar que a palavra “regular” poderia ser dividida em um pedaço com quatro caracteres e outro com três e assim, ambos os pedaços iriam satisfazer a expressão. Isso não acontece por que o processamento de uma expressão regular

sempre tende a identificar o maior número possível de caracteres que satisfaça a expressão. Este comportamento é vulgarmente chamado de “modo guloso” e pode ser mudado em situações onde tal comportamento não é desejado com uso de um metacaractere apropriado.

Às vezes, a quantidade que se deseja definir não tem um limite, mas apenas uma quantidade mínima. Para isso, basta fazer uso dos quantificadores com a quantidade mínima seguida de uma vírgula como mostrado no Exemplo 10.

Exemplo 10 – Expressão Regular com quantificadores de mínimo

```
[A-Za-z]{3,}
```

Fonte: Elaborado pelo autor

Essa expressão buscará todas as sequências de caracteres alfabéticos formadas por três ou mais caracteres. Quando se deseja procurar uma expressão em que um certo caractere, ou lista de caracteres, pode aparecer ou não, basta usar o valor zero no início do quantificador.

2.12.4 Parênteses

Os parênteses são usados para definir grupos de expressões e modificam a ordem de procedência das expressões regulares assim como fazem os parênteses em equações matemáticas.

2.12.5 Barra

O caractere barra (\) é usado em conjunto com outro caractere subsequente para formar um metacaractere do tipo barra-letra e apresenta significados especiais. O Quadro-02 apresenta os metacaracteres barra-letra mais utilizados e seu respectivo significado.

Quadro 03 – Metacarecteres barra-letra mais utilizados

Metacaractere Barra-Letra	Equivalência	Descrição
\n		Caractere de quebra de linha
\t		Caractere Tab.

\d	[0-9]	Caractere numérico
\s	[ ]	Caractere espaço
\w	.	Qualquer caractere (exceto quebra de linha)
\.	[.]	Caractere ponto (literal)
\b		Início ou fim de uma palavra
\N		Retrovisor (N é um número de 1 a 9)
\[		Caractere ‘abrir’ colchete
\]		Caractere ‘fechar’ colchete

Fonte: Elaborado pelo autor

O metacaractere \b indica o início ou o fim de uma palavra. Este metacaractere é muito útil para garantir que a busca de uma sequência de letras só aceite uma sequência alfabética que não apresente caracteres alfabéticos antes ou depois, dependendo de onde se coloca tal metacaractere.

2.12.6 Metacaracteres simplificados

Existem três metacaracteres em expressões regulares que simplificam os quantificadores {0,1},{0,} e {1,}. O Quadro-03 a seguir apresenta os metacaracteres e suas respectivas equivalências.

Quadro 04 – Metacaracteres quantificadores simplificados

Metacaractere	Equivalência	Descrição
?	{0,1}	Zero ou um
+	{1,}	Um ou mais
*	{0,}	Zero ou mais

Fonte: Elaborado pelo autor

Estes três metacaracteres apresentam exatamente a mesma função dos quantificadores formados por chaves, no entanto, seu uso torna as expressões mais simples, resumida e legível.

2.12.7 Metacaractere lógico

As expressões regulares apresentam o metacaractere | (barra vertical) que indica a lógica booleana OU. Quando duas ou mais expressões regulares são agrupadas com este metacaractere, a expressão regular obtida irá buscar toda e qualquer sequência de caracteres que satisfaça pelo menos uma das expressões individuais.

2.12.8 Decisão

Em análise de texto, é comum a necessidade de se efetuar a busca de uma determinada sequência de caracteres que, além de apresentar determinado padrão, seja precedida por outra sequência de caractere de determinado padrão.

Por exemplo: suponha-se que seja necessário efetuar uma busca de toda e qualquer palavra de um texto formada exclusivamente por caracteres alfabéticos e que tenha logo em seguida o caractere interrogação (?). No entanto, não é desejado selecionar o caractere interrogação, embora sua presença seja necessária para o correto atendimento da expressão regular.

Inicialmente, montamos uma expressão regular que seleciona toda e qualquer palavra formada exclusivamente por caracteres alfabéticos com a expressão regular “\b[A-Za-z]+\b”. Em seguida, basta implementar o comando de decisão que exige a ocorrência do caractere interrogação na frente da palavra. Para isso, utiliza-se a expressão “(?=[?])”.

A expressão completa resulta na seguinte expressão: “\b[A-Za-z]+(?=[?])”.

É possível resumir as expressões de decisão conforme mostra o Exemplo 11.

Exemplo 11 – Expressão Regular com teste de decisão

```
texto1(?=texto2)
```

Fonte: Elaborado pelo autor

Sendo o “texto1” a sequência de caracteres que se deseja selecionar e o “texto2” a sequência de caracteres que deve preceder a primeira sequência para que toda a expressão seja atendida.

Da mesma forma que é possível efetuar uma busca de uma determinada sequência de caracteres que preceda outra sequência, é possível efetuar a busca de uma sequência de caracteres que não seja precedida por outra determinada sequência, ou seja, existe o comando inverso. Este comando é apresentado no Exemplo 12.

Exemplo 12 – Expressão Regular com teste de decisão negado

```
texto1(?!texto2)
```

Fonte: Elaborado pelo autor

Este comando indica que será buscado qualquer ocorrência do “texto1” que não seja precedido por ‘texto2”

2.12.9 Retrovisores

Os retrovisores são expressões capazes de reutilizar uma sequência de caracteres anteriormente identificada no texto. Os retrovisores são muito úteis, por exemplo, para se efetuar buscas de palavras repetidas em um texto. Às vezes, é interessante fazer uma varredura em um texto para verificar se alguma palavra foi digitada repetidamente. Para isso, basta fazer a busca de uma palavra qualquer seguida de outra palavra igual. Isso só é possível com uso de retrovisores.

“Apenas como lembrete, algumas linguagens e programas, além da função de busca, têm a função de substituição. O retrovisor é muito útil nesse caso, para substituir ‘alguma coisa’ por ‘apenas uma parte dessa coisa’”. (JARGAS, 2016, p.51)

A expressão regular que faz uma busca de palavras que aparecem repetidamente em um texto é apresentada no Exemplo 13.

Exemplo 13 – Expressão Regular com retrovisores

```
(\b[A-Za-z0-9]+\b)\s\1
```

Fonte: Elaborado pelo autor

A primeira parte da expressão, que está dentro de parênteses, define qualquer sequência de caracteres delimitada por um início e fim de uma palavra e que seja formada apenas de caracteres alfabéticos ou numéricos. O sinal + indica que tal sequência pode ser formada por um ou mais caracteres. Uma vez encontrada tal sequência, é analisado se em seguida tem-se um caractere espaço seguido da mesma sequência anteriormente definida dentro dos parênteses.

A palavra anteriormente definida dentro dos parênteses é solicitada pelo metacaractere \1, que é um retrovisor. Esse retrovisor solicita toda a sequência anteriormente encontrada dentro do primeiro par de parênteses. Da mesma forma, para solicitar as sequências anteriormente encontradas dentro do segundo, terceiro até o nono par de parênteses, utiliza-se os metacaracteres \1, \2, até o \9, ou seja, uma expressão regular pode fazer uso de até 9 retrovisores.

Os retrovisores também se apresentam como uma poderosa ferramenta para efetuar a localização e substituição de textos em editores de texto que oferecem o uso de expressões regulares.

2.12.10 Quantificador não guloso

As expressões regulares são processadas de maneira a buscar sempre o maior número de caracteres que satisfaça a expressão. Este comportamento é denominado de “modo guloso”. Este comportamento pode ser melhor entendido a partir do seguinte exemplo: deseja-se analisar um texto de um código fonte html e identificar todas as partes do texto em negrito. Como em html um texto é codificado em negrito ao ser envolvido pelas tags e , é natural imaginar que a expressão regular que busca o texto “”, seguido de uma sequência de qualquer caractere em qualquer quantidade “.+” e finalizada com o texto “” será capaz de identificar as partes em negrito de um texto. No entanto, isso não ocorre.

Aplicando a expressão regular “.+” no texto “As Expressões Regulares são muito úteis na Análise e Busca de textos.” Será selecionada toda sequência de caracteres destacada com fundo amarelo.

Exemplo 14 – Expressão Regular com quantificador guloso

As Expressões Regulares são muito úteis na Análise e Busca de textos.

Fonte: Elaborado pelo autor

Isso ocorre porque toda essa sequência começa com “” e termina com “” conforme solicitado pela expressão regular, sendo uma só sequência. No entanto, o objetivo é gerar uma expressão regular que selecione as duas sequências de caracteres destacadas com fundo amarelo a seguir:

Exemplo 15 – Expressão Regular com quantificador não guloso

As Expressões Regulares são muito úteis na Análise e Busca de textos.

Fonte: Elaborado pelo autor

Para isso, é preciso fazer com que a expressão regular selecione a menor quantidade de caracteres possíveis que satisfaça a expressão e para isso, é preciso fazer uso do metacaractere interrogação (?) após o metacaractere quantificador. Com isso, a expressão correta seria “.+?”.

Ao adicionar o metacaractere (?) após um metacaractere quantificador, a execução da expressão regular será finalizada após o encontro da primeira ocorrência de uma sequência de texto que satisfaça a expressão regular. Esse quantificador não guloso pode ser combinado com quantificadores conforme mostra o Quadro-04.

Quadro 05 – Metacaracteres não gulosos

??	Após o metacaractere ? (0 ou 1)
*?	Após o metacaractere * (0 ou mais)
+?	Após o metacaractere + (1 ou mais)
{n,m}?	Após quantificadores (de n a m ocorrências)

Fonte: Elaborado pelo autor

2.12.11 Expressões regulares em linguagem de programação

Por se apresentar como uma excelente ferramenta de análise e busca de texto, as expressões regulares são frequentemente utilizadas em compiladores de diversas linguagens de programação.

“Várias linguagens de programação possuem suporte às expressões regulares, seja nativo, como módulo importável, como biblioteca carregável, como objeto instanciável, como... Ou seja, opções há várias”. (JARGAS, 2016, p.131)

A linguagem *Python*, por exemplo, além de oferecer de forma nativa muitos recursos voltados para manipulação e processamento de texto também oferece a possibilidade de utilizar expressões regulares a partir do carregamento de bibliotecas.

"*Python* possui um dos mais completos suportes às expressões regulares, com objetos e métodos já prontos para obter diversas informações sobre os casamentos" (JARGAS, 2016, p.186).

3 ESTUDO DE CASO

O presente capítulo apresenta um estudo de caso que busca analisar dados estruturados do Exame Nacional do Ensino Médio (ENEM) utilizando métodos estatísticos e técnicas de análise de dados.

Os cadernos de prova, os gabaritos, as informações sobre os itens, as notas das provas e os questionários socioeconômicos são compilados e disponibilizados no site do INEP de forma a não oferecerem identificação pessoal de qualquer um de seus candidatos para que assim possam ser disponibilizados na internet. “As notas das provas e as informações do questionário socioeconômico são disponibilizadas no formato de tabela em um arquivo “.csv “.

Também é disponibilizado um dicionário de variáveis no formato de planilha “.xlsx”, uma vez que os dados da tabela se encontram de forma codificada, pois se trata de uma grande quantidade de dados e sua codificação torna o arquivo mais compacto (consome menos quantidade de armazenamento), facilitando não só seu armazenamento como também seu processamento.

Neste estudo, foram utilizados os dados do ENEM de 2021 uma vez que os dados referentes ao ENEN de 2022 ainda não haviam sido disponibilizados. Nestes dados, estão disponibilizadas as notas de quatro avaliações objetivas e da redação além de diversas características socioeconômicas dos participantes. Através da aplicação de métodos estatísticos e análise de dados, buscou-se identificar padrões, tendências e relações entre as variáveis.

3.1 Metodologia

Para obter um levantamento sobre os métodos e aplicações de análise de dados foi realizada uma pesquisa bibliográfica no repositório da CAPES e no *Google* Acadêmico utilizando as palavras chaves “mineração de dados”, “análise de dados”, “aprendizado de máquina” e “dados educacionais”. Realizou-se recorte temporal na busca de artigos científicos publicados entre 2015 e 2022. Esta pesquisa teve como objetivo buscar embasamento teórico e conceitual sobre a análise e mineração de dados.

Durante o processo de pesquisa, foram analisados e revisados diversos artigos e publicações, buscando uma ampla compreensão dos conceitos,

ferramentas e metodologias relacionadas à análise e mineração de dados. A pesquisa bibliográfica desempenhou um papel fundamental na construção do embasamento teórico deste trabalho, fornecendo referências relevantes para a compreensão das abordagens utilizadas na análise de dados.

Os dados para análise deste trabalho foram obtidos diretamente do *site* oficial do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Por meio do acesso ao portal do INEP, foram baixados os arquivos referentes ao ENEM 2021, que contêm informações detalhadas sobre o desempenho dos estudantes, suas características socioeconômicas, além de outros dados relevantes.

A metodologia utilizada tem como objetivo identificar relações e padrões entre as características socioeconômicas dos candidatos e seus rendimentos finais no exame. Primeiramente, foi realizada uma exploração inicial dos dados, buscando obter uma visão geral. Em seguida, foram aplicadas técnicas estatísticas, como Regressão Linear, Teste X^2 e Árvore de Decisão para identificar padrões, relações e tendências nos dados.

Para realizar a análise descritiva, foram utilizadas medidas estatísticas, como média, desvio padrão, mediana, percentil, valor mínimo e valor máximo. Com isso, foi possível analisar características dos dados e obter uma visão geral do desempenho dos estudantes nas quatro áreas de conhecimento e na redação.

Posteriormente, foram aplicadas técnicas estatísticas para investigar as relações entre as variáveis socioeconômicas e o desempenho dos estudantes. Essas análises estatísticas permitiram obter parâmetros capazes de mensurar objetivamente como se dá a relação existente entre as variáveis analisadas.

Foram utilizados gráficos para representar os dados de forma visual tornando-os mais intuitivos e compreensíveis.

Por fim, foram geradas informações capazes de auxiliar estudos educacionais voltados para a compreensão do sistema educacional brasileiro. A aplicação de métodos estatísticos permitirá identificar padrões, relações e tendências, de forma objetiva.

3.1.1 Etapas da análise de dados

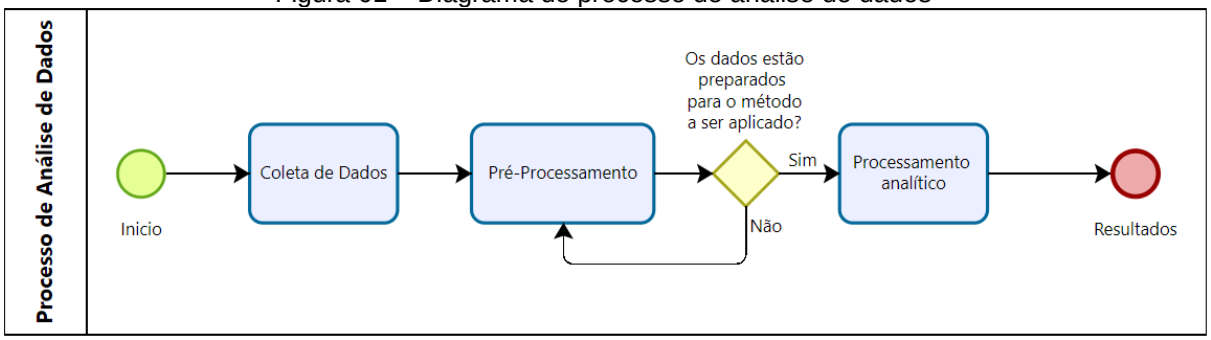
Conjuntos de dados brutos, em geral, apresentam dados faltantes, em excesso ou mesmo duplicados. Por isso, não se pode efetuar técnicas estatísticas

em dados brutos, sendo necessário primeiramente efetuar um tratamento prévio nos dados de forma a torná-los compatíveis com cada tipo de método que se deseja aplicar. Sendo assim, os dados analisados foram submetidos a um processo de três etapas: Coleta de dados, Pré-processamento e Processamento analítico.

> O fluxo de trabalho de um processo típico de Mineração de Dados contém as seguintes etapas: 1) Coleta de dados; 2) Extração de recursos e limpeza de dados (Pré-Processamento e Transformação) – para tornar os dados adequados para processamento; 3) Processamento analítico e algoritmos – projetar métodos analíticos eficazes para extrair informações e conhecimentos relevantes a partir dos dados processados (AGGARWAL, 2015 apud SOUZA, 2021, p.3).

A Figura 01 apresenta o diagrama do processo de análise de dados.

Figura 01 – Diagrama do processo de análise de dados



Fonte: Elaborado pelo autor

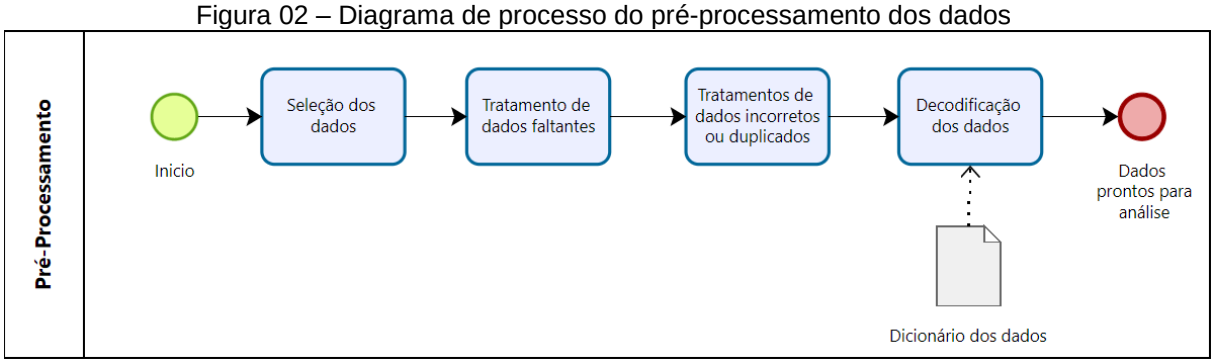
3.1.2 Coleta de dados

Os dados para análise deste trabalho foram obtidos diretamente do site oficial do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). A forma como os dados foram disponibilizados é do tipo “estruturado”. Os dados são organizados em uma única tabela de forma que cada coluna corresponde a um determinado atributo do participante e cada linha corresponde a um único participante.

3.1.3 Pré-processamento

O pré-processamento dos dados consiste em organizar os dados de forma a prepará-los para determinados testes posteriores. A Figura 02 apresenta o diagrama do pré-processamento dos dados.

Figura 02 – Diagrama de processo do pré-processamento dos dados



Fonte: Elaborado pelo autor

O pré-processamento se iniciou com a seleção dos dados. Os dados brutos apresentam muitos atributos, sendo necessário selecionar aqueles que correspondam aos objetivos do trabalho. O Quadro-05 apresenta os atributos selecionados, a descrição dos atributos e os tipos de dados para cada um deles.

Quadro 06 – Atributos Selecionados

Atributos	Descrição dos Atributos	Tipo de dado
NOTA_CIENCIAS_ NATURAIS	Nota da prova Objetiva de Ciências da Natureza e suas Tecnologias.	Numérico (0 a 1000)
NOTA_CIENCIAS_ HUMANAS	Nota da prova Objetiva de Ciências Humanas e suas Tecnologias.	Numérico (0 a 1000)
NOTA_LINGUAGENS_ CODIGOS	Nota da prova Objetiva de Linguagem e Códigos e suas Tecnologias.	Numérico (0 a 1000)
NOTA_MATEMATICA	Nota da prova Objetiva de Matemática e suas Tecnologias.	Numérico (0 a 1000)
NOTA_REDACAO	Nota da prova de Redação	Numérico (0 a 1000)
SEXO	Sexo do participante	Feminino Masculino
COR_RACA	Etnia do participante	Branca Preta Parda Amarela Indígena
FAIXA_ETARIA	Faixa etária em que se encontra a idade do participante no ano da aplicação do exame.	-Menor que 17 -17 -18 -19 -20 -21 -22 -23 -24 -25 -26 até 30 -31 até 35 -36 até 40 -41 até 45 -46 até 50 -51 até 55 -56 até 60

		-61 até 65 -66 até 70 -Maior que 70
TIPO_ESCOLA	Tipo de escola em que o participante estuda ou estudou o ensino médio	Pública Privada Não declarada
UF	Estado em que mora o participante	Sigla do Estado
ESCOLARIDADE_PAI	Escolaridade do pai do participante	-Nunca estudou. -Não completou a 4ª série/5º ano do Ensino Fundamental. -Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental. -Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio. -Completou o Ensino Médio, mas não completou a Faculdade. -Completou a Faculdade, mas não completou a Pós-graduação. -Completou a Pós-graduação.
ESCOLARIDADE_MAE	Escolaridade da mãe do participante	
RENDA_FAMILIAR	Renda familiar mensal do participante (R$)	Nenhuma Renda Até 1.100 De 1.100 até 1.650 De 1.650 até 2.200 De 2.200 até 2.750 De 2.750 até 3.300 De 3.300 até 4.400 De 4.400 até 5.500 De 5.500 até 6.600 De 6.600 até 7.700 De 7.700 até 8.800 De 8.800 até 9.900 De 9.900 até 11.000 De 11.000 até 13.200 De 13.200 até 16.500 De 16.500 até 22.000 Acima de 22.000.

Fonte: Elaborado pelo autor

Após selecionar os atributos, efetuou-se o tratamento de dados faltantes. Optou-se por eliminar todas as linhas da tabela que apresentavam dados nulos nos atributos em que seria efetuado algum método estatístico posterior. Para efetuar, por exemplo, uma análise de como a nota de redação está relacionada com a renda familiar foram retiradas as linhas da tabela que apresentavam dados nulos na nota de redação ou na renda familiar.

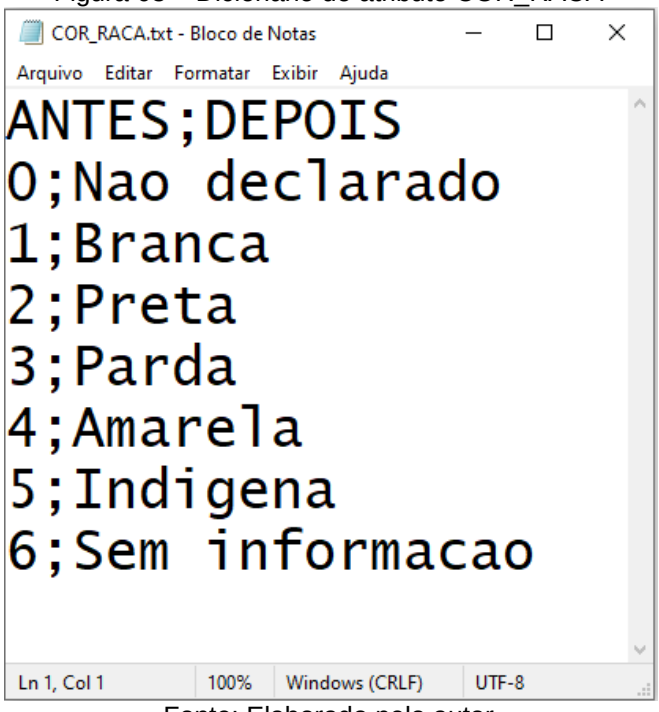
Em seguida, efetuou-se o tratamento de dados incorretos ou duplicados. Dados que apresentavam erros foram corrigidos, quando possível ou eliminados.

Por fim, efetuou-se a decodificação dos dados. Como os dados brutos são armazenados e disponibilizados em código, é preciso substituir cada código pelo texto correspondente. Para isso, foram montados arquivos.txt, a partir do dicionário de dados que fora disponibilizado pelo INEP junto com a tabela de dados.

Os arquivos.txt foram criados com a utilização de um programa feito em linguagem c# com uso de uma biblioteca de expressões regulares chamada “System.Text.RegularExpressions;”. Cada arquivo contém pares de texto que representam um dicionário de substituição. Com os arquivos montados, foi possível utilizar a linguagem *python* para substituir os dados de cada uma das colunas da tabela pelas informações correspondentes.

A Figura 03 apresenta o arquivo “COR_RACA.txt” do atributo ‘COR_RACA’ em que contem o código de cada dado e seu dado correspondente.

Figura 03 – Dicionário do atributo COR_RACA



Fonte: Elaborado pelo autor

3.1.4 Processamento Analítico

O processamento analítico consiste na aplicação de métodos em um conjunto de dados a fim de produzir informações relevantes. Este processo envolve a aplicação de algoritmos computacionais e técnicas estatísticas para explorar, modelar e analisar os dados. O processamento analítico pode incluir uma variedade

de técnicas, como análise descritiva, análise exploratória de dados, análise estatística, mineração de dados, modelagem preditiva, entre outros.

Inicialmente foi feita uma análise descritiva inicial. Foram feitos gráficos de frequência de cada atributo categórico, que representa uma variável socioeconômica do estudante, em função das notas de cada uma das provas ao mesmo tempo em que se calcularam a média, desvio padrão, mediana, valor mínimo, valor máximo e percentil dos mesmos. Também foi feita um teste de aderência comparando a distribuição das notas para cada uma das variáveis socioeconômicas a fim de se verificar se os dados apresentam uma distribuição normal. Com isso, é possível efetuar uma análise visual e obter uma ideia de como os dados são relacionados e com isso definir quais métodos podem ser aplicados para confirmar tais relações além de propor a execução de determinados métodos posteriormente.

Para construir os gráficos foi utilizada a linguagem *python* e as bibliotecas 'pandas', 'numpy', 'os' e 'matplotlib.pyplot'.

Após efetuada a análise descritiva, foram efetuados métodos estatísticos e métodos de aprendizagem de máquina com o objetivo de analisar relações e padrões entre diferentes variáveis além de criar modelos de previsão.

4 RESULTADOS

Este capítulo apresenta os resultados obtidos a partir da aplicação de diferentes métodos de análise de dados.

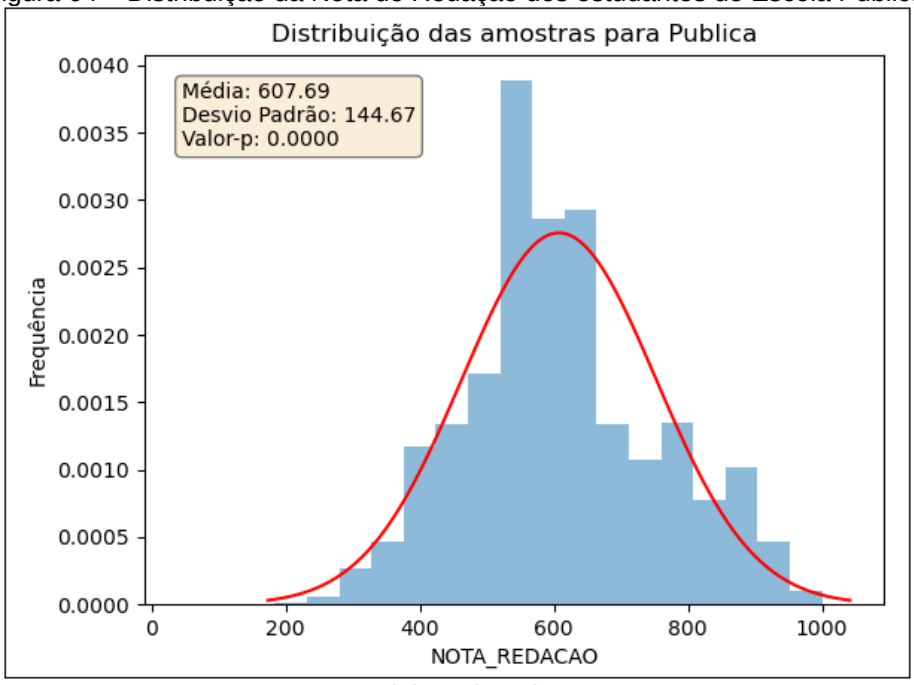
4.1 Teste de Normalidade

Para testar se a distribuição dos dados das notas de cada prova do ENEM apresenta semelhança significativa com a distribuição normal foi executado o algoritmo de teste de aderência de Kolmogorov-Smirnov em *python*.

Também foi implementado no algoritmo a geração de um gráfico de frequência dos dados junto com a curva normal esperada a fim de disponibilizar uma visualização gráfica do teste aplicado.

A Figura 04 apresenta o gráfico de frequência das notas de redação de todos os estudantes cujo tipo de escola em que estudam, ou estudaram, o ensino médio seja Pública junto com a curva normal esperada.

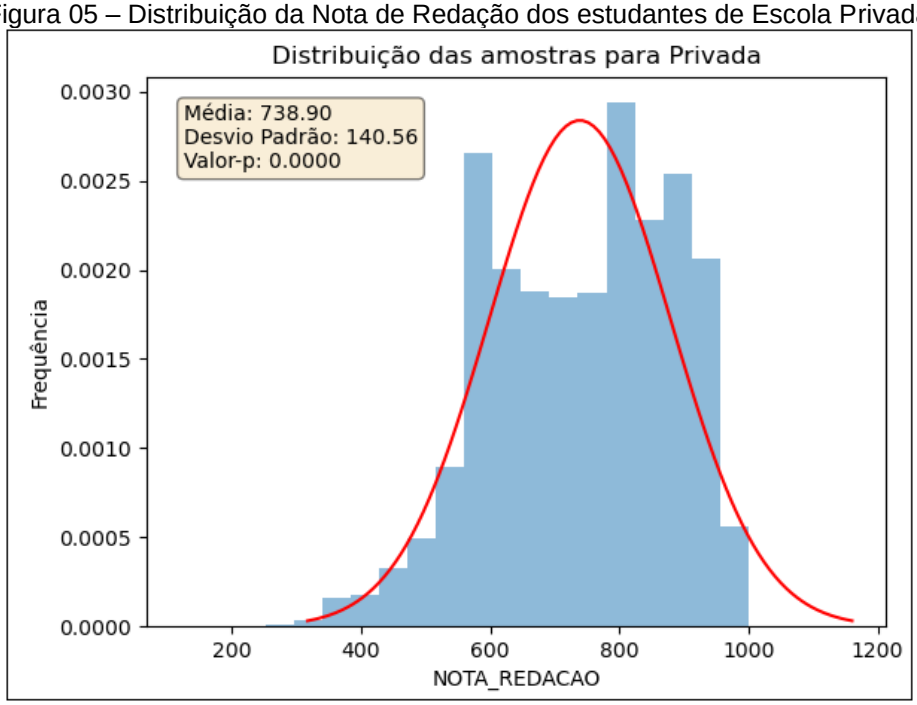
Figura 04 – Distribuição da Nota de Redação dos estudantes de Escola Pública



Fonte: Elaborado pelo autor

A Figura 05 apresenta o gráfico de frequência das notas de redação de todos os estudantes cujo tipo de escola em que estudam, ou estudaram, o ensino médio seja Privada junto com a curva normal esperada.

Figura 05 – Distribuição da Nota de Redação dos estudantes de Escola Privada



Fonte: Elaborado pelo autor

Como o valor *p* obtido no teste é próximo de zero, rejeita-se a hipótese de que os dados das notas seguem uma distribuição normal, ou seja, conclui-se que os dados não apresentam normalidade.

Este mesmo resultado foi obtido para todas as demais categorias.

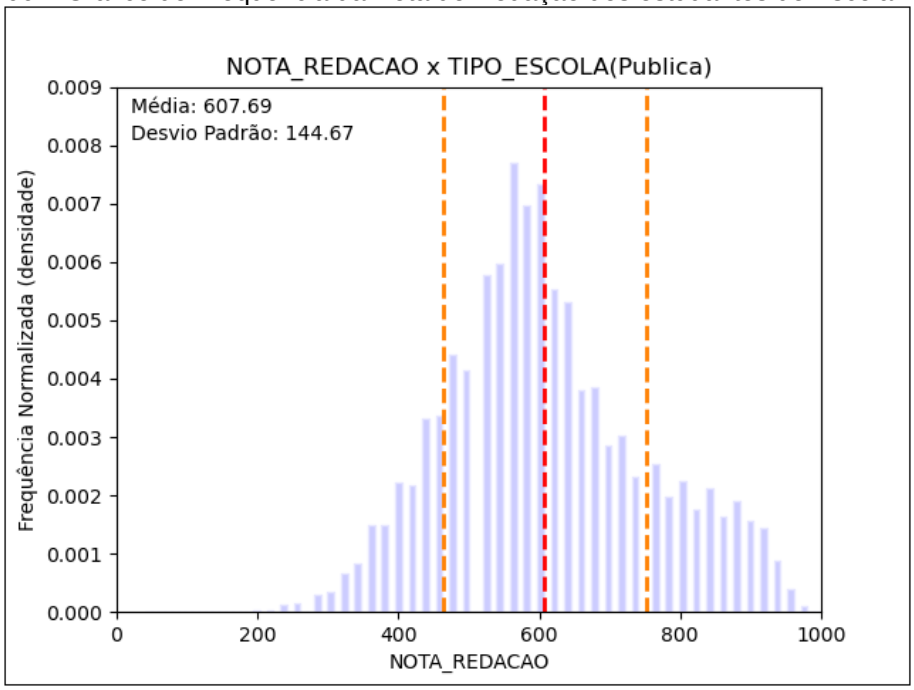
4.2 Análise de Frequência

Primeiramente, foram construídos gráficos de frequência das notas de cada uma das provas em função das diferentes variáveis socioeconômicas assim como o cálculo das médias e desvios padrões.

A Figura 06 apresenta o gráfico de frequência das notas de redação de todos os estudantes cujo tipo de escola em que estudam, ou estudaram, o ensino médio, seja pública.

A linha vertical pontilhada vermelha indica a média e as linhas laranjadas pontilhadas indicam a distância da média igual ao desvio padrão.

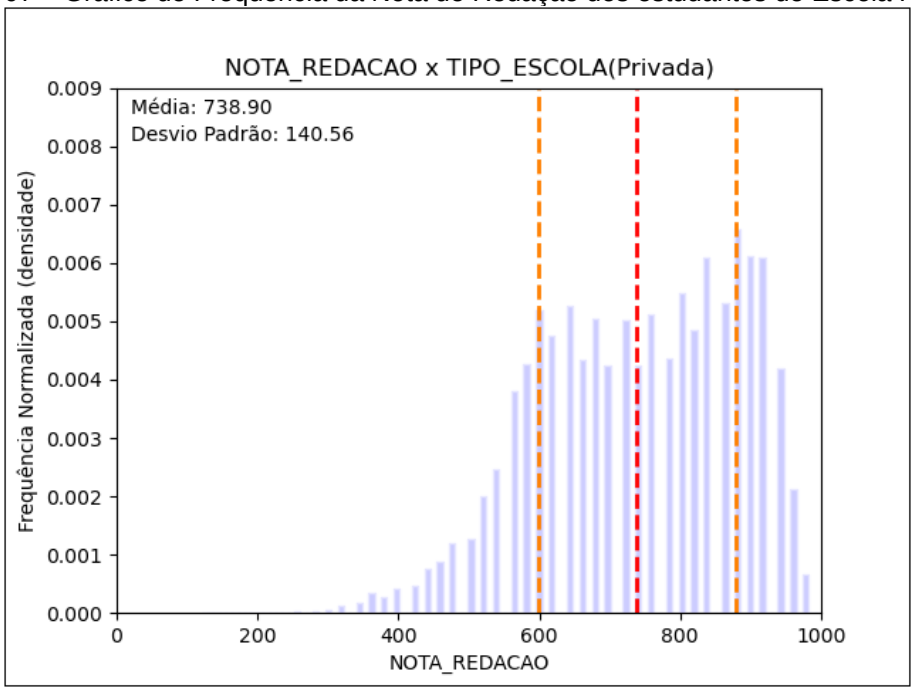
Figura 06 – Gráfico de Frequência da Nota de Redação dos estudantes de Escola Pública



Fonte: Elaborado pelo autor

A Figura 07 apresenta o gráfico de frequência das notas de redação de todos os estudantes cujo tipo de escola em que estudam, ou estudaram, o ensino médio, seja privada.

Figura 07 – Gráfico de Frequência da Nota de Redação dos estudantes de Escola Privada



Fonte: Elaborado pelo autor

O eixo X corresponde à nota de redação enquanto o eixo y corresponde à frequência normalizada (densidade) da quantidade de estudantes que atingiram cada uma das notas. O eixo Y foi normalizado para que fosse possível fazer comparações. A variável “tipo de escola”, por exemplo, apresenta muito mais alunos de escola pública do que de alunos de escola privada. Sendo assim, não seria possível fazer uma comparação de desempenho destes dois grupos se os gráficos não fossem normalizados.

É possível perceber visualmente que o grupo de alunos de escola privada apresenta melhor rendimento uma vez que as barras de frequência são situadas mais a direita do gráfico. Além disso, é possível perceber a diferença das médias de cada um dos grupos.

Essa característica também pode ser observada nas demais provas do exame conforme são mostradas na Figura-08.

A Tabela 01 apresenta a média, desvio padrão, número de amostra, valor mínimo, valor máximo e percentil das notas de redação de cada um dos grupos de estudantes formados pelas diversas categorias.

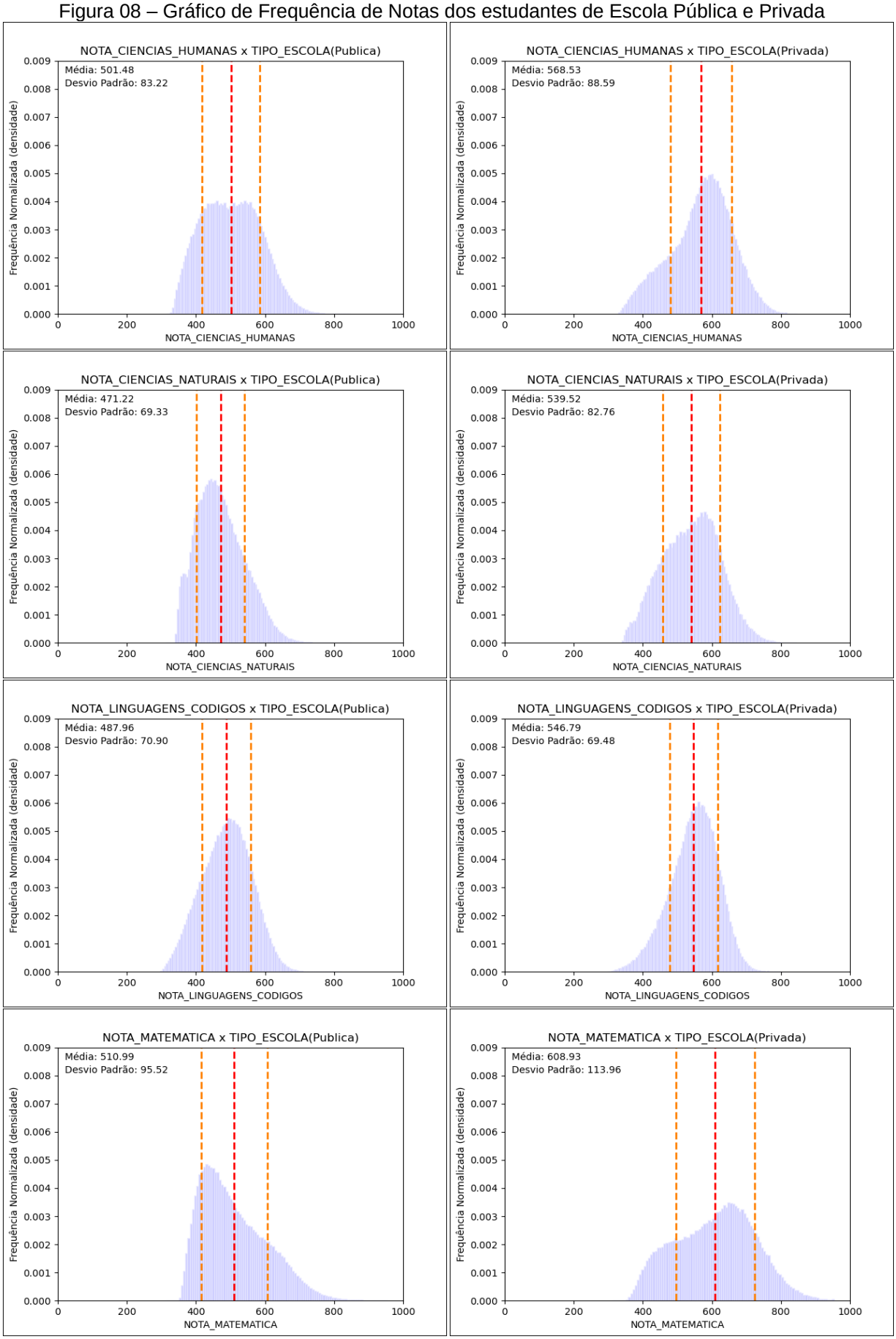
Ao analisar a média das notas em função dos atributos “Escolaridade da mãe”, “Escolaridade do pai” e “Renda familiar”, é possível perceber que há um aumento da média de acordo com o aumento da escolaridade e da renda.

Como estes atributos são categorias ordinais, suas categorias foram substituídas por números inteiros para que assim fosse possível aplicar o método de regressão linear.

4.3 Aplicação de Regressão Linear

A Figura 09 apresenta o gráfico obtido com a aplicação do método de regressão linear entre a Escolaridade da mãe e a média da nota de redação do estudante. O valor de R^2 = 97,3% indica que a média da nota de redação de cada grupo formado por cada uma das categorias de escolaridades da mãe apresenta forte relação linear.

Figura 08 – Gráfico de Frequência de Notas dos estudantes de Escola Pública e Privada



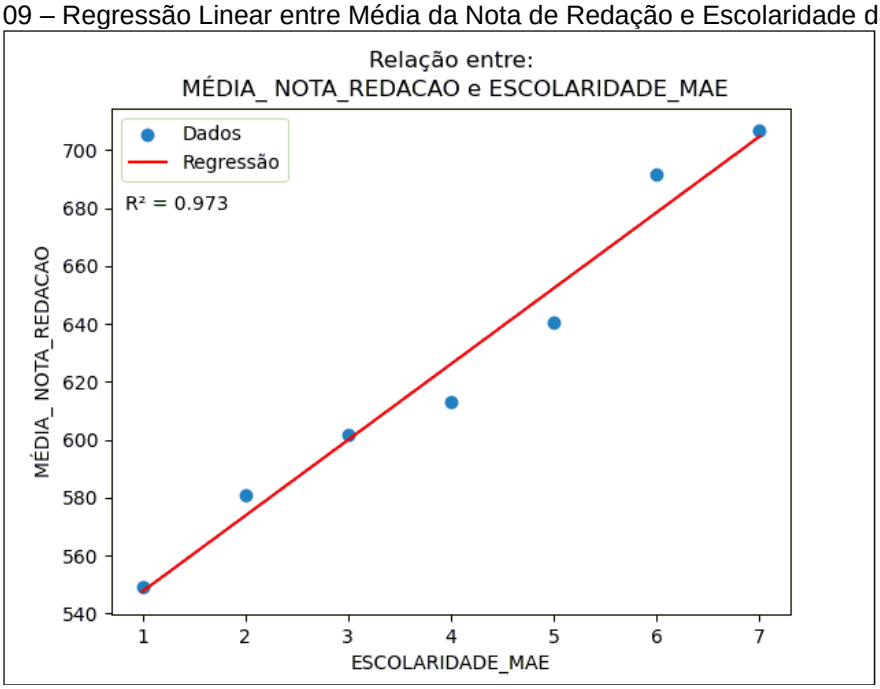
Fonte: Elaborado pelo autor

Tabela 01 – Dados estatísticos da Nota de Redação

CATEGORIAS	MÉDIA	DESVIO PADRÃO	NÚMERO DE OCORRÊNCIAS	VALOR MÍNIMO	PERCENTIL			VALOR MÁXIMO
					25	50	75	
SEXO (F)	649,52	154,08	1346711	40	540	620	760	1000
SEXO (M)	625,89	149,38	826820	40	520	600	720	1000
COR_RACA (Amarela)	635,74	151,97	43686	80	520	620	740	980
COR_RACA (Branca)	670,67	152,57	958750	40	560	660	800	1000
COR_RACA (Indigena)	560,15	137,04	9866	40	480	560	640	980
COR_RACA (Parda)	618,73	149,50	880915	40	520	600	720	1000
COR_RACA (Preta)	606,17	142,76	238470	40	520	580	680	980
TIPO_ESCOLA (Privada)	738,90	140,56	177894	120	620	760	860	1000
TIPO_ESCOLA (Publica)	607,69	144,67	640582	40	520	600	700	1000
ESCOLARIDADE_PAI (A)	564,22	138,94	72162	40	480	560	640	980
ESCOLARIDADE_PAI (B)	592,08	141,51	336605	40	500	580	660	980
ESCOLARIDADE_PAI (C)	615,17	146,02	272217	40	520	600	700	1000
ESCOLARIDADE_PAI (D)	628,23	147,60	238101	40	520	600	720	1000
ESCOLARIDADE_PAI (E)	651,58	149,39	665370	40	540	640	760	1000
ESCOLARIDADE_PAI (F)	705,65	148,78	242701	40	600	700	840	1000
ESCOLARIDADE_PAI (G)	724,40	147,53	180835	40	600	720	860	1000
ESCOLARIDADE_PAI (H)	600,23	142,32	165540	40	500	580	680	980
ESCOLARIDADE_MAE (A)	548,93	133,42	45020	40	460	540	620	980
ESCOLARIDADE_MAE (B)	580,72	138,03	233287	40	500	560	660	980
ESCOLARIDADE_MAE (C)	601,53	142,31	217775	40	500	580	680	1000
ESCOLARIDADE_MAE (D)	613,19	144,23	237339	40	520	600	700	980
ESCOLARIDADE_MAE (E)	640,40	147,94	774398	40	540	620	740	1000
ESCOLARIDADE_MAE (F)	691,44	151,14	315561	40	580	680	820	1000
ESCOLARIDADE_MAE (G)	706,57	151,42	304382	40	600	700	840	1000
ESCOLARIDADE_MAE (H)	574,95	141,41	45769	40	480	560	640	980
RENDA_FAMILIAR (A)	566,66	138,54	106374	40	480	560	640	980
RENDA_FAMILIAR (B)	591,30	141,97	522989	40	500	580	680	1000
RENDA_FAMILIAR (C)	614,55	143,08	336269	40	520	600	700	1000
RENDA_FAMILIAR (D)	627,68	144,53	270374	40	540	600	720	980
RENDA_FAMILIAR (E)	644,05	147,06	140529	40	540	620	740	1000
RENDA_FAMILIAR (F)	656,90	147,58	164596	40	560	640	760	980
RENDA_FAMILIAR (G)	672,97	148,50	144957	40	560	660	800	1000
RENDA_FAMILIAR (H)	689,19	148,25	116862	40	580	680	820	1000
RENDA_FAMILIAR (I)	702,09	148,30	74295	40	600	700	840	1000
RENDA_FAMILIAR (J)	711,04	147,31	46289	40	600	700	840	1000
RENDA_FAMILIAR (K)	717,53	146,11	36638	200	600	720	840	980
RENDA_FAMILIAR (L)	722,01	145,93	32131	200	600	720	840	1000
RENDA_FAMILIAR (M)	729,25	143,40	47233	180	620	740	860	1000
RENDA_FAMILIAR (N)	739,75	141,82	29640	200	620	740	860	1000
RENDA_FAMILIAR (O)	746,72	139,86	32469	220	640	760	860	1000
RENDA_FAMILIAR (P)	753,30	138,76	32427	120	640	760	880	1000
RENDA_FAMILIAR (Q)	760,47	135,55	39459	160	660	780	880	1000

Fonte: Elaborado pelo autor

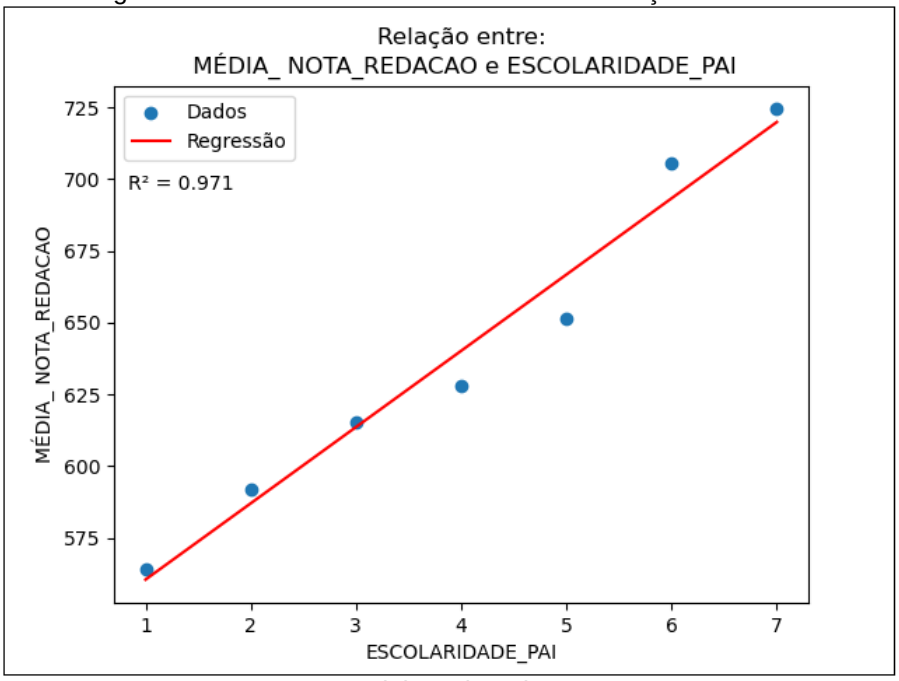
Figura 09 – Regressão Linear entre Média da Nota de Redação e Escolaridade da Mãe



Fonte: Elaborado pelo autor

A Figura 10 apresenta o gráfico obtido com a aplicação do método de regressão linear entre a Escolaridade do pai e a média da nota de redação do estudante. O valor de R^2 = 97,1% indica que a média da nota de redação de cada grupo formado por cada uma das categorias de escolaridades do pai apresenta forte relação linear.

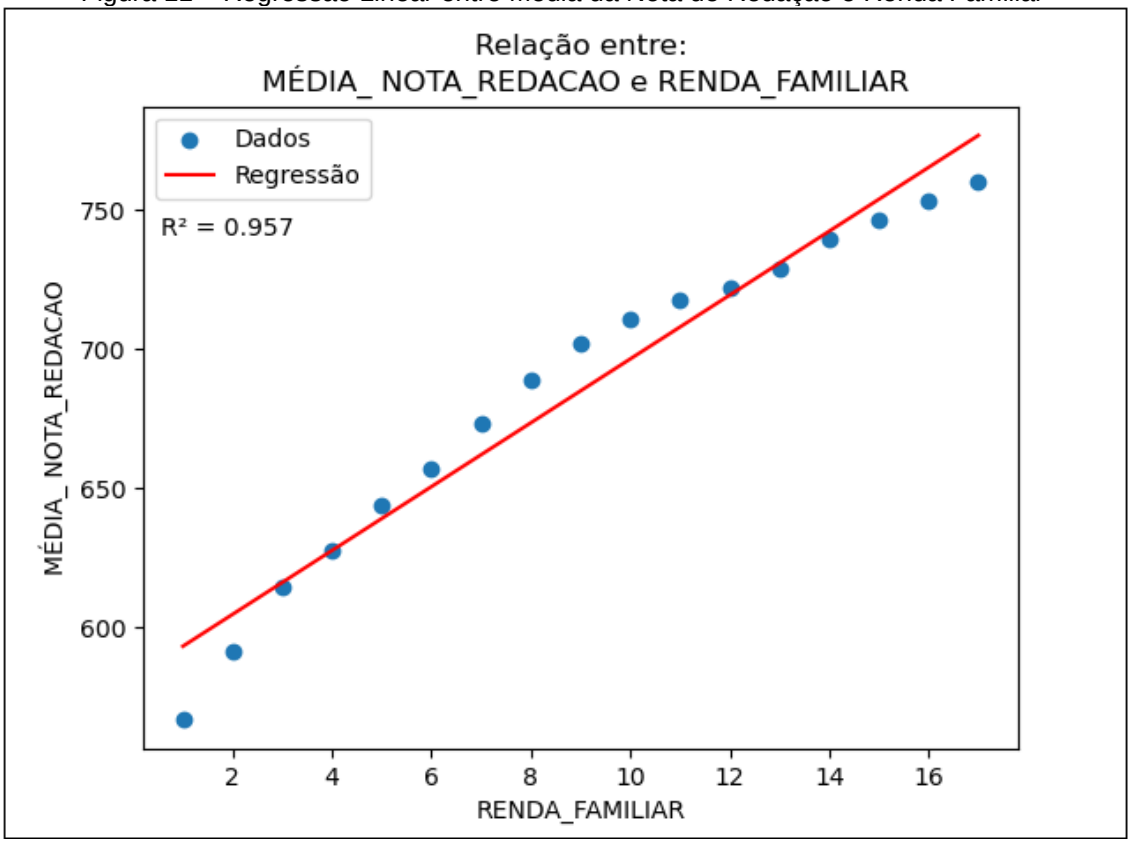
Figura 10 – Regressão Linear entre Média da Nota de Redação e Escolaridade do Pai



Fonte: Elaborado pelo autor

A Figura 11 apresenta o gráfico obtido com a aplicação do método de regressão linear entre a Renda familiar do estudante e a média da nota de redação do estudante. O valor de R^2 = 95,7% indica que a média da nota de redação de cada grupo formado por cada uma das categorias de Renda familiar apresenta forte relação linear.

Figura 11 – Regressão Linear entre Média da Nota de Redação e Renda Familiar



Fonte: Elaborado pelo autor

Esta forte relação também pode ser observada ao se analisar as notas das provas objetivas.

4.4 Mapa de calor

A Tabela 02 apresenta a média, o desvio padrão, o número de amostra, o valor mínimo, o valor máximo e o percentil das notas de redação de cada um dos grupos de estudantes formados pelo Estado em que residem. As linhas foram ordenadas de acordo com o valor da média.

Tabela 02 – Dados estatísticos da Nota de Redação por Estado

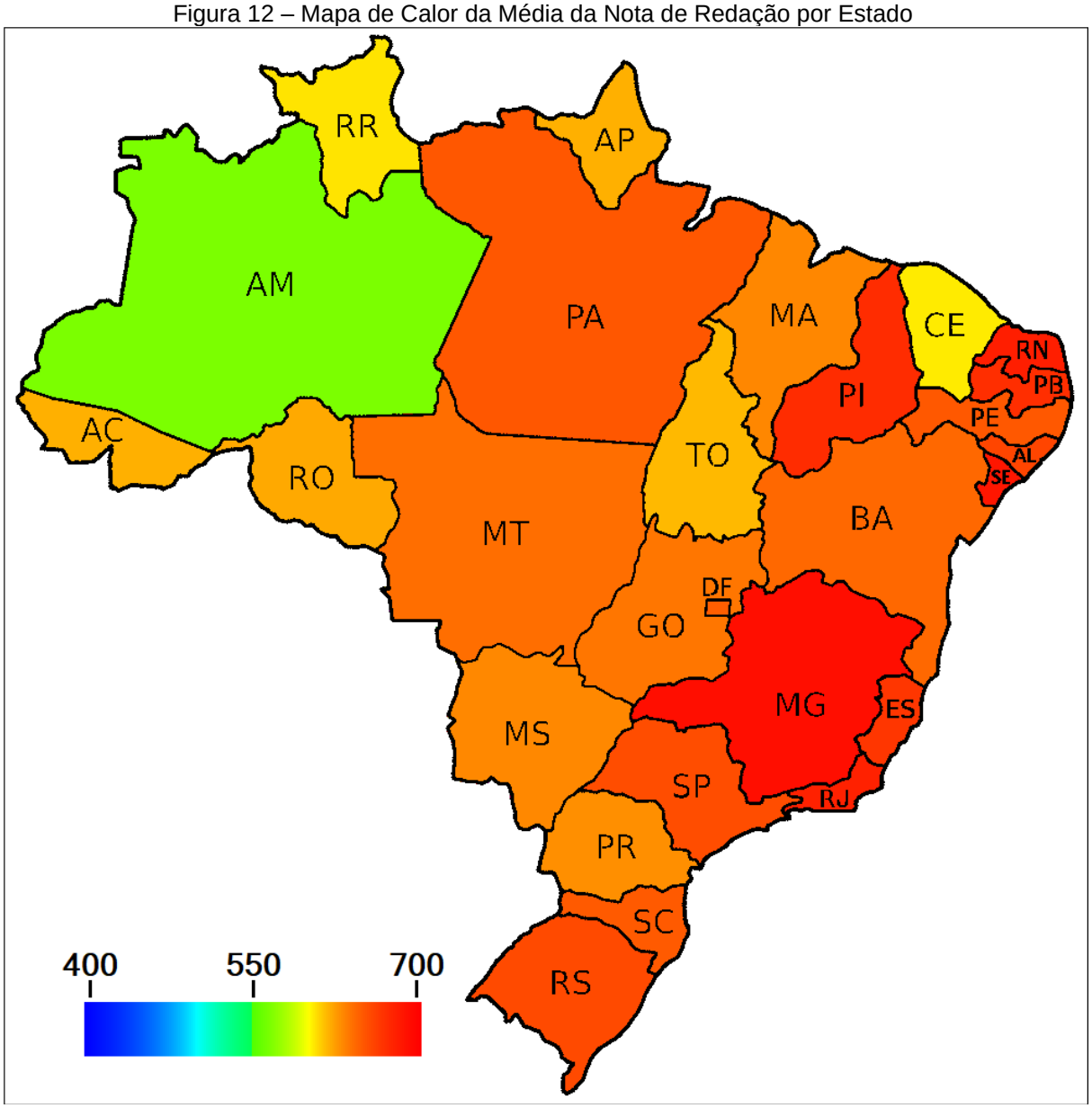
CATEGORIAS	MÉDIA	DESVIO PADRÃO	NÚMERO DE OCORRÊNCIAS	VALOR MÍNIMO	PERCENTIL			VALOR MÁXIMO
					25	50	75	
UF(AM)	564,92	138,98	14595	40	480	560	640	980
UF(CE)	604,03	165,80	53479	40	480	580	720	1000
UF(RR)	605,56	144,60	1161	240	500	600	700	960
UF(TO)	615,79	154,24	4358	60	500	600	720	980
UF(AC)	618,28	157,29	2170	80	520	600	720	980
UF(AP)	618,32	150,59	2040	200	520	600	700	980
UF(RO)	620,24	150,90	4715	40	520	600	720	980
UF(PR)	628,21	139,83	32408	60	540	600	720	980
UF(MS)	630,44	148,42	7839	160	520	600	740	980
UF(MA)	630,63	160,83	15675	40	520	600	760	980
UF(GO)	636,70	161,79	23767	40	520	620	760	1000
UF(MT)	639,61	156,24	9242	80	540	620	760	980
UF(BA)	642,16	156,95	33898	40	540	620	760	1000
UF(DF)	646,05	144,32	14175	140	540	620	740	980
UF(SC)	647,92	141,28	19515	120	560	620	740	980
UF(PE)	648,75	159,31	28692	60	540	620	780	980
UF(PA)	649,05	161,71	20276	40	540	620	780	980
UF(SP)	653,64	142,37	111229	80	560	640	760	1000
UF(RS)	654,87	146,97	27454	40	560	640	760	980
UF(AL)	657,40	158,73	8385	40	540	640	780	980
UF(ES)	664,87	155,38	13553	40	560	640	800	980
UF(PB)	669,61	165,11	12998	40	540	660	820	980
UF(PI)	670,51	173,46	10927	40	540	660	820	980
UF(RJ)	677,08	156,66	43397	40	560	660	820	1000
UF(RN)	678,96	158,62	10719	80	560	660	820	980
UF(SE)	683,98	167,15	6542	140	560	680	840	980
UF(MG)	689,82	160,77	49458	120	560	680	840	1000

Fonte: Elaborado pelo autor

Para apresentar visualmente a média das notas dos estudantes de acordo com cada Estado foram construídos Mapas de Calor, pois assim é possível perceber visualmente o desempenho médio dos estudantes de cada um dos Estados do Brasil.

A Figura 12 apresenta o mapa de calor correspondente à média da nota de redação de cada um dos Estados brasileiros.

Figura 12 – Mapa de Calor da Média da Nota de Redação por Estado



Fonte: Elaborado pelo autor

É possível perceber que o Estado de Minas Gerais (MG) apresenta o melhor rendimento médio na prova de redação enquanto o Estado do Amazonas (AM) apresenta o pior rendimento.

O Estado de Minas Gerais (MG) apresentou melhor rendimento médio não só na prova de redação como também nas demais provas.

A Tabela 03 apresenta a média, o desvio padrão, o número de amostra, o valor mínimo, o valor máximo e o percentil das notas da prova de Ciências Humanas de cada um dos grupos de estudantes formados pelo Estado em que residem. As linhas foram ordenadas de acordo com o valor da média.

Tabela 03 – Dados estatísticos da Nota de Ciências Humanas por Estado

CATEGORIAS	MÉDIA	DESVIO PADRÃO	NÚMERO DE OCORRÊNCIAS	VALOR MÍNIMO	PERCENTIL			VALOR MÁXIMO
					25	50	75	
UF(AM)	478,77	81,44	14595	328,9	415,25	467,30	536,40	805,6
UF(CE)	491,97	87,34	53479	328,9	422,10	482,60	555,15	846,9
UF(MA)	492,82	86,36	15675	328,9	423,45	483,80	556,60	811
UF(TO)	494,15	84,59	4358	328,9	426,30	484,70	555,90	832,5
UF(AP)	498,00	85,18	2040	330,1	431,50	494,50	561,23	756,2
UF(RO)	499,72	82,07	4715	331,3	435,20	498,50	559,75	792,7
UF(PA)	501,81	87,31	20276	328,9	431,60	496,00	565,50	819,1
UF(AC)	503,42	81,40	2170	328,9	440,20	503,70	563,20	749,8
UF(RR)	508,83	82,34	1161	337,9	443,20	507,90	568,50	790,5
UF(PI)	509,41	91,92	10927	330,5	435,55	504,30	575,35	827
UF(AL)	510,02	88,50	8385	328,9	439,10	507,80	575,60	811
UF(PE)	510,88	89,88	28692	324,5	439,00	508,80	576,20	832,5
UF(BA)	513,37	91,46	33898	326,3	438,73	512,05	582,20	832,5
UF(PB)	513,47	91,64	12998	325,4	439,80	510,90	581,10	814,3
UF(GO)	514,76	90,30	23767	328,9	442,20	513,80	581,05	846,9
UF(MT)	515,61	86,26	9242	328,9	447,60	517,65	579,40	824
UF(MS)	515,82	87,76	7839	328,9	446,40	516,30	581,30	818,7
UF(SE)	516,22	94,41	6542	330,3	438,70	515,65	585,70	816,1
UF(RN)	530,84	91,82	10719	328,9	456,50	535,60	599,50	803,2
UF(ES)	530,93	89,05	13553	329	462,90	536,20	596,00	819,1
UF(PR)	535,41	87,01	32408	328,9	470,50	542,80	598,20	824
UF(DF)	538,23	91,02	14175	328,9	469,30	543,10	603,60	824
UF(RS)	541,58	86,99	27454	328,9	477,80	548,80	603,80	846,9
UF(SC)	541,84	85,87	19515	330	479,55	549,90	602,20	846,9
UF(RJ)	542,62	90,30	43397	328,9	474,70	550,20	608,10	846,9
UF(SP)	545,92	85,61	111229	328,9	485,20	553,10	606,00	846,9
UF(MG)	547,67	92,14	49458	328,9	479,13	555,10	614,40	846,9

Fonte: Elaborado pelo autor

A Tabela 04 apresenta a média, o desvio padrão, o número de amostra, o valor mínimo, o valor máximo e o percentil das notas da prova de Ciências da Natureza de cada um dos grupos de estudantes formados pelo Estado em que residem. As linhas foram ordenadas de acordo com o valor da média.

Tabela 04 – Dados estatísticos da Nota de Ciências da Natureza por Estado

CATEGORIAS	MÉDIA	DESVIO PADRÃO	NÚMERO DE OCORRÊNCIAS	VALOR MÍNIMO	PERCENTIL			VALOR MÁXIMO
					25	50	75	
UF(AM)	454,27	66,93	14595	338,5	404,65	444,10	492,60	790,7
UF(CE)	465,40	73,75	53479	339,7	410,50	453,60	508,90	815,1
UF(AP)	467,15	69,60	2040	340,7	414,50	459,95	510,40	721,6
UF(MA)	467,49	70,89	15675	332,8	414,40	458,00	512,50	818

UF(AC)	470,08	70,23	2170	341,6	416,60	462,10	516,33	743,3
UF(TO)	471,22	72,33	4358	342	415,50	462,55	517,20	788,2
UF(PA)	473,47	73,13	20276	336,2	418,50	464,00	519,90	810,1
UF(RO)	475,34	69,30	4715	335,5	423,35	468,30	520,80	743
UF(RR)	480,84	68,63	1161	344,1	430,10	474,20	529,00	729,7
UF(AL)	481,33	74,47	8385	341,7	425,20	472,40	532,00	844,5
UF(PI)	483,61	81,42	10927	340,8	421,70	471,50	535,65	835,1
UF(PE)	484,99	78,20	28692	324,5	424,90	475,60	538,50	799,4
UF(BA)	485,05	76,04	33898	315,3	427,40	475,90	536,60	819,4
UF(PB)	485,92	79,18	12998	329,7	425,60	476,10	539,50	808,5
UF(GO)	485,98	79,76	23767	341,5	425,40	475,10	538,85	822
UF(MT)	487,93	74,88	9242	341,4	430,50	481,20	539,40	796,6
UF(MS)	489,38	78,68	7839	341,5	429,20	479,40	541,95	837,2
UF(SE)	489,82	81,34	6542	342,8	427,70	479,00	545,80	808,1
UF(RN)	500,04	81,03	10719	335,7	437,20	492,30	558,30	834,4
UF(ES)	503,85	81,01	13553	343,1	440,90	497,60	562,30	839
UF(PR)	504,79	79,36	32408	340	443,80	499,15	562,32	802
UF(RS)	505,46	76,91	27454	340,9	447,00	501,70	561,00	818
UF(RJ)	506,93	81,18	43397	312,3	444,20	501,60	566,90	820,8
UF(DF)	507,00	82,25	14175	341	442,40	500,10	566,70	808,4
UF(SP)	512,21	81,66	111229	333,9	449,30	507,50	571,40	836,9
UF(SC)	513,36	77,27	19515	315,5	455,30	511,50	569,80	813,8
UF(MG)	514,91	83,78	49458	340,7	450,00	511,00	577,40	848,7

Fonte: Elaborado pelo autor

A Tabela 05 apresenta a média, o desvio padrão, o número de amostra, o valor mínimo, o valor máximo e o percentil das notas da prova de Linguagens e Códigos de cada um dos grupos de estudantes formados pelo Estado em que residem. As linhas foram ordenadas de acordo com o valor da média.

Tabela 05 – Dados estatísticos da Nota de Linguagens e Códigos por Estado

CATEGORIAS	MÉDIA	DESVIO PADRÃO	NÚMERO DE OCORRÊNCIAS	VALOR MÍNIMO	PERCENTIL			VALOR MÁXIMO
					25	50	75	
UF(AM)	466,83	73,95	14595	298	410,60	464,90	519,00	723,6
UF(MA)	476,90	73,79	15675	299,5	422,20	476,00	529,30	761,4
UF(TO)	479,59	73,27	4358	306,1	425,33	478,80	531,08	705,5
UF(AP)	480,08	75,19	2040	308,3	424,80	477,90	533,60	731,8
UF(PA)	480,15	75,52	20276	300,4	423,80	479,50	533,50	761,4
UF(CE)	481,15	74,45	53479	298,3	426,80	481,60	533,10	760,5
UF(RO)	484,71	69,92	4715	298	433,90	485,50	534,55	725,8
UF(AC)	488,83	70,73	2170	304	438,40	491,85	539,15	685,1
UF(PI)	492,35	79,16	10927	299,6	434,20	492,20	547,50	746
UF(AL)	493,46	74,33	8385	299,7	440,20	494,00	547,50	725,2
UF(PB)	494,64	76,01	12998	298,4	439,50	495,30	548,40	763,6

UF(MT)	495,26	72,99	9242	298	443,00	496,50	547,50	793,5
UF(RR)	495,58	70,01	1161	311,5	448,80	497,90	543,10	687,9
UF(SE)	496,89	79,56	6542	301	437,53	499,10	554,20	743,4
UF(BA)	497,10	76,26	33898	298,3	442,00	499,00	552,20	781,6
UF(PE)	498,83	74,25	28692	302,3	446,80	499,50	550,50	777
UF(GO)	499,61	74,87	23767	301,2	446,80	500,10	551,40	784,5
UF(MS)	501,72	72,93	7839	305,8	450,00	503,10	554,30	747,4
UF(RN)	511,34	75,63	10719	301,5	458,10	515,60	565,75	820,5
UF(ES)	513,02	71,62	13553	300,8	463,90	514,80	563,00	792,4
UF(PR)	517,85	70,60	32408	301,9	470,90	521,40	568,30	774,5
UF(SC)	519,16	71,86	19515	298,2	470,70	524,10	570,90	756,4
UF(RS)	523,25	71,50	27454	298	475,40	527,00	574,68	767,5
UF(RJ)	524,48	72,74	43397	301,3	476,20	528,70	576,50	777,2
UF(DF)	525,46	72,93	14175	305	476,90	529,40	577,30	788,6
UF(MG)	528,09	74,30	49458	301,4	478,60	532,80	580,80	790,8
UF(SP)	530,95	67,75	111229	301,7	487,60	534,70	578,10	799,3

Fonte: Elaborado pelo autor

A Tabela 06 apresenta a média, o desvio padrão, o número de amostra, o valor mínimo, o valor máximo e o percentil das notas da prova de Matemática de cada um dos grupos de estudantes formados pelo Estado em que residem. As linhas foram ordenadas de acordo com o valor da média.

Tabela 06 - Dados estatísticos da Nota de Matemática por Estado

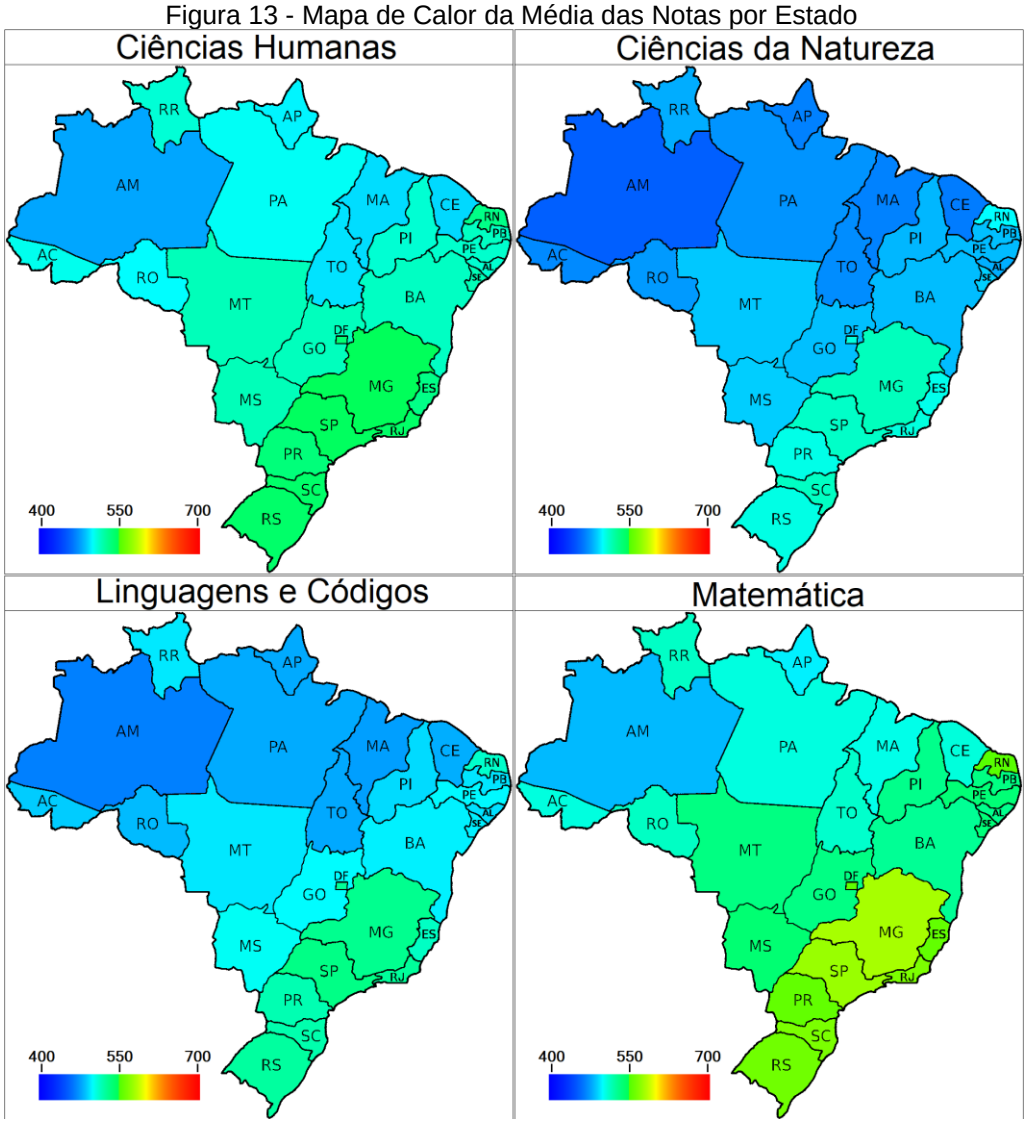
CATEGORIAS	MÉDIA	DESVIO PADRÃO	NÚMERO DE OCORRÊNCIAS	VALOR MÍNIMO	PERCENTIL			VALOR MÁXIMO
					25	50	75	
UF(AM)	484,38	88,08	14595	321	419,10	461,50	529,05	953,1
UF(AP)	497,28	89,36	2040	345,2	427,53	480,55	548,40	861,4
UF(MA)	504,47	98,56	15675	325	427,70	480,90	564,70	920,5
UF(PA)	506,52	97,18	20276	314,3	429,80	484,80	567,50	913,3
UF(AC)	506,73	90,98	2170	334,2	434,40	488,70	566,98	831,6
UF(CE)	507,26	103,18	53479	353,8	427,20	480,30	569,30	953,1
UF(TO)	511,04	97,99	4358	353,8	432,30	488,60	576,85	884,1
UF(RR)	512,64	95,84	1161	353,8	435,10	494,50	576,60	866,3
UF(RO)	513,17	94,45	4715	322,2	438,00	494,00	576,15	889,8
UF(BA)	526,27	104,81	33898	311,8	440,60	506,20	600,28	953,1
UF(AL)	529,25	106,82	8385	338,8	440,90	510,90	605,40	948,5
UF(PI)	529,53	114,29	10927	346,7	437,40	503,60	605,90	953,1
UF(GO)	531,78	107,95	23767	336,5	442,60	511,80	608,60	953,1
UF(PB)	532,22	110,21	12998	353,8	442,23	509,20	608,40	944,1
UF(MT)	532,54	103,93	9242	344,8	448,30	515,80	607,30	944,1
UF(SE)	535,68	108,32	6542	318,6	446,03	517,25	613,78	953,1
UF(PE)	536,55	110,33	28692	340,8	445,90	516,65	616,40	953,1
UF(MS)	538,09	108,04	7839	343,2	449,10	519,80	613,40	948,5

UF(RN)	551,42	113,20	10719	326,3	456,95	536,60	635,80	944,1
UF(DF)	554,14	112,96	14175	341,9	460,70	540,30	637,30	953,1
UF(PR)	555,52	109,83	32408	335,1	464,80	544,20	637,10	953,1
UF(ES)	555,70	112,63	13553	345,1	461,00	544,80	638,90	953,1
UF(RS)	561,00	109,16	27454	335	470,10	554,40	643,10	953,1
UF(SC)	565,16	109,55	19515	327,7	474,50	559,80	646,90	953,1
UF(RJ)	565,46	114,06	43397	312,9	469,40	557,80	651,70	953,1
UF(SP)	575,14	110,22	111229	317,6	484,40	573,40	656,40	953,1
UF(MG)	579,15	117,09	49458	332,5	480,90	577,00	667,50	953,1

Fonte: Elaborado pelo autor

A Figura 13 apresenta quatro mapas de calor correspondentes à média das notas de Ciências Humanas, Ciências da Natureza, Linguagens e códigos e Matemática, correspondentes a cada um dos Estados brasileiros.

Figura 13 - Mapa de Calor da Média das Notas por Estado



Fonte: Elaborado pelo autor

Embora os rendimentos médios dos Estados apresentem diferenças entre si, pode-se notar que as notas máximas e mínimas são muito próximas, ou seja, em todos os Estados houve candidatos que apresentaram resultados muito baixos e muito altos em todas as provas.

4.5 Aplicação do método X²

Como foi observado anteriormente, o resultado dos estudantes de escola privada apresentou um rendimento médio melhor que o resultado dos estudantes de escola pública. Da mesma forma, percebeu-se um melhor rendimento médio dos estudantes cuja renda familiar é mais alta.

A Tabela 07 apresenta os dados dos grupos formados pelas categorias Tipo de Escola e Renda Familiar:

Tabela 07 – Frequência da Renda familiar x tipo de Escola

		RENDA_FAMILIAR																	TOTAL
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
Tipo de Escola	Privada	1163	6578	9137	12512	9436	14252	16019	17655	12946	8719	7702	7200	12073	7931	9637	10460	14474	177894
	Publica	39635	190964	111685	87595	44881	49383	39614	27193	15244	8666	6187	4691	5711	3063	2740	1941	1389	640582
TOTAL		40798	197542	120822	100107	54317	63635	55633	44848	28190	17385	13889	11891	17784	10994	12377	12401	15863	818476

Fonte: Elaborado pelo autor

Cada uma das letras de A até Q representam uma determinada faixa de renda. O Quadro 07 apresenta cada uma das rendas de acordo com cada uma das letras.

Quadro 07 – Relação entre as Letras e a Renda

Letra	Renda mensal (R$)
A	Nenhuma Renda
B	Até 1.100
C	De 1.100 até 1.650
D	De 1.650 até 2.200
E	De 2.200 até 2.750
F	De 2.750 até 3.300
G	De 3.300 até 4.400
H	De 4.400 até 5.500
I	De 5.500 até 6.600
J	De 6.600 até 7.700
K	De 7.700 até 8.800
L	De 8.800 até 9.900

M	De 9.900 até 11.000
N	De 11.000 até 13.200
O	De 13.200 até 16.500
P	De 16.500 até 22.000
Q	Acima de 22.000.

Fonte: Elaborado pelo autor

É possível perceber a partir dos dados que o número de candidatos de escola privada cresce ao mesmo tempo em que candidatos de escola pública diminui conforme cresce a renda familiar.

A Tabela 08 apresenta os mesmos dados, mas de forma relativa, ou seja, mostra a porcentagem de alunos de escola pública e privada para cada uma das rendas.

Tabela 08 – Frequência normalizada da Renda familiar x tipo de Escola

		RENDA_FAMILIAR																TOTAL
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
Tipo de Escola	Privada	2,9%	3,3%	7,6%	12,5%	17,4%	22,4%	28,8%	39,4%	45,9%	50,2%	55,5%	60,5%	67,9%	72,1%	77,9%	84,3%	91,2%	21,7%
	Publica	97,1%	96,7%	92,4%	87,5%	82,6%	77,6%	71,2%	60,6%	54,1%	49,8%	44,5%	39,5%	32,1%	27,9%	22,1%	15,7%	8,8%	78,3%

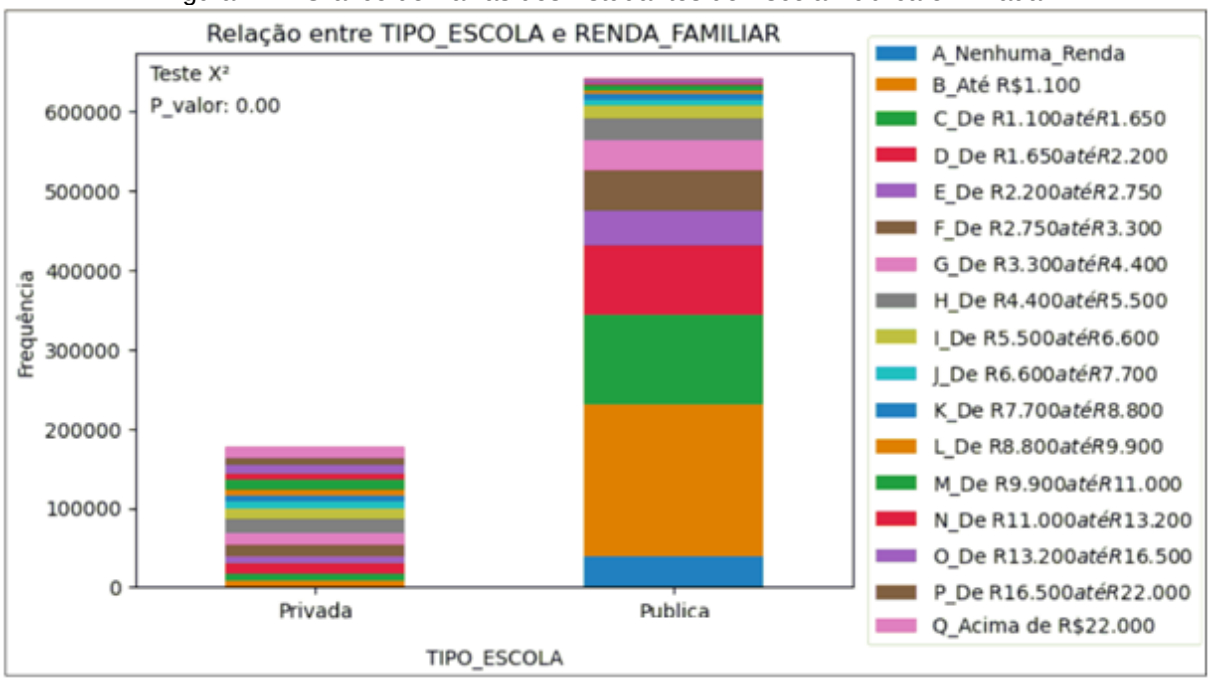
Fonte: Elaborado pelo autor

Para a renda mais baixa (A) apenas 2,9% dos estudantes são de escola privada enquanto para a renda mais alta (Q) tal índice é de 91,2%. Essa relação indica claramente que existe uma relação entre o fato de o estudante ser de escola pública ou privada e sua renda familiar.

Partindo da suposição de que o acesso à escola privada está relacionado à renda familiar, foi aplicado o método X^2 para confirmar essa relação, ou seja, de que realmente existe relação entre as categorias do tipo de escola do estudante (pública ou privada) e a renda familiar do mesmo.

A Figura 14 apresenta um gráfico de barras para os dois grupos de estudantes (de escola pública e de escola privada) em que cada barra é dividida em camadas de diferentes cores sendo cada cor correspondente a uma determinada renda familiar. Também foi calculado um P_valor = 0,0 ao aplicar o teste X^2, indicando que existe uma relação entre os dois atributos analisados.

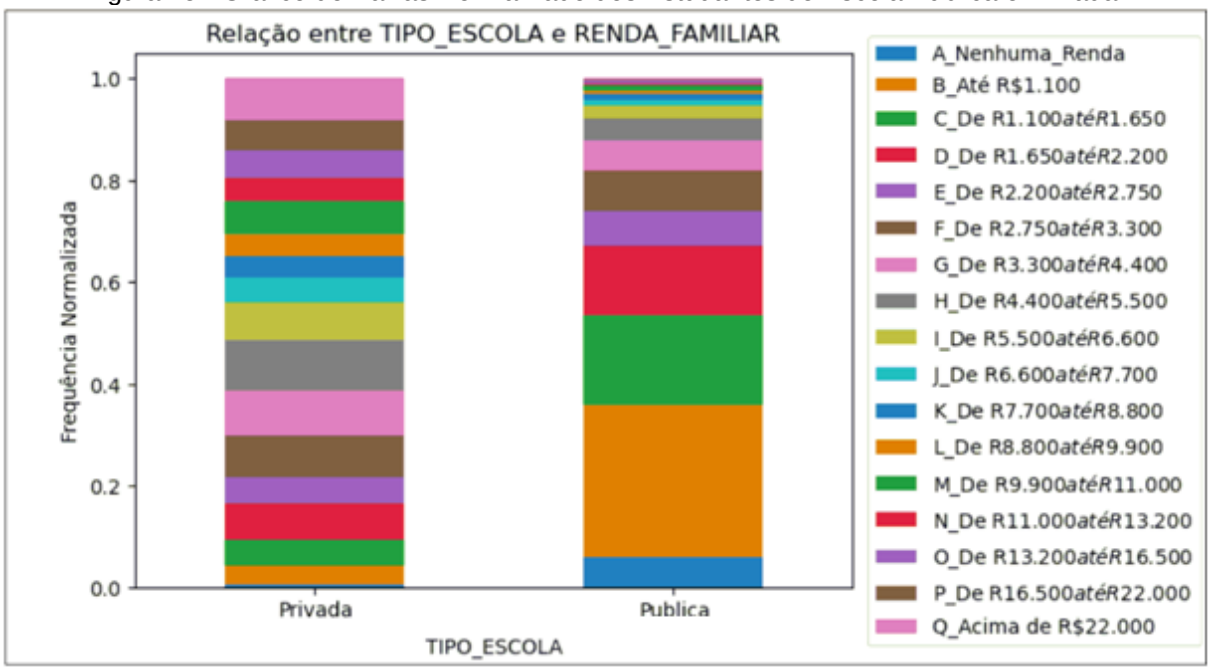
Figura 14 – Gráfico de Barras dos Estudantes de Escola Pública e Privada



Fonte: Elaborado pelo autor

Como existe um maior número de estudantes de escola pública que estudantes de escola privada, um novo gráfico foi gerado com o eixo y normalizado conforme mostra a Figura 15, pois assim é possível comparar como a renda familiar se distribui em cada uma das duas categorias do tipo de escola.

Figura 15 - Gráfico de Barras Normalizado dos Estudantes de Escola Pública e Privada



Fonte: Elaborado pelo autor

Analisando visualmente, é possível perceber que a barra correspondente ao grupo de estudantes de escola privada apresenta rendas familiares distribuídas mais próximas de uma distribuição uniforme. Já a barra correspondente ao grupo de estudantes de escola pública apresenta uma quantidade muito maior de estudantes de baixa renda e poucos estudantes de alta renda, ou seja, uma distribuição não uniforme.

4.6 Árvore de Decisão "*LightGBM*"

Para criar um modelo de previsão das notas do ENEM a partir das variáveis socioeconômicas dos estudantes foi implementado o método de árvore de decisão. Esse modelo foi utilizado devido à sua capacidade de lidar com múltiplas variáveis.

Segundo Russell e Norvig (2013, p.811), "A indução de árvores de decisão é uma das formas mais simples, e ainda assim mais bem sucedidas, de aprendizagem de máquina".

O método de árvore de decisão funciona por meio de uma estrutura hierárquica de ramificações e nós. Seu algoritmo avalia as características socioeconômicas relevantes em cada etapa, identificando os fatores mais importantes que influenciam o desempenho acadêmico.

Este método é particularmente útil quando se deseja formar um modelo de previsão. Além disso, é possível gerar informação quais variáveis socioeconômicas mais influenciam na nota final.

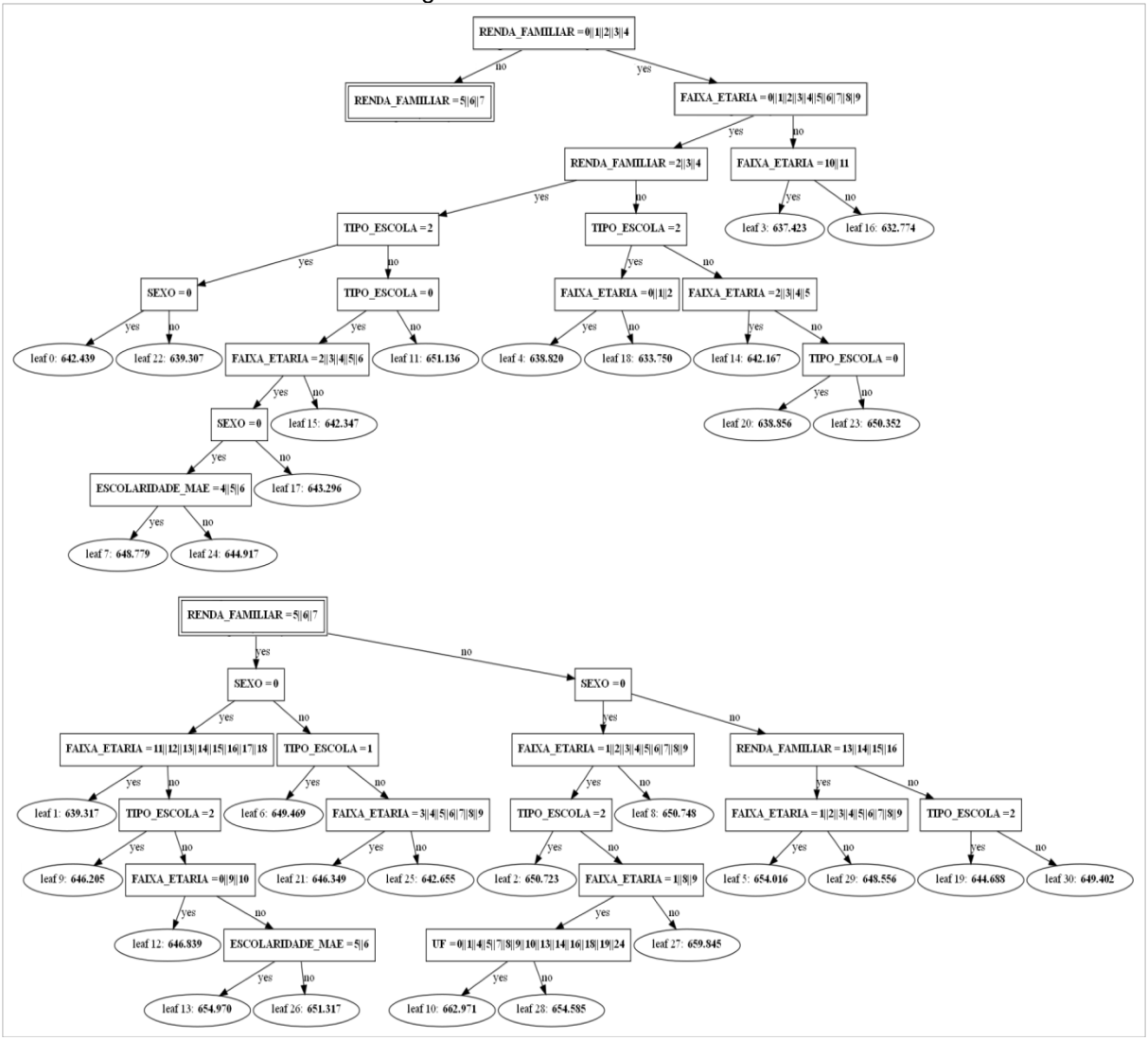
Para aplicar o método de árvore de decisão LightGBM foi utilizada a linguagem *python* e as bibliotecas "lightgbm" e "sklearn".

A variável dependente aplicada foi a NOTA_REDACAO e as variáveis independentes selecionadas foram: FAIXA_ETARIA, SEXO, COR_RACA, TIPO_ESCOLA, UF, ESCOLARIDADE_PAI, ESCOLARIDADE_MAE e RENDA_FAMILIAR.

Antes de efetuar o treinamento, os dados foram divididos em duas partes. Uma primeira parte, contendo 80% dos dados foi separada para aplicar no processo de aprendizagem da máquina e os 20% restantes foram separados para testar a eficiência do modelo gerado ao final do aprendizado.

A Figura 16 apresenta o modelo de árvore de decisão gerado.

Figura 16 – Árvore de Decisão



Fonte: Elaborado pelo autor

Após efetuar o aprendizado de máquina, foi avaliado o modelo gerado a partir dos dados de teste. Foi obtido o valor de Erro Quadrático Médio (MSE) igual a 18543 e um Coeficiente de Determinação (R^2) igual a 20,91%.

Este valor de R^2 é muito baixo. Portanto, pode-se dizer que o modelo de previsão obtido não se enquadra bem nos dados aplicados. Além disso, é possível perceber que todos os valores previstos são muito próximos da média geral de todos os estudantes analisados, que é de 644, o que supõe que o modelo gerado agrega pouca informação além de simplesmente considerar a média.

O mesmo método foi aplicado nas demais notas. A Tabela 09 apresenta o Erro Quadrático Médio (MSE) e um Coeficiente de Determinação (R^2) obtido para cada uma das notas.

Tabela 09 – Resultado da aplicação do método Árvore de Decisão

NOTAS	MSE	R^2
NOTA_REDACAO	18543	20,91%
NOTA_CIENCIAS_NATURAIS	4817	26,18%
NOTA_CIENCIAS_HUMANAS	6700	22,06%
NOTA_LINGUAGENS_CODIGOS	4455	24,70%
NOTA_MATEMATICA	8732	29,35%

Fonte: Elaborado pelo autor

As notas de matemática apresentaram o melhor R^2, com um valor de 29,35%. No entanto, este valor ainda é muito baixo e por isso, o modelo obtido não se enquadra bem nos dados.

Segundo Russell e Norvig (2013, p.812), "Para uma grande variedade de problemas, o formato da árvore de decisão gera um resultado agradável e conciso. Mas algumas funções não podem ser representadas de forma concisa".

4.7 Árvore de Decisão "*Standard Decision Tree*"

Como o modelo de árvore de decisão anterior não apresentou um bom resultado a partir dos dados analisados, optou-se por aplicar outro modelo de árvore de decisão. Para aplicar o método, foi utilizada a linguagem *python* e as bibliotecas *sklearn.tree* e *matplotlib*.

Criaram-se cinco variáveis binárias relacionadas às notas de cada prova de forma a dividir as notas em dois grupos: notas acima do quartil de 75% e notas abaixo do quartil de 75%. Assim foi possível aplicar o método considerando tais variáveis como dependentes e com isso buscou-se analisar quais as variáveis socioeconômicas relacionadas às melhores notas.

A Tabela 10 apresenta os valores das notas de cada uma das provas que divide o terceiro e o quarto percentil, ou seja, o quartil 75% de cada uma das notas.

Tabela 10 – Percentil 75% das notas das provas

Tipo de Prova	**Percentil 75%**
Ciências Naturais	549,9
Ciências Humanas	595,3
Linguagens e Códigos	563,6
Matemática	620,4
Redação	760,0

Fonte: Elaborado pelo autor

As variáveis independentes consideradas foram 'FAIXA_ETARIA', 'SEXO','COR_RACA', 'TIPO_ESCOLA', 'UF', 'ESCOLARIDADE_PAI', 'ESCOLARIDADE_MAE' e 'RENDA_FAMILIAR'. No entanto, essa quantidade de variáveis gerou uma árvore de decisão muito complexa e sua representação visual gerada tornou-se ilegível. Sendo assim, optou-se por diminuir o número de variáveis independentes para três, sendo estas: 'SEXO', 'COR_RACA' e 'TIPO_ESCOLA'

O método de árvore de decisão foi aplicado cinco vezes, uma vez para cada uma das notas. A Tabela 11 apresenta a precisão de cada um dos modelos.

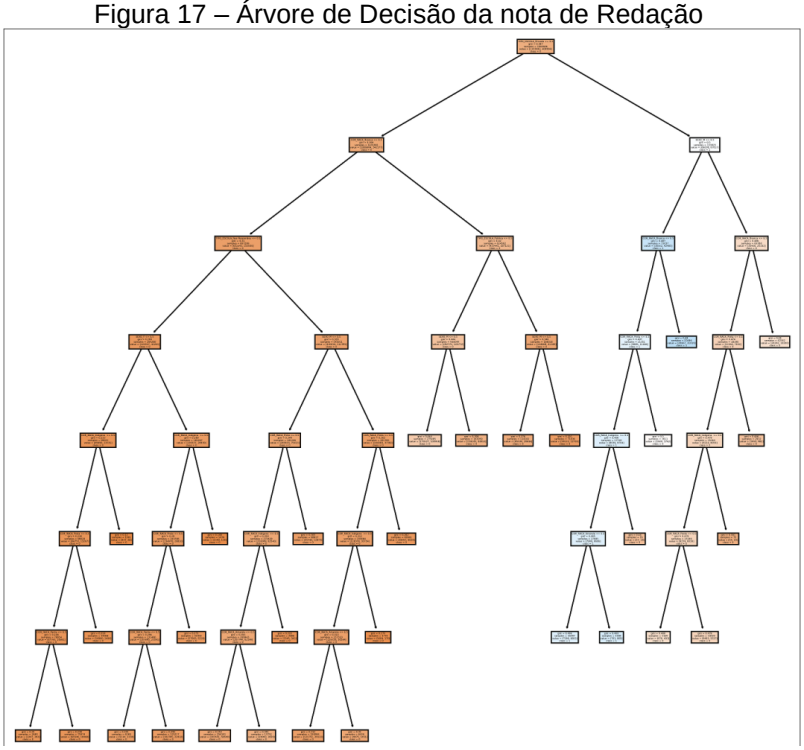
Tabela 11 – Precisão do Módelo

Tipo de Prova	Precisão do Modelo
Ciências Naturais	75,42%
Ciências Humanas	75,08%
Linguagens e Códigos	75,14%
Matemática	75,69%
Redação	74,61%

Fonte: Elaborado pelo autor

A Figura 17 apresenta o modelo da Árvore de Decisão gerada a partir dos dados das notas de redação.

Figura 17 – Árvore de Decisão da nota de Redação

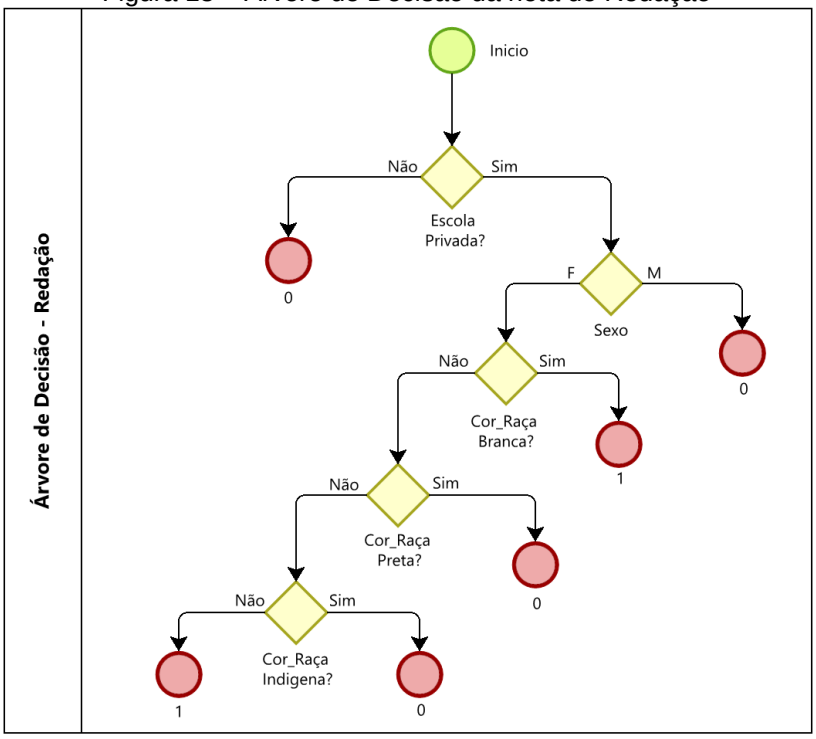


Fonte: Elaborado pelo autor usando a biblioteca ‘matplotlib’

A partir deste modelo de árvore de decisão, foi desenvolvido um modelo mais simplificado, removendo-se todos os ramos de um mesmo nó que levariam ao mesmo resultado no final. Assim, obteve-se uma representação para a árvore de decisão muito mais simples mantendo o resultado final de cada situação socioeconômica analisada.

A Figura 18 apresenta uma representação da árvore de decisão correspondente aos dados das notas de redação.

Figura 18 – Árvore de Decisão da nota de Redação



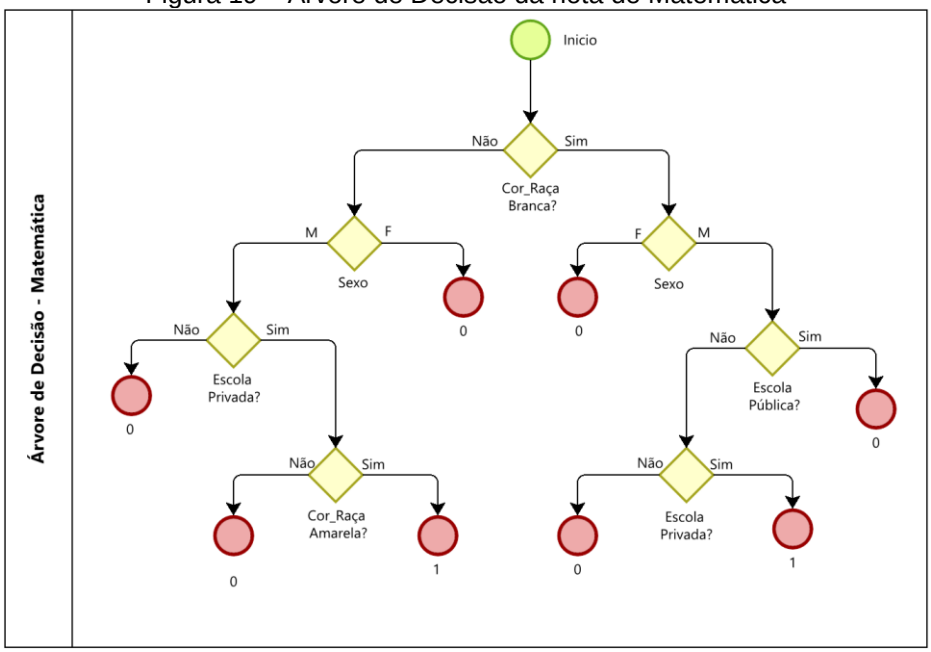
Fonte: Elaborado pelo autor usando o software Bizage

A partir do modelo apresentado, é possível perceber que o primeiro parâmetro considerado foi o tipo de escola em que o estudante estudou. Se a escola não for do tipo privada, então é previsto que o estudante não terá sua nota de redação no quartil superior.

Além disso, caso o estudante seja de escola privada, também precisa ser do sexo feminino e se declarar de cor branca ou amarela para que a previsão de sua nota esteja no quartil superir.

A Figura 19 apresenta uma representação da árvore de decisão correspondente aos dados das notas de matemática.

Figura 19 – Árvore de Decisão da nota de Matemática



Fonte: Elaborado pelo autor usando o software Bizage

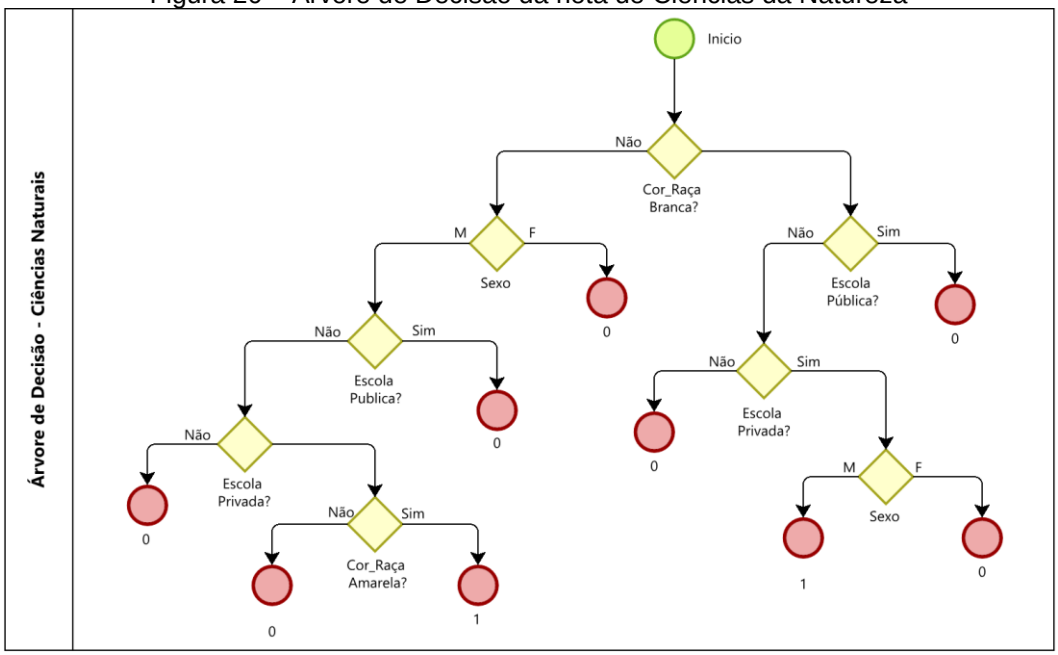
A partir do modelo apresentado, é possível perceber que o primeiro parâmetro considerado foi se a cor declarada pelo estudante é branca. Em caso negativo, apenas os estudantes do sexo masculino, de escola privada e de cor amarela é que são previstos a atingir uma nota correspondente ao quartil superior. Em caso positivo, o resultado também é o mesmo que o anterior, ou seja, apenas os estudantes do sexo masculino, de escola privada é que são previstos a atingir uma nota correspondente ao quartil superior.

A Figura 20 apresenta uma representação da árvore de decisão correspondente aos dados das notas de Ciências da Natureza

A partir do modelo apresentado, é possível perceber que o primeiro parâmetro considerado foi se a cor declarada pelo estudante é branca. Em caso negativo, apenas os estudantes do sexo masculino, de escola privada e de cor amarela é que são previstos a atingir uma nota correspondente ao quartil superior. Em caso positivo, o resultado também é o mesmo que o anterior, ou seja, apenas os estudantes do sexo masculino, de escola privada é que são previstos a atingir uma nota correspondente ao quartil superior.

As árvores de decisão das provas de Linguagens e Códigos e de Ciências Humanas apresentaram resultado negativo em todas as saídas, ou seja, sua previsão é de que todos os estudantes não atingem o quartil superior.

Figura 20 – Árvore de Decisão da nota de Ciências da Natureza



Fonte: Elaborado pelo autor usando o software Bizage

Os modelos de árvore de decisão apresentados, oferecem uma previsão da nota do aluno a partir das variáveis socioeconômicas. Além disso, é possível estimar quais variáveis apresentam maior influência na previsão. No entanto, essa estimativa não é absoluta uma vez que variáveis de igual importância não podem ser representadas paralelamente e assim, uma delas é necessariamente selecionada durante o aprendizado da máquina para ser utilizada primeiro num determinado nó de decisão.

Segundo Russell e Norvig (2013, p.816), "observamos que há um perigo de superinterpretar a árvore que o algoritmo seleciona. Quando existem diversas variáveis de importância similar, a escolha entre elas é um tanto arbitrária".

No entanto, pode-se entender que as decisões que aparecem mais acima na árvore de decisão apresentam importância igual ou superior às decisões que aparecem mais a baixo.

Pode ser observada em todas as ocorrências positivas de previsão a atingir uma nota correspondente ao quartil superior que tais ocorrências correspondem necessariamente de estudantes cuja cor (raça) é branca ou amarela e que também tenham cursado o ensino médio em uma escola privada.

O sexo também se apresentou como condição necessária para que a previsão a atingir uma nota correspondente ao quartil superior seja positiva. No entanto, houve diferença no sexo dependendo das provas analisadas. Na análise da

prova de redação, apresentou-se como condição necessária o estudante ser do sexo feminino enquanto nas provas de Ciências Naturais e Matemática a condição necessária é de o estudante ser do sexo masculino.

5 CONCLUSÃO

Os métodos estatísticos e de aprendizado de máquina são ferramentas capazes de gerar informações úteis a partir de dados brutos. A aplicação de métodos estatísticos nos dados do ENEM de 2021 permitiu obter uma compreensão inicial dos fatores que influenciam o desempenho dos estudantes, tais como: os dados demográficos, renda familiar, cor (raça), sexo, escolaridade dos pais e tipo de escola em que o estudante cursou o ensino médio. A partir de técnicas de aprendizado de máquina foi possível construir modelos preditivos. Tais modelos gerados foram capazes de prever quais características socioeconômicas dos estudantes influenciam as melhores notas com uma precisão significativa.

Após a aplicação do método de Regressão Linear, constatou-se que o nível de escolaridade dos pais e a renda familiar estão diretamente relacionados com a média das notas de todas as provas do ENEM, ou seja, ao se dividir os participantes em grupos de acordo com a escolaridade dos pais ou de acordo com a renda familiar, a média das notas de cada grupo apresentou um valor proporcional à escolaridade ou renda anteriormente citada.

Pode-se observar também que as médias das notas de cada um dos Estados do Brasil apresentaram valores diferentes, sendo que a maior média refere-se ao Estado de Minas Gerais o a menor média refere-se ao Estado do Amazonas. No entanto, em todos os Estados continham notas muito altas e muito baixas em todas as provas do exame.

Com a aplicação do método X^2 foi possível confirmar a hipótese de que a renda familiar do participante está relacionada com o tipo de escola (pública ou privada), pois a quantidade relativa de alunos de escola privada mostrou-se ser proporcional ao valor da renda familiar.

O método de Árvore de Decisão *LightGBM* foi aplicado aos dados com o objetivo de se obter uma previsão das notas a partir das variáveis socioeconômicos, no entanto, tal método resultou em um modelo de previsão que não se enquadra aos dados, uma vez que a precisão do modelo é muito baixa e os valores previstos são muito próximos entre si, ou seja, o modelo não acrescenta informação significativa se comparado com a simples previsão de que toda nota será igual à média geral das notas.

O método de Árvore de Decisão *Standard Decision Tree* foi aplicado aos dados com o objetivo de se obter um modelo de previsão que mostre quais as características socioeconômicas dos participantes que apresentaram correlação com as notas pertencentes ao quartil superior em cada uma das provas. Constatou-se que as tais características são: cor ‘branca’ e ‘amarela’ para a variável ‘Cor_Raça’ e ‘escola privada’ para a variável ‘Tipo_de_Escola’.

A análise de dados do ENEM utilizou métodos estatísticos e aprendizado de máquina. Tais métodos se mostraram como uma abordagem promissora para compreender e aprimorar a avaliação educacional. Os conhecimentos gerados por este trabalho sugerem novos estudos educacionais que podem ser utilizados como auxílio na tomada de decisão sobre o direcionamento de recursos e na Fórmulação de políticas educacionais mais efetivas. Espera-se que esse trabalho inspire novas investigações na área educacional, impulsionando avanços e inovações na educação.

REFERÊNCIAS BIBLIOGRÁFICAS

ARANHA, Christian; PASSOS, Emmanuel. A tecnologia de mineração de textos. **RESI - Revista Eletrônica de Sistemas de Informação**, v. 5, n. 2, 2006. DOI: https://doi.org/10.21529/RESI.2006.0502001. Disponível em: http://www.periodicosibepes.org.br/index.php/reinfo/article/view/171. Acesso em: 10 dev. 2023.

BARBETTA, Pedro Alberto; REIS, Marcelo Menezes; BORNIA, Antonio Cezar. **Estatística para concursos de engenharia e informática**. 3ª Edição.São Paulo: Atlas, 2010.

BITENCOURT, Wanderci Alves; SILVA, Diego Mello; XAVIER, Gláucia do Carmo. Pode a inteligência artificial apoiar ações contra evasão escolar universitária? **Ensaio: avaliação e políticas públicas em educação**, Rio de Janeiro, v. 30, n. 116, p. 669-694, jul./set. 2002. DOI: https://doi.org/10.1590/ S0104-403620220003002854. Disponível em: https://revistas.cesgranrio.org.br/ index.php/ensaio/article/view/2854. Acesso em: 10 fev. 2023.

CARVALHO, André Carlos Ponce de Leon Ferreira de. Inteligência artificial: riscos, benefícios e uso responsável. **Estudos Avançados**, São Paulo, v. 35, n.101, p. 21-36, 2021. DOI: https://doi.org/10.1590/s0103-4014.2021.35101. 003. Disponível em: https://www.revistas.usp.br/eav/article/view/185002. Acesso em: 10 fev. 2023.

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. **Introdução à mineração de dados:** conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016. 376 p.

COSTA, Lucas José da *et al*. Análise de métodos de detecção e reconhecimento de faces utilizando visão computacional e algoritmos de aprendizado de máquina. **Colloquium Exactarum**, v. 13, n.2, 2021. Disponível em: https://revistas.unoeste.br/index.php/ce/article/view/4145. Acesso em: 10 fev. 2023.

CURTY, Renata Gonçalves; SERAFIM, Jucenir da Silva. A formação em ciência de dados: uma análise preliminar do panorama estadunidense. **Informação & Informação**, v. 21, n. 2, p. 307–331. DOI: 10.5433/1981-8920. 2016v21n2p307. Disponível em: https://ojs.uel.br/revistas/uel/index.php/ informacao/article/view/27945. Acesso em: 10 fev. 2023.

DUARTE, Alan; NEGÓCIO, Ramon de Vasconcelos. Todos são iguais perante o algoritmo? uma resposta cultural do direito à discriminação algorítmica. **Direito Público**, v. 18, n. 100, 2021. DOI: https://doi.org/10.11117/rdp.v18i100. 5869. Disponível em: https://www.portaldeperiodicos.idp.edu.br/direitopublico/ article/view/5869. Acesso em: 10 fev. 2023.

FALCÃO, João V. R.; MOREIRA, Vinicius de A.; SANTOS, Flavia A. O.; RAMOS, Celso de A. Redes Neurais Deep Learning com TensorFlow. **Revista Eletrônica Científica de Ciência da Computação**. 2019. Disponível em: https://revistas.unifenas.br/index.php/RE3C/article/view/232 Acessado em 05 maio 2023.

GUIMARÃES, André José Ribeiro; MENDES Junior, Ricardo; FREITAS, Maria do Carmo Duarte. Requisitos para a ciência de dados: analisando anúncios de vagas de emprego com mineração de texto. **RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação**, n. 46, p. 54-70, 2022. Disponível em: https://dialnet.unirioja.es/servlet/articulo?codigo=8568720. Acesso em: 10 fev. 2023.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Microdados do Enem 2021**. Brasília: Inep, 2022. Disponível em: < https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>. Acesso em: 19 abr. 2023.

JARGAS, Aurelio Marinho. **Expressões regulares**: Uma abordagem divertida. 5ª Edição. São Paulo: Novatec, 2016.

LUDERMIR, Teresa Bernarda. Inteligência artificial e aprendizado de máquina: estado atual e tendências. **Estudos Avançados**, v. 35, n. 101, p. 85-94, 2021. DOI: https://doi.org/10.1590/s0103-4014.2021.35101.007. Disponível em: https://www.revistas.usp.br/eav/article/view/185035. Acesso em 10 fev. 2023.

METELO, Marcelo; BERNARDINO, Jorge; PEDROSA, Isabel. Avaliação de ferramentas open source para data science usando a metodologia OSSpal. **RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação**, n. E46, p. 588-607, nov. 2022. Disponível em: https://www.proquest.com/docview/ 2647406894. Acesso em 10 fev. 2023.

OLIVEIRA, Caue Gomes de; BARWALDT, Regina; LUCCA, Giancarlo. Análise do desempenho de pessoas com deficiência que prestaram o exame nacional do ensino médio - ENEM. **Tear: Revista de Educação, Ciência e Tecnologia**, Canoas, v. 9, n. 1, 2020. DOI: https://doi.org/10.35819/tear.v9.n1.a4038. Disponível em: https://periodicos.ifrs.edu.br/index.php/tear/ article/view/4038. Acesso em: 10 fev. 2023.

PIERSON, Lillian. **Data science para leigos**. Tradução Eveline Machado. Rio de Janeiro: Alta Books, 2019. 384 p.

RUSSELL, Stuart; NORVIG, Peter. Inteligência Artificial. Tradução de Regina Célia Simille. Rio de Janeiro: Editora Campus-Elsevier, 2013.

SONNENSTRAHL, Thiago Siqueira; BERNARDI, Giliane; PERTILE, Solange. Análise de interações do ambiente virtual de aprendizagem para predição de evasão em cursos no ensino a distância. **Revista EaD em Foco**, v. 11, n.1, 2021. DOI: https://doi.org/10.18264/eadf.v11i1.1463. Disponível em: https://eademfoco.cecierj.edu.br/index.php/Revista/article/view/1463. Acesso em: 10 fev. 2023.

SOUZA, Ewerton Pacheco de *et al*. Aplicações do Deep Learning para diagnóstico de doenças e identificação de insetos vetores. **Saúde em Debate**: pesquisa translacional em saúde coletiva: da bancada ao SUS, Rio de Janeiro, v. 43, n. especial 2, p. 147-154, nov. 2019. Disponível em: https://www.saudeemdebate.org.br/sed/article/view/2495. Acesso em 10 fev. 2023.

SOUZA, Vanessa Faria de. Mineração de dados educacionais com aprendizagem de máquina. **Revista Educar Mais**, v. 5, n. 4, p. 766–787, 2021. DOI: https://doi.org/10.15536/reducarmais.5.2021.2417. Disponível em: https://periodicos.ifsul.edu.br/index.php/educarmais/article/view/2417. Acesso em: 10 fev. 2023.

VIEGAS, Anderson. Cresce o acesso da pessoa com deficiência ao ensino superior no país. **G1.globo.com**, Campo Grande, 10 jun. 2016. Disponível em: http://glo.bo/1TZxKa6. Acesso em: 13 jan. 2023.