**UNIVERSIDADE DO ESTADO** 

**DO RIO DE JANEIRO**

****

**INSTITUTO POLITÉCNICO**

**GRADUAÇÃO EM**

**ENGENHARIA DA COMPUTAÇÃO**

Jonathan da Conceição Vieira

Uso do aprendizado de máquina para predição da nota de redação de um candidato do ENEM.

Nova Friburgo

2021

UNIVERSIDADE DO ESTADO DO RIO DE JANEIRO

INSTITUTO POLITÉCNICO DO RIO DE JANEIRO

CURSO DE ENGENHARIA DE COMPUTAÇÃO

Reitor: Ricardo Lodi Ribeiro

Vice-Reitor: Mário Sérgio Alves Carneiro

Diretor do Instituto Politécnico: Ângelo Modaini Calvão

Coordenadora de Curso: Silvia Mara da Costa Campos

Banca Avaliadora Composta por: Profa. Dra. Silvia Mara da Costa Campos (Orientadora)

Prof. Dr. Guilherme de Melo Baptista Domingues

Prof. Dr. Roberto Pinheiro Domingos

CATALOGAÇÃO NA FONTE

UERJ/REDE SIRIUS/BIBLIOTECA CTC/E

D111 Vieira, Jonathan da Conceição

Uso do aprendizado de máquina para predição da nota de redação de um candidato do ENEM / Jonathan da Conceição Vieira. - 2021. 60 f.

Orientador: Silvia Mara da Costa Campos

Trabalho de Conclusão de Curso apresentado à Universidade do Estado do Rio de Janeiro, Instituto Politécnico, para obtenção do grau de bacharel em Engenharia da Computação.

1. Aprendizado de Máquina - Monografias. 2. Ciências de dados - Monografias. I. da Costa Campos, Silvia Mara. II. Universidade do Estado do Rio de Janeiro. Instituto Politécnico. III. Título.

CDU 004.41

Endereço: UERJ - IPRJ

Caixa Postal 97282

CEP 28614-090 - Nova Friburgo – RJ – Brasil.

Este trabalho nos termos da legislação que resguarda os direitos autorais é considerado de propriedade da Universidade do Estado do Rio de Janeiro (UERJ). É permitida a transcrição parcial de partes do trabalho, ou mencioná-lo, para comentários e citações, desde que sem propósitos comerciais e que seja feita a referência bibliográfica completa.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Jonathan da Conceição Vieira

**UNIVERSIDADE DO ESTADO** 

**DO RIO DE JANEIRO**

****

**INSTITUTO POLITÉCNICO**

**GRADUAÇÃO EM**

**ENGENHARIA DA COMPUTAÇÃO**

Jonathan da Conceição Vieira

Uso do aprendizado de máquina para predição da nota de redação de um candidato do ENEM.

Trabalho de conclusão de curso apresentado como pré-requisito para obtenção do título de Engenheiro de Computação, ao Depar- tamento de Modelagem Computacional, do Instituto Politécnico, da Universidade do Estado do Rio de Janeiro.

Orientadora: Profa. Dra. Silvia Mara da Costa Campos

Nova Friburgo

2021

Jonathan da Conceição Vieira

Uso do aprendizado de máquina para predição da nota de redação de um candidato do ENEM.

Trabalho de conclusão de curso apresentado como pré-requisito para obtenção do título de Engenheiro de Computação, ao Depar- tamento de Modelagem Computacional, do Instituto Politécnico, da Universidade do Estado do Rio de Janeiro

BANCA EXAMINADORA:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Profa. Dra. Silvia Mara da Costa Campos (Orientadora)

Instituto Politécnico - UERJ

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Dr. Guilherme de Melo Baptista Domingues

Instituto Politécnico - UERJ

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof. Dr. Roberto Pinheiro Domingos

Instituto Politécnico - UERJ

**AGRADECIMENTOS**

Agradeço primeiramente, a Deus, por ter me dado saúde e sabedoria para chegar até aqui. Agradeço à minha família por todos os anos de apoio e amor. Sthefany Dionísio, minha namorada, por sempre acreditar em mim, estar ao meu lado e me incentivar em momentos de reclamações, dúvidas e frustrações, ao seu lado eu nunca perdi a fé. Aos meus amigos pela motivação, pelos conselhos e risadas compartilhadas nesta etapa tão desafiadora da vida acadêmica.

À professora Silvia Campos, que aceitou colaborar comigo na ideia de criar um projeto de ciência de dados e me ajudou nessa área na qual eu pretendo seguir carreira, muito obrigado por ter me ajudado nesse pontapé inicial.

E finalmente, agradeço à Universidade do Estado do Rio de Janeiro e a todos da instituição que de alguma forma contribuíram para minha formação.

*“Tudo tem o seu tempo determinado, e há tempo para todo o propósito debaixo do céu”*

Eclesiastes 3

**RESUMO**

VIEIRA, J. *Uso do aprendizado de máquina para predição da nota de redação de um candidato do ENEM.* 2021. 60 f. Trabalho de Conclusão de Curso (Engenharia de Computação) - Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2021.

No presente trabalho foi realizada uma análise exploratória dos microdados do ENEM, cujo objetivo principal é elucidar alguns panoramas da educação no Brasil, principalmente no estado do Rio de Janeiro e no município de Nova Friburgo, além disso foram propostos duas modelagens de aprendizado de máquina capazes de predizer a nota de redação de um candidato a partir dos seus dados socioeconômicos. O desenvolvimento dos modelos se deu através da linguagem de programação *python* no ambiente de desenvolvimento *jupyter notebook*, dentre as principais bibliotecas que foram usadas tanto para análise exploratória quanto para a modelagem do aprendizado de máquina, pode-se citar: *pandas*, *numpy*, *scikit-learn*, *seaborn* e *matplotlib*. O trabalho seguiu o passo a passo proposto pela Microsoft para projetos de ciências de dados de equipe. Durante a fase de exploração dos dados foram analisadas as variáveis relacionadas ao número de inscritos por estado, por município do Rio de Janeiro, idade, sexo, gestação, raça, escolaridade do pai e da mãe, e renda familiar. Já no decorrer da fase de modelagem foram aplicadas técnicas de engenharia de atributos como: tratamento de *outliers*, criação de novos atributos, *one-hot-encoding*, *ordinal-encoding*, dimensionamento, importância do recurso e ajuste dos hiperparâmetros. As métricas definidas para avaliação do modelo foram a raiz quadrada do erro médio e o erro absoluto médio.

**ABSTRACT**

VIEIRA, J. *An exploratory analysis and the use of machine learning on ENEM microdata.* 2021. 58 f. Trabalho de Conclusão de Curso (Engenharia de Computação) - Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2021.

In the present work, an exploratory analysis of the ENEM microdata was carried out, whose main objective is to elucidate some views of education in Brazil, mainly in the state of Rio de Janeiro, and the city of Nova Friburgo, in addition, were proposed two machine learning models ables to predict a candidate’s essay note from their socioeconomic data. The development of the models took place through the python programming language in the jupyter notebook development environment, among the main libraries that were used both for exploratory analysis and for machine learning modeling, we can mention: pandas, numpy, scikit- learn, seaborn and matplotlib. The work followed the step by step proposed by microsoft for team data science projects, during the data exploration phase, the variables related to the number of candidates per state, per city of Rio de Janeiro, age, sex, pregnancy, ethnic group, education of father and mother, and family income, already during the modeling phase, features engineering techniques were applied, such as: handling outliers, creation of new features, one-hot-encoding, ordinal-encoding, scaling, feature importance and hyperparameter optmization. The metrics defined for evaluating the model were the root mean squared error and the mean absolute error.

**LISTA DE FIGURAS**

[Figura 1 - Data never sleeps](#_rsgnvn4hq22r) ……………………………………………………………………….**13**

[Figura 2 - Data Science Lifecycle](#_fkuam3ox8sbv) ………………………………………………………………….**19**

[Figura 3 - What data scientists spend the most time doing](#_vnu44qvmnalj) …………..………………………...**21**

[Figura 4 - One-hot-encoding](#_p49d5mii5rql) ……………………………………………………………………….**23**

[Figura 5 - Ordinal-encoding](#_aix12522xd42) ………………………………………………………………………..**23**

[Figura 6 - Regressão linear](#_6ap5q92eadb) ………………………………………………………………………...**27**

[Figura 7 - Conjunto de microdados do ENEM 2019](#_7z28sdqss48x) …………………………………………….**29**

[Figura 8 - Conjunto de microdados do ENEM filtrado](#_1sylsrspdg3d) …………………………………………..**30**

[Figura 9 - Número de inscritos por estado](#_vqfa2kku5xe8) ……………………………………………………….**31**

[Figura 10 - Número de inscritos por município do Rio de Janeiro](#_x9wzdsqpvosi) ……………………………..**33**

[Figura 11 - Distribuição dos candidatos por idade](#_esa6iex0mldq) ……………………………………………….**34**

[Figura 12 - Distribuição dos candidatos por Sexo](#_ik9wroe7ipdo) ……………………………………………….**34**

[Figura 13 - Notas máximas por sexo](#_muz3vzvuedvn) ……………………………………………………………...**36**

[Figura 13 - Notas minimas por sexo](#_vapav43iyiys) ……………………………………………………………….**37**

[Figura 14 - Mapa dos municípios do Rio de Janeiro](#_mig7isgenuet) …………………………………………….**38**

[Figura 15 - Médias das notas de ciências da natureza](#_evqxso4jdcl3) ………………………………………….**38**

[Figura 16 - Médias das notas de ciências Humanas](#_lacc9sb6ed7v) …………………………………………….**39**

[Figura 17 - Médias das notas de linguagens e códigos](#_v2lld9imnc0e) ………………………………………...**39**

[Figura 18 - Médias das notas de matemática](#_sz8jxp6j9zyl) …………………………………………………….**40**

[Figura 19 - Médias das notas de redação](#_dt5iavy90xk7) ………………………………………………………..**40**

[Figura 20 - Distribuição racial dos candidatos](#_wrk11zvhpecy) …………………………………………………...**41**

[Figura 21 - Média das notas por raça no estado do Rio de Janeiro](#_8a4u1nc24ay4) …………………………..**41**

[Figura 22 - Média das notas por raça no município de Nova Friburgo](#_z3hh7ibdoh0v) ………………………..**42**

[Figura 23 - Média das notas por escolaridade dos pais](#_y915zn3sfnlu) ………………………………………..**43**

[Figura 24 - Média das notas por escolaridade dos pais no vestibular da UERJ](#_tmb0lna0pxqs) ……………..**44**

[Figura 25 - Média das notas por renda familiar](#_hnncilos5mel) ………………………………………………….**45**

[Figura 26 - Média das notas por renda familiar no vestibular da UERJ](#_5bzltr1up8xt) ……………………….**46**

[Figura 27 - Abertura de um arquivo csv](#_b0ueggpob1id) …………………………………………………………..**47**

[Figura 28 - Seleção de atributos](#_a13rdwxy1ih3) …………………………………………………………………...**48**

[Figura 29 - Gráfico boxplot para visualização dos outliers](#_5qv3p14q4j6n) ……………………………………..**52**

[Figura 30 - Gráfico boxplot após o tratamento dos outliers](#_33znnw7356ft) …………………………………….**52**

[Figura 31 - Permutation Importance](#_56wbzv793r3g) ……………………………………………………………….**54**

[Figura 32 - Tabela de dados estatísticos](#_cqg14w1axq7w) ………………………………………………………….**57**

[Figura 33 - Tabela de predições do modelo](#_1amjcvw6z435) ……………………………………………………...**57**

**LISTA DE ABREVIATURAS E SIGLAS**

CRISP-DM *Cross Industry Standard Process for Data Mining*

ENEM Exame Nacional do Ensino Médio

Fies Fundo de Financiamento Estudantil

IA Inteligência Artificial

KDD *Knowledge-discovery in databases*

Libras Língua Brasileira de Sinais

MAE *Mean Absolute Error*

ProUni Programa Universidade para todos

RMSE *Root Mean Squared Error*

SISU Sistema de Seleção Unificada

UERJ Universidade do Estado do Rio de Janeiro

**SUMÁRIO**

1. [**INTRODUÇÃO**](#_zih1xzwirrc) **…………………………………………………………………………. 12**
2. [Objetivo](#_2dyx825t29k7) …………………………………………………………………………………. 14
3. [Descrição do trabalho](#_ud8ayzny5tbf) …………………………………………………………………. 14
4. [Organização do trabalho](#_ontepttquhlz) ………………………………………………………………. 15

**2.** [**FUNDAMENTAÇÃO TEÓRICA**](#_pro3yg2bxx8y) **………………………………………………………. 17**

1. [Estatística](#_l446ase0em6c) ……………………………………………………………………………….. 17
2. [Aprendizado de máquina](#_413xxzq2pf3i) ……………………………………………………………… 17
3. [Etapas de um projeto de ciência de dados](#_yjkhr0diudxy) …………………………………………. 19
4. [Entendimento do negócio](#_30i9ic1wje6c) ……………………………………………………………… 20
5. [Aquisição e Entendimento dos dados](#_pjp75gmjdmyg) ……………………………………………….. 20
6. [Modelagem](#_onsavn2jt1h6) ……………………………………………………………………………… 22
7. [Implantação e validação por parte do cliente](#_clb7a2ciqqe2) ………………………………………. 26

**3.** [**DEFINIÇÕES DO APRENDIZADO DE MÁQUINA**](#_ap4o4pm1ov80) **……………………………….... 27**

**4.** [**METODOLOGIA**](#_n55d3vpuej7v) **……………………………………………………………………….. 29**

1. [Entendimento do negócio](#_4de4shvrk2fk) …………………………………………………………….. 29
2. [Aquisição e entendimento dos dados](#_qvuhmuemm9e) ……………………………………………….. 29
3. [Análise exploratória](#_3f7556n7ffhq) ……………………………………………………………………. 30
4. [Análise de inscritos por estado](#_tpccdj9gqiid) ……………………………………………………….. 31
5. [Análise de inscritos dos municípios do Rio de Janeiro](#_44zxwt12hii0) ……………………………. 32
6. [Análise da idade dos candidatos](#_2bkjxdapcn0i) …………………………………………………….. 33
7. [Análise do sexo dos candidatos](#_s3c363w0how9) …………………………………………………….... 34
8. [Análise das candidatas gestantes](#_ltpoe75xlfr7) …………………………………………………… 35
9. [Análise das notas por sexo](#_o6h8imjc319g) …………………………………………………………… 35
10. [Análise racial dos candidatos](#_4e6eidpvl6rb) ………………………………………………………… 41
11. [Análise das médias por escolaridade do pai e da mãe](#_mdk5474t8j0n) ……………………………. 43
12. [Análise das médias por renda mensal familiar](#_c39dpxp83v60) ……………………………………… 44

4.2.2 [Limpeza](#_bpj0we7juwh8) …………………………………………………………………………………. 47

4.2.3 [Modelagem](#_a6a7numbrjec) ……………………………………………...……………………………… 50

**5.** [**RESULTADOS**](#_gcz5rvlzk7sx) **………………………………………………………………………….. 56**

**6.** [**CONCLUSÃO**](#_7h2orfgwggar) **……………………………………………………………………...…… 58**

[**REFERÊNCIAS**](#_qtoz6ooq582h) **……………………………………………………………………………………. 59**

# INTRODUÇÃO

Criado em 1998 pelo governo federal do Brasil, o Exame Nacional do Ensino Médio (ENEM) inicialmente teve como principal finalidade avaliar o desempenho escolar dos estudantes após a conclusão da educação básica, e durante onze anos esse foi o seu único e exclusivo propósito. A partir de 2009, o ENEM, além de ser um processo de avaliação dos alunos concluintes do Ensino Médio, passou também a ser uma forma de acesso ao ensino superior no Brasil [1].

Dados de 2020, segundo o site do SISU (Sistema de Seleção Unificada, sistema informatizado do ministério da Educação, onde instituições públicas de ensino superior oferecem vagas para candidatos participantes do ENEM), mostram que centenas de milhares de vagas são ofertadas em mais de 130 instituições nacionais, e em 50 instituições de educação superior portuguesa [2][3]. As notas obtidas nos exames também podem ser utilizadas para programas governamentais, como o Programa Universidade para Todos (ProUni) e ao Fundo de Financiamento Estudantil (Fies) [3]. Em 2020 foram 5.893.369 inscritos em todo o país[4]. Durante o ano em questão, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) implementou pela primeira vez, soluções como provas em braile e ampliada, ajuda para leitura e transcrição e tradutor-intérprete em Língua Brasileira de Sinais (Libras) para contemplar pessoas com deficiência [5].

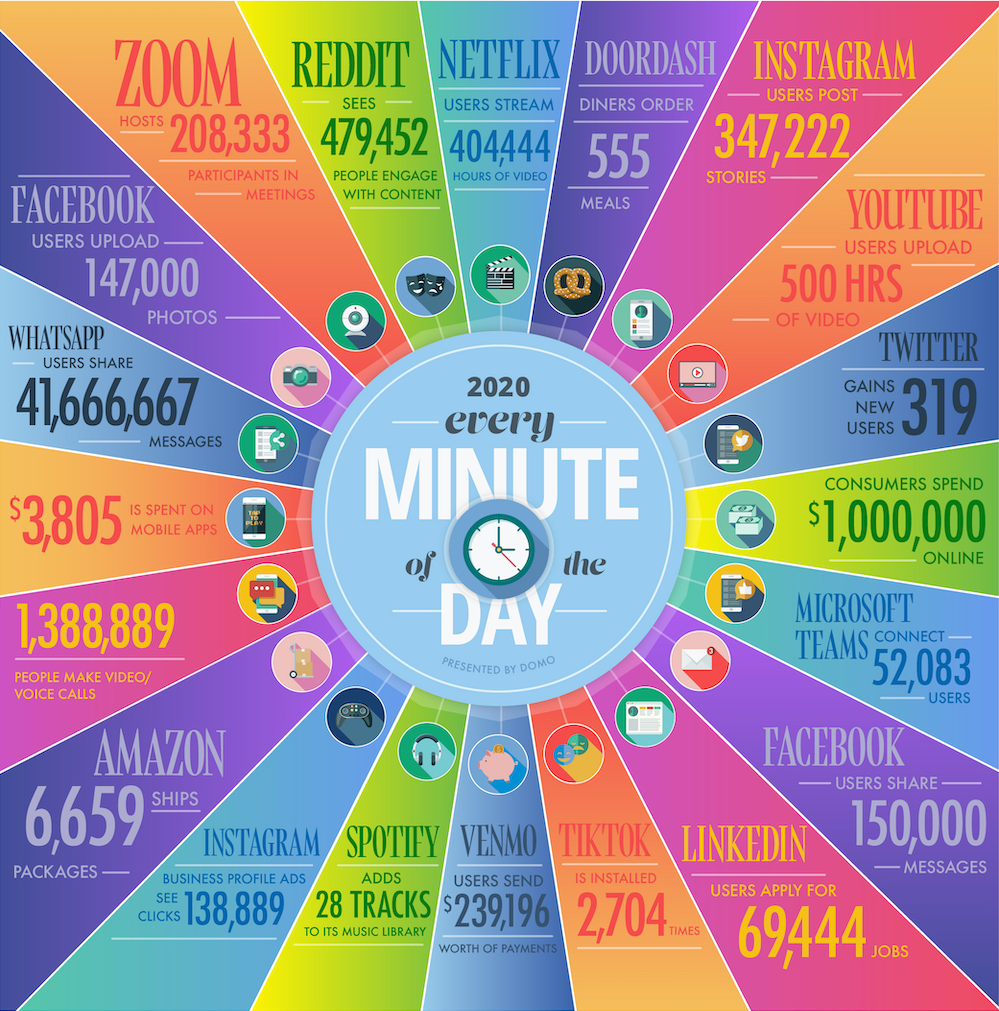
A Prova do ENEM dá-se em dois dias e consiste de uma redação, na qual os avaliados devem desenvolver uma dissertação-argumentativa a partir de uma situação específica, e 180 questões objetivas divididas em quatro áreas de conhecimento, sendo elas: Ciências da natureza, Ciências Humanas, Linguagens e Códigos, e Matemática [3].

Em passos um pouco mais acelerados que o ENEM, o ramo da Inteligência Artificial vem crescendo a cada ano que passa; só para exemplificar, em 2018, 32 empresas que utilizam inteligência artificial (IA) em seus principais negócios superaram a marca de avaliação de US$ 1 bilhão, de acordo com a pesquisa feita pela empresa CB Insights [6]. As empresas presentes na pesquisa vão de segmentos como prevenção a fraude bancária a softwares de recrutamento.

Em um artigo publicado pelo canal *MIT Sloan Management Review*, é destacado a junção de *big data* e inteligência artificial, o que ele chama de “o desenvolvimento mais importante que está moldando o futuro de como as empresas agregam valor aos negócios a partir de seus dados e atributos analíticos” [7]. Não é difícil perceber o motivo pelo qual IA e *big data* crescem juntos; *big data* (grandes coleções de dados) cresce diariamente. Um estudo publicado pela DOMO, retratado abaixo na figura 1, nos faz ter uma dimensão da quantidade de dados gerados pelas grandes empresas [8], e são justamente os dados que impulsionam a IA, são essas grandes coleções que possibilitam que as aplicações de aprendizado de máquina - um ramo da IA - aprendam de forma rápida e independente.

É fácil perceber a vantagem competitiva que uma boa análise de dados pode levar, ainda mais em um ambiente onde existem muito mais dados sendo gerados do que as mãos humanas têm a capacidade de analisar, e é nesse meio que técnicas como aprendizado de máquina, análises preditivas e visualização dos dados cresce, apoiando e melhorando a precisão e a velocidade das tomadas de decisão.

# Figura 1 - *Data never sleeps*



Legenda: Infográfico que mostra a quantidade de dados gerados por minuto do dia por grandes empresas.

Fonte: (https://www.domo.com/learn/infographic/data-never-sleeps-8).

Nesse contexto foi desenvolvido esse trabalho, onde procurou-se evidenciar alguns cenários da educação brasileira através dos dados fornecidos pelo governo [9], além disso é apresentado uma modelagem computacional capaz de predizer a nota de redação do candidato com base nos questionários, em especial o questionário socioeconômico, respondidos pelo próprio candidato durante a realização da prova.

A modelagem foi desenvolvida no ambiente *jupyter notebook* através da linguagem de programação *python* com auxílio de algumas bibliotecas, são elas: *pandas*, *numpy*, *scikit-learn*, *seaborn* e *matplotlib*. Durante a modelagem este trabalho seguiu o modelo de projeto proposto pela Microsoft para projeto de ciências de dados de equipe, e dentre as técnicas que aqui foram utilizadas para engenharia de atributos, pode-se citar: tratamento de *outliers*, criação de novos atributos, *one-hot-encoding*, *ordinal-encoding*, dimensionamento, importância do recurso e ajuste dos hiperparâmetros.

## Objetivo

Este trabalho tem como objetivo analisar e explorar as variáveis que compõem os microdados do ENEM fornecidos pelo site do Inep [9]. Por meio desses dados espera-se elucidar alguns panoramas da educação no Brasil, principalmente no estado do Rio de Janeiro e no município de Nova Friburgo e, através do questionário socioeconômicos dos participantes, propor um modelo de aprendizado de máquina capaz de prever as notas das redações dos participantes. Esse trabalho também busca trazer à tona as etapas essenciais em um projeto de ciência de dados.

## Descrição do trabalho

Este trabalho foi desenvolvido em três etapas:

A primeira etapa consistiu na importação dos dados brutos disponibilizados pelo site governamental do Inep [9] e realizada uma análise exploratória de algumas variáveis de interesse para o autor; são elas: estado, município, idade, sexo, raça, além das variáveis socioeconômicas como: renda mensal, escolaridade dos pais e ocupação dos pais. Nessa etapa buscou-se comparar dados dos estudantes de Nova Friburgo com dados de estudantes do estado do Rio de Janeiro, também houve uma breve comparação dos dados disponibilizados pelo ENEM, com os dos dados fornecidos pela Universidade do Estado do Rio de Janeiro (UERJ).

Na segunda etapa foi desenvolvida uma modelagem de aprendizado de máquina a partir dos dados estudados anteriormente, porém com a adição de novos atributos do questionário socioeconômico, como: quantidade de pessoas que moram atualmente na residência do candidato e perguntas referentes a habitação e bens domésticos. Nessa fase foi feita uma breve limpeza dos dados, e a aplicação de tratamentos suficientes apenas para possibilitar a predição do algoritmo.

Na terceira etapa foi desenvolvida uma segunda modelagem de aprendizado de máquina; contudo, houve uma maior dedicação a limpeza, processamento, organização e manipulação dos mesmos dados trabalhos na fase anterior. Ao final foram analisados e comparados os resultados obtidos pelo modelo apresentado no passo dois, e o modelo desenvolvido no passo três.

É importante ressaltar que o conjunto de dados fornecidos pelo governo estão em conformidade com a lei geral de proteção dos dados; logo, esse estudo não possui nenhuma informação pessoal como nome, identidade ou qualquer outro campo que possa identificar o candidato.

## Organização do trabalho

Este trabalho foi organizado em seis capítulos.

O primeiro capítulo tem como objetivo apresentar ao leitor, o que é o ENEM, bem como contextualizar a importância do aprendizado de máquina nos dias de hoje.

O segundo capítulo encontra-se dividido em três subcapítulos, sendo o primeiro uma breve abordagem aos termos estatísticos que estão contidos no trabalho; o segundo detalha melhor o que é o aprendizado de máquina e os seus grupos, enquanto o terceiro subcapítulo trata das etapas de um projeto de ciências de dados; este subcapítulo subdivide-se ainda em quatro outros tópicos, estes referentes ao ciclo de vida de um projeto de ciências de dados. São eles: entendimento do negócio, aquisição e entendimento dos dados, modelagem, e implantação e validação por parte do cliente.

O terceiro capítulo apresenta as definições do modelo de aprendizado de máquina utilizado no trabalho, assim como a linguagem de programação, o ambiente de desenvolvimento e as métricas usadas para avaliação de desempenho dos modelos.

O quarto capítulo contém a metodologia que foi utilizada para criação dos modelos. O intuito deste capítulo é apresentar como o trabalho percorreu as etapas de um projeto de ciências de dados descritas no capítulo dois. Aqui é apresentada uma longa análise exploratória dos dados e seus resultados, as técnicas de limpeza utilizadas e a modelagem do aprendizado de máquina.

O quinto capítulo expõe e analisa os resultados obtidos com os modelos, e fica a cargo do sexto capítulo apresentar as conclusões obtidas ao fim do trabalho.

# FUNDAMENTAÇÃO TEÓRICA

## Estatística

Segundo o artigo: Aplicações da Estatística nas Ciências, de Nilo Antonio e Maria Cristina [10], a Estatística é descrita como a “ciência que utiliza-se das teorias probabilísticas para explicar a frequência da ocorrência de eventos, tanto em estudos observacionais quanto em experimentos para modelar a aleatoriedade e a incerteza de forma a estimar ou possibilitar a previsão de fenômenos futuros”.

Sendo parte essencial do estudo de Ciência de Dados, ao lado da computação e de um bom entendimento de negócios, esse trabalho em alguns pontos se utiliza de termos estatísticos básicos, na qual vale a pena relembrar, antes de prosseguir:

* Moda: O valor mais frequente. Normalmente utilizada para variáveis nominais.
* Média aritmética simples: soma dos valores dividida pelo número de elementos.
* Mediana: é o valor central do conjunto de valores quando ordenados de forma crescente. Em caso de número par, será a média dos dois valores centrais da amostra.
* Variância: é a distância quadrática média em relação à média.
* Desvio-padrão: é uma medida de dispersão em torno da média da variável. Medida pela raiz quadrada da variância.

## Aprendizado de máquina

Aprendizado de máquina é um subcampo da Ciência de Inteligência Artificial; o seu intuito é replicar habilidades inteligentes de humanos em máquinas. Nesse subcampo da IA, a principal questão é de como fazer com que as máquinas sejam capazes de “aprender”, e ao ler “aprender” subentende-se como inferência indutiva, onde a máquina, através da observação de alguns exemplos, consiga predizer eventos futuros [11].

O Aprendizado de máquina (*Machine learning*) pode ser dividido em três grupos principais; são eles: Supervisionado, Não Supervisionado e por Reforço.

* Supervisionado: em modelos de aprendizado de máquina supervisionado, as variáveis alvos, ou rótulos (variáveis que queremos prever) são conhecidos para cada exemplo apresentado ao algoritmo; isto é, para cada entrada de dados, sabemos previamente qual é a saída esperada.

O algoritmo é apresentado a uma série de registros (exemplos ou instâncias), com as saídas desejadas, e espera-se que ele seja capaz de aprender uma regra onde, ao ser exposto para novos exemplos, ele consiga predizer corretamente qual é a saída.

Neste modelo, existem as seguintes subdivisões:

* Classificação: Quando a variável alvo é discreta.

Exemplo: Detecção de fraudes (sim ou não)

* Regressão: Quando a variável alvo é contínua. O objetivo não é prever uma classe à qual o rótulo faz parte, mas sim predizer um valor.

Exemplo: Predizer a nota de um aluno.

* Séries temporais: O rótulo é contínuo, porém dependente do tempo.

Exemplo: Rendimento da bolsa de valores.

* Não supervisionado: Os exemplos submetidos ao algoritmo não possuem rótulos, e é esperado que através das semelhanças dos seus atributos ele consiga agrupá-los em *clusters*. Depois de feita essa organização, geralmente é preciso um estudo para determinar qual o significado de cada agrupamento no contexto em que o problema está sendo discutido.

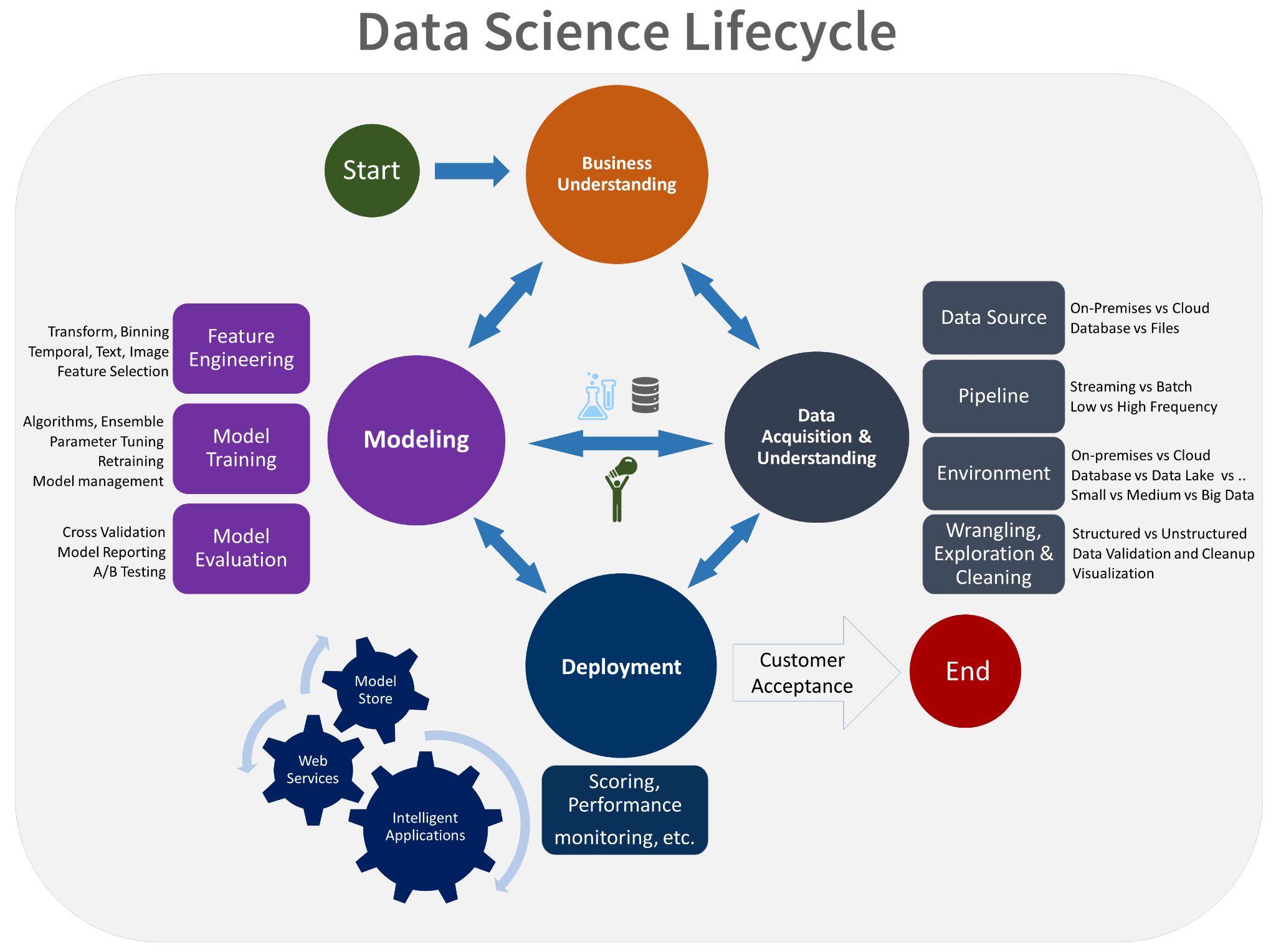
Exemplo: Sistemas de recomendação.

* Por reforço: neste aprendizado o algoritmo não conhece a resposta certa, através dos exemplos em que ele foi apresentado; é formulada uma hipótese, essa hipótese então recebe um sinal de recompensa ou punição, caso ela seja boa ou ruim.

Exemplo: Jogos e robótica, processamento de linguagem natural [12].

## Etapas de um projeto de ciência de dados

# Figura 2 - *Data Science Lifecycle*



Legenda: Ciclo de vida do processo de ciências de dados de equipe.

Fonte: (https://docs.microsoft.com/en-us/azure/architecture/data-science-process/lifecycle).

Um projeto de ciência de dados compreende algumas etapas, que não necessariamente devem ser seguidas à risca; as organizações têm total liberdade para criar e definir o seu próprio processo personalizado. Este trabalho baseou-se no modelo proposto pela documentação fornecida pela Microsoft [13], representado pela figura 2, pois acredita-se que ele facilita a compreensão das etapas a serem seguidas. Vale salientar que este processo não é o único disponível no mercado, existem ainda o CRISP-DM (Cross Industry Standard Process for Data Mining ou Processo Padrão para Data Mining) e o KDD (Knowledge-discovery in databases ou Descoberta de Conhecimento em Banco de Dados)

### Entendimento do negócio

A primeira etapa descrita no artigo “O Ciclo de vida do processo de Ciências de Dados de equipe” fornecido pela Microsoft [13] de um projeto de ciência de dados, passa pelo entendimento do negócio. É preciso compreender e identificar qual problema deve ser resolvido e quais objetivos devem ser alcançados.

Cinco tipos de perguntas podem ser respondidas utilizando ciência de dados e o aprendizado de máquina:

* Quanto custa ou quantos? (regressão)
* Qual categoria? (classificação)
* Qual grupo? (*clustering*)
* Isso é estranho? (detecção de anomalias)
* Qual opção deve ser escolhida? (recomendação)

Um bom exercício é determinar quais dessas perguntas o modelo a ser desenvolvido deve responder. Ainda nessa etapa do projeto é necessário identificar as fontes de dados, ou seja, localizar os dados relevantes e as principais variáveis que o modelo precisa inferir; são elas que irão ajudar a responder os objetivos propostos pelo projeto. As métricas associadas a tais variáveis irão determinar o sucesso ou não do modelo de aprendizado de máquina. Um exemplo de uma variável alvo seria a nota de um candidato do ENEM.

### Aquisição e Entendimento dos dados

Após a definição do principal objetivo do projeto, e de se ter mapeado as principais fontes de dados, a próxima etapa consiste em construir um conjunto de dados (*dataset*) limpo e de alta qualidade. Uma expressão em inglês atribuída ao técnico da IBM George Fuechsel diz, “Garbage In, Garbage out”, referindo-se ao fato de entradas pobres, produzirem saídas pobres. No contexto do aprendizado de máquina isso pode ser interpretado como: caso os conjuntos de dados de entrada tenham uma baixa qualidade, o resultado proposto pelo modelo também terá uma baixa qualidade.

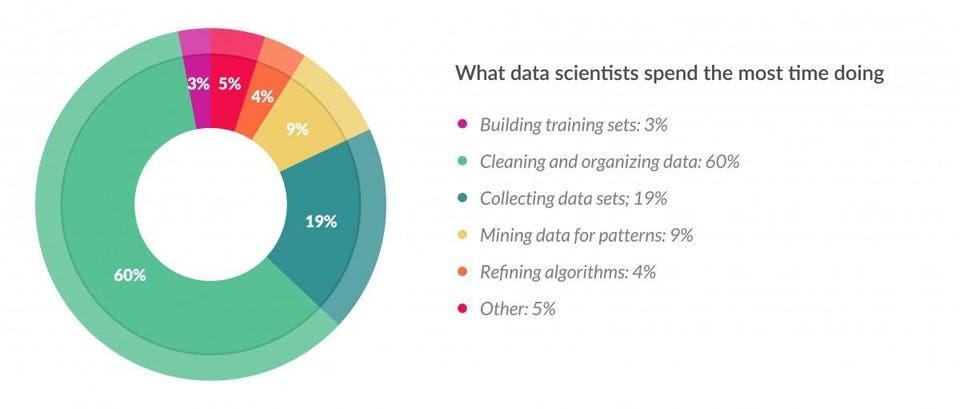
* Exploração, limpeza & estruturação dos dados

E como construir um conjunto de dados de alta qualidade?

Antes de tudo é necessário dedicar grande parte do tempo aqui, tempo para explorar, limpar, pré-processar e organizar os seus dados, além de adquirir um amplo conhecimento dos dados que estão sendo trabalhados. Em sua maioria os conjuntos de dados que refletem situações do mundo real possuem valores ausentes, ruídos ou outras inconformidades e isso precisa ser tratado. Nesta etapa de exploração, é de grande ajuda a visualização dos dados através de gráficos ou tabelas para fornecer informações sobre a qualidade dos dados.

Segundo a Forbes, esta etapa consome cerca de 60% do tempo do projeto, dada sua importância; em contrapartida, a maioria dos cientistas de dados julgam ser a parte menos agradável do trabalho [14], a figura 3 detalha as tarefas onde cientistas de dados gastam mais tempo.

# Figura 3 - *What data scientists spend the most time doing*



Legenda: Gráfico que mostra quais tarefas mais consomem o tempo dos cientistas de dados.

Fonte: (https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-  
least-enjoyable-data-science-task-survey-says/?sh=b3f90bd6f637).

Não menos importante que a análise exploratória dos dados, a definição de uma arquitetura de solução para o consumo de novos dados, como parte de um processo de aprendizado de máquina contínuo, também é realizada nesta fase do projeto.

### Modelagem

Nessa fase é onde define-se o conjunto de dados ideal para o modelo de aprendizado de máquina. O modelo tem como base dados históricos e aprende a partir deles para então conseguir predizer a saída esperada. Os dados de entrada compreendem atributos (*features*), que em sua maioria se encontram na forma de colunas estruturadas. Os algoritmos de aprendizado de máquina tem como pré-requisito algumas características específicas, daí surge a necessidade da Engenharia de atributos, ou *feature engineering.*

A Engenharia de atributos além de melhorar o desempenho desses modelos, ajudam a preparar o conjunto de dados adequado para o algoritmo de aprendizado de máquina escolhido.

Algumas técnicas compreendem o universo da Engenharia de atributos, como: Manipulação de *outliers*, criação de novos atributos, *One-hot-encoding*, dimensionamento, entre outros. Essas técnicas requerem um pouco de experiência e expertise, além de *insights* obtidos durante a fase de exploração.

* Manipulação de *outliers*: *Outlier*, trata-se de um ponto que está significamente distante da estrutura linear descrita pelos dados, que acaba influenciando a estimação, conduzindo a modelos ajustados impróprios [15]. Algumas técnicas matemáticas podem ser usadas para detecção dos *outliers*, e tratar esses pontos pode potencializar o desempenho do modelo final.
* Criação de novos atributos: Ainda que a maior parte dos atributos brutos possam ser utilizados diretamente para treinar um modelo, por vezes é necessário criar novos atributos (*features*) para um conjunto de dados de treinamento mais avançado.
* *One-hot-encoding*: Alguns algoritmos de machine learning não compreendem dados categóricos, por isso é necessário converter esses dados para um formato numérico sem que nenhum tipo de informação seja perdida. O método *one-hot-encoding* distribui os valores de uma coluna categórica nominal para várias outras colunas atribuindo 0 ou 1 a elas, de forma que esses valores binários representam a relação entre as colunas agrupadas e codificadas. A figura 4 exemplifica esse processo.

# Figura 4 - *One-hot-encoding*

Legenda: Exemplificação da técnica de *One-hot-encoding* com uma coluna relacionada a animais.

Fonte: (https://towardsdatascience.com/stop-one-hot-encoding-your-categorical-variables-  
bbb0fba89809).

* *Ordinal-encoding*: o *ordinal-encoding* tem o mesmo propósito que o *one-hot-encoding*, só que para variáveis categóricas ordinais, variáveis cujos valores expressam atributos que seguem algum tipo de ordem ou grau, como por exemplo: ensino fundamental, ensino médio e faculdade; Nesse caso a informação se traduziria para valores numéricos que representam essa ordem ou grau. A figura 5 exemplifica esse processo.

# Figura 5 - *Ordinal-encoding*

|  |  |
| --- | --- |
| **Escolaridade** | **Ordinal-encoding** |
| Ensino fundamental | 0 |
| Ensino médio | 1 |
| faculdade | 2 |

Legenda: Exemplificação da técnica de *ordinal-encoding* com uma coluna relacionada à Escolaridade.

Fonte: O autor, 2021.

* Dimensionamento: Geralmente os atributos numéricos do conjunto de dados não estão contidos no mesmo intervalo, e esse intervalo difere de um recurso para o outro; o que acontece é que no mundo real, uma coluna que representa a idade, tem um intervalo diferente para uma coluna que representa a renda, por exemplo. Mas, tratando-se de aprendizado de máquina, como comparar essas duas colunas? É justamente por isso que existe o dimensionamento, após a aplicação dessa técnica os atributos contínuos tornam-se exatamente iguais nesse aspecto.

Comumente são utilizadas duas formas de dimensionamento:

* Normalização ou Normalização *min-max* faz com que todos os valores sejam postos em um intervalo fixo entre 0 e 1, esse tipo de dimensionamento não modifica a distribuição do recurso, e devido à diminuição dos desvios padrões, os efeitos dos *outliers* aumentam, sendo assim, é interessante que antes da normalização os *outliers* sejam tratados.

* Padronização ou Normalização escore-z, esse tipo de dimensionamento é feito com base no desvio padrão: caso o desvio padrão dos atributos forem diferentes, o seu intervalo também irá diferir, isso diminui o efeito de *outliers*.

Foram descritas apenas as técnicas utilizadas neste trabalho, porém existem algumas outras que vale ressaltar, como: *Binning*, agrupamento e separação de atributos e a transformação logarítmica [16].

Um outro estágio da modelagem que faz parte do processo de um projeto de Ciência de Dados, é a seleção de atributos ou *feature selection*. A seleção de atributos é um conceito importante em aprendizado de máquina, pois além de impactar no custo computacional, selecionar corretamente os atributos utilizados pelo modelo pode afetar o desempenho final (a deleção de atributos é chamada de redução de dimensionalidade). Existem algumas técnicas que são de fácil implementação e que provêm bons resultados. São elas:

* Matriz de correlação com mapa de calor: A correlação determina o grau de intensidade da relação dos atributos entre si e da relação com a variável alvo. A correlação pode ser negativa ou positiva, sendo negativa uma relação inversamente proporcional; ou seja, a medida que o valor do recurso aumenta, o valor da variável alvo diminui, e sendo positiva representada por uma relação diretamente proporcional; isso é, quando o valor do recurso aumenta, o valor da variável alvo também aumenta. É comumente utilizado o gráfico de mapa de calor para facilitar a identificação e visualização dessas correlações.
* Importância do recurso (*Feature Importance*): Modelos de aprendizado de máquinas específicos (modelos de regressão e classificação baseados em árvores) são capazes de estimar para cada recurso do seu conjunto de dados a *feature* *importance*, ou importância do recurso. Essa técnica confere uma pontuação para os atributos utilizados para estimar a variável alvo de acordo com a sua utilidade; ao aferir a pontuação que cada recurso recebeu, pode-se tomar uma decisão sobre a redução ou não da dimensionalidade [17].

Ainda na fase de modelagem é realizado o treinamento do modelo em si; essa fase compreende os passos abaixo:

* Separação dos dados de entrada de forma aleatória em dois conjuntos, um de treino e um de teste.
* Criação de um ou mais modelos utilizando os dados de treinamento.
* Ajustes e definições dos hiperparâmetros dos modelos; antes de se treinar um modelo, existem parâmetros que podem ser ajustados manualmente, em modelos baseados em árvores por exemplo, pode-se ajustar o número de árvores a ser utilizado pelo algoritmo. Existem duas técnicas comumente usadas para auxiliar na otimização desses parâmetros, o *grid search* e o *random search*;
* O *grid search* consiste em testar exaustivamente todas as possíveis combinações dos hiperparâmetros e em seguida fornecer os hiperparâmetros que alcançaram o menor erro.
* O *random search* testa aleatoriamente combinações de parâmetros previamente fornecidos e, dentre os informados, seleciona aqueles que obtiveram o menor erro. [18]
* Avaliação dos modelos, ao desenvolver um modelo de aprendizado de máquina; no final, você espera que as respostas fornecidas sanem as dúvidas do problema proposto. Portanto, após a aplicação das técnicas de modelagem, o algoritmo é treinado, e deve-se verificar se o seu desempenho em fornecer as respostas foi satisfatório ou não; para isso utiliza-se o conjunto de testes.

Existem certas métricas que auxiliam na avaliação do modelo proposto, essas métricas podem variar de modelo para modelo.

* Definição do resultado; em outras palavras, dentre os modelos criados, qual apresenta a melhor avaliação de acordo com as métricas definidas.

### Implantação e validação por parte do cliente

As duas últimas fases do projeto incluem a implantação e a validação por parte do cliente; isso significa que, caso você tenha atingido bons resultados na modelagem, o seu modelo está pronto para ser exposto a problemas reais, e consumido em outras aplicações, como por exemplo: dashboards, sites, planilhas, dentre outros.

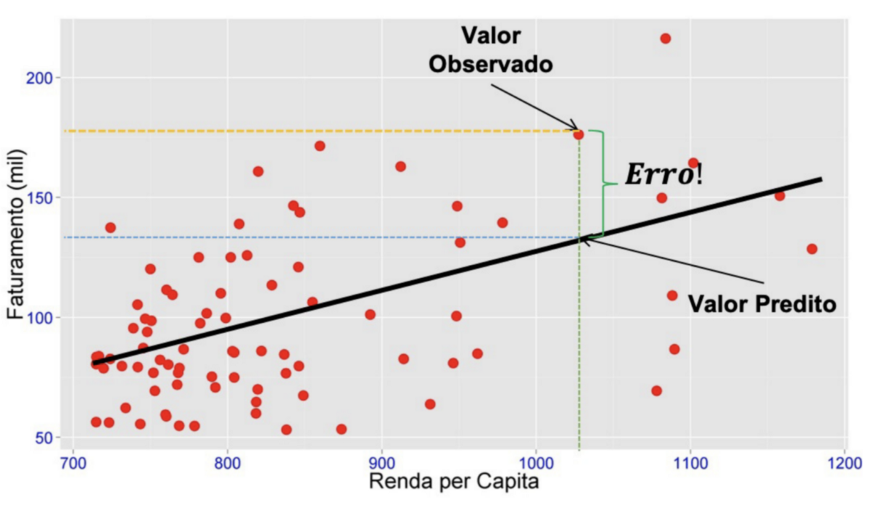
Nessas aplicações o seu modelo será apresentado a novos dados, e espera-se que ele consiga predizer tão bem quanto foi durante a fase de modelagem. Cabe ao cliente aferir se o algoritmo e o pipeline de consumo de dados implementados satisfazem a sua necessidade.

# DEFINIÇÕES DO APRENDIZADO DE MÁQUINA

O modelo de aprendizado de máquina que veio a ser desenvolvido a fim de inferir a nota de um participante do ENEM através do questionário socioeconômico, é um modelo supervisionado e de regressão, uma vez que a variável que deseja-se predizer é a nota de uma prova, uma variável contínua.

Modelos de Regressão consistem na construção de uma equação que melhor representa a relação entre as variáveis dependentes e uma ou mais variáveis independentes. A partir disso, o modelo consegue predizer um valor numérico contínuo o mais próximo possível do valor esperado para a variável de saída quando submetido a novos valores nas suas variáveis de entrada, sendo a diferença entre esses valores uma medida de erro do algoritmo, como pode-se observar na figura 6.

# Figura 6 -Regressão linear



Legenda: Exemplo do cálculo de erro para um exemplo na regressão linear.

Fonte: (ESCOVEDO & KOSHIYAMA, 2020).

Quanto a avaliação ficou definido que a acurácia do modelo será medida através da raiz quadrada do erro médio ou RMSE (*Root Mean Squared Error*) que representa a raiz quadrática média das diferenças entre os valores reais e os valores estimados. Outra medida definida para avaliação do modelo é o erro absoluto médio ou MAE (*Mean Absolute Error*) que é o cálculo dos erros entre os valores reais e os valores estimados.

Por fim, foi definida a linguagem *python* para a modelagem do aprendizado de máquina, e o *jupyter notebook* (interface gráfica prática, amigável e gratuita capaz de compilar trechos de códigos em navegadores web) como ambiente de desenvolvimento. A linguagem *python* foi escolhida para o desenvolvimento por ser uma linguagem muito popular no universo de ciências de dados, que além de contar com uma comunidade ampla e bastante suporte, também contém enormes bibliotecas capazes de facilitar o desenvolvimento de uma modelagem de aprendizado de máquina; dentre elas, pode-se citar:

* *Pandas* para manipulação dos dados em formatos de colunas estruturadas.
* *Numpy* capaz de manipular arrays multidimensionais e realizar cálculos de forma eficiente.
* *Seaborn* e *Matplotlib* são bibliotecas poderosas para visualização de dados.
* *Scikit-learn* uma das mais conhecidas bibliotecas de aprendizado de máquina capaz de possibilitar a aplicação de diversas técnicas de modelagem de forma simples, rápida e eficiente.

Todas as bibliotecas citadas foram utilizadas no desenvolvimento deste trabalho.

# METODOLOGIA

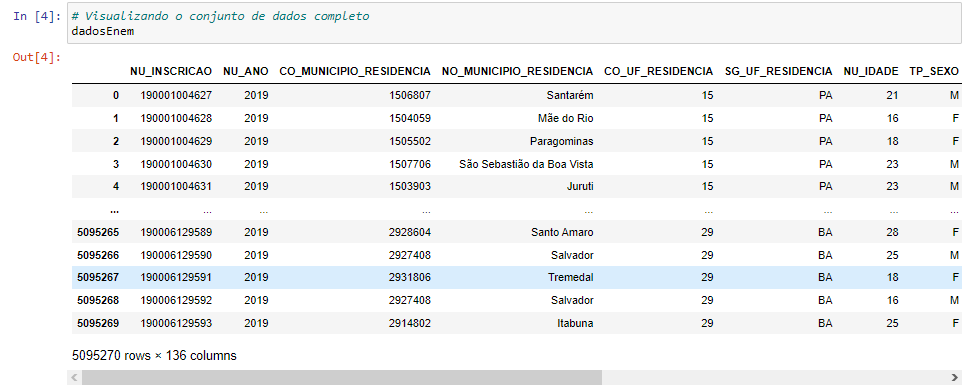
## Entendimento do negócio

Este trabalho não trata uma questão de negócio em si, mas ainda sim, antes de efetivamente se debruçar nos dados, conversas entre o autor e sua orientadora foram importantes para compreensão e identificação do problema; a partir dessas conversas que surgiu a proposta de analisar os dados da educação brasileira e posteriormente desenvolver uma modelagem de aprendizado de máquina. Essas conversas, que em projetos de grandes empresas se equiparariam a uma reunião de negócios, foram importantes para definição da linguagem de programação a ser usada durante o projeto, o ambiente de desenvolvimento, localização do conjunto de dados e definição das métricas de avaliação.

## Aquisição e entendimento dos dados

Este trabalho utilizou o conjuntos de dados “Microdados do Enem 2019 (Atualizado em 20/05/2021)” e também do seu “Dicionário de Variáveis - ENEM 2019” que na concepção do estudo era o mais atual disponibilizado pelo portal do Inep [9].

# Figura 7 -Conjunto de microdados do ENEM 2019



Legenda: Amostra de alguns registros dos microdados do ENEM 2019.

Fonte: O autor, 2021.

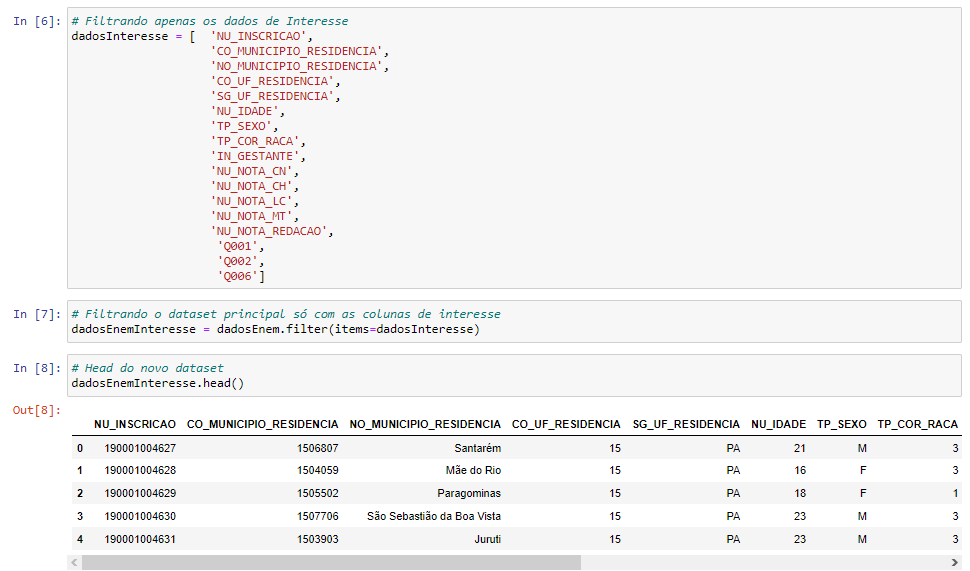
Verificou-se que o mesmo possui 136 colunas, e 5.095.270 registros. As variáveis que o compõem podem ser melhor compreendidas verificando-se o dicionário. Cada registro do conjunto de dados representa um candidato do ENEM 2019.

### Análise exploratória

Especialmente durante a análise exploratória, as análises serão acompanhadas dos resultados e de uma breve discussão sobre os mesmos.

Como seria muito difícil a manipulação de 136 atributos (*features*), e visando evitar um custo computacional muito alto, antes de efetivamente fazer as primeiras análises, filtrou-se o conjunto inicial selecionando apenas os atributos de interesse do autor.

# Figura 8 -Conjunto de microdados do ENEM filtrado



Legenda: Alguns registros dos microdados do ENEM filtrado por dados de interesse do autor.

Fonte: O autor, 2021.

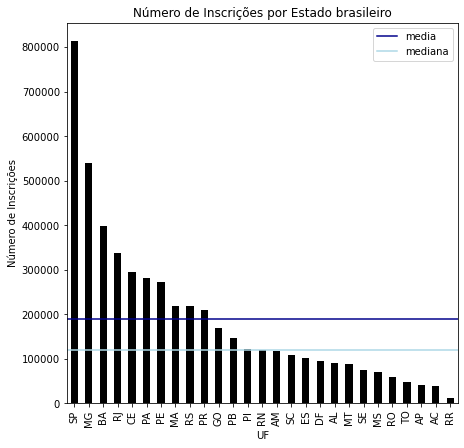
Os atributos selecionados representam:

* Número da inscrição
* Código do município de residência
* Nome do município de residência
* Código da Unidade da Federação de residência
* Sigla da Unidade da Federação de residência
* Idade
* Sexo
* Raça
* Indicador de gestante
* Nota da prova de Ciências da Natureza
* Nota da prova de Ciências Humanas
* Nota da prova de Linguagens e Códigos
* Nota da prova de Matemática
* Nota da prova de Redação
* Até que série seu pai, ou o homem responsável por você, estudou?
* Até que série sua mãe, ou a mulher responsável por você, estudou?
* Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)

#### Análise de inscritos por estado

Para realizar essa análise, foi utilizado somente o recurso referente a sigla da unidade da federação de residência do candidato; foi realizada a contagem de candidatos por estado e utilizado um gráfico de barras para facilitar a visualização.

# Figura 9 -Número de inscritos por estado



Legenda: Gráfico apresentando o número de inscritos no ENEM por estado brasileiro.

Fonte: O autor, 2021.

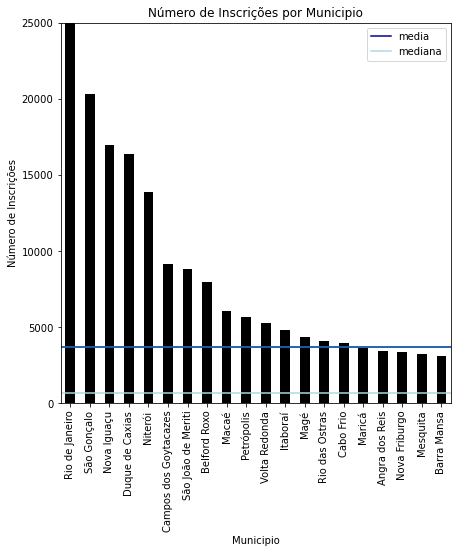
Nesta análise pode-se perceber que São Paulo é o estado do Brasil com a maior quantidade de inscritos, com cerca de 813.772 inscrições, Roraima é o estado com a menor quantidade de inscritos, com apenas 12.947 inscrições. O Rio de Janeiro é o quarto maior, com 338.710 inscrições. Três dentre os cinco estados com maior número de inscritos são da região sudeste (São Paulo, Minas Gerais e Rio de Janeiro), em contrapartida dos cinco estados com o menor número de inscritos, todos são pertencentes à região Norte (Rondônia, Tocantins, Amapá, Acre e Roraima).

#### Análise de inscritos dos municípios do Rio de Janeiro

Utilizou-se a informação que descreve o nome do município de residência dos candidatos para se ter uma visão de todos os municípios do Rio de Janeiro. A capital do estado foi responsável por 137.765 inscrições, um número quase sete vezes maior que o segundo município, São Gonçalo, com 20.305 inscrições. O município com o menor número de inscritos foi São José de Ubá com apenas 97.

Nova Friburgo foi o 18º município com 3.347 estudantes, dentre as 92 localidades que tiveram ao menos um candidato, esse valor representa aproximadamente 1% do total de candidatos de todo o Estado; a média de inscritos por município do Rio de Janeiro foi de 3.681,63. Abaixo encontra-se a figura 10, um gráfico feito para facilitar a visualização dos 20 municípios do Rio de Janeiro com maior número de inscrições, por questões puramente estéticas optou-se por limitar o eixo *y* a 25.000, mas vale ressaltar novamente, que a capital do estado obteve um total de 137.765 candidatos inscritos.

# Figura 10 -Número de inscritos por município do Rio de Janeiro



Legenda: Gráfico apresentando a média e o número de inscritos no ENEM por município do Rio de Janeiro,

Fonte: O autor, 2021

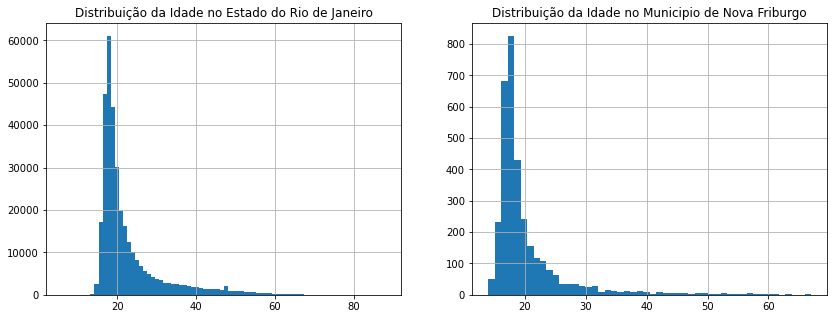
#### Análise da idade dos candidatos

A idade com maior número de inscritos no estado do Rio de Janeiro foi 18 anos: 61.059 candidatos declararam essa idade. Curiosamente, houveram registros de dois inscritos bem abaixo da média, sendo um com 12 anos, e um com 9 anos, além de três idosos com 83, 86 e 88 anos de idade respectivamente.

Em Nova Friburgo a idade com maior número de registros também foi 18 anos, com 826 candidatos. A menor idade registrada foi 14 anos para um dos candidatos e a maior foi 67.

Abaixo na figura 11, dois histogramas mostram melhor a distribuição da idade no estado do Rio de Janeiro e no município de Nova Friburgo.

# Figura 11 -Distribuição dos candidatos por idade



Legenda: Histogramas apresentando a distribuição da idade dos candidatos do ENEM no estado do Rio de Janeiro e no município de Nova Friburgo.

Fonte: O autor, 2021.

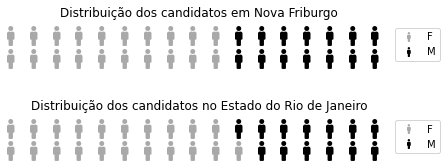
#### Análise do sexo dos candidatos

59,78% dos inscritos no estado do Rio de Janeiro são do sexo Feminino, isso representa 202.488 inscritos; candidatos do sexo masculino são 136.222, isso representa 40,22% do total.

Em Nova Friburgo essa porcentagem é bem próxima, foram 1948 candidatas do sexo Femino e 1399 candidatos do sexo masculino. Isso corresponde a aproximadamente 58,20% e 41,80% do total de candidatos do município, respectivamente.

Em uma sala de prova com 34 inscritos, a distribuição mais provável de se encontrar em Nova Friburgo seria aproximadamente de 20 mulheres para cada 14 homens, enquanto no Rio de Janeiro essa distribuição seria de aproximadamente de 21 mulheres para cada 13 homens, a figura 12 ilustra isso.

# Figura 12 -Distribuição dos candidatos por Sexo



Legenda: Gráficos exemplificando a distribuição dos candidatos do ENEM por sexo no municipio de Nova Friburgo e no estado do Rio de Janeiro.

Fonte: O autor, 2021.

#### Análise das candidatas gestantes

82 candidatas fizeram a prova em período de gestação no estado do Rio de Janeiro, isso corresponde a 0,04% do total de candidatas do sexo feminino do estado. Já no municipio de Nova Fribugo apenas uma candidata declarou gestação durante a realização da prova, ela representa 0,05% do total de candidatas do sexo feminino de Nova Friburgo.

#### Análise das notas por sexo

Ao realizar essa análise, verificou-se que as seguintes variáveis continham valores nulos. (Descrição da variável, variável e quantidade de registros nulos).

Para o conjunto de dados de candidatos do Rio de Janeiro:

* Nota da prova de Ciências da Natureza NU\_NOTA\_CN 93.786
* Nota da prova de Ciências Humanas NU\_NOTA\_CH 79.431
* Nota da prova de Linguagens e Códigos NU\_NOTA\_LC 79.431
* Nota da prova de Matemática NU\_NOTA\_MT 93.786
* Nota da prova de Redação NU\_NOTA\_REDACAO 79.431

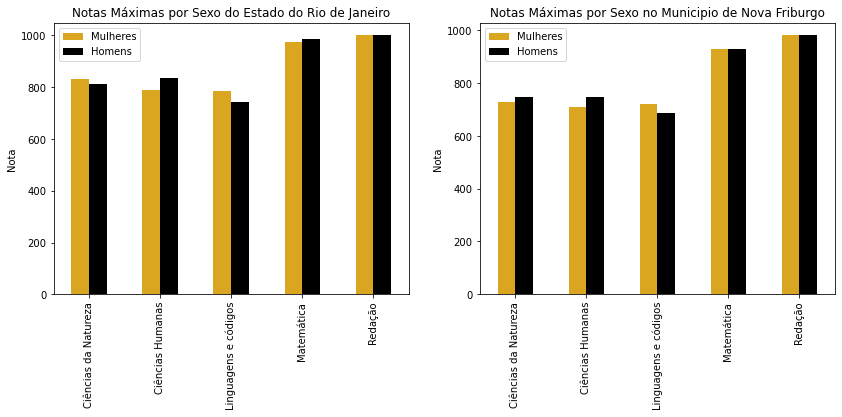
Para o conjunto de dados de candidatos de Nova Friburgo:

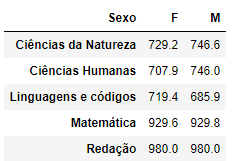
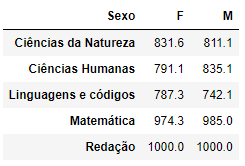
* Nota da prova de Ciências da Natureza NU\_NOTA\_CN 765
* Nota da prova de Ciências Humanas NU\_NOTA\_CH 630
* Nota da prova de Linguagens e Códigos NU\_NOTA\_LC 630
* Nota da prova de Matemática NU\_NOTA\_MT 765
* Nota da prova de Redação NU\_NOTA\_REDACAO 630

Para ambos os conjuntos de dados, os valores foram excluídos.

* Nota máxima por sexo

# Figura 13 -Notas máximas por sexo



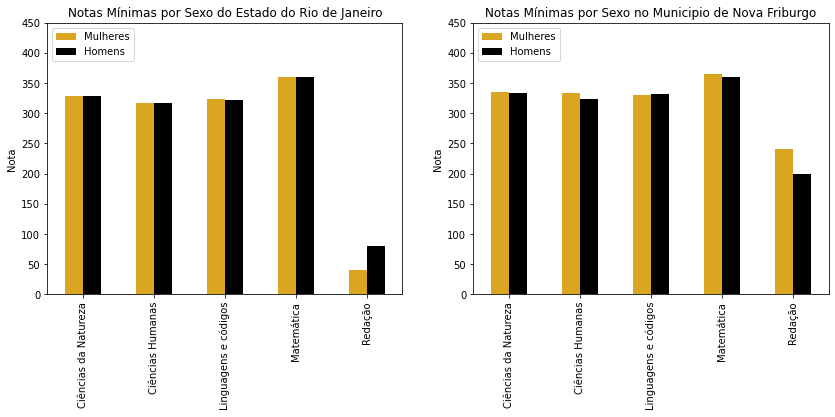


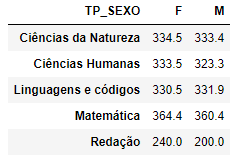
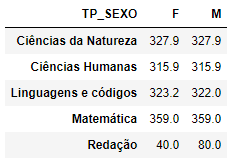
Legenda: Gráficos e tabelas apresentando as notas máximas do ENEM por sexo no estado do Rio de Janeiro e no municipio de Nova Friburgo.

Fonte: O autor, 2021.

* Nota minima por sexo (só foram consideradas as notas minimas maiores do que zero)

# Figura 13 -Notas minimas por sexo





Legenda: Gráficos e tabelas apresentando as notas minimas do ENEM por sexo no estado do Rio de Janeiro e no municipio de Nova Friburgo.

Fonte: O autor, 2021.

* Média das Notas por Sexo e por Municipio

# Figura 14 -Mapa dos municípios do Rio de Janeiro



Legenda: Localização do município de Nova Friburgo.

Fonte: (Abreu, 2006; Acervo Fundação CEPERJ).

* Ciências da Natureza

Na prova de ciências da natureza, as médias das notas masculinas são maiores que a média das notas femininas na maioria dos municípios do estado, inclusive no município de Nova Friburgo, como mostra a figura 15.

# Figura 15 -Médias das notas de ciências da natureza

Legenda: Gráfico representando a média das notas do ENEM na prova de ciências da natureza para o sexo feminino e masculino nos municipios do estado do Rio de Janeiro.

Fonte: O autor, 2021.

* Ciências Humanas

Ao analisar a figura 16, verifica-se que na prova de ciências humanas houve um equilíbrio maior com relação à prova de ciências da natureza, inclusive no município de Nova Friburgo. Porém apesar de alguns municípios estarem em situações de igualdade, ainda há muitos municípios que as notas masculinas superaram as notas femininas.

# Figura 16 -Médias das notas de ciências Humanas

Legenda: Gráfico representando a média das notas do ENEM na prova notas de ciências Humanas para o sexo feminino e masculino nos municipios do estado do Rio de Janeiro.

Fonte: O autor, 2021.

* Linguagens e Códigos

Talvez a média de notas mais equilibradas seja na prova de linguagens e códigos, tanto em Nova Friburgo, quanto em vários outros municípios do estado, ambos os sexos apresentam médias bem próximas, isso torna-se mais evidente ao visualizar o gráfico da figura 17.

# Figura 17 -Médias das notas de linguagens e códigos

Legenda: Gráfico representando a média das notas do ENEM na prova de linguagens e códigos para o sexo feminino e masculino nos municipios do estado do Rio de Janeiro.

Fonte: O autor, 2021.

* Matemática

Em matemática, os municipios no geral tiveram candidatos do sexo masculino com notas superiores que as candidatas do sexo feminino. Em Nova Friburgo a diferença foi mínima, mas houve, como mostra a figura 18.

# Figura 18 -Médias das notas de matemática

Legenda: Gráfico representando a média das notas do ENEM na prova de matemática para o sexo feminino e masculino nos municipios do estado do Rio de Janeiro.

Fonte: O autor, 2021.

* Redação

Ao analisar o gráfico da figura 19, percebe-se que ao contrário das notas de matématica, em redação as candidatas do sexo feminino foram superiores na maioria dos municipios, porém em Nova Friburgo as notas foram semelhantes.

# Figura 19 -Médias das notas de redação

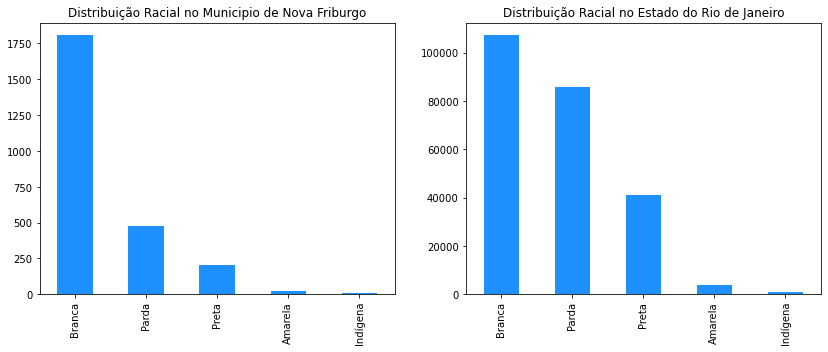
Legenda: Gráfico representando a média das notas do ENEM na prova de redação para o sexo feminino e masculino nos municipios do estado do Rio de Janeiro.

Fonte: O autor, 2021.

#### Análise racial dos candidatos

Dados do censo de 2010 mostram que a maioria da população do estado do Rio de Janeiro é branca [19], e isso se confirma ao ver a distribuição racial através dos gráficos da figura 20, tanto no estado, quanto no município de Nova Friburgo.

# Figura 20 -Distribuição racial dos candidatos



Legenda: Gráfico representando a distribuição racial do ENEM no município de Nova Friburgo e no estado do Rio de Janeiro.

Fonte: O autor, 2021.

# Figura 21 -Média das notas por raça no estado do Rio de Janeiro

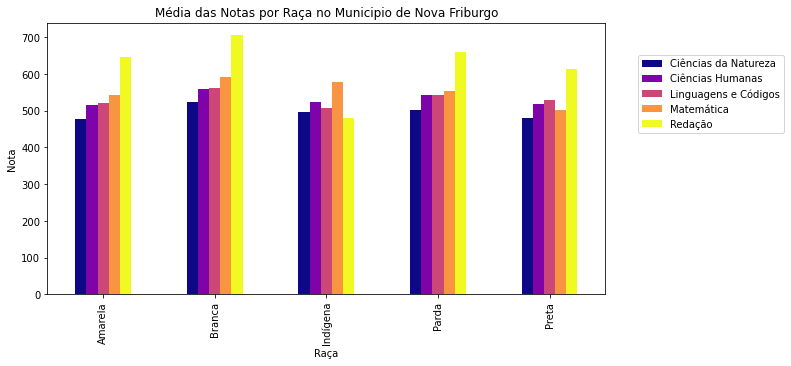
Legenda: Gráfico que detalha o desempenho nas provas do ENEM por raça no estado do Rio de Janeiro

Fonte: O autor, 2021.

Ao observar as médias das notas das provas no estado do Rio de Janeiro na figura 21, percebe-se um padrão entre amarelos, indígenas e pardos; os candidatos que declararam essas raças obtiveram as maiores notas em redação, seguidas de linguagens e códigos, matemática, ciências humanas e ciências da natureza. Pretos comparados às três raças citadas anteriormente, tiveram em sua maioria uma nota de matemática pior que a nota de ciências humanas, enquanto brancos nesse mesmo conjunto de questões obtiveram a segunda melhor média.

A figura 22 mostra que em Nova Friburgo, o melhor desempenho de todas as raças, com exceção dos indígenas, continuou sendo na prova de redação, enquanto o pior desempenho foi na prova de ciências da natureza. As médias de amarelos, brancos e pardos seguiram a mesma tendência, entre redação e ciências da natureza, as médias, da maior para menor, foram, matemática, linguagens e códigos e ciências humanas. Os candidatos declarados pretos não foram tão bem em matemática, e no geral tiveram notas de linguagens e códigos maiores que ciências humanas. Indígenas, tiveram o melhor desempenho em matemática, e em ordem decrescente as médias foram: ciências humanas, linguagens e códigos, ciências da natureza, e redação.

# Figura 22 -Média das notas por raça no município de Nova Friburgo



Legenda: Gráfico que detalha o desempenho nas provas do ENEM por raça no município de Nova Friburgo.

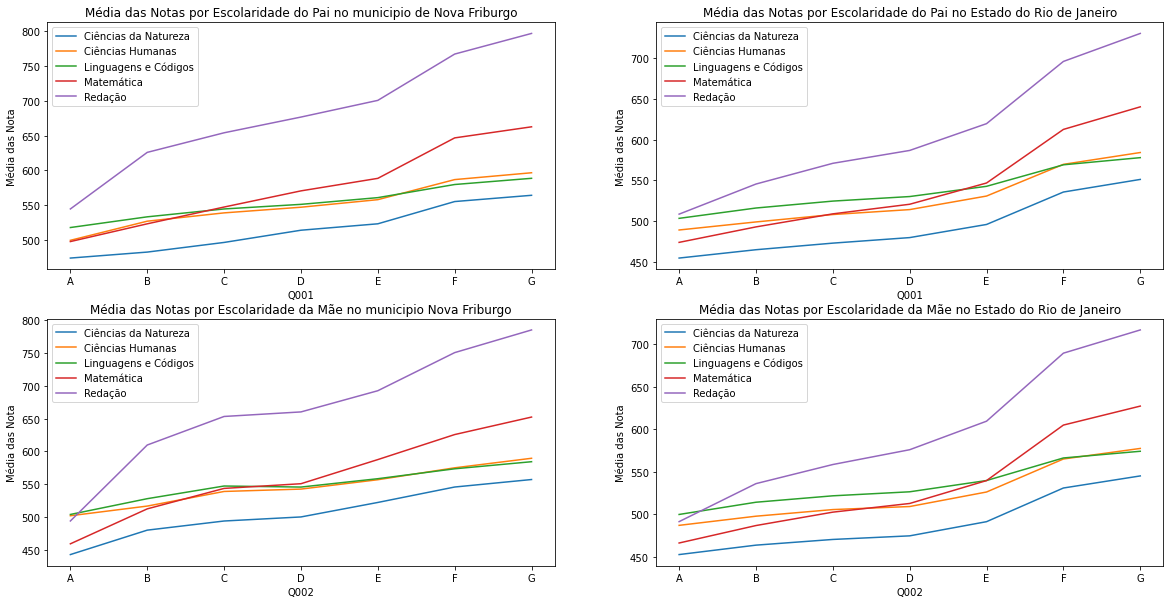
Fonte: O autor, 2021.

#### Análise das médias por escolaridade do pai e da mãe

As duas primeiras perguntas do questionário socioeconômico se referem à escolaridade do pai e da mãe, sendo as opções:

* A - Nunca estudou
* B - Não completou a 4ª série/5º ano do Ensino Fundamental
* C - Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental
* D - Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio
* E - Completou o Ensino Médio, mas não completou a Faculdade
* F - Completou a Faculdade, mas não completou a Pós-graduação
* G - Completou a Pós-graduação
* H - Não sei

# Figura 23 -Média das notas por escolaridade dos pais



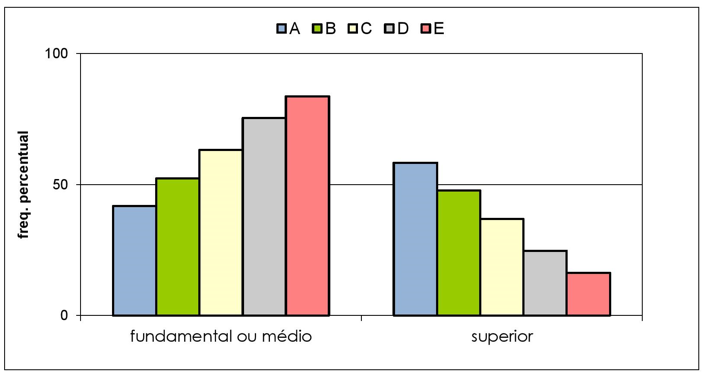
Legenda: Gráfico representando a média das notas do ENEM por escolaridade do pai e da mãe no município de Nova Friburgo e no estado do Rio de Janeiro.

Fonte: O autor, 2021.

Ao observar a análise da média das notas por escolaridade do pai e da mãe (figura 32), nota-se que a média das notas de todas as provas aumentaram, conforme a escolaridade do pai e da mãe aumentam, tanto para os candidatos do Rio de Janeiro, quanto para os candidatos de Nova Friburgo. Para essa análise foram retirados os candidatos que responderam essa questão com a letra H - Não sei.

Ao comparar, com dados fornecidos pela Universidade do Estado do Rio de Janeiro (UERJ) [20], percebe-se comportamento semelhante, candidatos com pais com ensino superior, em sua maioria obtiveram nota A no primeiro exame de qualificação. Por outro lado, candidatos com pais somente com ensino fundamental ou médio, em sua maioria obtiveram nota E no primeiro exame de qualificação.

# Figura 24 -Média das notas por escolaridade dos pais no vestibular da UERJ



Legenda: Gráfico representando as notas do primeiro exame de qualificação da Universidade do Estado do Rio de Janeiro (UERJ) por escolaridade dos pais.

Fonte: (https://www.vestibular.uerj.br/wp-content/uploads/2020/03/2019-SOCIO-PRIMEIRO-  
EXAME.pdf)

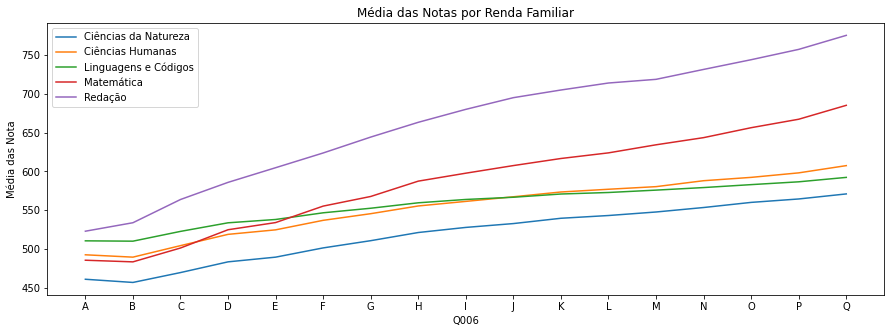
#### Análise das médias por renda mensal familiar

A sexta questão do questionário socioeconômico refere-se a renda mensal familiar, essa questão possui as seguintes opções:

* A - Nenhuma renda.
* B - Até R $998,00.
* C - De R $998,01 até R $1.497,00.
* D - De R $1.497,01 até R $1.996,00.
* E - De R $1.996,01 até R $2.495,00.
* F - De R $2.495,01 até R $2.994,00.
* G - De R $2.994,01 até R $3.992,00.
* H - De R $3.992,01 até R $4.990,00.
* I - De R $4.990,01 até R $5.988,00.
* J - De R $5.988,01 até R $6.986,00.
* K - De R $6.986,01 até R $7.984,00.
* L - De R $7.984,01 até R $8.982,00.
* M - De R $8.982,01 até R $9.980,00.
* N - De R $9.980,01 até R $11.976,00.
* O - De R $11.976,01 até R $14.970,00.
* P - De R $14.970,01 até R $19.960,00.
* Q - Mais de R $19.960,00.

Ao realizar essa análise percebeu-se que a média das notas dos inscritos aumentou conforme a renda aumentava, o gráfico na figura 25 retrata isso. Essa análise foi feita somente para o estado do Rio de Janeiro.

# Figura 25 -Média das notas por renda familiar

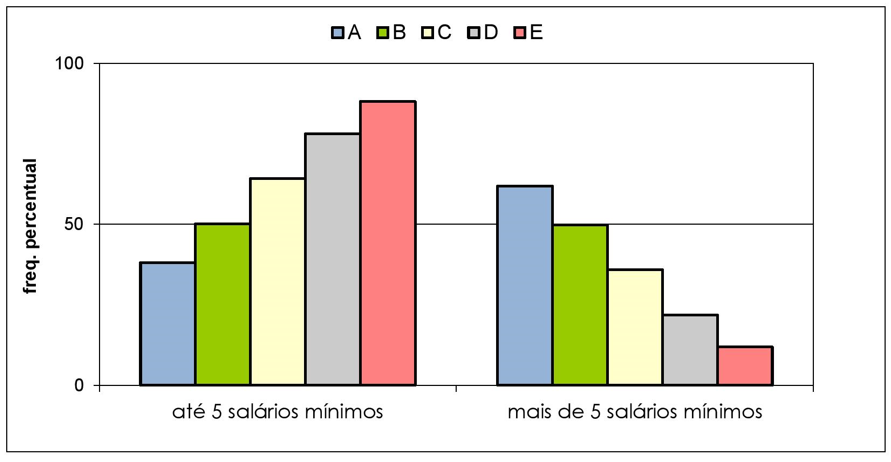


Legenda: Gráfico que mostra a evolução média das notas do ENEM de acordo com a renda familiar dos candidatos..

Fonte: O autor, 2021.

Ao comparar com os dados fornecidos pela UERJ, o comportamento também se assemelha, assim como nas análises de escolaridade; candidatos que possuem maior renda familiar mensal, em sua maior parte obtiveram notas A, enquanto candidatos com renda familiar mensal menor, em sua maior parte obtiveram notas E.

# Figura 26 -Média das notas por renda familiar no vestibular da UERJ



Legenda: Gráfico representando a frequência percentual das notas do primeiro exame de qualificação da Universidade do Estado do Rio de Janeiro (UERJ) por renda familiar dos candidatos.

Fonte: (https://www.vestibular.uerj.br/wp-content/uploads/2020/03/2019-SOCIO-PRIMEIRO-  
EXAME.pdf)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Aqui conclui-se a etapa da análise exploratória, esta etapa ela é dita como essencial num projeto de ciências de dados, pois ela permite que o cientista de dados consiga compreender a natureza dos dados e como eles estão distribuídos, o intuito desse passo foi de descobrir padrões, revelar características, comportamentos estranhos e tendências das principais variáveis do conjunto de dados do ENEM. Um exemplo de como a exploração das variáveis ajudou a maximizar o desempenho do modelo que será apresentado a seguir, foi quando ao notar a importância da escolaridade do pai e da mãe e da renda familiar sobre a média das notas dedicou-se um maior esforço durante as fases de limpeza e modelagem nesses atributos, explicado mais adiante.

As próximas etapas do projeto consistem de dois modelos de aprendizado de máquina, um no qual foram aplicadas técnicas suficientes apenas para que o algoritmo escolhido consiga inferir um resultado, enquanto no segundo modelo, foram aplicadas técnicas um pouco mais avançadas, para que ao final do processo possa-se comparar os resultados e verificar se as aplicações das técnicas aqui descritas, surtem efeito quanto ao desempenho.

Sendo assim, as fases que seguem a análise exploratória, serão explicadas como elas foram diferentemente utilizadas no primeiro, chamado de “*baseline*” e no segundo modelo denominado de “Modelo Final”.

É importante destacar que as fases que se seguem assim como todo o processo de um projeto de ciências de dados proposto pela microsoft não é um processo engessado, então neste trabalho, primeiro foi desenvolvido o modelo nomeado de “*baseline*”, e após isso ele foi constantemente revisitado e treinado diversas vezes para aplicações das técnicas de limpeza e modelagem até chegar em resultado mais satisfatório, e esse resultado foi nomeado de “Modelo final”. Um exemplo claro dessa parte do processo ocorreu durante a estimação do *permutation importante*, onde primeiro treinou-se um esboço do “Modelo final” com todas as técnicas de limpeza e modelagem aplicadas, estimou-se a importância de cada atributo, foi feito uma seleção de acordo com o valor de cada importância, e o modelo foi então retreinado somente com os atributos tidos como mais importantes.

### Limpeza

Após a fase de exploração, iniciou-se no projeto a fase de limpeza do conjunto de dados.

Em ambos os modelos os primeiros passos foram semelhantes:

Carregou-se o arquivo novamente:

# Figura 27 -Abertura de um arquivo csv

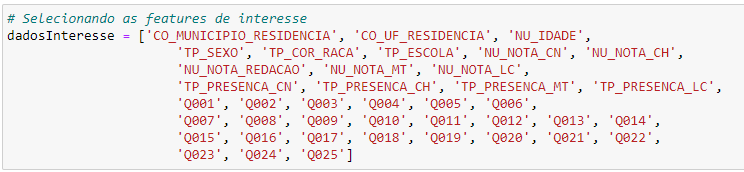
# 

Legenda: Função utilizando a biblioteca *pandas* para leitura e carregamento dos microdados do ENEM.

Fonte: O autor, 2021.

e após isso foram selecionadas apenas os atributos de interesse do autor:

# Figura 28 -Seleção de atributos



Legenda: Seleção dos atributos de interesse do conjunto de dados do ENEM.

Fonte: O autor, 2021.

os atributos selecionados representam:

* Número da inscrição
* Código do município de residência
* Idade
* Sexo
* Raça
* Tipo de escola do Ensino Médio
* Nota da prova de Ciências da Natureza
* Nota da prova de Ciências Humanas
* Nota da prova de Redação
* Nota da prova de Matemática
* Nota da prova de Linguagens e Códigos
* Presença na prova objetiva de Ciências da Natureza
* Presença na prova objetiva de Ciências Humanas
* Presença na prova objetiva de Linguagens e Códigos
* Presença na prova objetiva de Matemática
* Até que série seu pai, ou o homem responsável por você, estudou?
* Até que série sua mãe, ou a mulher responsável por você, estudou?
* A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. (Se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele).
* A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você. (Se ela não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela).
* Incluindo você, quantas pessoas moram atualmente em sua residência?
* Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)
* Em sua residência trabalha empregado(a) doméstico(a)?
* Na sua residência tem banheiro?
* Na sua residência tem quartos para dormir?
* Na sua residência tem carro?
* Na sua residência tem motocicleta?
* Na sua residência tem geladeira?
* Na sua residência tem freezer (independente ou segunda porta da geladeira)?
* Na sua residência tem máquina de lavar roupa? (o tanquinho NÃO deve ser considerado)
* Na sua residência tem máquina de secar roupa (independente ou em conjunto com a máquina de lavar roupa)?
* Na sua residência tem forno micro-ondas?
* Na sua residência tem máquina de lavar louça?
* Na sua residência tem aspirador de pó?
* Na sua residência tem televisão em cores?
* Na sua residência tem aparelho de DVD?
* Na sua residência tem TV por assinatura?
* Na sua residência tem telefone celular?
* Na sua residência tem telefone fixo?
* Na sua residência tem computador?
* Na sua residência tem acesso à Internet?

O conjunto de dados inicial foi então filtrado apenas para conter os atributos de interesse.

A partir desses atributos, filtrou-se ainda mais o conjunto de dados, só que dessa vez a nível de linhas, e não colunas, primeiro foram selecionados apenas os registros de candidatos que estiveram presentes nas provas, e após isso, o eliminou-se os candidatos que tiraram zero em alguma das provas.

Verificou-se então se o conjunto de dados possuía algum valor nulo, e foi encontrado apenas um valor nulo no recurso NU\_NOTA\_REDACAO, que representa a nota de redação do candidato, como só foi encontrado um único valor, esse registro foi deletado em ambos os modelos, caso fossem encontrados mais registros com valores nulos, uma opção seria substituir valores nulos pela média da coluna ou pela mediana.

Para o modelo nomeado de *baseline*, o tratamento de valores faltantes acabou aqui, para o modelo final, explorou-se um pouco mais a limpeza.

Verificou-se que os atributos selecionados referentes a: raça, escola, escolaridade do pai, escolaridade da mãe, profissão do pai e profissão da mãe possuem no questionário a opção “Não declaro”, “Não respondeu” ou “Não sei”. Para os atributos referentes a raça e escola o tratamento escolhido foi a substituição desses valores nulos pela moda, para os atributos referentes a escolaridade e profissão dos pais foi analisado quantos candidatos não sabiam a escolaridade de ambos os pais, por acreditar que esses candidatos não trazem informações relevantes para o modelo, visto que em ambos os atributos haveriam dados que foram tidos como faltantes; A partir da análise foram encontrados 63.171 ou 1.8% dos candidatos que não sabiam a escolaridade de nenhum dos pais, e 110.974 ou 3.1% que não sabiam a ocupação de nenhum dos pais. Como a porcentagem de ambas as análises são baixas, deletou-se esses registros do modelo final.

## Modelagem

* Criação de novos atributos:

Essa técnica foi aplicada somente no “modelo final”.

Durante a fase de limpeza dos dados foram removidos os registros que não continham informações sobre a escolaridade do pai e da mãe, e ocupação do pai e da mãe, porém ainda podem existir registros que contém informações de apenas um dos pais, seja ele do pai, ou da mãe, e para tratar esses possíveis valores ausentes, foram criados dois novos atributos a partir desses quatro, o primeiro recurso é a maior escolaridade entre o pai e a mãe, um recurso que contém apenas o registro da maior escolaridade registrada entre o pai e a mãe do candidato, e o segundo recurso é a maior ocupação, um recurso semelhante ao anterior, onde os valores referem-se apenas a maior profissão registrada entre o pai e a mãe, dessa forma caso um candidato respondesse “não sei” para escolaridade/ocupação de um dos pais, o modelo só irá considerar a escolaridade/ocupação do outro. Após esse tratamento, os atributos utilizados para formar as duas novas colunas foram retirados.

O terceiro e último recurso criado foi feito a partir das colunas que trazem informações sobre a renda mensal familiar do candidato e a quantidade de moradores na residência, a partir da divisão de uma coluna pela outra, foi possível obter a informação da renda mínima por morador da residência. Como esses dois atributos utilizados para criar a nova coluna de renda mínima por morador, não haviam valores faltantes, os dois atributos foram mantidos no modelo.

* Tratamento de *outliers*

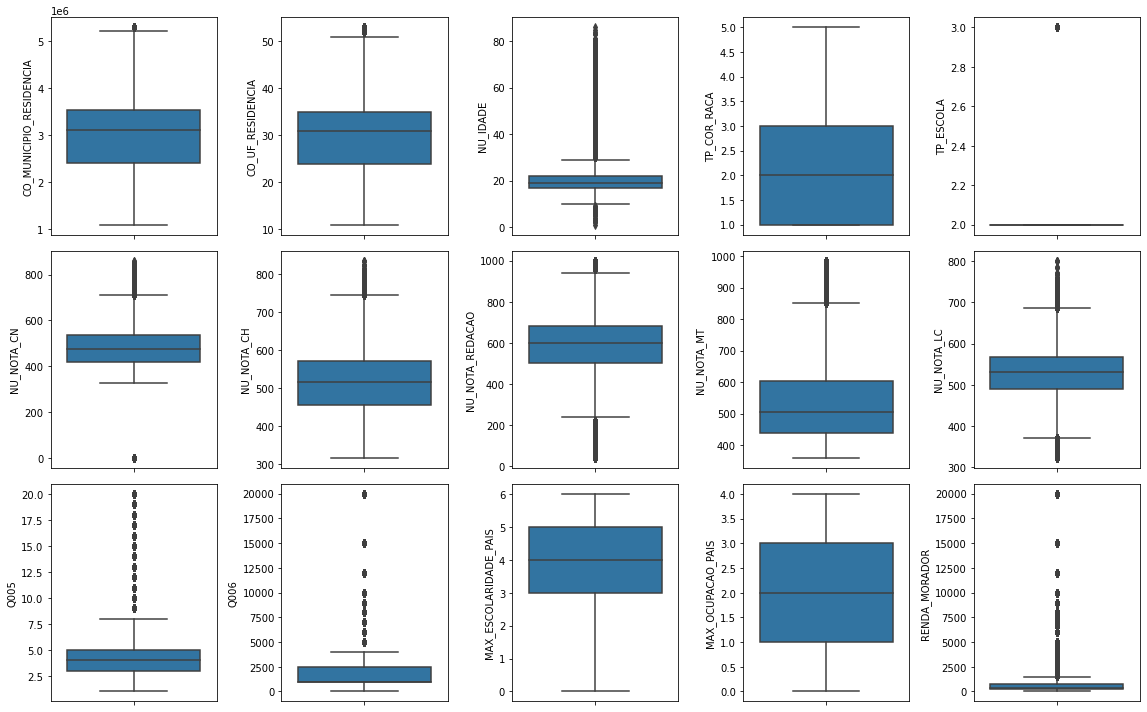
Essa técnica foi aplicada somente no “modelo final”.

Primeiramente, utilizou-se o gráfico *boxplot* (figura 29) para verificar os *outliers* nos atributos numéricos. *Boxplot* é um gráfico no qual pode-se observar a distribuição dos *outliers* nos dados, assim como a distribuição dos dados através dos intervalos interquartis. A caixa presente no gráfico, representa onde os dados estão localizados, sendo os extremos da caixa o primeiro e o terceiro quartis (de baixo para cima), a linha dentro dessa caixa representa o segundo quartil. As linhas abaixo e acima desta caixa indicam o mínimo e o máximo do atributo respectivamente, não levando em consideração os *outliers*. Os pontos após essas duas linhas são os *outliers.*

Percebeu-se então, que alguns atributos numéricos referem-se a informações nominais, são os casos das colunas: CO\_UF\_RESIDENCIA, CO\_MUNICIPIO\_ RESIDENCIA, TP\_COR\_RACA, TP\_ESCOLA. Esses atributos, apesar de serem representados numericamente, não passaram pelo procedimento.

Para então tratar os *outliers* das colunas numéricas restantes do conjunto de dados aplicou-se uma função em *python* capaz de identificar um *outlier*, e atualizar o seu valor para os valores máximos ou mínimos dos intervalos interquartis. A fim de verificar o resultado final, utilizou-se novamente o gráfico *boxplot* (figura 30).

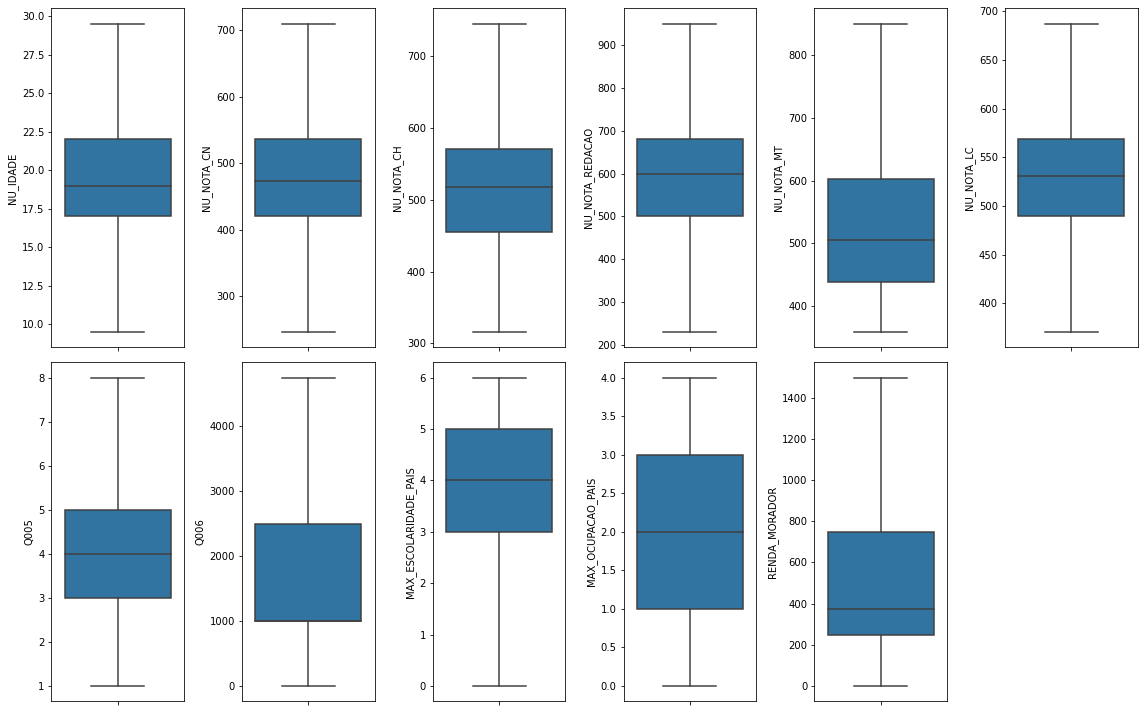
# Figura 29 -Gráfico *boxplot* para visualização dos *outliers*



Legenda: Gráfico que permite a visualização da distribuição dos dadosnos atributos numéricos.

Fonte: O autor, 2021.

# Figura 30 -Gráfico *boxplot* após o tratamento dos *outliers*



Legenda: Gráfico que permite a visualização da distribuição dos dadosnos atributos numéricos após feito o tratamento dos *outliers*.

Fonte: O autor, 2021.

* *One-hot-encoding* & *Ordinal-encoding*

Essa técnica foi aplicada em ambos os modelos.

Os conjuntos de dados de ambos os modelos possuem muitos atributos que não são numéricos; as questões do questionário socioeconômico por exemplo, a maioria são respostas que vão da letra A à F. Portanto, para aplicar o algoritmo de regressão faz-se necessário traduzir esses atributos categóricos nominais e ordinais, para valores numéricos.

Em ambos modelos o procedimento realizado seguiu o mesmo passo-a-passo; primeiro foram identificadas e separadas as variáveis em três grupos: nominais, ordinais, e numéricas; para as variáveis nominais aplicou-se o *one-hot-encoding*, e para as variáveis ordinais aplicou-se o *ordinal-encoding*, ambos descritos na seção 2.3.3 deste trabalho. As variáveis numéricas não necessitam de nenhum tipo de *encoding*.

* Dimensionamento

Essa técnica foi aplicada somente no “modelo final”.

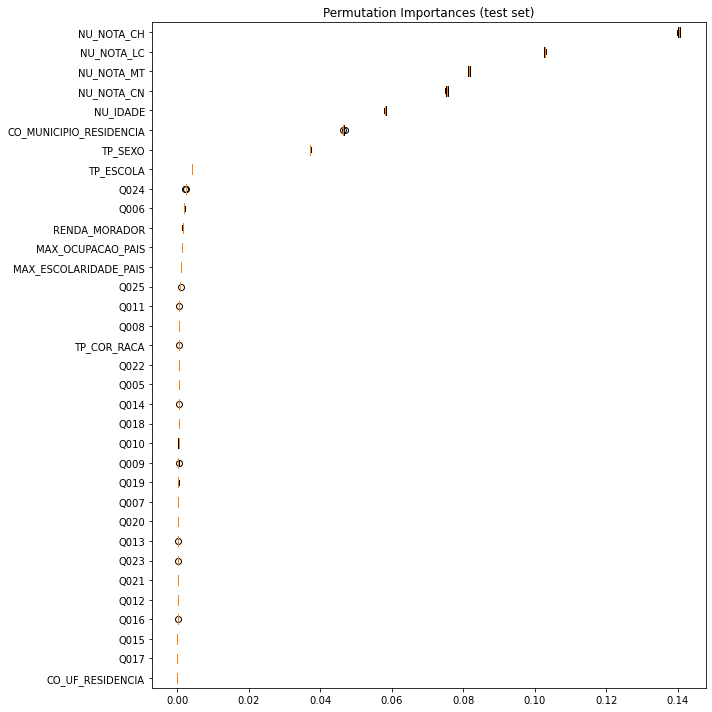
Optou-se por aplicar a técnica de dimensionamento por normalização *min-max*, descrita na seção 2.3.3 deste trabalho, visto que anteriormente houve o tratamento dos *outliers*.

* Seleção de atributos

Essa técnica foi aplicada somente no “modelo final”.

Para melhorar a seleção de atributos do modelo, foi utilizado o *permutation importance*, uma variação do *feature importance*, mas que possui o mesmo intuito. Devido o modelo possuir muitas *features* categóricas nominais que foram distribuídas em várias outras *features* após a aplicação do *one-hot-encoding*, aplicar a importância do recurso no modelo, retornaria a pontuação ganha por cada feature após essa distribuição, enquanto ao utilizar o *permutation importance*, o algoritmo é capaz de estimar uma pontuação para a *feature* no seu formato original, desconsiderando as modificações feitas após a aplicação do *encoding*. Abaixo um gráfico (figura 31) mostra quais os atributos foram considerados mais importantes na hora de estimar a nota de um aluno do ENEM.

# Figura 31 - *Permutation Importance*



Legenda: Gráfico representando as informações referentes ao *permutation importance* no conjunto de dados de teste.

Fonte: O autor, 2021.

Dos 34 atributos iniciais, foram selecionados apenas os 20 tidos como mais importantes, isso foi feito com intuito de diminuir o custo computacional e melhorar a acurácia do modelo.

* Ajustes e definições dos hiperparâmetros dos modelos

Essa técnica foi aplicada somente no “modelo final”.

Utilizou-se a técnica conhecida como *random search*, por considerar que a mesma teria um custo computacional menor que o *grid search*, para auxiliar na otimização dos hiperparâmetros do modelo. Para uma melhor compreensão dos parâmetros do algoritmo, utilizou-se a documentação fornecida pelos desenvolvedores [21].

# RESULTADOS

Como os resultados da análise exploratória já foram apresentados ao longo do capítulo 4, esse capítulo se limitará apresentar os resultados obtidos da modelagem do aprendizado de máquina.

Após a fase de modelagem de ambos os modelos, eles foram separados de forma aleatória em dois conjuntos, sendo um de treino e um de teste, os modelos foram treinados com o primeiro conjunto e testados com dados do segundo conjunto, avaliou-se então a acurácia dos dois modelos conforme as métricas descritas na seção 3 deste trabalho.

Abaixo os resultados de cada modelo.

Resultado do modelo denominado como *“baseline”* :

* Raiz quadrada do erro médio (RMSE): 139.98
* Erro absoluto médio (MAE): 108.66

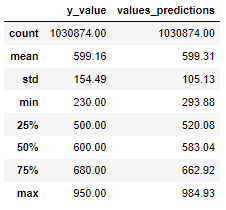
Modelo intitulado de “modelo final” :

* Raiz quadrada do erro médio (RMSE): 112.84
* Erro absoluto médio (MAE): 87.87

Do primeiro para o segundo modelo observou-se uma melhora significativa, obteve-se uma diminuição de 27,14 pontos no erro de raiz quadrático médio e 20,79 pontos no erro absoluto. Essa diminuição pode ser atribuída aos tratamentos dados ao longo do processo de modelagem, o “modelo final”recebeu muito mais técnicas, como uma limpeza mais assertiva, criação de novos atributos, tratamento de *outliers*, dentre outros, quando comparado ao modelo *“baseline”*, onde o processo de modelagem se resumiu a uma limpeza simplista e a aplicação das técnicas de *one-hot-encoding* e *ordinal-encoding*.

A seguir são apresentados dados estatísticos do conjunto de variáveis alvo, e do conjunto de valores previstos pelo modelo.

# Figura 32 -Tabela de dados estatísticos



Legenda: Tabela com os dados estatísticos do conjunto de dados da variável alvo, e do conjunto de dados dos valores previstos pelo modelo.

Fonte: O autor, 2021.

Ao comparar dados estatísticos da variável alvo e dados estatísticos dos valores previstos pelo modelo (figura 32), percebe-se que o desvio padrão da variável alvo é alto, cerca de 154,49 pontos, esse dado representa o quão longe cada instância se encontra com relação à média, essa média no conjunto de dados da variável alvo é muito próximo da média dos valores previstos, diferença de apenas 0,16 pontos, e o desvio padrão do conjunto de predições é menor, cerca de 105,13 pontos. Além da média, os quartis dos dois conjuntos também são relativamente próximos. Abaixo, na figura 33, apresenta-se alguns exemplos de valores para a variável alvo no conjunto de teste, e ao lado o quanto o modelo previu.

# Figura 33 -Tabela de predições do modelo



Legenda: Tabela com os valores reais da variável alvo e dos valores previstos pelo “modelo final” e a diferença entre ambos os valores, o erro.

Fonte: O autor, 2021.

# CONCLUSÃO

A principal conclusão após as análises exploratórias é de que a educação no Brasil ainda é muito desequilibrada e não atinge todas as camadas da população. Dados da UERJ corroboram o fato de que candidatos com pais mais escolarizados e com maiores rendas familiares mensais possuem médias melhores do que outros candidatos, provando que o sistema público de educação ainda peca ao levar a educação para populações de baixa renda.

Quanto aos modelos de aprendizado de máquina, a clara redução das métricas do modelo na qual foram aplicadas mais técnicas de modelagem de aprendizado de máquina evidenciam a importância do processo no conjunto de dados antes do treinamento do algoritmo em si.

E apesar das métricas de erro utilizadas apresentaram magnitudes que a princípio pareçam altas (erro absoluto médio: 87,87 e raiz quadrada do erro médio: 112.84), ao comparar os dados estatísticos do conjunto das variáveis alvos, e do conjunto de dados previstos pelo modelo (figura 32) pode-se concluir que o conjunto de dados de predições tem uma distribuição de dados semelhante ao conjunto de dados de respostas, isso evidencia-se ao analisar alguns exemplos (figura 33), onde apesar de existirem valores previstos que diferem bastante dos valores reais, também nota-se a presença de valores previstos bem próximos dos valores que deveriam ser.

Todavia, o modelo denominado “modelo final” foi então considerado como somente aceitável, talvez ele não seja o modelo ideal para estar em um ambiente de produção (um site que oferecesse o serviço de predição da nota a partir de dados socioeconômicos dos candidatos, por exemplo), contudo ele também não é um modelo a ser descartado. Este trabalho mostrou como as técnicas de modelagem apuraram a sua acurácia, e as técnicas aqui aplicadas podem servir como base para futuros trabalhos neste mesmo conjunto de dados. Uma sugestão para a obtenção de um melhor resultado, seria enriquecer o conjunto de dados com dados não pertencentes aos microdados do ENEM (dados do índice de desenvolvimento humano dos municípios, por exemplo) e explorar melhor algumas técnicas de modelagem, como a otimização dos hiperparâmetros; uma outra opção seria avaliar a troca do algoritmo de regressão, por outros, como algoritmos de classificação, isto é, transformar o problema em um problema de classificação, onde a variável alvo em vez de ser uma variável contínua, seria um rótulo, por exemplo: um intervalo de nota 500-550, representaria uma nota boa, ou mesmo avaliar o uso de redes neurais para a resolução do problema.

# REFERÊNCIAS

1. SILVEIRA, L. F. *Exame Nacional do Ensino Médio (Enem): Uma análise crítica*. scielo, 2021. (https://doi.org/10.1590/S1806-11173710001). Acesso em: 16 de setembro de 2021.
2. SISU. *O QUE É*. (https://sisu.mec.gov.br/#/#oquee). Acesso em: 16 de setembro de 2021.
3. GOV. *Exame Nacional do Ensino Médio (ENEM)*. (https://www.gov.br/inep/pt- br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem). Acesso em: 16 de setembro de 2021.
4. GOV. *Divulgados os resultados finais do exame.* (https://www.gov.br/inep/ pt-br/assuntos/noticias/enem/divulgados-os-resultados-finais-do-exame). Acesso em: 16 de setembro de 2021.
5. GOV. *ENEM 2020 é marco da acessibilidade na educação.* (https: //www.gov.br/inep/pt-br/assuntos/noticias/enem/enem-2020-e-marco-da-acessibilidade-na-educacao). Acesso em: 16 de setembro de 2021.
6. CBINSIGHTS. *The increasingly crowded AI Unicord Club*. (https:// www.cbinsights.com/research/ai-unicorn-club/). Acesso em: 17 de setembro de 2021.
7. MIT SLOAN MANAGEMENT REVIEW. *How big data is empowering AI and machine learning at scale*. (https://sloanreview.mit.edu/article/how-big-data-is- empowering-ai-and-machine-learning-at-scale/). Acesso em: 17 de setembro de 2021.
8. DOMO. *Data Never Sleeps 8.0*. (https://www.domo.com/learn/infographic/ data-never-sleeps-8). Acesso em: 17 de setembro de 2021.
9. GOV. *ENEM. (*https://www.gov.br/inep/pt-br/acesso-a-informacao/dados- abertos/microdados/enem). Acesso em: 18 de setembro de 2021.
10. SAMPAIO, S. A. N.; DANELON, M. T. C. C. *Aplicações da Estatística nas Ciências. aedb, 2021. (*https://www.aedb.br/wp-content/uploads/2015/05/ 64.pdf*)*. Acesso em: 19 de setembro de 2021.
11. Rätsch, G. J. F. M. L. o. t. M. P. S. *A brief introduction into machine learning*. csc villanova, 2021. (http://www.csc.villanova.edu/~tway/courses/mse2400/ s2016/handouts/Ratsch%20-%20Brief%20Intro%20into%20Machine%20Learning.pdf). Acesso em: 19 de setembro de 2021.
12. LUDEMIR, B. T. *Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. scielo, 2021*. (https://doi.org/10.1590/s0103-4014.2021.35101. 007). Acesso em: 19 de setembro de 2021.
13. MICROSOFT. *The Team Data Science Process lifecycle. (*https://docs. microsoft.com/en-us/azure/architecture/data-science-process/lifecycle*).* Acesso em: 19 de setembro de 2021.
14. FORBES. *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says.* (https://www.forbes.com/sites/gilpress/ 2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=b3f90bd6f637) Acesso em: 19 de setembro de 2021.
15. RODRIGUES, S. C. A. *Modelo de regressão linear e suas aplicações*. 2012, Universidade da Beira Interior.
16. TOWARDSDATASCIENCE. *Fundamental Techniques of Feature Engineering for Machine Learning*. (https://towardsdatascience.com/feature-engineering -for-machine-learning-3a5e293a5114). Acesso em: 20 de setembro de 2021.
17. TOWARDSDATASCIENCE. *Feature Selection Techniques in Machine Learning with Python*. (https://towardsdatascience.com/feature-selection- techniques-in-machine-learning-with-python-f24e7da3f36e) Acesso em: 20 de setembro de 2021.
18. MEDIUM. *Otimizando os hiperparâmetros*. (https://medium.com/data-hackers/ otimizando-os-hiperparâmetros-621de5e9be37) Acesso em: 20 de setembro de 2021.
19. SIDRA. *População residente por cor ou raça e religião.* (https://sidra.ibge. gov.br/tabela/2094#/n1/all/n2/all/n3/all/v/1000093/p/last%201/c86/allxt/c133/0/d/v1000093%201/l/v,p+c86,t+c133/resultado) . Acesso em: 22 de setembro de 2021.
20. UERJ. *Vestibular estadual 2019 1º exame de qualificação.* (https:// www.vestibular.uerj.br/wp-content/uploads/2020/03/2019-SOCIO-PRIMEIRO-EXAME.pdf). Acesso em: 22 de setembro de 2021.
21. XGBOOST. *XGBoost Parameters.* (https://xgboost.readthedocs.io/en/latest/ parameter.html). Acesso em: 23 de setembro de 2021.