UNIVERSIDADE FEDERAL DE ALAGOAS

INSTITUTO DE COMPUTAÇÃO

MANOELA CASSIA DOS SANTOS

Orientador: Prof. Dr. Bruno Almeida Pimentel

DESENVOLVIMENTO DE MODELOS PARA ESTIMAR PRINCIPAIS CARACTERÍSTICAS SOCIOECONÔMICAS PARA AS NOTAS DE REDAÇÃO E MATEMÁTICA PARA ENEM 2018

Maceió, AL

2020

UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO

MANOELA CASSIA DOS SANTOS

DESENVOLVIMENTO DE MODELOS PARA ESTIMAR PRINCIPAIS CARACTERÍSTICAS SOCIOECONÔMICAS PARA AS NOTAS DE REDAÇÃO E MATEMÁTICA PARA ENEM 2018

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Prof. Dr. Bruno Almeida Pimentel

Maceió, AL
2020

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Taciana Sousa dos Santos – CRB-4 – 2062

S237d Santos, Manoela Cassia dos.
Desenvolvimento de modelos para estimar principais características socioeconômicas para as notas de redação e matemática para ENEM 2018 / Manoela Cassia dos Santos. – 2021.
41, [3] f. : il. color.

Orientador: Bruno Almeida Pimentel.
Monografia (Trabalho de Conclusão de Curso em Ciência da Computação) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2020.

Bibliografia: f. 38-39.
Apêndice: f. 41-[44].

1. Análise preditiva. 2. Análise de dados. 3. Aprendizagem de máquina. 4. Mineração de dados (Computação) . I. Título.

CDU: 004

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Instituto de Computação – IC
Campus A. C. Simões - Av. Lourival de Melo Mota, BL 12
Tabuleiro do Martins, Maceió/AL - CEP: 57.072-970
Telefone: (082) 3214-1401

Trabalho de Conclusão de Curso – TCC

Formulário de Avaliação

Curso: Bacharel em Ciência da Computação

Nome do Aluno

MANOELA CASSIA DOS SANTOS

Nº de Matrícula

13210308

Título do TCC (Tema)

Desenvolvimento de Modelos para Estimar Principais Características Socioeconômicas para as Notas de Redação e Matemática para Enem 2018

Banca Examinadora

Bruno Almeida Pimentel Nome do Orientador	 Assinatura
Lucas Benevides Viana de Amorim Nome do Professor	Lucas Amorim Assinatura
Rafael de Amorim Silva Nome do Professor	Rafael A. Silva Assinatura

Data da Defesa

10 / 08 / 2020

Nota Obtida

8,0 (oito)

Observações

Coordenador do Curso
De Acordo

Roberta V. V. Lopes
Assinatura

Errata

Manoela Cassia dos Santos, **DESENVOLVIMENTO DE MODELOS PARA ESTIMAR PRINCIPAIS CARACTERÍSTICAS SOCIOECONÔMICAS PARA AS NOTAS DE REDAÇÃO E MATEMÁTICA PARA ENEM 2018** . nº de páginas. Monografia- , Universidade Federal de Alagoas, Maceió, 2020.

Página	Linha	Onde se lê	Leia-se
16	10		

Dedico este trabalho a minha mãe que sempre acreditou e me deu apoio nesta jornada.

Agradecimentos

Agradeço a todos que de alguma forma me ajudaram, com conhecimento ou motivação. Amigos, famíliares, orientador, este trabalho também é extensão de vocês.

"A ciência de hoje é a tecnologia de amanhã."
Edward Teller

Resumo

Análise preditiva é uma técnica analítica avançada que usa dados, algoritmos e Aprendizagem de Máquina para antecipar tendências e fazer projeções nos negócios. Utilizando como base os dados coletados pelo Exame Nacional do Ensino Médio (ENEM) do ano de 2018, esse trabalho tem como proposta analisar os dados socioeconômicos dos participantes, a fim de aprimorar modelos regressores para prever as notas de Redação e de Matemática dos participantes. Foram desenvolvidos modelos preditivos utilizando algoritmos de Aprendizagem de Máquina (Rede Neural, Random Forest, Árvore de Decisão, Regressão Linear). Após aplicação métricas de avaliação de modelos preditivos, foi constatado que o modelo Rede Neural teve melhor desempenho na predição das notas de Matemática, enquanto Regressão Linear na predição das nota de Redação. Destacaram-se algumas características socioeconômicas sobre tipo de escola, sexo, computador, ocupação do pai, renda familiar mensal, influentes na predição das notas de Redação e Matemática.

Palavras-chave: Ciência de Dados. Análise de Dados. Aprendizado de Máquina. Análise Preditiva. Mineração de Dados. ENEM.

Abstract

Predictive analytics is an advanced analytical technique that uses data, algorithms and Machine Learning to anticipate trends and make business projections. Using the data collected by the Exame Nacional do Ensino Médio (ENEM) in 2018, this work aims to analyze the socioeconomic data of the participants, in order to improve regression models to predict the Writing and Mathematics scores of the participants. Predictive models were developed using Machine Learning algorithms (Neural Network, Random Forest, Decision Tree, Linear Regression). After applying metrics to evaluate predictive models, it was found that the Neural Network model performed better in predicting math scores, while Linear Regression in predicting writing scores. Some socioeconomic features on type of school, sex, computer, father's occupation, monthly family income, which are influential in the prediction of Writing and Mathematics scores, stood out.

Keywords: Data Science. Data Analysis. Machine Learning. Predictive Analysis. Data Mining. ENEM Predictive Analysis.

Lista de Ilustrações

Lista de Abreviaturas e Siglas

ABNT	Associação Brasileira de Normas Técnicas
IC	Instituto de Computação
UFAL	Universidade Federal de Alagoas
ENEM	Exame Nacional do Ensino Médio
TCC	Trabalho de Conclusão de Curso
BI	Business Intelligence
AM	Aprendizagem de Máquina
MD	Mineração de dados
DICS	Dados, Informação, Conhecimento e Sabedoria
KDD	knowledge discovery in databases
DW	Data Warehouse
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MSE	Mean Squared Error
MAE	Mean Absolute Error
CV	Cross Validation
MLP	Multi-layer Perceptron
R^2	Coeficiente de Determinação

Sumário

1 Introdução

A organização de mecanismos de avaliação da educação básica, com objetivos de melhorar e criar prioridades na melhoria da qualidade do ensino nacional, é responsabilidade da União definida pela lei das Diretrizes e Base da Educação Nacional (LDB). O Exame Nacional do Ensino Médio (ENEM) é um destes mecanismos, criado em 1998 pelo instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) órgão vinculado ao Ministério da Educação.

O papel inicial do ENEM era avaliar o desempenho dos alunos quando terminavam a educação básica, após reformulação em 2009, este exame passou a ser realizado não apenas como forma de avaliação, mas também por alunos, na etapa final da educação básica, que almejam uma vaga nos principais programas de Educação Superior: Sistema de Seleção Unificada (SISU), Programa Universidade pra Todos (ProUni), Fundo de Financiamento ao Estudante do Ensino Superior (Fies). CASSIANI, SILVA e PIERSON (2016)

Vários trabalhos foram produzidos sobre o ENEM ao londo de sua trajetória. O INEP vem produzindo, a cada nova edição, inúmeros documentos (relatórios pedagógicos) e bases de dados (microdados), que compreendem os resultados, conceitos e análises decorrentes deste sistema de avaliação da educação nacional.CASSIANI, SILVA e PIERSON (2016). O trecho a seguir descreve com palavras do INEP (2005, P. 8) a importância deste exame.

> A análise dos resultados do desempenho dos participantes do Enem permite a identificação de lacunas em seu aprendizado e, também, das potencialidades que ele apresenta ao final da escolaridade básica

Os autores CASSIANI, SILVA e PIERSON (2016) deixam claro que mesmo com todos esse dados, a maioria dos trabalhos encontrados são voltados em classificar, alunos, instituições, não em criar mecanismo ou descobrir insight que possam ajudar os professores com a prática diária. Este trabalho visa, descobrir insight que possam ser usados para melhorar a qualidade da educação nacional, utilizando os microdados do ENEM 2018, aplicando técnicas de Aprendizagem de Máquina estimando resultados e capturando características com maiores relevâncias dos participantes.

1.1 Justificativa

Em 2018, cerca de 6.774.891 estudantes de todo o país se inscreveram para fazer esse o ENEM segundo INEP. O foco deste trabalho será utilizar os microdados do ENEM 2018 (contém os dados destes 6.774.891 participante) para implementar modelos de Aprendizagem de Máquina

que possam revelar tendências, prever resultados sobre estes estudantes para as notas de Redação e Matemática, que auxiliem na tomada de decisão e direcionamento de políticas públicas e privadas de todos Brasil. Serão deixados três questionamentos sobre esses participantes com relação aos dados socioeconômicos: (1) Ter computador na residência influencia a nota de Redação e de Matemática? (2) A renda do participante influencia as notas de Redação e de Matemática? (3) A idade do participante tem muita influencia das notas de Redação e Matemática? Perguntas que serão respondidas ao decorrer deste trabalho.

1.2 Objetivos

Analisar os dados do ENEM 2018 para extrair informações relevantes para a sociedade aplicando modelos de Aprendizagem de Máquina que possam aprender e mostrar características socioeconômicas relevantes ao estimar as notas de Redação e Matemática.

1.2.1 Objetivo Principal

Implementar modelos preditivos regressores para prever as notas de Redação e Matemática do ENEM 2018, analisar as características socioeconômicas dos participantes que mais influenciaram os modelos preditivos; responder os questionamentos apresentado na justificativa.

1.2.2 Objetivos Específicos

- Criar, modelar e executar modelos preditivos para um conjunto de dados de amostragem simples de participantes residentes em todo o Brasil do ENEM 2018.
- Analisar eficiência dos modelos preditores, utilizando métricas para modelos regressores.
- Analisar características socioeconômicas utilizando a propriedade feature importance presente nos modelos Random Forest e Árvore de Decisão.
- Utilizar permutations importance para inspecionar os modelos e obter as características mais importantes para os modelos Random Forest, Árvore de Decisão, Rede Neural, Regressão Linear.

1.3 Organização do Trabalho

Serão apresentados os principais conceitos envolvendo a temática deste trabalho, seguido pelo estudo de caso, e por fim, uma análise geral dos resultados encontrados.

1.3.1 Estrutura da Monografia

Capítulo 1, serão descritos os principais pontos do trabalho; **Capítulo 2**, serão definidos dos principais conceitos relacionado ao tema abordado; **Capítulo 3**, exposição dos principais pontos da metodologia utilizada para estimação das notas de Redação e Matemática; **Capítulo 4**, apresentação, interpretação e analise dos resultados obtidos para estimação das notas de Matemática e Redação do ENEM 2018; **Capítulo 5** esclarecimento final acerca dos resultados coletados pela análise preditiva do ENEM 2018 e propostas para continuidade do estudo.

2 Embasamento Teórico

Apresentação de trabalhos relacionados, explanação dos conceitos envolvendo Aprendizagem de Máquinas, alguns modelos de aprendizagem sobre o paradigma de aprendizagem supervisionada e principais formas de validação de modelos preditivos.

2.1 Trabalhos Relacionados

A monografia Gerald et al. (2014) intitulada "Pesquisa em Educação Matemática: desafios à prática docente"trata do tema ENEM sobre as questões de Matemática, fazendo uma análise sobre a perspectiva de um professor; fala sobre metodologias e estratégias de ensinar matemática; investigação de como ensinar matemática e os alunos gostarem da disciplina; utilização de jogos que permitissem o entendimento de conceitos matemáticos de forma lúdica, simples e eficiente.

O trabalho de Gerald et al. (2014) apresentou diferenças no estudo abordado neste trabalho, pois não foram aplicadas técnicas de Aprendizagem de Máquina. O estudo dele foi focado em aplicação prática à docência, não foram capturadas nem estudadas características socioeconômica dos participantes.

2.2 Aprendizagem de Máquina - AM

Aprendizagem de Máquina é uma área muito popular atualmente, seus conceitos e aplicações vão muito além de máquinas capazes de realizar tarefas de humanos, ela está presente no nosso cotidiano com tarefas simples como filtragem de emails entre “spam” e “não-spam”; sugestões de palavras em um editor de texto; anúncios em lojas e sites; ou em tarefas mais elaboradas como desenvolver sistema capaz de estimar o risco de um paciente desenvolver câncer de próstata (SILVA, 2012). Segundo Géron (2019) “Aprendizado de Máquina é o campo de estudo que dá aos computadores a habilidade de aprender sem ser explicitamente programado [Arthur Samuel, 1959]". Um dos paradigmas mais abordados em AM é o aprendizado supervisionamento.

2.2.1 Aprendizado Supervisionado

No aprendizado supervisionado, os dados de treinamento são aplicando aos algoritmos também incluem as soluções desejadas. Assim é possível que os modelos criem mecanismos que possa ser usados posteriormente para classificar, ou prever resultados em bases de dados. Géron (2019)

2.2.1.1 Regressão Linear

O objetivo da regressão linear é estabelecer uma relação direta entre uma variável de resposta (variável resultado ou dependente) e uma variável preditora (variável covariante ou independente) Chen (2018). Para representar a relação entre uma variável dependente (y) e uma variável independente (x), usamos o modelo

$$Y = \beta_0 + \beta_1 \cdot x$$

Quando implementamos um modelo de Regressão Linear, queremos encontrar os valores dos parâmetros β_0 e β_1 que melhor representam o relacionamento entre as variáveis.

2.2.1.2 Árvore de Decisão

Segundo Freitas e Rozenberg (2002) uma Árvore de Decisão é uma estrutura de representação de conhecimento, onde ramos e nós são organizado em forma de árvore tal que:

1. cada nó interno (não folha) é rotulado com o nome de um dos atributos do preditor;
2. os ramos que saem de um nó interno são rotulados com valores do atributo naquele nó;
3. cada nó folha é rotulado com uma classe (um valor do atributo objetivo).

Árvore de Decisão é um modelo preditivo, simples e eficiente, que pode ser usado para classificação e estimação de valores. Sua estratégia é organizar os dados de forma que aumentem a probabilidade de alcançar o objetivo Rokach (2008).

2.2.1.3 Random Forest

Random Forest é um conjunto de árvores de decisão geradas por subconjuntos do conjunto original, onde esses subconjuntos são escolhidos aleatoriamente. Este algoritmo pode ser usado para classificar, quando o objetivo for agrupar elementos por características semelhantes (cor do cabelo, idade, sexo), ou regressão, quando o objetivo é encontrar um valor (o preço de um carro, notas em um exame).Natingga (2017)

A "floresta" que ele cria é uma combinação de árvores de decisão, na maioria dos casos treinados com o método de bagging. A ideia principal do método de bagging é que a combinação dos modelos de aprendizado aumenta o resultado geral. Essa combinação de modelos, torna random forest um algoritmo muito mais poderoso do que Árvore de Decisão.

2.2.1.4 Rede Neural

Haykin (2007) diz que a Rede Neural na sua forma mais básica é uma máquina que é projetada para modelar a maneira como o cérebro realiza uma tarefa particular ou função de

interesse. Mostra ainda dois aspectos em que as Redes Neurais artificiais se assemelha ao cérebro. Haykin (2007)

1. O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem.

2. Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenas o conhecimento adquirido.

Algoritmos de redes neurais utilizam, como forma de aprendizagem, as alterações dos pesos sinápticos dos neurônios, de forma ordenada, visando alcançar um objetivo predefinido. Haykin (2007)

2.3 Validação de Modelos

Uma vez que os modelos já foram criados e executados, é necessários avaliar se resultados encontrados podem ser realmente aplicados em produção.

2.3.1 Validação Cruzada - CV

A Validação Cruzada funciona com a separação do conjunto total, em dados de teste e dados de treinamento, a fim de simular uma previsão fora da amostra. McKinney (2019, P. 20) A abordagem básica (k-fold), consiste em dividir o conjunto de dados em k grupos, onde os (k-1) grupos são usados para treinamento e 1 grupo usado para validação, este procedimento é repetido por k iterações. Haykin (2007, P. 244) A figura a seguir ilustra o funcionamento do Validação Cruzada k-fold com k = 5 grupos total de dados (All Data), onde (K-1) grupos são os dados de treinamento (Training data), 1 grupo com dados de validação (Test data), em K-iteração.

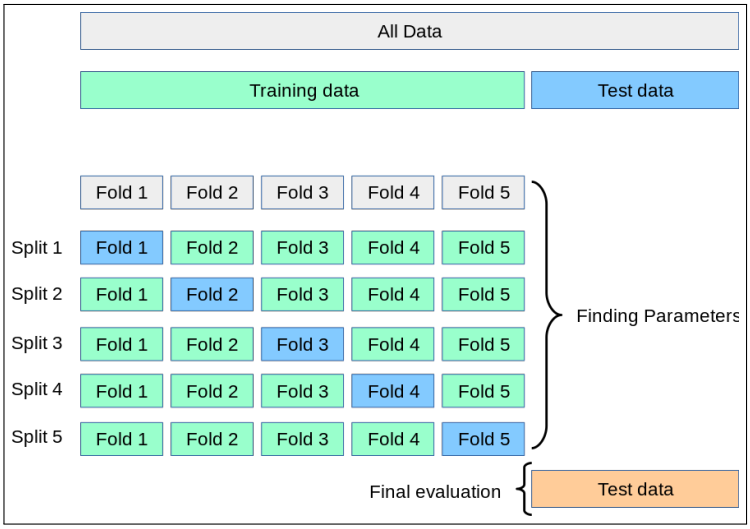

Figura 2.1 – Validação Cruzada K-Fold, (SCIKIT-LEARN, 2020)

2.3.2 Erro Máximo

Essa métrica calcula o pior erro absoluto, diferença em módulo entre as saídas originais e finais, para cada amostra do conjunto de dados observado.Reda e Shafique (2018)

Se $\hat{y}_i$ é o valor previsto da i-ª amostra, e y_i é o valor verdadeiro correspondente, o erro máximo é definido como:

$$MaxError(y, \widehat{y}) = max\left(|y_i - \widehat{y_i}|\right)$$

2.3.3 MAE - Erro Médio Absoluto

Erro médio absoluto calcula somatório das diferenças absolutas entre os valores reais e estimados para cada amostra, dividido pelo número total de amostras do conjunto. Lima e Oliveira (2016, P. 20).

Se $\widehat{y_i}$ é o valor previsto da i-ª amostra, n o número total de amostras, y_i é o valor verdadeiro correspondente, então o MAE será:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \widehat{y_i}|}{n}$$

2.3.4 MSE - Erro Médio ao Quadrado

Erro médio ao quadrado é o somatório das diferenças entre os valores reais e os valores estimados de cada amostra, elevados ao quadrado e dividido pelo número de observações (total de amostras) do conjunto Lima e Oliveira (2016, P. 22).

Se $\widehat{y_i}$ é o valor previsto da i-ª amostra, n o número total de amostras, y_i é o valor verdadeiro correspondente, então o MSE será:

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{n}$$

2.3.5 R² - Coeficiente de Determinação

O coeficiente de determinação (R^2) é uma métrica de regressão comum. Geralmente seu resultado está entre 0 e 1, representando o percentual da variância do alvo (y) com o qual os atributos x contribuem. Quanto mais próximo de 1, menor é o percentual de variação que não é explicado da variação total.Harrison (2019). Segundo Fumio (2000, P. 21) R² pode ser negativo quando se utiliza da fórmula para modelos regressores com variáveis não constantes mostrada a seguir. Onde $\widehat{y_i}$ é o valor previsto da i-ª amostra, n é o número total de amostras e y_i é o valor verdadeiro correspondente.

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n} (e_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y_i})^2} \right) Onde\ \bar{y_i} = \sum_{i=1}^{n} (y_i),\ e_i = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

Isso ocorre porque, sem o benefício de um intercepto, a regressão pode ser pior que a média da amostra em termos de rastreamento do dependente variável Fumio (2000) .

2.3.6 Features Importances

Uma das vantagens dos modelos Random Forest e Árvore de Decisão é que eles possuem a propriedade Feature Importances que fornece que determina como a dados foram distribuídos nos nós de decisão. Essas importâncias podem ajudar a informar a seleção de features mais relevantes de um conjunto de dados. Molin (2019)

A ideia desse algoritmo é diminuir a impureza do do nó, em modelos regressores essa redução ocorre reduzindo a variância aplicando o Erro Médio ao Quadrado (MSE). Então para cada árvore: primeiro, se calula o ganho de informação (GI) do nó, que equivale a impureza desse nó subtraída da soma ponderada das impurezas dos nós filhos; segundo, a importância da feature (IF) é cacula dividindo o GI do nó pelo somatório do GI de todos os nós da árvore; terceiro, o IF é normalizado, isso se dá dividindo o IF do nó pelo somatório do IF de todos os nós da árvore. Para o Random Forest este procedimento é calculado para todas as árvore, somando o IF de cada uma e depois divido pelo número de árvores do modelo.

2.3.7 Permutation Importances

Fornece as importâncias das features, uma técnica de inspeção de modelo, podendo ser usada para qualquer estimador treinado com dados tabulares (pois a técnica trata a manipulação destas tabelas). Essa técnica se beneficia de ser independente do modelo e pode ser calculada várias vezes com diferentes permutações das features para um mesmo modelo (BREIMAN, 2001).

A ideia dela é calcular o score (R^2, MSE, MAE) de um estimador treinado quando uma feature não está disponível, então saberíamos sua relevância no conjunto. O procedimento não é realmente retirar a feature, pois, seria necessário treinar o estimador a cada vez que uma feature fosse removida, então para não haver novo treinamento para cada feature seus valores são permutando, produzindo uma espécie de ruído, então é calculado o score novamente do modelo, a diferença entre o score original e o score com ruído é a importância da feature.

3 Estudo de caso

A motivação deste projeto é desenvolver uma Análise Preditiva sobre o conjunto de dados do ENEM 2018, disponibilizados no portal do INEP (INEP, 2020). Dentro da análise encontrar fatores socioeconômicos que mais influenciaram nas notas de Redação e Matemática dos candidatos do ENEM do ano de 2018.

3.1 Visão Geral do Caso

Os microdados do ENEM 2018 englobam as provas, os gabaritos, informações sobre os itens (questões), as notas e o questionário socioeconômico respondido pelos participantes inscritos. Para execução do projeto serão abordados os dados referente as notas e o questionário socioeconômico dos participantes. Os dados do questionário são questões que variam de Q001 até Q027, questões de múltipla escolha, variando de no máximo (A - P) alternativas, um conjunto de dados muito rico e consideravelmente volumoso. Mais informações sobre os dados do questionário ou os tipos de respostas podem ser visto no Dicionário de Microdados ENEM 2018.

3.2 Análise do Caso

Para fazer a análise preditiva sobre o conjunto de dados do ENEM, adotaremos o modelo de projeto com as seguintes etapas;

1. Entender o problema
2. Obter os dados que queremos que nosso modelo trabalhe
3. Pré-processamento
    - Pré-processamento:

      Limpeza

      Integração

      Seleção

      Transformação
4. Explorar os dados
5. Modelagem
    - Desenvolvimento e execução dos modelos

6. Validação dos resultados
    - Aplicação de Métricas de avaliação para modelos preditores.
7. Execução e Otimização

3.2.1 Entender o problema

Nessa etapa temos que definir o que queremos desenvolver, ou o que precisamos pra solucionar o problema. Definir os tópicos, requisitos e abordagem que utilizaremos. Neste projeto foi adotado uma abordagem Nacional (df_nacional), com participantes residentes em todo Brasil, para criar modelos preditivos, aplicar nessa base e analisar se os resultados que foram encontrados.

3.2.2 Obter os dados

Etapa crucial em qualquer desenvolvimento de projeto Data Science, pois, o conjunto de dados de fontes seguras, trazem mais confiabilidade aos resultados encontrados. O conjunto de dados que foi utilizado é encontrado no portal do INEP (INEP, 2020), na página de microdados, ENEM, ano de 2018.

3.2.3 Pré-processamento

É constituído pelas etapas de Limpeza, Seleção e Transformação. Nesses processos ocorrem remoção de dados redundantes, ausentes, duplicados, desnecessários, conversão de tipos, ou seja, prepararam os dados para as próximas etapas. Na

Limpeza, foram removidos os dados ausentes com (Not a Number - NaN).

Seleção, Todos esses dados foram escolhidos com o objetivo de representar a maior quantidade de participantes, então, questões mais específica para grupos isolados, ou dados com sentido redundantes não foram selecionados. Foram selecionados dados sobre participantes (NU_INSCRICAO, SG_UF_RESIDENCIA, NU_IDADE, TP_SEXO, TP_ESTADO_CIVIL, TP_COR_RACA, IN_TREINEIRO, TP_ANO_CONCLUIU, TP_ESCOLA, TP_ENSINO); dados da escola (TP_DEPENDENCIA_ADM_ESC, TP_LOCALIZACAO_ESC) ; dados da prova objetiva (NU_NOTA_CN, NU_NOTA_CH, NU_NOTA_LC, NU_NOTA_MT); dados da redação (NU_NOTA_REDACAO); e todos dados do questionário socioeconômico: (Q001, Q002, Q003, Q004, Q005, Q006, Q007, Q008, Q009, Q010, Q011, Q012, Q013, Q014, Q015, Q016, Q017, Q018, Q019, Q020, Q021, Q022, Q023, Q024, Q025, Q026, Q027). Mais informações e descrição sobre os esse dados podem ser visto no dicionário de dados em apêndice.

Transformação, foi utilizado uma sequência na transformações nos dados numéricos (*numeric_features*) no qual foi aplicado o transformador numérico (*numeric_tranformer*) aplicando o método de *StandardScaler()*; Nos dados categóricos (*categorical_features*) foi aplicado

o transformador categórico (*categorical_transformer*) aplicando o método de padronização *OneHotEncoder()*.

3.2.4 Entender o conjunto de dados

Os gráficos a seguir foram plotados com dados de uma amostra de 70.000 participantes colhidos dos microdados do ENEM 2018. Os gráficos mostra como de alguma features que queremos saber sua importância da predição das notas de Redação e de Matemática pro ENEM 2018.

3.2.4.1 Explorando as features: idades e sexo do participantes

O gráfico a seguir mostra a quantidade de participante por grupos etários de acordo com seu gênero. Segundo o gráfico, a maior quantidade de participantes está nos grupos entre (16 e 17 anos) e (18 e 19 anos) do sexo feminino, seguidos pelo mesmo grupo etário, porém do sexo masculino. O grupo menos representativo é formado de participantes com mais de 60 anos, vale notar que existem mais participante no grupo (50 e 60 anos) do que no grupo (14 e 15 anos).

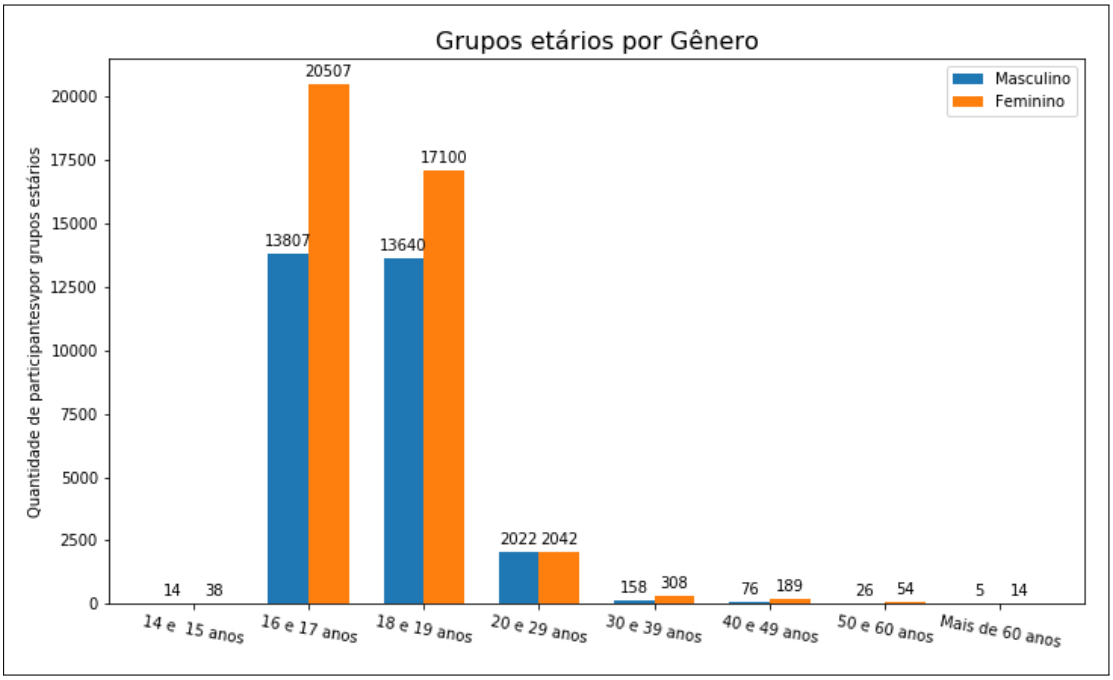

Figura 3.1 – Gráfico: Quantidade gênero de participante do ENEM 2018 separados em grupos etários

3.2.4.2 Exploração dos dados da questão Q006

A quantidade de participante entre os grupos B e o grupos C representa a maior parcela dos participantes do ENEM 2018. Se fosse somado todos os participantes do grupos F até Q o número de participantes ainda seria menor que a soma dos grupos A, B, C e D.

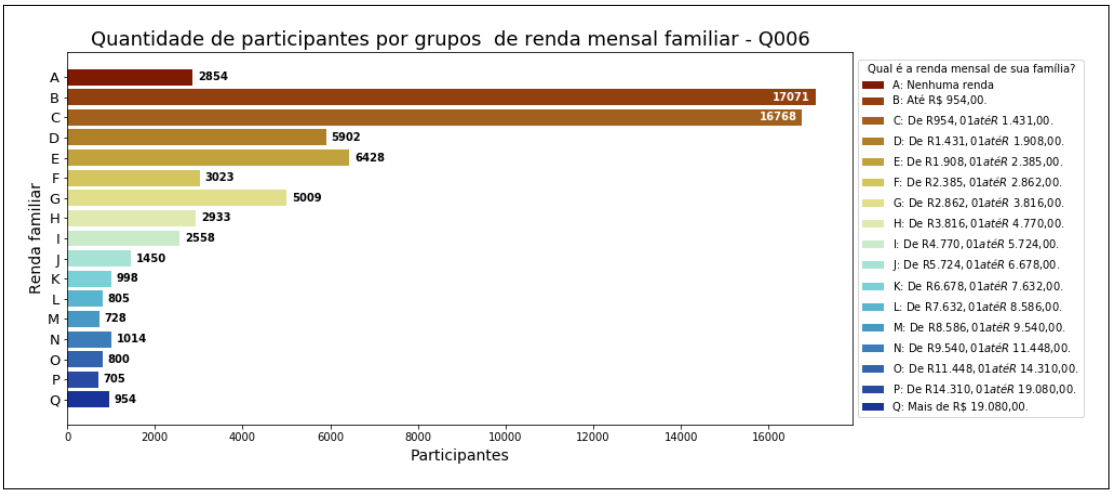

Figura 3.2 – Gráfico: Participantes separados de acordo com o grupo de renda familiar mensal

3.2.4.3 Exploração dos dados da questão Q024

A quantidade de participantes que, no ENEM 2018, responderam não ter computador em sua residência foi cerca de 42% (29.404) participante de um total de 70 mil com essa mesma realidade, enquanto 58% destes participantes têm pelo menos 1 computador em sua residência.

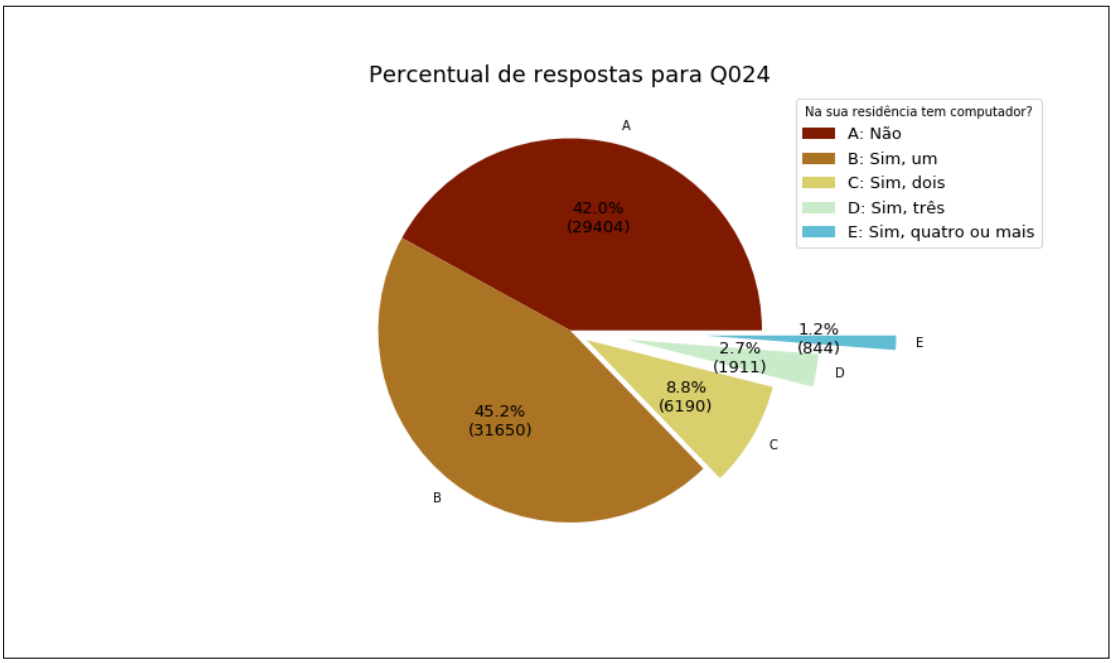

Figura 3.3 – Gráfico: Percentual de participantes de acordo com a resposta da questão Q024 do questionário socioeconômico

3.2.5 Modelagem

Construção de modelos preditivos utilizando quatro algoritmos de AM (Random forest, Árvore de Decisão, Regressão Linear, Rede Neural). O objetivo é encontrar as notas e com base

na estratégia de cada modelo, capturar a importância das features e saber quais tiveram mais relevância na predição das notas de Matemática e Redação.

Os dados foram separados em features alvos (NU_NOTA_REDACAO, NU_NOTA_MT) respectivamente representando o alvo na predição da nota de Redação e o alvo para predição da nota de Matemática. Das features capturadas na seleção não foram incluídas treinamento as seguintes: features referentes as notas (NU_NOTA_CN, NU_NOTA_CH, NU_NOTA_LC, NU_NOTA_MT, NU_NOTA_REDACAO); features de identificação e localização (NU_INSCRICAO, SG_UF_RESIDENCIA), pois o objetivo é encontrar as features socioeconômicas mais importantes, sem especificar por localização ou identificar o participante, uma ideia geral no sentido nacional.

3.2.6 Validação do modelos

Serão aplicadas as métricas de validação para modelos regressores (Erro Máximo, MAE, MSE, Coeficiente de Determinação - R^2) para analisar o desempenho de cada modelo (Random Forest, Árvore de Decisão, Rede Neural e Regressão Linear) na estimação das notas de Redação e Matemática ao aplicar Validação Cruzada. Em Seguida serão capturas propriedade relacionadas as importâncias das features (Features Importances e Permutation Importances) para os modelos preditivos.

3.2.7 Concluir desenvolvimento e comunicar os resultados

Responsável por criar mecanismo para apresentar os resultados encontrados, um trabalho de visualização dos resultados. Seja com gráficos, tabelas. Foram desenvolvidos gráficos boxplots para visualizar a distribuição das notas, gráfico de barras para expor importâncias das features.

4 Resultados

No capítulo anterior os modelos preditivos foram executados, neste capítulo iremos analisar os resultados obtidos: predição das notas, métricas de avaliação dos modelos e importâncias das features com auxílio de gráficos para melhor visualização. Primeiramente com as notas de Matemática e posteriormente para as notas de Redação dos participantes do ENEM 2018.

4.1 Análise das Notas de Matemática ENEM 2018

Apresentaremos os resultados obtidos para a predição das notas de matemática. Plotaremos gráficos para ver o comportamentos das notas, métricas de validação, Features Impotances, Permutation Importances.

4.1.1 Notas

Para visualizar se existe ou não equivalência entre conjuntos de dados é muito comum utilizar o boxplot. Com a finalidade de observar essas variações o gráfico abaixo mostra as notas obtidas por cada um dos modelos preditivos (Random Forest, Árvore de Decisão, Rede Neural e Regressão Linear) e também a nota real, obtidas dos microdados do ENEM 2018, para os mesmo participantes.

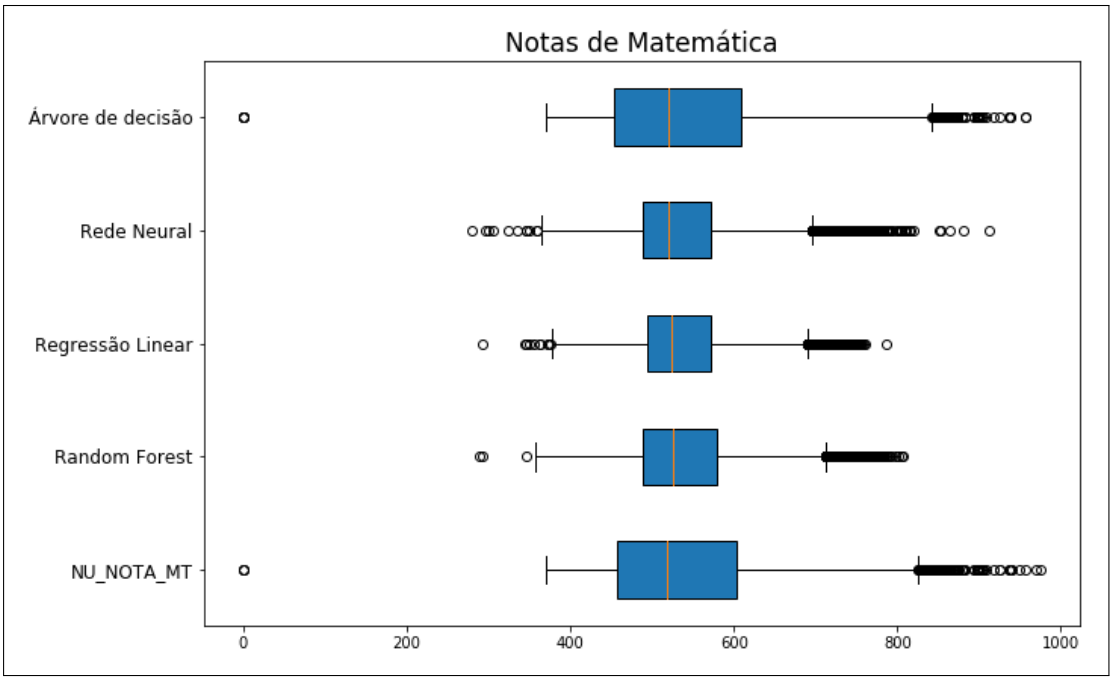

Figura 4.1 – Distribuição das notas de Matemática

Podemos notar uma variação entre eles, a caixa representa 50% de todos os valores observados, concentrando-se na tendência central dos valores entre 400 e pouco mais de 600 pontos, as medianas apresentaram valores similares, mas a distribuição entre o primeiro quartil e o terceiro (o tamanho das caixas) variou bastante entre os modelos, vale observar que o boxplot do modelo de Árvore de Decisão teve bastante semelhança com o boxplot das notas reais dos participantes para as notas de matemática.

4.1.2 Métricas

Na etapa do projeto chamada de validação dos dados, são aplicada métricas que nos dirão como os nossos modelos se comportam com os dados, se tiveram bom desempenho em prever os resultados. Nesta etapa iremos analisar estes resultados.

Figura 4.2 – Métricas para modelos regressores

As métricas de avaliação (Erro Máximo, MAE, MSE) calculam o erro do conjunto, cada uma com sua fórmula, mas elas têm em comum é que quanto maior for o valor do erro, pior o desempenho do modelo ao estimar a nota de Matemática, enquanto o coeficiente de determinação

(R^2) tem comportamento diferente, quanto mais próximo de 1 melhor o despenho do modelo em estimar a nota, quanto menor pior o despenho do modelo.

O modelo Árvore de Decisão apresentou pior resultado para todas as métricas de avaliação; Regressão Linear teve os melhores desempenhos para as métricas MAE, MSE e R^2; Random Forest teve o melhor desempenho para a métrica Erro Máximo. De acordo com as métricas de validação aplicadas, o modelo Regressão Linear teve o melhor comportamento ao estimar a nota de Matemática.

4.1.3 Features importance

Fornece as importâncias das features baseados em suas impurezas. Quanto maior, mais importante a feature. Isso nos mostra a importância que cada feature (característica) teve para a predição das notas pelo modelo.Essa é uma propriedade presente nos modelos Random Forest e Árvore de decisão.

4.1.3.1 Features Importances: Random Forest

As features comas maiores importâncias para o modelos Random Forest estão listadas e descritas a seguir:

Tabela 4.1 – Feature Importances Random Forest: tabela mostra as features mais importantes para o modelo Random Forest na predição da nota de Matemática.

Importância	**Feature**	**Descrição**
15.4	TP_ESCOLA_3	Escola privada
3.9	TP_SEXO_0	Sexo feminino
3.9	TP_DEPENDENCIA_ADM_ESC_3	Dependência administrativa da escola municipal
3.0	Q024_B	Residência possui 1 computador
1.6	Q008_C	Residência com 2 banheiros
1.6	TP_ESTADO_CIVIL_0.0	Solteiro(a)
1.2	Q005_5	5 pessoas moram na residência
1.2	TP_SEXO_1	Sexo masculino
1.2	Q002_F	A mãe completou a faculdade, mas não completou a pós-graduação
1.1	TP_COR_RACA_2	Cor/raça preta

Podemos notar que features foram sobre características bem diversificadas, escolaridade dos pais, sexo do participante, tipo de ensino, quantidade de pessoas numa casa, cor.

4.1.3.2 Features Importances: Árvore de Decisão

Para o modelo Árvore de Decisão as features com as maiores importâncias nacional são listadas e descritas abaixo.

Tabela 4.2 – Feature Importances Árvore de Decisão: tabela mostra as features mais importantes para o modelo Random Forest na predição da nota de Matemática.

Importância	**Feature**	**Descrição**
19.4	TP_ESCOLA_3	Escola privada
3.7	TP_SEXO_0	Sexo feminino
2.9	Q024_B	1 computador na residência
2.1	Q008_C	Residência com 2 banheiros
1.8	TP_ESTADO_CIVIL_0.0	Solteiro(a)
1.3	Q005_5	5 pessoas moram na residência do participante
1.2	Q002_F	A mãe completou a faculdade, mas não completou a pós-graduação
1.2	Q001_F	O pai completou a faculdade, mas não completou a pós-graduação
1.2	Q005_4	4 pessoas moram na residência do participante
1.1	Q022_E	Na residência existem 4 ou mais celulares

No modelo Árvore de Decisão aconteceu poucas mudanças em relação ao Random Forest, entre elas, a presença das feature Q022_F e Q002_F, e não conter a feature TP_DPENDENCIA_ADM_ESC_3 entre as 10 features com mais importância.

4.1.4 Permutation importances

Analisar os resultados encontrando e assim saber quais características dos participantes mais influenciaram os modelos. A seguir temos um gráfico que mostra a importância médias da features para os modelos.

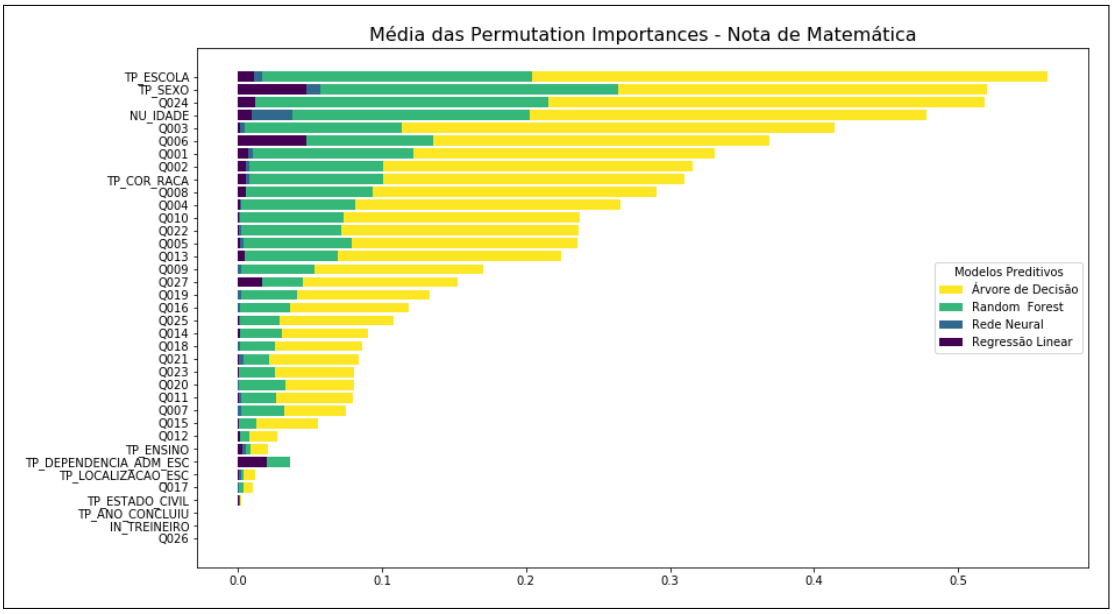

Figura 4.3 – Média Permutation Importances: Mostra a importâncias de cada features para estimar nota de Matemática

As feature mais importantes para foram sobre o tipo de escola do participante (TP_ESCOLA), sexo (TP_SEXO), sobre ter computador (Q024), idade (NU_IDADE), ocupação do pai (Q003) , renda familiar mensal (Q006), escolaridade dos pais(Q001, Q002), cor/raça do participante (TP_COR_RACA), banheiro na residência (Q008), ocupação da mãe (Q004). Mais detalhes sobre a descrição das features consultar dicionário no apêndice deste trabalho.

4.1.5 Conclusões

Conseguimos encontrar muitas informações importantes com essa análise, pudemos ver um pouco das notas previstas pelos modelos em comparação com as notas originais, com os boxplots ficou bem mais fácil analisar essas variações e notar a semelhança do boxplot de Árvore de Decisão e Notas reais de matemática.

Posteriormente aplicamos as métricas de avaliação dos modelos, e com base nela a Rede Neural teve maior pontuação nesse quesito. Com as features importances dos modelos Random Forest e Árvore de Decisão pudemos ter resultados interessantes e bem diversificados sobre as features.

Vimos que as permutations importances mediram as importância das features para todos os modelos e nos mostraram características que influenciaram os modelos independente das ordem de associação. Com isso concluímos uma parte do objetivo central deste trabalho, que foi analisar os principais pontos sobre a predição das notas de Matemática.

4.2 Análise Notas de Redação ENEM 2018

Nesta etapa do projeto serão mostrado os principais resultados obtidos com a predição das notas de Redação, Métricas de validação de modelos regressores, Feature Importance, Permutation Importance.

4.2.1 Notas

Para analisar como as notas obtidas pelos modelos variaram, iremos aplicar o boxplot, isso irá nos mostrar visualmente essa variação.

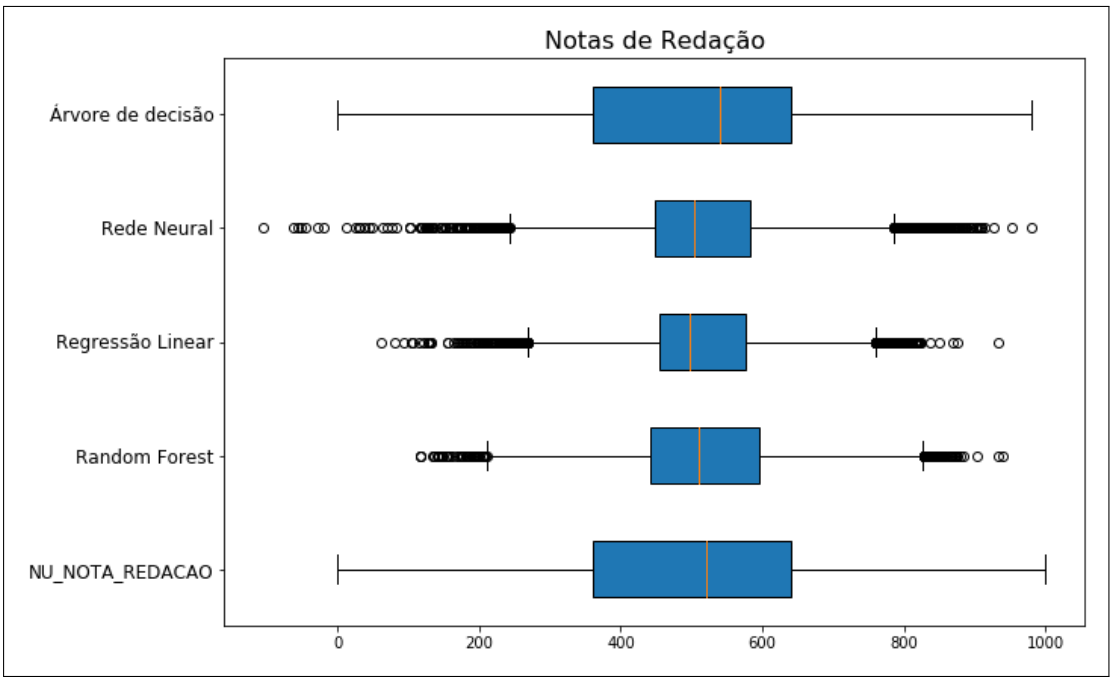

Figura 4.4 – Boxplot - Nota de Redação

Já falamos anteriormente como são dispostas as notas no boxplot, sabemos que a caixa representa 50% de todos os valores plotados. Se olharmos o gráfico, veremos que na maioria deles as caixas estão entre pouco menos de 400 e 600 e pouco mais de pontos.

Os valores que estão antes da lateais esquerda dos gráfico representam 25% dos valores mais baixos dos conjuntos, ou seja 25% dessas notas foram menores que 400 ponto. Os valores que estão depois das laterais direita representam os 25% valores mais altos do conjunto, o que diz que cerca de 25% das notas foram maiores que 600 pontos. A linha vertical presentes nas caixas representam a mediana de cada boxplot.

4.2.2 Métricas de Avaliação de Modelos Regressores

Principais resultados obtidos com métricas de avaliação dos modelos preditivos serão apresentados com os gráficos a seguir.

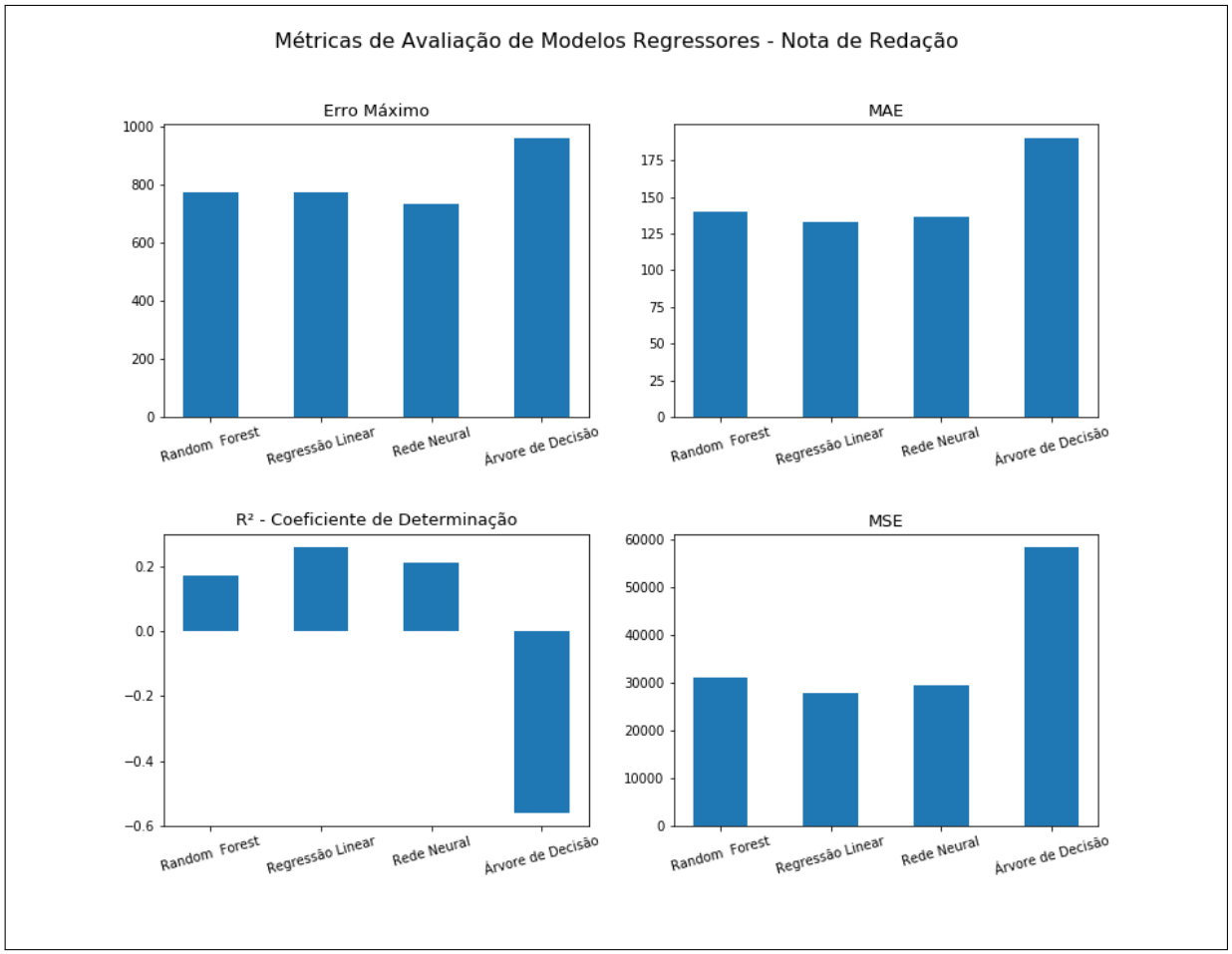

Figura 4.5 – Métricas para modelos regressores - Nota de Redação

Para as métricas Erro Máximo, MAE e MSE, quanto maior forem seus valores pior avaliados são os modelos, com isso vemos que o modelo Árvore de Decisão teve o pior comportamento segundo essas três métricas, Rede neural teve o melhor desempenho em Erro Máximo, mas Regressão linear teve melhor performance em MAE e MSE.

A métrica R^2 ou coeficiente de determinação se comporta um pouco diferente das outras, pois quanto mais distante de 1 for a resposta pior será o desempenho do modelo segundo essa métrica. Logo, temos que o pior resultado foi para Árvore de Decisão e melhor para Regressão Linear.

4.2.3 Features Importances

Os resultados de Features Importances foram muito semelhantes aos resultados encontrados para as notas de Matemática. Para o Random Forest o que variou foi a Q009_D sobre a quantidade de quarto na residência do participante.

Tabela 4.3 – Feature Importances: tabela mostra as features mais importantes para o modelo Random Forest na predição da nota de Redação.

Importância	**Feature**	**Descrição**
17.6	TP_DEPENDENCIA_ADM_ESC_3	Dependência administrativa da escola municipal
5.8	TP_SEXO_0	Sexo feminino
1.6	Q024_B	1 computador na residência
1.3	Q005_5	5 pessoas moram na residência
1.2	Q005_4	4 pessoas moram na residência
1.1	TP_COR_RACA_4	Cor/raça amarela
1.1	Q001_F	O pai completou a faculdade, mas não concluiu a pós-graduação.
1.1	Q009_D	A residência possui 3 quartos
1.1	Q005_6	6 pessoas moram na residência
1.1	Q022_D	Existem 3 celulares na residência.

As características mais importantes pro Random Forest foram administração da escola, sexo feminino, ter um computador na residência, quantidade de pessoas na residência (4 6 pessoas), cor/raça amarela, educação do pai, possui 3 quarto e 3 celulares na residência.

Tabela 4.4 – Feature Importances: tabela mostra as features mais importantes para o modelo Árvore de Decisão na predição da nota de Redação.

Importância	**Feature**	**Descrição**
17.5	TP_DEPENDENCIA_ADM_ESC_3	Dependência administrativa da escola municipal
5.8	TP_SEXO_0	Sexo feminino
1.6	Q024_B	1 computador na residência
1.3	Q005_4	4 pessoas moram na residência
1.2	Q005_5	5 pessoas moram na residência
1.2	Q009_D	A residência possui 3 quartos
1.2	TP_COR_RACA_4	Cor/raça amarela
1.1	Q005_6	6 pessoas moram na residência
1.1	Q022_D	Existem 3 celulares na residência.
1.1	Q006_D	Renda familiar mensal varia de R$ 1.431,01 até R$ 1.908,00.

As features mais importantes estavam relacionadas com tipo de administração da escola municipal, sexo feminino, ter um computador na residência, quantidade de pessoas na residência entre 4 e 6, residência possui três quarto e três celulares no total, cor/raça amarela e renda familiar mensal.

4.2.4 Permutation Importances

Os resultados de Permutation Importances foram bem mais interessantes com as notas de redação, deixando claro que a nota de Códigos e Linguagens tem muita relevância para as notas

de Redação. Podemos ver que o resultado de maior importância foi praticamente unânime para todos os modelos, exceto para Árvore de Decisão.

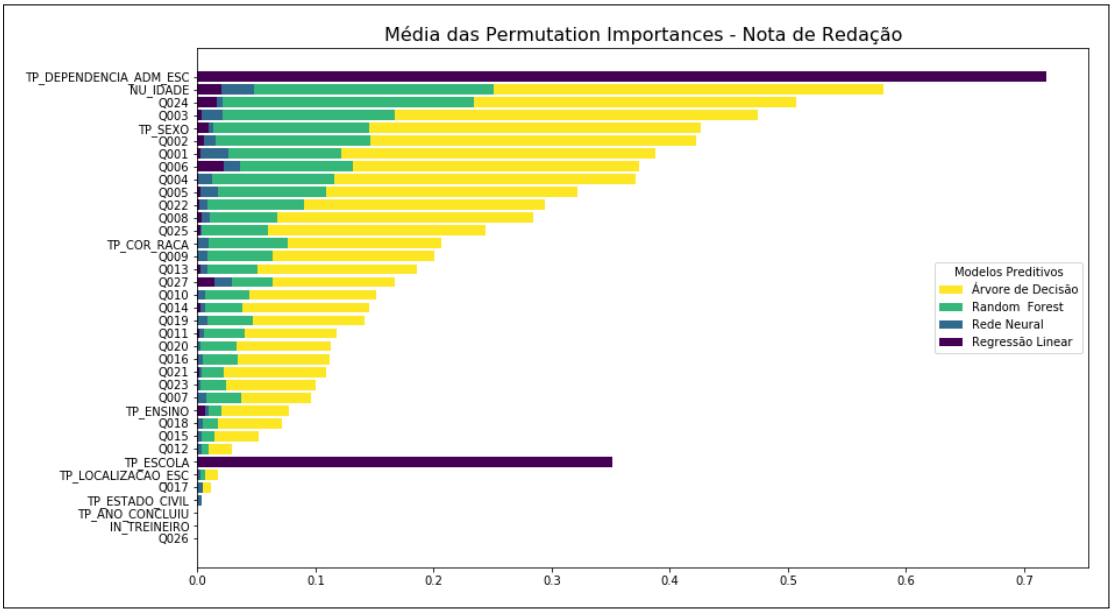

Figura 4.6 – Média Permutation Importances: Mostra as importâncias das features para estimar a nota de Redação

Para a nota de Redação as features mais importantes variaram um pouco, foram elas sobre o tipo de administração da escola (TP_DEPENDENCIA_ESC), idade (NU_IDADE), computador na residência (Q024), ocupação do pai (Q003), sexo (TP_SEXO), escolaridade dos pais(Q001, Q002), renda familiar mensal (Q006), ocupação da mãe (Q004), quantidade de pessoas na residência (Q005), um observação é quanto a feature tipo escola (TP_ESCOLA) que alta importância para o modelo Regressão Linear. Para saber mais sobre a descrição das features, consultar dicionário de dado em apêndice neste trabalho.

4.2.5 Conclusão

Foram usados boxplots para analisar as variações nas notas de redação encontradas pelos modelos preditivos, mas nada novo foi identificado, já havíamos mostrado isso com as notas de matemática, foi possível ver que com redação não teve grandes mudanças.

Posteriormente aplicamos as métricas de avaliação dos modelos, e com base nela podemos ver que a Regressão Linear teve maior pontuação neste quesito. Com as features importances dos modelos Random Forest e Árvore de Decisão tivemos resultados interessantes e diversificado sobre as features que influenciaram esses modelos.

Vimos que as permutations importances mediram as importância das features para todos os modelos e nos mostraram características que influenciaram os modelos.

Com isso pudemos ver que o ano de conclusão e idade tiveram , além das outras notas, muita influência nos modelos. Com o término desta etapa estamos finalizando o objetivo central deste trabalho.

5 Considerações Finais

Percebe-se que para a utilização de Data Science é necessário possuir uma gama enorme de competências dentro de diversas áreas do conhecimento, tais como: programação, matemática, estatística, Análise de Dados, entre outras. Por conta disto a conclusão de uma aplicação final é algo extremamente demorado e que demanda bastante empenho por conta do cientista. No escopo deste projeto, todos os temas mencionados tiveram de ser estudados com a finalidade de incrementar o desenvolvimento da aplicação final.

5.1 Conclusão

Foi apresentado, no capítulo 3 deste trabalho, a metodologia abordada e os principais pontos desta metodologia (definição, coleta, exploração, modelagem, validação e otimização/comunicação). Foram apresentados gráficos sobre o quantitativo de participante de acordo com a renda mensal familiar (Q006), Q024 (sobre o participante ter computador, e a quantidade caso tenha) e sobre a idade dos participantes.

Posteriormente, no capítulo 4, foi apresentado os resultados para as notas de Matemática e Redação encontradas pelos modelos preditivos, métricas de avaliação (Erro Máximo, MAE, MSE e R^2) mostrando que o modelo implementado usando Rede Neural teve um desempenho geral melhor que outros modelos na análise das notas de Matemática, já em Redação, o modelo Regressão Linear foi melhor.

Para as importâncias das features tivemos em Features Importances, características sobre o tipo de ensino, sexo, tipo de administração da escola, participante ter um computador (Q024), dois banheiros na residência (Q008); Permutation Importances com as características socioeconômicas que mais influenciaram nas notas de Matemática e Redação, tipo escola, sexo, computador (Q024), idade, ocupação do pai (Q003), renda familiar mensal (Q006), escolaridade do pais (Q001, Q002), cor/raça, banheiro na residência (Q008). Com uma ressalva na característica (tipo de administração da escola) que teve muita importância para a nota de Redação, enquanto pra Matemática isso não ocorreu.

As features sobre computador na residência (Q024), renda familiar mensal (Q006) e idade (NU_IDADE) dos participantes, estiveram entre as features com mais importância para maioria dos modelos, tanto pra Redação, quanto para Matemática. Então podemos concluir que essas features tiveram sim muita influencia para as notas de Matemática e Redação.

5.2 Trabalhos Futuros

Tendo em vista que características socioeconômica dos participantes, idade, sexo e renda tiveram uma certa influência nos modelos. Algumas proposta de pesquisas a serem desenvolvidas a partir deste, podem ser as seguintes:

1. Investigar e obter mais detalhes para montar um perfil mais detalhado de como essas características influenciam as notas.
2. Analisar usando apenas os modelos que obtiveram melhor performances após aplicação das métricas de avaliação dos modelos regressores.
3. Uma análise preditiva com base em um questionamento extraído a partir de uma análise exploratória, por exemplo, explorando os participantes em grupos por renda, sexo, idade.

Os microdados do ENEM oferece muitos caminhos a serem explorados, utilizando Data Science, Machine learning e muita paciência, ótimas respostas podem ser encontradas e utilizadas para melhorar a educação de muitos estudantes brasileiros.

Referências

BREIMAN, L. *Random Forests*. [S.l.: s.n.], 2001. 5- 32 p. (Machine Learning 45).

CASSIANI, S.; SILVA, H. D.; PIERSON, A. *OLHARES PARA O ENEM NA EDUCAÇÃO CIENTÍFICA E TECNOLÓGICA*. JUNQUEIRA & MARIN, 2016. ISBN 9788582030257. Disponível em: <https://books.google.com.br/books?id=q7R2DwAAQBAJ>.

CHEN, D. *Análise de dados com Python e Pandas*. NOVATEC, 2018. ISBN 9788575226995. Disponível em: <https://books.google.com.br/books?id=ILFwDwAAQBAJ>.

FREITAS, A.; ROZENBERG, G. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer, 2002. (Natural Computing Series). ISBN 9783540433316. Disponível em: <https://books.google.com.br/books?id=KkdZlfQJvbYC>.

FUMIO, H. *Econometrics*. [S.l.]: Princeton University Press, 2000. ISBN 0-691·01 018·8.

GERALD, L. M. A.; SCADELAI, L. M.; BOLZAN, W. J.; PEREIRA, J. R. *Pesquisa Em Educação Matemática*. Faculdade de Educação de São Luis, 2014. ISBN 9788575227510. Disponível em: <https://books.google.com.br/books?id=9u1xDwAAQBAJ>.

GÉRON, A. *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. Alta Books, 2019. ISBN 9788550809021. Disponível em: <https://books.google.com.br/books?id=Z0mvDwAAQBAJ>.

HARRISON, M. *Machine Learning – Guia de Referência Rápida: Trabalhando com dados estruturados em Python*. Novatec Editora, 2019. ISBN 9788575228180. Disponível em: <https://books.google.com.br/books?id=VvXADwAAQBAJ>.

HAYKIN, S. *Redes Neurais: Princípios e Prática*. Artmed, 2007. ISBN 9788577800865. Disponível em: <https://books.google.com.br/books?id=bhMwDwAAQBAJ>.

INEP, B. *EXAME NACIONAL DO ENSINO MÉDIO (ENEM) Fundamentação Teórico-Metodológica*. Inep/MEC – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2005. Disponível em: <http://portal.inep.gov.br/documents/186968/484421/ENEM+-+Exame+Nacional+do+Ensino+M\%C3\%A9dio+fundamenta\%C3\%A7\%C3\%A3o+te\%C3\%B3rico-metodol\%C3\%B3gica/449eea9e-d904-4a99-9f98-da804f3c91f5?version=1.1>.

INEP, S. O. *Microdados INEP*. 2020. Acessado em: 13 junho 2020. Disponível em: <http://inep.gov.br/microdados>.

LIMA, V. de; OLIVEIRA, P. *Previsão de Demanda*. Editora Baraúna, 2016. ISBN 9788543705217. Disponível em: <https://books.google.com.br/books?id=Xi9FDQAAQBAJ>.

MCKINNEY, W. *Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython*. Novatec Editora, 2019. ISBN 9788575227510. Disponível em: <https://books.google.com.br/books?id=4hmWDwAAQBAJ>.

MOLIN, S. *Hands-On Data Analysis with Pandas: Efficiently perform data collection, wrangling, analysis, and visualization using Python*. Packt Publishing, 2019. ISBN 9781789612806. Disponível em: <https://books.google.com.br/books?id=buGlDwAAQBAJ>.

NATINGGA, D. *Data Science Algorithms in a Week*. Packt Publishing, 2017. ISBN 9781787282742. Disponível em: <https://books.google.com.br/books?id=UJZGDwAAQBAJ>.

REDA, S.; SHAFIQUE, M. *Approximate Circuits: Methodologies and CAD*. Springer International Publishing, 2018. ISBN 9783319993225. Disponível em: <https://books.google.com.br/books?id=Drh9DwAAQBAJ>.

ROKACH, L. *Data Mining with Decision Trees: Theory and Applications*. World Scientific, 2008. (Series in machine perception and artificial intelligence). ISBN 9789812771728. Disponível em: <https://books.google.com.br/books?id=GlKIIR78OxkC>.

SCIKIT-LEARN, S. O. *Logo Scikit-Learn*. 2020. Acessado em: 13 junho 2020. Disponível em: <https://scikit-learn.org/stable/_images/scikit-learn-logo-notext.png>.

SILVA, T. Desenvolvimento e validação de nanogramas para estimativa de risco para cancer de próstata em população basileira. *Tese(doutorado) - Faculdade de Medicina da Universidade de São Paulo*, p. PaginaInicial–PaginaFinal, 2012.

Apêndices

APÊNDICE A – Dicionário de Variáveis Resumido - Enem 2018

DICIONÁRIO DE VARIÁVEIS - ENEM 2018

NOME DA VARIÁVEL	Descrição	Variáveis Categóricas		Tipo
		Categoria	Descrição	
DADOS DO PARTICIPANTE				
NU_INSCRICAO	Número de inscrição1			Numérica
SG_UF_RESIDENCIA	Sigla da Unidade da Federação de residência			Alfanumérica
NU_IDADE	Idade2			Numérica
TP_SEXO	Sexo	M	Masculino	Alfanumérica
		F	Feminino	
TP_ESTADO_CIVIL	Estado Civil	0	Solteiro(a)	Numérica
		1	Casado(a)/Mora com companheiro(a)	
		2	Divorciado(a)/Desquitado(a)/Separado(a)	
		3	Viúvo(a)	
TP_COR_RACA	Cor/raça	0	Não declarado	Numérica
		1	Branca	
		2	Preta	
		3	Parda	
		4	Amarela	
		5	Indígena	
TP_ANO_CONCLUIU	Ano de Conclusão do Ensino Médio	0	Não informado	Numérica
		1	2017	
		2	2016	
		3	2015	
		4	2014	
		5	2013	
		6	2012	
		7	2011	
		8	2010	
		9	2009	
		10	2008	
		11	2007	
		12	Antes de 2007	
TP_ESCOLA	Tipo de escola do Ensino Médio	1	Não Respondeu	Numérica
		2	Pública	
		3	Privada	
		4	Exterior	
TP_ENSINO	Tipo de instituição que concluiu ou concluirá o Ensino Médio	1	Ensino Regular	Numérica
		2	Educação Especial - Modalidade Substitutiva	
		3	Educação de Jovens e Adultos	
IN_TREINEIRO	Indica se o inscrito fez a prova com intuito de apenas treinar seus conhecimentos3	1	Sim	Numérica
		0	Não	
DADOS DA ESCOLA				
TP_DEPENDENCIA_ADM_ESC	Dependência administrativa (Escola)	1	Federal	Numérica
		2	Estadual	
		3	Municipal	
		4	Privada	
TP_LOCALIZACAO_ESC	Localização (Escola)	1	Urbana	Numérica
		2	Rural	
DADOS DA PROVA OBJETIVA				
NU_NOTA_CN	Nota da prova de Ciências da Natureza			Numérica
NU_NOTA_CH	Nota da prova de Ciências Humanas			Numérica
NU_NOTA_LC	Nota da prova de Linguagens e Códigos			Numérica
NU_NOTA_MT	Nota da prova de Matemática			Numérica
DADOS DA REDAÇÃO				
NU_NOTA_REDACAO	Nota da prova de redação			Numérica
DADOS DO QUESTIONÁRIO SOCIOECONÔMICO				
Q001	Até que série seu pai, ou o homem responsável por você, estudou?	A	Nunca estudou.	Alfanumérica
		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.	
Q002	Até que série sua mãe, ou a mulher	A	Nunca estudou.	Alfanumérica
		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	

Q002	responsável por você, estudou?	E	Completou o Ensino Médio, mas não completou a Faculdade.	Alfanumérica
		F	Completou a Faculdade, mas não completou a Pós-graduação.	
		G	Completou a Pós-graduação.	
		H	Não sei.	
Q003	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	Grupo 1: Lavrador, agricultor sem empregados, bóia fria, criador	Alfanumérica
		B	Grupo 2: Diarista, empregado doméstico, cuidador de idosos,	
		C	Grupo 3: Padeiro, cozinheiro industrial ou em restaurantes,	
		D	Grupo 4: Professor (de ensino fundamental ou médio, idioma,	
		E	Grupo 5: Médico, engenheiro, dentista, psicólogo, economista,	
		F	Não sei.	
Q004	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).	A	Grupo 1: Lavradora, agricultora sem empregados, bóia fria,	Numérica
		B	Grupo 2: Diarista, empregada doméstica, cuidadora de idosos,	
		C	Grupo 3: Padeira, cozinheira industrial ou em restaurantes,	
		D	Grupo 4: Professora (de ensino fundamental ou médio, idioma,	
		E	Grupo 5: Médica, engenheira, dentista, psicóloga, economista,	
		F	Não sei.	
Q005	Incluindo você, quantas pessoas moram atualmente em sua residência?	1	1, pois moro sozinho(a).	Numérica
		2	2	
		3	3	
		4	4	
		5	5	
		6	6	
		7	7	
		8	8	
		9	9	
		10	10	
		11	11	
		12	12	
		13	13	
		14	14	
		15	15	
		16	16	
		17	17	
		18	18	
		19	19	
		20	20	
Q006	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)	A	Nenhuma renda.	Alfanumérica
		B	Até R$ 954,00.	
		C	De R$ 954,01 até R$ 1.431,00.	
		D	De R$ 1.431,01 até R$ 1.908,00.	
		E	De R$ 1.908,01 até R$ 2.385,00.	
		F	De R$ 2.385,01 até R$ 2.862,00.	
		G	De R$ 2.862,01 até R$ 3.816,00.	
		H	De R$ 3.816,01 até R$ 4.770,00.	
		I	De R$ 4.770,01 até R$ 5.724,00.	
		J	De R$ 5.724,01 até R$ 6.678,00.	
		K	De R$ 6.678,01 até R$ 7.632,00.	
		L	De R$ 7.632,01 até R$ 8.586,00.	
		M	De R$ 8.586,01 até R$ 9.540,00.	
		N	De R$ 9.540,01 até R$ 11.448,00.	
		O	De R$ 11.448,01 até R$ 14.310,00.	
		P	De R$ 14.310,01 até R$ 19.080,00.	
		Q	Mais de R$ 19.080,00.	
Q007	Em sua residência trabalha empregado(a) doméstico(a)?	A	Não.	Alfanumérica
		B	Sim, um ou dois dias por semana.	
		C	Sim, três ou quatro dias por semana.	
		D	Sim, pelo menos cinco dias por semana.	
Q008	Na sua residência tem banheiro?	A	Não.	Alfanumérica
		B	Sim, um.	
		C	Sim, dois.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q009	Na sua residência tem quartos para dormir?	A	Não.	Alfanumérica
		B	Sim, um.	
		C	Sim, dois.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q010	Na sua residência tem carro?	A	Não.	Alfanumérica
		B	Sim, um.	
		C	Sim, dois.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q011	Na sua residência tem motocicleta?	A	Não.	Alfanumérica
		B	Sim, uma.	
		C	Sim, duas.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q012	Na sua residência tem geladeira?	A	Não.	Alfanumérica
		B	Sim, uma.	

Q012	Na sua residência tem geladeira?	C	Sim, duas.	Alfanumérica
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q013	Na sua residência tem freezer (independente ou segunda porta da geladeira)?	A	Não.	Alfanumérica
		B	Sim, um.	
		C	Sim, dois.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q014	Na sua residência tem máquina de lavar roupa? (o tanquinho NÃO deve ser considerado)	A	Não.	Alfanumérica
		B	Sim, uma.	
		C	Sim, duas.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q015	Na sua residência tem máquina de secar roupa (independente ou em conjunto com a máquina de lavar roupa)?	A	Não.	Alfanumérica
		B	Sim, uma.	
		C	Sim, duas.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q016	Na sua residência tem forno micro-ondas?	A	Não.	Alfanumérica
		B	Sim, um.	
		C	Sim, dois.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q017	Na sua residência tem máquina de lavar louça?	A	Não.	Alfanumérica
		B	Sim, uma.	
		C	Sim, duas.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q018	Na sua residência tem aspirador de pó?	A	Não.	Alfanumérica
		B	Sim.	
Q019	Na sua residência tem televisão em cores?	A	Não.	Alfanumérica
		B	Sim, uma.	
		C	Sim, duas.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q020	Na sua residência tem aparelho de DVD?	A	Não.	Alfanumérica
		B	Sim.	
Q021	Na sua residência tem TV por assinatura?	A	Não.	Alfanumérica
		B	Sim.	
Q022	Na sua residência tem telefone celular?	A	Não.	Alfanumérica
		B	Sim, um.	
		C	Sim, dois.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q023	Na sua residência tem telefone fixo?	A	Não.	Alfanumérica
		B	Sim.	
Q024	Na sua residência tem computador?	A	Não.	Alfanumérica
		B	Sim, um.	
		C	Sim, dois.	
		D	Sim, três.	
		E	Sim, quatro ou mais.	
Q025	Na sua residência tem acesso à Internet?	A	Não.	Alfanumérica
		B	Sim.	
Q026	Você já concluiu ou está concluindo o Ensino Médio?	A	Já concluí o Ensino Médio.	Alfanumérica
		B	Estou cursando e concluirei o Ensino Médio em 2018.	
		C	Estou cursando e concluirei o Ensino Médio após 2018.	
		D	Não concluí e não estou cursando o Ensino Médio.	
Q027	Em que tipo de escola você frequentou o Ensino Médio?	A	Somente em escola pública.	Alfanumérica
		B	Parte em escola pública e parte em escola privada SEM bolsa de estudo integral.	
		C	Parte em escola pública e parte em escola privada COM bolsa de estudo integral.	
		D	Somente em escola privada SEM bolsa de estudo integral.	
		E	Somente em escola privada COM bolsa de estudo integral.	
		F	Não frequentei a escola	

1. Referente ao Enem 2017, trata-se de uma máscara e não o seu número de inscrição original no Enem. O mesmo NU_INSCRICAO para anos diferentes não identifica o mesmo participante no evento, não permite o acesso aos dados cadastrais como nome, endereço, RG etc, nem identifica o mesmo participante em microdados de pesquisas diferentes.

2. Idade do inscrito em 31/12/2017. Idades inferiores a 10 anos e superiores a 100 anos estão com o campo vazio na base.

3. Foi considerado treineiro o inscrito que tinha menos de 18 anos no primeiro dia de realização do exame (05/11/2017) e que não havia concluído o ensino médio e não o concluiria em 2017.

4. Número gerado como identificador da escola no Censo Escolar da Educação Básica.

5. Segundo o Edital do Enem 2017, no ato da inscrição o participante deveria informar a condição especial ou específica que motiva o atendimento.

6. Segundo o Edital do Enem 2017, no ato da inscrição o participante poderia declarar-se travesti ou transexual e solicitar o atendimento pelo Nome Social.

7. A opção de certificação só é apresentada para participantes concluintes com idade mínima de 18 anos, conforme Edital do Enem 2017.

8. As 45 primeiras posições deste campo são referentes as respectivas respostas. O asterisco (*) indica dupla marcação e o ponto (.) resposta em branco.

9. As 45 primeiras posições deste campo são referentes as respectivas respostas, das quais as 5 primeiras correspondem a parte de língua estrangeira. O asterisco (*) indica dupla marcação e o ponto (.) resposta em branco.

10. As 45 primeiras posições deste campo são referentes ao respectivo gabarito

11. As 50 primeiras posições deste campo são referentes ao respectivo gabarito, das quais, para as 10 primeiras, as 5 primeiras correspondem à prova de Lingua Inglesa e as outras 5 à prova de Lingua Espanhola.