ANÁLISE DE DADOS COLETADOS NO ENEM 2019 – O QUE REVELAM OS DADOS?

INTRODUÇÃO

O Exame Nacional do Ensino Médio (Enem) tem como objetivo analisar o desempenho daqueles que participaram da prova, em determinadas áreas de conhecimento, e como os participantes demonstraram domínio desse conteúdo após a conclusão do ensino médio. O INEP disponibiliza as informações referentes as provas realizadas, onde para esse presente estudo, são utilizados os dados referentes ao ano de 2019. São disponibilizados provas, gabaritos, informações sobre os candidatos como: notas e os questionários que são respondidos para poder se inscrever no exame, como a avaliação social do candidato.

O questionário para avaliação social do candidato foi disponibilizado com um total de 25 questões e traz questões sobre seu nível socioeconômico, família, educação e trabalho.

As provas do ENEM são compostas por 180 questões, e uma redação com tema que é modificado a cada ano, e a aplicação é feita em dois dias, onde o primeiro dia os participantes realizaram as provas de Linguagens, Códigos e suas tecnologias e Redação e de Ciências Humanas e suas tecnologias e, no segundo, as provas de Ciências da Natureza e suas tecnologias e Matemática e suas tecnologias. O quadro 1 apresenta a descrição das áreas de conhecimento e componentes curriculares do Enem.

Quadro 1 – Descrição das Áreas de Conhecimento e Componentes Curriculares do Enem.

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

As provas do ano de 2019 foram aplicadas nos dias 3 e 10 de novembro de 2019. A segunda aplicação do Enem 2019, por sua vez, ocorreu nos dias 10 e 11 de dezembro de 2019, para Pessoas Privadas de Liberdade e Jovens sob Medida Socioeducativa que inclua privação de liberdade – PPL, bem como para os participantes com direito à reaplicação.

A prova pode ser adaptada para participantes surdos e deficientes auditivos, onde os mesmos podem solicitar durante a inscrição o recurso para atendimento especializado videoprova em Libras. Esses participantes receberam, além do caderno de questões impresso, vídeos contendo a tradução correspondente na Língua Brasileira de Sinais - Libras. Também foram disponibilizados os cadernos ampliados, superampliados, ledor (correspondente ao caderno em braille) e a videoprova, conforme a condição de acessibilidade necessária ao participante.

Os arquivos disponibilizados no site do Inep estão em formato ".csv", que facilita o seu manuseio pelo usuário, ao tornar sua utilização mais intuitiva e imediata. O arquivo principal (MICRODADOS_ENEM_2019) traz as informações gerais sobre a realização das provas, a caracterização do participante e da escola que ele declarou ter frequentado, e as notas das provas objetivas e da redação.

Atualmente, a grande disponibilidade de dados nos temos permite utilizar os conhecimentos de Estatística, afim de se tirar conclusões a respeito de diversas áreas. A coleta, tratamento e análise quantitativa dos dados torna possível a criação de estudos direcionados, onde é possível a extração de resultados e conclusões significativas para o problema em questão.

Desse modo, no presente estudo, buscamos associar os dados disponíveis do Enem do ano de 2019, a técnicas de análise exploratória de dados com o intuito de encontrar respostas que associem o desempenho do candidato de acordo com a sua realidade social, como a renda familiar, o tipo de escola frequentada (pública ou particular), geolocalização, etc.

Os dados foram analisados através da criação de um *Jupyter Notebook*, desenvolvido na linguagem *Python*, onde as principais bibliotecas usadas foram *Numpy, Pandas, Matplotlib, Seaborn e Sklearn*.

O QUE OS DADOS MOSTRAM?

O Enem foi idealizado para avaliar as habilidades e competências dos alunos, além de facilitar no acesso a vagas oferecidas pelas Instituições de Ensino Superior nos mais variados locais do nosso país.

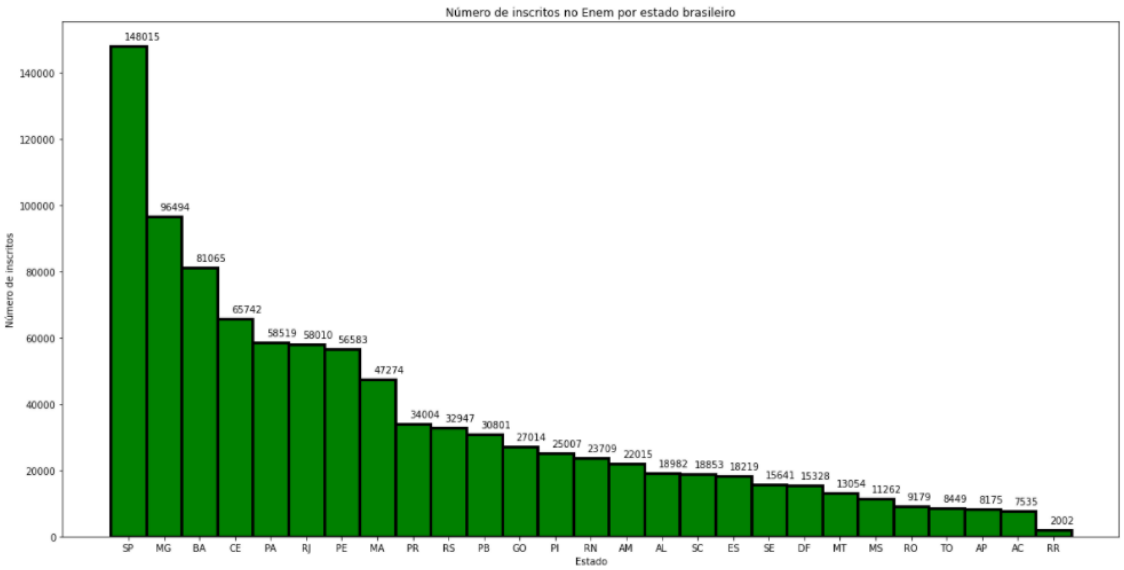
O presente estudo buscou encontrar padrões entre os dados disponibilizados pelo INEP, que contem informações sobre os candidatos, de acordou com o preenchimento durante a inscrição para realização da prova do Enem 2019.

O arquivo em formato *.csv* possuía 5.095.271 linhas de dados, onde após algumas manipulações, como a remoção de dados duplicados e de dados nulos, resultou em um total de 2.292.871. As linhas de dados possuíam 136 colunas, com as mais variadas informações, mas para poder otimizar o estudo, foram separadas as seguintes variáveis:

- Número de inscrição do candidato;
- Estado onde o candidato vive;
- O tipo da escola do candidato (particular/publica);
- A nota da prova de Ciências da Natureza;
- A nota da prova de Ciências Humanas;
- A nota da prova de Linguagens e Códigos;
- A nota da prova de Matemática;
- A nota da prova de Redação;
- A renda familiar declarada durante a inscrição;

A primeira visualização foi sobre a quantidade de inscritos de acordo com o estado brasileiro no qual o candidato vive. A figura 1 mostra a distribuição da quantidade de inscritos de acordo como estado.

Figura 1 – Distribuição do número de inscritos pelos estados brasileiros



Pelo gráfico de barras acima, pode-se observar que a maioria dos candidatos inscritos no Enem 2019 foram dos estados de São Paulo, Minas Gerais e Bahia, logo, os dados correspondentes a esses três estados influenciam predominantemente nos resultados das análises.

Estabelecendo agora uma avaliação sobre as notas obtidas em cada prova individual, temos que algumas medidas estatísticas podem ser tiradas desses dados. O quadro 1 mostra algumas medidas, como o valor máximo, mínimo, a média das notas de cada prova, entre outros.

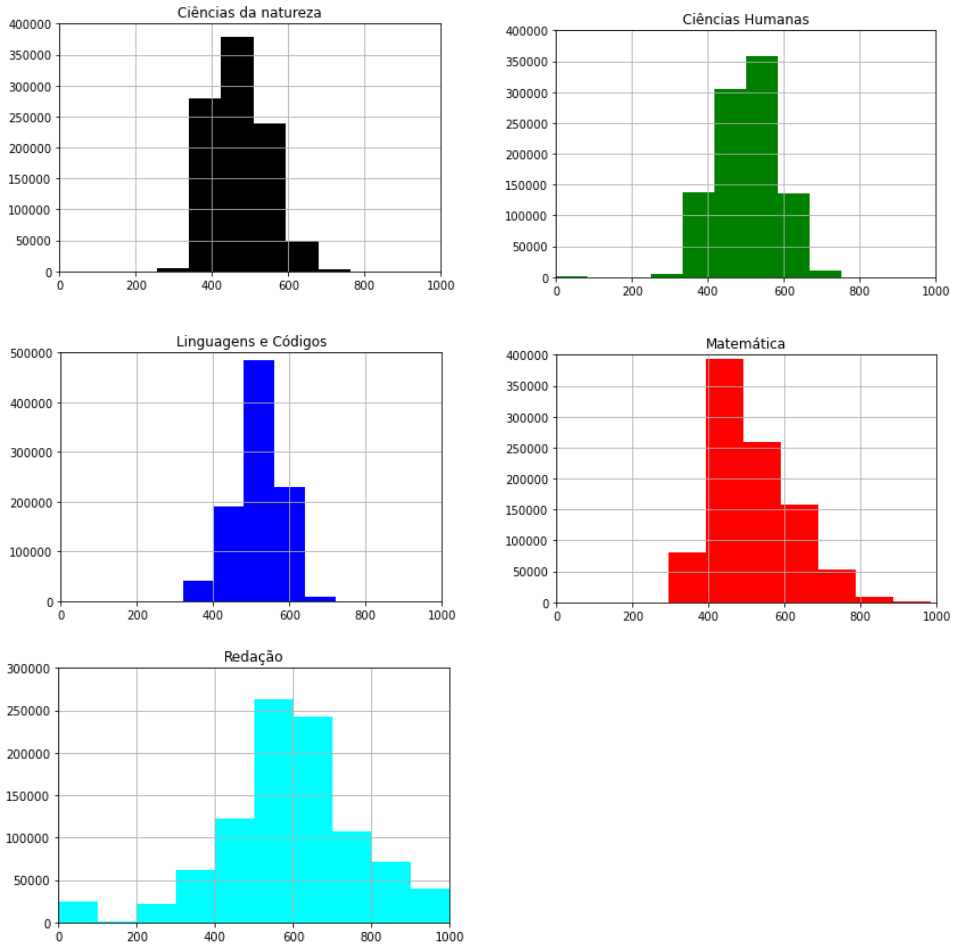
Quadro 1 – A estatística sobre os dados das notas das provas do Enem 2019

	CIENCIAS_NATUREZA	CIENCIAS_HUMANAS	LINGUAGENS_CODIGOS	MATEMATICA	REDACAO
count	953878.000000	953878.000000	953878.000000	953878.000000	953878.000000
mean	471.154960	504.111415	518.310507	514.425074	580.692017
std	72.233556	77.908424	60.969581	102.992083	177.619281
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	414.500000	447.400000	482.600000	432.500000	500.000000
50%	463.700000	507.600000	523.700000	493.200000	580.000000
75%	523.100000	560.900000	561.000000	582.600000	680.000000
max	847.800000	835.100000	801.700000	985.000000	1000.000000

Pelos dados estatísticos apresentados acima, pode-se verificar que a única prova que atingiu a nota máxima (1000) foi a de redação, porém existiram pontuações próximas a máxima na prova de Matemática, que foi 985. Além disso, pode-se observar que a maior média entre as notas também é para a prova de redação, enquanto que a menor média ficou na prova de Ciências da Natureza.

A partir desse quadro, buscou-se então uma compreensão mais aprofundada a respeito da distribuição das notas das 5 provas diferentes do Enem, onde cada uma avalia uma competência diferente acerca do candidato. Dessa forma, temos que a figura 2 apresenta os histogramas para as provas de Ciências da Natureza, Ciências Humanas, Linguagens e Códigos, Matemática e Redação.

Figura 2 – Histogramas para avaliar a distribuição das notas nas provas do Enem 2019



De um modo geral, o histograma acima nos mostra a distribuição das notas das 5 provas aplicadas no Enem de 2019 em relação aos candidatos que as fizeram. Quanto mais a direita o gráfico estiver posicionado, melhores são as pontuações dos participantes em cada uma das provas. Existe uma tendência nos três primeiros gráficos em relação as notas medianas, entre o intervalo de 400 a 600. Já os dois últimos gráficos, que correspondem as provas de Matemática e de Redação, as notas estão posicionadas em sua maioria à direita do gráfico, após o valor de 400 de pontuação na prova.

Posteriormente buscou-se estabelecer uma comparação entre as notas dos candidatos nas 5 provas individualmente em relação aos estados onde os mesmos moram e que foram inseridos durante a inscrição. Dessa forma, as figuras 3, 4, 5, 6 e 7 apresentam os boxplot para que se possa visualizar a tendencia das notas de acordo com o estado, assim como torna possível a visualização de possíveis pontos extremos à maioria dos dados, que chamamos de outliers.

Figura 3 – Boxplot para as notas da prova de Ciências da Natureza por estado

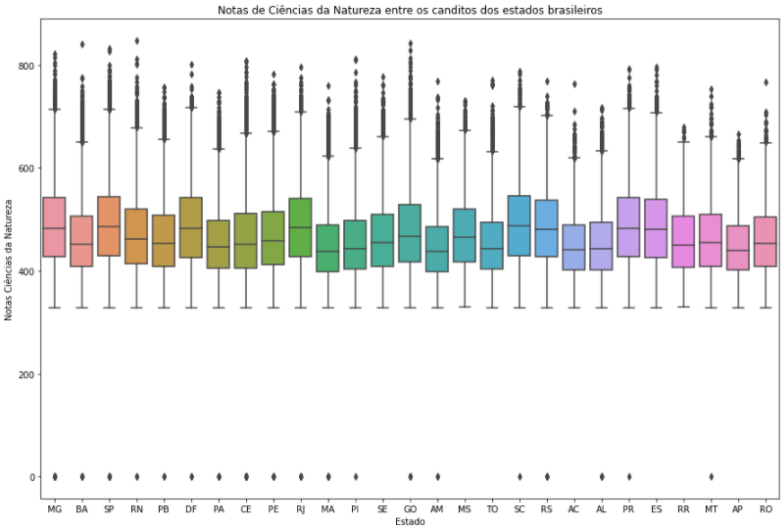


Figura 4 – Boxplot para as notas da prova de Ciências Humanas por estado

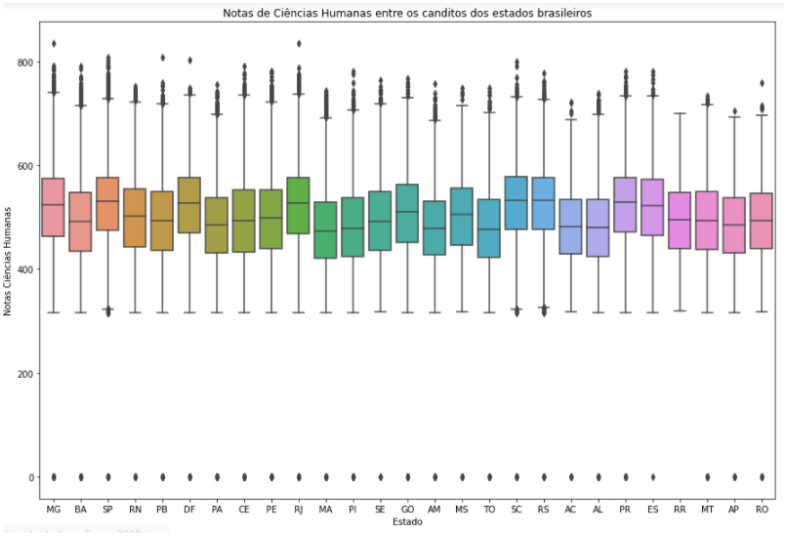


Figura 5 – Boxplot para as notas da prova de Linguagens e Códigos por estado

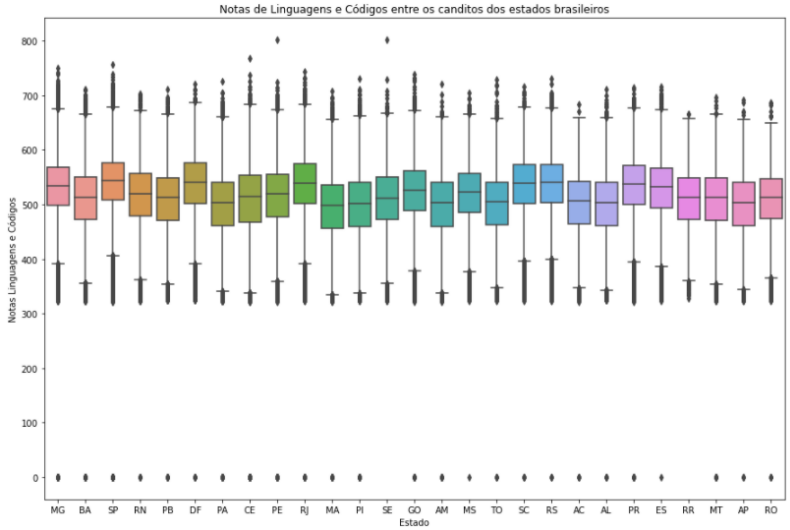


Figura 6 – Boxplot para as notas da prova de Matemática por estado

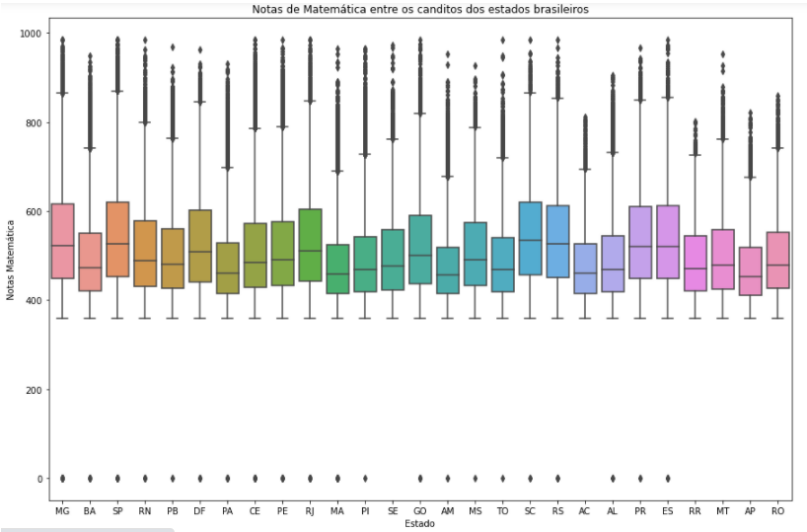
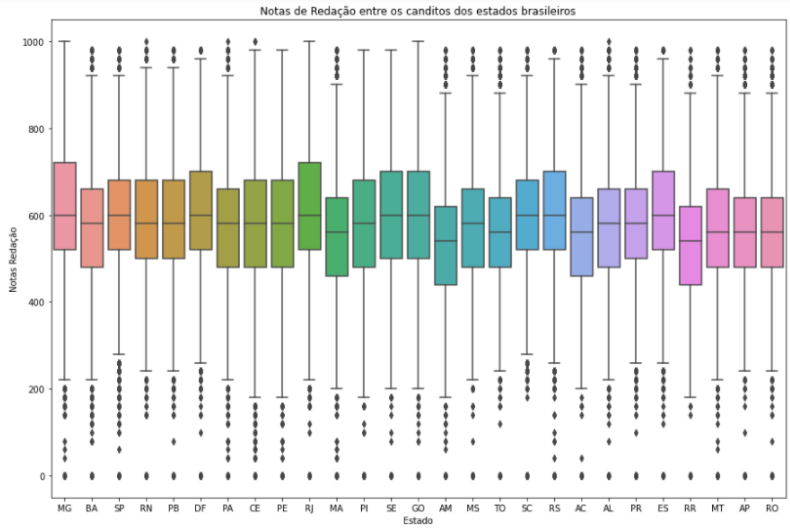


Figura 7 – Boxplot para as notas da prova de Redação por estado



Os boxplots acima mostra que em todas as provas, independente do estado que pertence o candidato, existem valores outliers, que são valores que fogem do padrão determinado por aplicações estatísticas para um intervalo de valores. Apenas para a prova de Redação é que existem notas entre 0 e 300. Os comportamentos dos valores variam de estado para estado, não possuindo uma relação entre si.

Dois boxplots também foram criados para dimensionar a variação das notas de duas provas específicas: matemática e redação, de acordo com a renda declarada pelo candidato durante a inscrição. A figuras 8 e 9 apresentam esses gráficos.

Figura 9 – Boxplot para as notas da prova de Matemática por Renda declarada

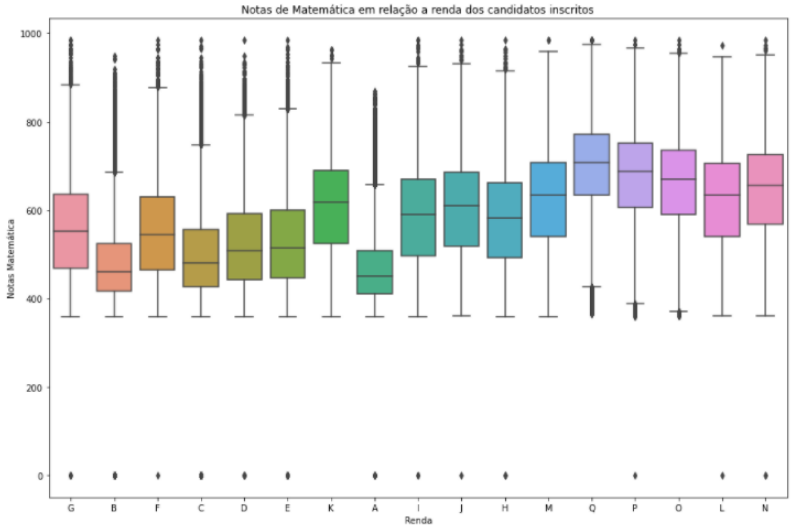
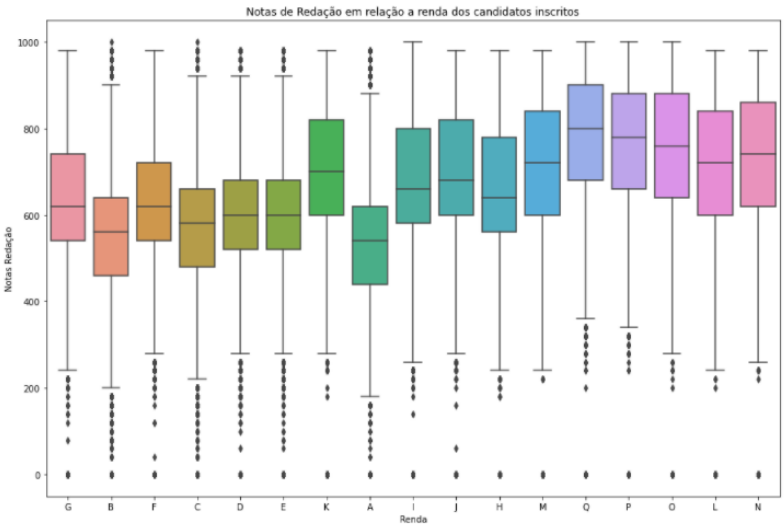


Figura 9 – Boxplot para as notas da prova de Redação por Renda declarada



Para uma análise aprofundada, foi criado os dois boxplot acima relacionando as notas das provas de Matemática e de Redação em relação as rendas familiares declaradas pelos candidatos no momento da inscrição para a prova do Enem 2019.

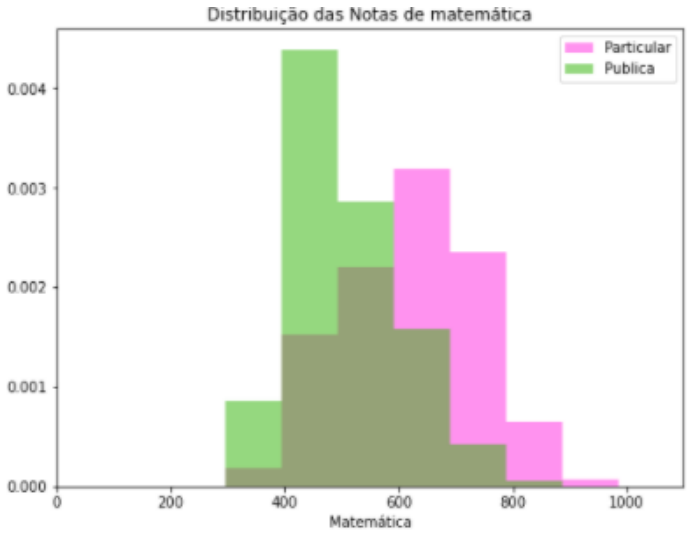
As classes estão definidas da seguinte forma:

- A Nenhuma renda.
- 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.

A média das notas da prova de matemática das classes de J à Q estão localizadas acima de 600. As demais médias estão abaixo de 600. A média das notas da prova de redação das classes de A à E estão localizadas abaixo de 600. As demais médias estão acima de 600. Essas observações mostram que quanto maior a renda familiar declarada pelo candidato durante a inscrição na prova do Enem 2019, maior a probabilidade de suas notas serem maiores.

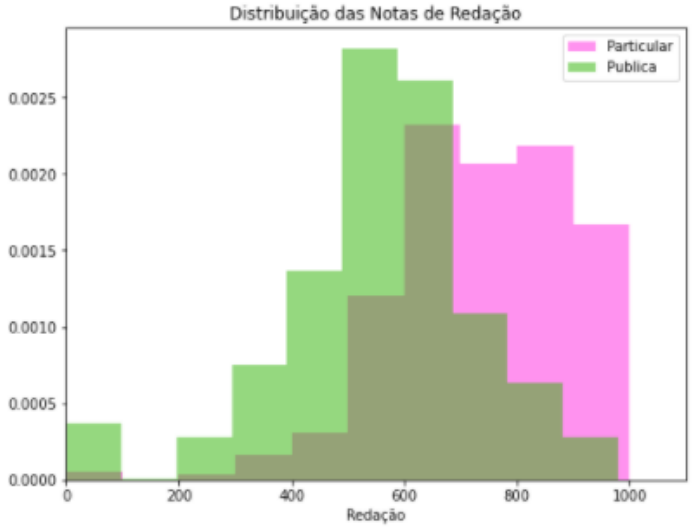
Após essas informações, buscou-se observar a distribuições das notas das provas de matemática e de redação sobre o tipo de escola que o candidato inseriu durante a inscrição, de modo a verificar se existe alguma influencia do tipo de escola sobre a nota do candidato. As figuras 10 e 11 apresentam os histogramas.

Figura 10 – Histograma de distribuição da nota na prova de Matemática para escolas particulares e públicas



Observando o histograma acima pode-se concluir que os candidatos que declararam durante a inscrição para realizar a prova do Enem de 2019 que são oriundos de escolas particulares apresentam uma maior distribuição de suas notas localizadas entre o intervalo de 500 a 800. A distribuição das notas dos candidatos de escola pública estão centrais ao gráfico, onde sua maioria está entre o intervalo de 400 a 700.

Figura 11 – Histograma de distribuição da nota na prova de Redação para escolas particulares e públicas



Observando o histograma acima pode-se concluir que os candidatos que declararam durante a inscrição para realizar a prova do Enem de 2019 que são oriundos de escolas particulares apresentam uma maior distribuição de suas notas localizadas a direita do gráfico, região onde estão localizadas as melhores notas na prova de Redação.

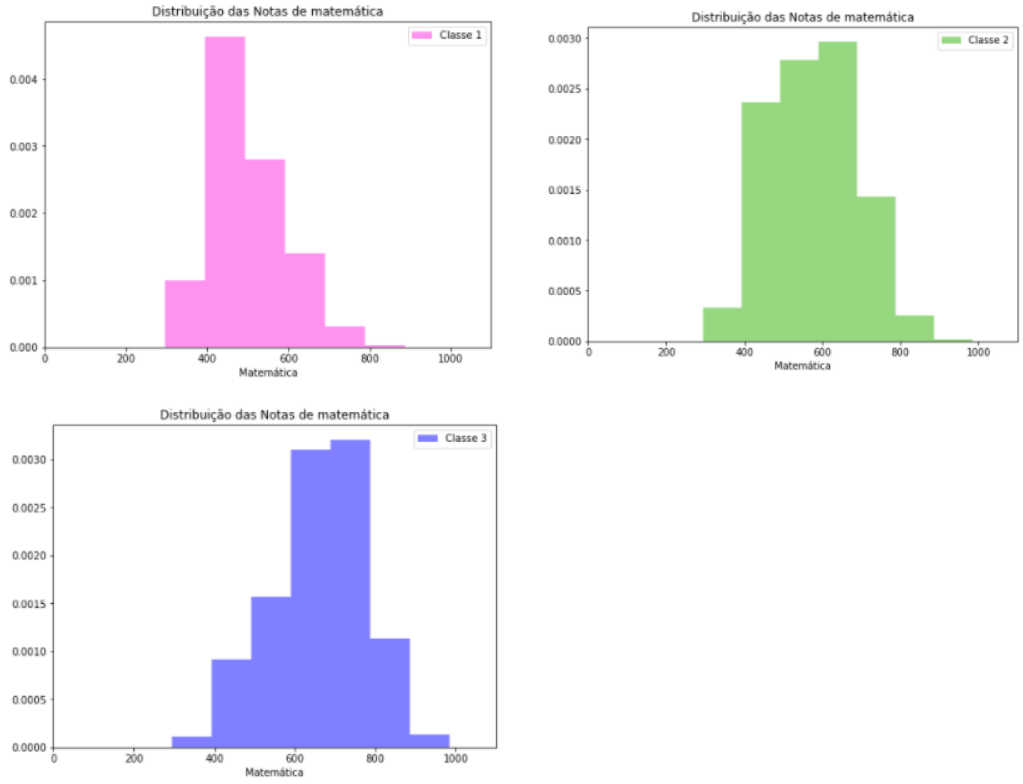
A distribuição das notas dos candidatos de escola pública estão centrais ao gráfico, onde sua maioria está entre o intervalo de 400 a 800.

Para otimizar os dados relativos as rendas familiares inseridas pelos candidatos no ato da inscrição para o Enem 2019, separou-se as mesmas em três classes, onde:

- Classe 1 = Renda familiar categorizada entre A e F;
- Classe 2 = Renda familiar categorizada entre G e L;
- Classe 3 = Renda familiar categorizada entre M e Q.

Com isso, foram criados histogramas para observar a distribuição das notas dos candidatos nas provas de matemática e de redação de acordo com a separação entre as classes, que diz respeito a renda familiar declarada pelo candidato. As figuras 12 e 13 são os histogramas para as provas de matemática e redação de acordo com a classe.

Figura 12 – Histogramas com distribuição das notas de matemáticas para as 3 classes

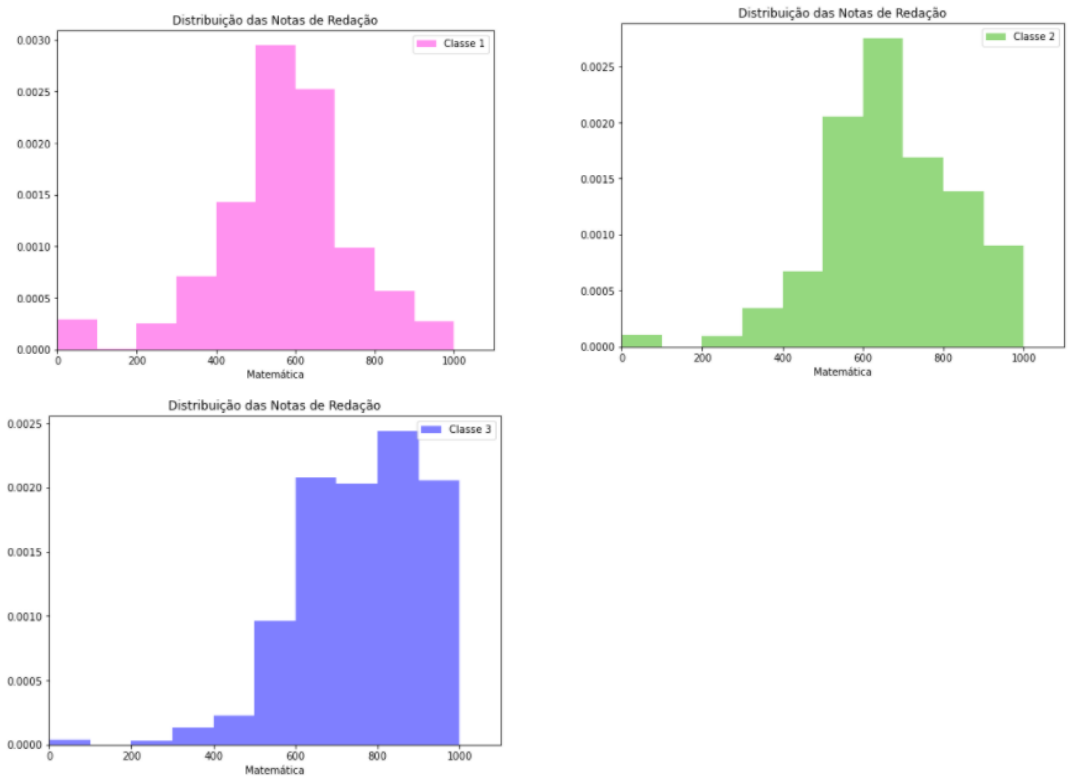


A classe 1 corresponde aos candidatos que informação a renda familiar inferior a R$ 2994,00. A classe 2 corresponde aos candidatos que informação a renda familiar superior a R$ 2994,00 e inferior a R$ 8982,00. A classe 3 corresponde aos candidatos que informação a renda familiar superior a R$ 8982,00 ou possui renda familiar acima de R$ 19960,00.

Os histogramas acima mostram a distribuição das notas na prova de matemática para os participantes entre as três classes criadas para limitar a renda familiar declarada durante a inscrição do candidato para a prova do Enem 2019. Pode-se observar que a distribuição que está localizada mais a direita dos gráficos correspondem a obtenção das melhores notas. Nesse caso, a classe onde a maioria das notas estão localizadas mais a direita do gráfico corresponde a classe 3, onde a renda familiar declarada pelo candidato é superior a R$ 8982,00 ou possui renda familiar acima de R$ 19960,00.

A distribuição das notas da classe 2 está em sua maioria localizada entre o intervalo das notas de 400 a 800, enquanto que os candidatos que em sua inscrição declarou que a renda familiar é inferior a R$ 2994,00 tem a maioria de suas notas na prova de matemática entre 400 e 600.

Figura 13 – Histogramas com distribuição das notas de Redação para as 3 classes

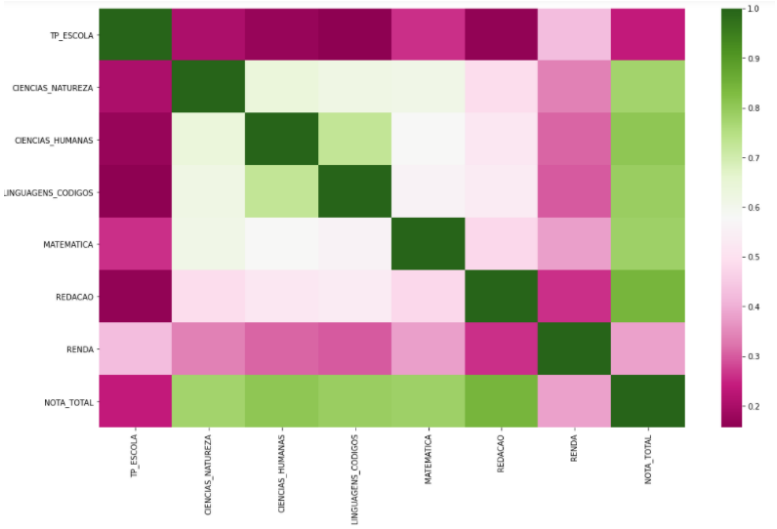


Em relação aos histogramas das notas das redações para as classes criadas, pode-se observar que as Classes 2 e 3 possuem a distribuição mais localizada a direita do gráfico, referente as notas maiores na prova em questão. A maior concentração das notas referentes a classe 3 está no intervalo entre 600 a 1000. Para a classe 1 a maior concentração das notas da prova de redação está no intervalo entre 400 e 800. De forma

geral, quanto maior a renda familiar declarada pelos participantes, observa-se uma probabilidade de tirar notas maiores.

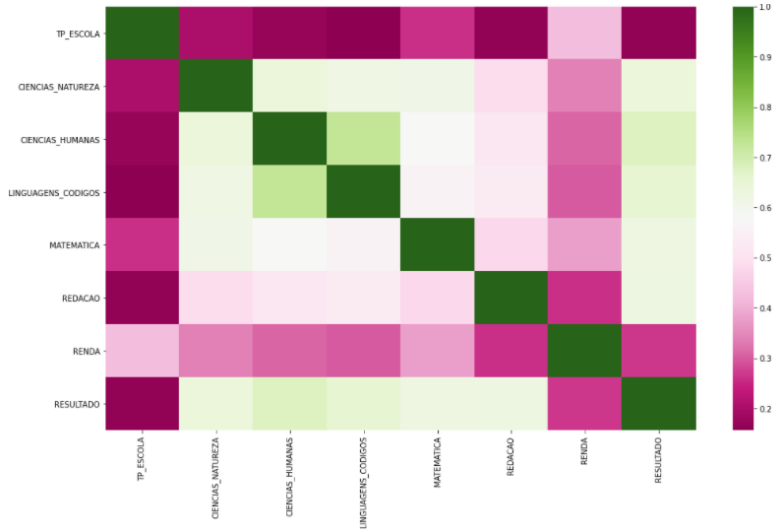
Após essas etapas anteriores, tentou-se estabelecer uma correlação entre os dados disponíveis de notas individuais das 5 provas do Enem, a nota total do candidato na prova fazendo um somatório das 5 provas e dividindo por 5 (média), e o tipo de escola que o candidato frequentou, a renda declarada pelo candidato afim de se determinar se existe uma relação entre essas variáveis. Um mapa de calor foi criado para verificar essa correlação, e o mesmo é apresentado na figura 14.

Figura 14 – Mapa de calor para verificar correlação entre as variáveis escolhidas



Buscou-se estabelecer uma correlação entre a Renda e o Tipo de escola, que foi inserida pelo candidato durante a inscrição para a prova do Enem 2019, com a nota final obtida pelo candidato, porém verificou-se que existe pouca correlação entre as variáveis. Dessa forma, criou-se uma separação entre os candidatos que tiveram notas acima de 510 e os que ficaram abaixo desse valor. De modo a verificar se o aluno que ultrapassou esse valor tem relação a determinada Renda ou Tipo de escola. A figura 15 agora apresenta o mapa de calor considerando o resultado do candidato, acima ou abaixo de 510.

Figura 15 – Mapa de calor para verificar correlação entre as variáveis escolhidas



A avaliação do mapa de calor acima mostra que também não existe uma forte correlação entre o resultado total da prova do Enem de 2019, com o tipo de renda ou o tipo de escola que o candidato declarou durante a inscrição.

Separou-se a base de dados em 70% para treinamento e 30% para testar o modelo. O modelo irá aplicar os algoritmos de classificação em 70% da base e depois com os 30%, onde os mesmos foram fornecidos ao modelo sem a coluna alvo, para que o modelo possa tentar prever o resultado, e assim medir o nível de acurácia do modelo. Os modelos utilizados foram KNN e Random Florest. Os resultados são mostrados nos quadros 2 e 3 abaixo.

Quadro 2 – Resultados obtidos com o algoritmo KNN

	precision	recall	f1-score	support
0	0.56	0.47	0.51	142033
1	0.55	0.63	0.58	144131
accuracy			0.55	286164
macro avg	0.55	0.55	0.55	286164
weighted avg	0.55	0.55	0.55	286164

Quadro 3 – Resultados obtidos com o algoritmo Random Florest

	precision	recall	f1-score	support
0	0.57	0.91	0.70	142033
1	0.79	0.32	0.46	144131
accuracy			0.62	286164
macro avg	0.68	0.62	0.58	286164
weighted avg	0.68	0.62	0.58	286164

Em relação aos dois algoritmos utilizados, observou-se que o Random Florest apresentou uma acurácia superior ao KNN. Isso pode ser explicado pelo fato do primeiro ser mais robusto, em comparação com o segundo. Outro ponto a se observar é que a

acurácia de ambos os métodos foi abaixo de 70%. Provavelmente a falta de correlação entre as variáveis resultam nessa dificuldade de o modelo encontrar uma previsão satisfatória. Uma avaliação mais aprofundada, com outros métodos de previsão, seria uma alternativa para trabalhos futuros. Apesar de muitos dados disponíveis, tornou-se difícil de determinar um modelo que pudesse ter uma alta acurácia e prever o comportamento do tipo de escola, a renda, como influenciadoras na nota final do candidato.