PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Matheus Araújo Tomaz

Análise da Relação Entre o Perfil Socioeconômico dos Inscritos do ENEM 2018 Com o Acesso Ao Ensino Superior Gratuito no Brasil

Belo Horizonte
2020

Matheus Araújo Tomaz

ANÁLISE DA RELAÇÃO ENTRE O PERFIL SOCIOECONÔMICO DOS INSCRITOS DO ENEM 2018 COM O ACESSO AO ENSINO SUPERIOR GRATUITO NO BRASIL

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2020

SUMÁRIO

1. Introdução

1.1. Contextualização

Com o intuito de avaliar o desempenho dos estudantes concluintes do ensino médio e posteriormente, democratizar o acesso ao ensino superior no Brasil, o Enem (Exame Nacional do Ensino Médio) foi criado em 1998. A partir de 2009 medidas governamentais estimularam o uso do ENEM não apenas como um processo de avaliação do Ensino Médio, mas como forma de acesso ao ensino superior no Brasil. (Silveira et al. 2015).

Também com o objetivo de democratizar o acesso ao ensino superior, o Governo Federal vem implementando nos últimos anos, políticas públicas educacionais de inserção socioeducacional no ambiente universitário, como forma de contribuir para o desenvolvimento do panorama educacional no Brasil (Arruda, 2020).

Como exemplo de políticas públicas educacionais, podemos citar o ProUni (Programa Universidade para Todos) e o SISU (Sistema de Seleção Unificada), que são programas governamentais de ação afirmativa, que visam à inserção de estudantes aprovados no Enem.

1.2. O problema proposto

Por quê?

Ao falar sobre a educação no Brasil, diferentes estudos realçam a relação direta em que os resultados escolares dos alunos têm com as condições socioeconômicas dos pais (Travitzki et al., 2011). As variáveis mais examinadas com relação ao desempenho escolar dos estudantes, são o nível de rendimento familiar, o nível de escolaridade dos pais e a raça/cor (Travitzki et al, 2016).

Para quem?

Entende-se que a metodologia e as análises realizadas neste estudo podem ser replicadas e utilizadas, respectivamente, para futuras análises mais detalhadas. O mesmo, inclusive, poderia ser utilizado para estudos de outras áreas de conhecimento, como por exemplo, estudos sociais sobre o acesso ao ensino superior gratuito no Brasil e devem ser aprimoradas e estudadas com mais profundidade para

que se tenha melhores resultados e que possam auxiliar outros estudos, inclusive de cunho social.

O quê?

Este estudo teve como objetivo realizar uma análise e disponibilizar a metodologia aplicada, para que outros estudos possam utilizar e aprimorar o estudo sobre como o perfil socioeconômico dos inscritos pode influenciar no acesso ao ensino superior gratuito. Tendo também como objetivo, desenvolver um modelo estatístico com o intuito de encontrar e evidenciar a relação que o perfil socioeconômico tem com o acesso ao ensino superior gratuito, com base nos dados.

Onde?

Os dados coletados e analisados, representam o perfil socioeconômico dos inscritos do ENEM, assim como também todos os cursos disponibilizados no PROUNI, e que abrange todo o Brasil.

Quando?

Os dados analisados correspondem ao ENEM e o PROUNI do ano de 2018.

2. Coleta de Dados

Os dados tratados neste estudo foram coletados a partir de fontes públicas, disponibilizadas na internet. Vale lembrar que os dados aqui utilizados foram inseridos nestes repositórios por usuários e podem não ser oficiais.

Contudo, como citado, o objetivo deste trabalho é disponibilizar a metodologia aplicada, para que outros estudos possam replicar, inclusive com dados oficiais.

Abaixo seguem os links das bases utilizadas e dos metadados das mesmas:

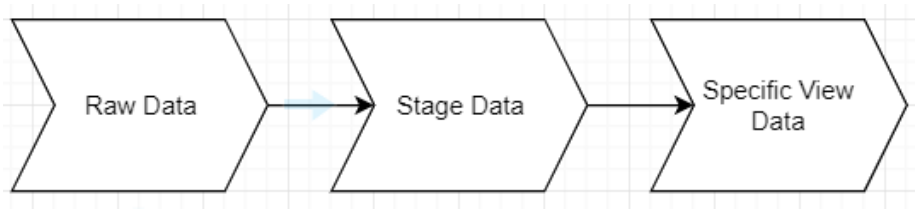
- **Metadados**
  https://drive.google.com/file/d/17HJPtZpqFEYj47-_puAaD39szL9pPXLD/view
  (Repositório pessoal)

- **ENEM 2018** (5.513.747 registros)
  https://www.kaggle.com/ffmenezes/microdados-enem-2018 (Kaggle)
  https://drive.google.com/file/d/1pNZKbv_3fh6HUJV28CWDGNVuZYMyfBQB/view?usp=sharing (Repositório pessoal)

- **PROUNI 2018** (41.447 registros)
  https://brasil.io/dataset/cursos-prouni/cursos/ (Brasil IO)
  https://drive.google.com/file/d/1fiG443cxmsBBBiMFaJLWjTTWSu6lyL4l/view?usp=sharing (Repositório Pessoal)
- **Cursos agrupados por área de conhecimento** (362 registros)
  https://drive.google.com/file/d/1lCaLFtZpmzJAKp93Mfh23ayiHMjuK8qc/view?usp=sharing (Repositório Pessoal)

3. Processamento/Tratamento de Dados

Para realizar a análise deste estudo, o tratamento dos dados foi feito obedecendo três camadas:



A primeira camada (*Raw data*) corresponde aos dados brutos. São dados ingeridos da forma como constam nas fontes de origem, para que sejam tratados na camada intermediária (*stage*).

A segunda camada (*Stage data*) corresponde aos dados tratados após passar pelo processo de limpeza dos dados, onde são retirados os valores faltantes, os dados imprecisos e etc. Os dados aqui já estão prontos para análises.

Na terceira e última camada (*Specific view data*) estão os dados específicos para cada aplicação, como no caso deste estudo, os dados prontos para o processo de treinamento do modelo.

3.1. Metodologia

Nessa subseção, será explicado como foi feito o tratamento dos dados com base no fluxo e camadas descritas. Para melhor entendimento, os passos foram divididos nas camadas citadas e o passo a passo das transformações realizadas, serão explicadas coluna por coluna.

3.1.1. Raw data

Utilizando os *datasets* citados no capitulo 2 (Coleta de dados), foram selecionadas as colunas que continham informações necessárias para o desenvolvimento deste estudo:

3.1.1.1. ENEM 2018

- SG_UF_RESIDENCIA
- NU_IDADE
- TP_SEXO
- TP_ESTADO_CIVIL
- TP_COR_RACA
- TP_ANO_CONCLUIU
- TP_ESCOLA
- TP_ENSINO
- IN_TREINEIRO
- NU_NOTA_CN
- NU_NOTA_CH
- NU_NOTA_LC
- NU_NOTA_MT
- NU_NOTA_REDACAO
- Q005
- Q006
- Q027

3.1.1.2. PROUNI 2018

- nome
- nota_integral_ampla
- nota_integral_cota

3.1.1.3. Cursos agrupados por área de conhecimento

- AREA DE CONHECIMENTO
- CURSO

3.1.2. Stage data

Para este estágio do tratamento dos dados, foram utilizadas as colunas definidas na etapa anterior para a construção da tabela final, pronta para a análise. Segue todos os passos utilizados para a obtenção da mesma:

3.1.2.1. ENEM 2018

Para as colunas a seguir foi realizado o “de para” com base nos metadados da tabela:

- *TP_ESTADO_CIVIL*

  0 → Solteiro

  1 → Casado

  2 → Divorciado

  3 → Viuvo

  Para os valores faltantes foi atribuído o valor “Não Informado”.

- *TP_COR_RACA*

  0 → Não Declarado

  1 → Branca

  2 → Preta

  3 → Parda

  4 → Amarela

  5 → Indígena

- *TP_ANO_CONCLUIU*

  0 → Não Informado

  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*

  1 → Não Respondeu

  2 → Publica

3 → Privada

4 → Exterior

Para esta coluna, foi identificado um alto número de valores “Não Respondeu”, o que é considerado como dado faltante. Porém, para estes registros, podemos utilizar a coluna *Q027*, que corresponde às respostas do questionário socioeconômico do ENEM para a pergunta: “Em que tipo de escola você frequentou no ensino médio?”.

Logo, para os registros faltantes da coluna *TP_ESCOLA*, utilizamos o seguinte “de para” com os valores da coluna *Q027*:

Se o valor da coluna *Q027* é igual a “A” → Publica

Se o valor da coluna *Q027* é igual a “D” ou “E” → Privada

Se o valor da coluna *Q027* é igual a “B” ou “C” → Publica/Privada

Se o valor da coluna *Q027* é igual a “F” → Não frequentou a escola

- *RENDA*

  Realizado um “de para” a partir da coluna *Q006*:

  “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

- *Q006*

  Com o intuito de ordenar a renda, foi feito o “de para”:

  “A” → 0
  “B” → 1
  “C” → 2
  “D” → 3
  “E” → 4
  “F” → 5
  “G” → 6
  “H” → 7
  “I” → 8
  “J” → 9
  “K” → 10
  “L” → 11
  “M” → 12
  “N” → 13
  “O” → 14
  “P” → 15
  “Q” → 16

- *REGIAO*

  Realizado um “de para” a partir da coluna SG_UF_RESIDENCIA:

  Se “RS”, “SC” ou “PR” → SUL

  Se “SP”, “MG”, “RJ” ou “ES” → SUDESTE

  Se “MT”, “MS”, “GO” ou “DF” → CENTROESTE

  Se “AM”, “RO”, “RR”, “AC”, “AP”, “PA” ou “TO” → NORTE

  Se “MA”, “PB”, “PI”, “CE”, “RN”, “BA”, “PE”, “SE” ou “AL” → NORDESTE

Para as colunas *“NU_NOTA_CH”, “NU_NOTA_LC”, “NU_NOTA_MT”, “NU_NOTA_REDACAO”, “IN_TREINEIRO”* e “*Q005*” foram mantidos os valores.

A coluna *NOTA_TOTAL* é resultante da soma das colunas *“NU_NOTA_CH”, “NU_NOTA_LC”, “NU_NOTA_MT”, “NU_NOTA_REDACAO”.*

Como último passo no tratamento dos dados da tabela ENEM 2018, foi realizado um filtro, considerando apenas os registros que não tivessem valores nulos para as colunas *“NU_NOTA_CH”, “NU_IDADE”* e *“NU_NOTA_MT”,* e um filtro considerando apenas os inscritos que não seriam treineiros, a partir da coluna “*IN_TREINEIROS*”, resultando em uma tabela com 3.434.726 de registros.

3.1.2.2. PROUNI NOTAS

Esta tabela tem um total de 8 registros, um para cada área de conhecimento.

- *AREA DE CONHECIMENTO*

Para obter esta coluna foi cruzada a coluna “nome” da tabela “PROUNI 2018” com a coluna “CURSO” da tabela “Cursos agrupados por área de conhecimento”, após retirar os registros duplicados da última tabela.

Foi necessário tratar o nome de alguns cursos, como mostrado na planilha: https://drive.google.com/file/d/1NtXgcXMlrljCthiF-o8guVRfoKXMUy_C/view

- *nota_ampla_media*

Média da coluna “nota_integral_ampla” da tabela PROUNI 2018

- *nota_ampla_min*

Menor valor da coluna “nota_integral_ampla” da tabela PROUNI 2018

- *nota_ampla_25*

Maior nota do primeiro quartil da coluna “nota_integral_ampla” da tabela PROUNI 2018

- *nota_ampla_50*

Maior nota do segundo quartil da coluna “nota_integral_ampla” da tabela PROUNI 2018

- *nota_ampla_75*

Maior nota do terceiro quartil da coluna “nota_integral_ampla” da tabela PROUNI 2018

- *nota_ampla_max*

Maior nota da coluna “nota_integral_ampla” da tabela PROUNI 2018

- *nota_cotas_media*

Média da coluna “nota_integral_cota” da tabela PROUNI 2018

- *nota_cotas_min*

Menor valor da coluna “nota_integral_cota” da tabela PROUNI 2018

- *nota_cotas_25*

  Maior nota do primeiro quartil da coluna “nota_integral_cota” da tabela PROUNI 2018

- *nota_cotas_50*

  Maior nota do segundo quartil da coluna “nota_integral_cota” da tabela PROUNI 2018

- *nota_cotas_75*

  Maior nota do terceiro quartil da coluna “nota_integral_cota” da tabela PROUNI 2018

- *nota_cotas_max*

  Maior nota da coluna “nota_integral_cota” da tabela PROUNI 2018

3.1.2.3. TABELA FINAL

Como último passo desta etapa, foi realizado a unificação das duas tabelas acima, adotando as seguintes premissas:

- Foram considerados as notas de corte do SISU e as notas de corte das bolsas integrais do PROUNI como notas de corte de todo o ensino gratuito.
- Para acesso ao PROUNI, neste estudo, não foi considerado o critério de seleção por renda.
- Como este estudo não conseguiu acesso à base de dados do SISU, e no geral, as notas de corte do SISU são maiores que as notas de corte do PROUNI, foram adotas as notas de corte do PROUNI como notas de corte de todo o ensino gratuito, ou seja, caso a nota de um inscrito no ENEM for maior que a nota de corte do curso do PROUNI, ele teria acesso ao ensino superior gratuito, pois teria garantido acesso ao menos no PROUNI.
- Como os valores de nota para as vagas destinadas às cotas são maiores que as notas de ampla concorrência, para este estudo, foram consideradas apenas as notas de ampla concorrência para corte.
- O maior valor do segundo quartil de ampla concorrência (nota_ampla_50) foi considerado como a nota de corte para que um escrito tivesse uma alta probabilidade de ser aprovado em algum ensino gratuito.

- Foi considerado que um inscrito no ENEM tem baixa probabilidade de acesso ao ensino gratuito se o mesmo obtivesse uma nota total maior que a menor nota de ampla concorrência (*nota_ampla_min*) e uma nota total menor que a maior nota do segundo quartil de ampla concorrência (*nota_ampla_50*).
- Não foi considerado, neste estudo, diferentes pesos das notas de cada prova do ENEM (Ciências da Natureza, Ciências Humanas, Matemática, Linguagem e Códigos, e Redação) para cada curso específico, como ocorre no processo do SISU e PROUNI.

Considerando estas premissas, foram realizadas as comparações pelas notas totais dos inscritos do ENEM com as faixas de notas de corte de cada área de conhecimento (PROUNI NOTAS), obtidas a partir do agrupamento das notas de cortes dos cursos por área de conhecimento, como mostra a tabela a seguir:

Área de Conhecimento	Menor nota de corte	Maior nota de corte do segundo quartil (50%)
Ciências Agrárias	450.0	607.05
Ciências Biológicas	450.0	623.64
Ciências Exatas e da Terra	450.0	598.89
Ciências Humanas	450.0	565.57
Ciências Sociais Aplicadas	450.0	568.70
Ciências da Saúde	450.0	601.11
Engenharias	519.0	572.74
Lingüística-Letras e Artes	450.0	575.34

DE PARA

- Se a coluna “NOTA_TOTAL” da tabela ENEM 2018 for menor que a **Menor nota de corte**, então o inscrito tem probabilidade NULA de acesso ao ensino superior gratuito.
- Se a coluna “NOTA_TOTAL” da tabela ENEM 2018 for maior que a **Menor nota de corte** e menor que a **Maior nota de corte do segundo quartil (50%)**, então o inscrito tem probabilidade BAIXA de acesso ao ensino superior gratuito
- Se a coluna “NOTA_TOTAL” da tabela ENEM 2018 for maior que a **Maior nota de corte do segundo quartil (50%)**, então o inscrito tem probabilidade ALTA de acesso ao ensino superior gratuito.

Para cada área de conhecimento foi criado uma coluna na tabela final com o valor “0” para NULA, “1” para BAIXA e “2” para ALTA, seguindo os critérios acima.

Também foi criada a coluna “*Prob_Total*”, que indica a probabilidade total de um inscrito ter acesso ao ensino superior gratuito, independente da área de conhecimento.

Para isso, foi considerado que se o inscrito tem probabilidade ALTA em ao menos uma área de conhecimento, indicaria que ele tem ALTA probabilidade no total, pois teria acesso ao ensino superior gratuito pelo menos naquela área de conhecimento.

Para o inscrito que tem probabilidade NULA em todas as áreas de conhecimento, foi considerado que o mesmo tem probabilidade NULA no total.

Para o inscrito que não atendeu as duas condições abordadas anteriormente, ou seja, que não tem probabilidade ALTA em nenhuma área de conhecimento, porém tem probabilidade BAIXA em ao menos uma área de conhecimento, foi considerado que o mesmo tem probabilidade BAIXA no total.

Após as comparações e critérios definidos acima, temos as seguintes colunas da tabela final:

- *NO_MUNICIPIO_RESIDENCIA*
  Nome do município de residência
- *SG_UF_RESIDENCIA*
  Sigla UF de residência
- *NU_IDADE*
  Idade do inscrito
- *TP_SEXO*
  Gênero do inscrito
- *TP_ESTADO_CIVIL*
  Estado civil do inscrito
- *TP_COR_RACA*
  Etnia do inscrito
- *TP_ANO_CONCLUIU*
  Ano de conclusão do ensino médio
- *TP_ESCOLA*
  Tipo de escola que estudou

- *NU_NOTA_CN*

  Nota da prova de ciência das naturezas
- *NU_NOTA_CH*

  Nota da prova de ciências humanas
- *NU_NOTA_LC*

  Nota da prova de Linguagem e Códigos
- *NU_NOTA_MT*

  Nota da prova de matemática
- *NU_NOTA_REDACAO*

  Nota da redação
- *Q005*

  (Questionário socioeconômico) Quantidade de pessoas que moram na mesma residência
- *Q006*

  (Questionário socioeconômico) Renda mensal da família (ordem de salários mínimos
- *Q027*

  (Questionário socioeconômico) Tipo de escola frequentada no ensino médio
- *RENDA*

  Renda mensal da família
- *NOTA_TOTAL*

  Soma de todas as notas das provas
- *REGIAO*

  Região do Brasil onde reside
- *Ciencias Agrarias*

  Probabilidade de acesso ao ensino superior gratuito em Ciências Agrárias
- *Ciencias Biologicas*

  Probabilidade de acesso ao ensino superior gratuito em Ciências Biológicas
- *Ciencias Exatas*

  Probabilidade de acesso ao ensino superior gratuito em Ciências Exatas
- *Ciencias Humanas*

  Probabilidade de acesso ao ensino superior gratuito em Ciências Humanas
- *Ciencias Sociais*

  Probabilidade de acesso ao ensino superior gratuito em Ciências Sociais
- *Ciencias da Saúde*

  Probabilidade de acesso ao ensino superior gratuito em Ciências da Saúde
- *Engenharias*

  Probabilidade de acesso ao ensino superior gratuito em Engenharias

- *Linguística-Letras e Artes*

  Probabilidade de acesso ao ensino superior gratuito em Linguística-Letras e Artes
- *Prob_Total*

  Probabilidade total de acesso ao ensino superior gratuito

Segue link para a base final:

https://drive.google.com/file/d/1tnZdqrFVl-9u8OuDoKhREQk_QQ7GR1jF/view?usp=sharing

3.1.3. Specific View Data

Como dito anteriormente, para este estudo, a camada *Specific View Data* abrange uma tabela pronta para treino do modelo de predição. Este modelo tem objetivo de encontrar as variáveis mais fortes que influenciam na hora de predizer se um inscrito terá probabilidade alta, baixa ou nula de ter acesso ao ensino superior gratuito.

Neste modelo, foi abordado apenas a probabilidade total "*Prob_total*" de um inscrito ter acesso ao ensino superior gratuito, de modo que não foi considerado a probabilidade por área de conhecimento.

Também não foram consideradas as notas que cada aluno tirou em cada prova do ENEM, e nem a nota total, pois essas variáveis estão fortemente correlacionadas com a variável resultante, uma vez que um simples "de para" das nota já poderia responder quem teria probabilidade alta ou não de acesso ao ensino superior gratuito. Como citado, o objetivo deste modelo é apontar o que o perfil socioeconômico de um inscrito pode influenciar no seu acesso ao ensino superior gratuito.

Foram excluídas também, da base de treino do modelo, a coluna com o nome do município de cada inscrito (por conter um alto grau de granularidade na informação, como por exemplo vários nomes de municípios únicos, o que poderiam causar um *overfiting* do modelo), a coluna com a sigla UF de residência (uma vez que optamos por utilizar a região do Brasil onde o inscrito reside, o que está fortemente correlacionado com o estado de residência), e por fim, as coluna *Q006* e *Q027*, que contém informações redundantes quando comparadas com as colunas *RENDA* e *TP_ESCOLA*, respectivamente.

Logo, para o treino deste modelo, foram consideradas as colunas abaixo:

- *NU_IDADE*

  Idade do inscrito

- *TP_SEXO*

  Gênero do inscrito
- *TP_ESTADO_CIVIL*

  Estado civil do inscrito
- *TP_COR_RACA*

  Etnia do inscrito
- *TP_ANO_CONCLUIU*

  Ano de conclusão do ensino médio
- *TP_ESCOLA*

  Tipo de escola que estudou
- *Q005*

  (Questionário socioeconômico) Quantidade de pessoas que moram na mesma residência
- *RENDA*

  Renda mensal da família
- *REGIAO*

  Região do Brasil onde reside
- *Prob_Total*

  Probabilidade total de acesso ao ensino superior gratuito

Após a definição das colunas utilizados para treinamento do modelo, foi realizado um filtro para considerar apenas os registros que contivessem a informação do ano de conclusão do inscrito no ensino médio, buscando por todos os registros da coluna *TP_ANO_CONCLUIU* que fossem diferente de “Não Informado”, resultando em uma tabela com 2.048.859 registros.

Com os dados filtrados, foi realizado um tratamento para as variáveis categóricas, transformando-as em variáveis *dummies*, que consiste em transformar variáveis não numéricas em numéricas, sem perder a informação e sem estabelecer pesos diferentes caso criássemos uma ordem ao optar trocar os valores por “1, 2, 3...”, por exemplo. Para realizar esta transformação é necessário criar uma coluna para cada valor de uma coluna não numérica, e para um registro em questão, marcar qual dessas colunas corresponde ao valor da coluna tratada, de forma binária (0 e 1), como é mostrado no exemplo abaixo:

> Para a coluna *TP_ESTADO_CIVIL*, neste exemplo, temos os valores “Solteiro”, “Casado” e “Divorciado”. Logo, após o tratamento, teríamos 3 colunas resultantes desta transformação:

"*TP_ESTADO_CIVIL_Solteiro*", "*TP_ESTADO_CIVIL_Casado*" e "*TP_ESTADO_CIVIL_Divorciado*". Para um registro da tabela que contenha o valor "Casado", por exemplo, a coluna "TP_ESTADO_CIVIL_Casado " receberia o valor "1", que corresponde a "Verdadeiro" e as outras colunas citadas receberiam "0", que corresponde a "Falso", como mostra a imagem a seguir:

Coluna TP_ESTADO_CIVIL antes do tratamento:

INSCRITO	TP_ESTADO_CIVIL
xpto	Casado

Após o tratamento:

INSCRITO	TP_ESTADO_CIVIL_Solteiro	TP_ESTADO_CIVIL_Casado	TP_ESTADO_CIVIL_Divorciado
xpto	0	1	0

Desta forma, não perdemos a informação e transformamos as mesmas variáveis que seriam não numéricas, em numéricas.

No entanto, é importante ressaltar, que esta técnica foi utilizada pois iremos utilizar de técnicas de predição que exigem o uso de variáveis numéricas, como algoritmos de regressão.

Como último passo, para melhor classificarmos os inscritos, foi considerado que aqueles que tivessem probabilidade ALTA teriam acesso ao ensino superior gratuito e aqueles que tivessem probabilidade BAIXA ou NULA, não teriam acesso ao ensino superior gratuito.

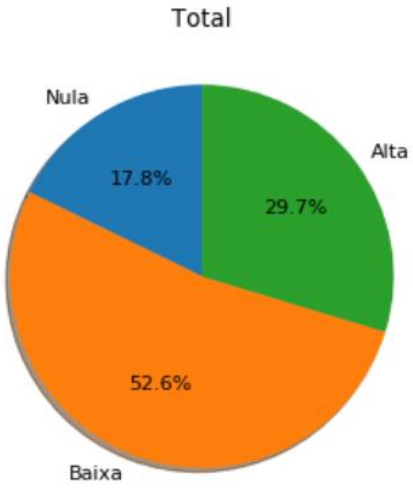
4. Análise e Exploração dos Dados

Nesta etapa, para análise dos dados, foi utilizada a tabela resultante da camada 2 (*Stage data*), citada no capítulo 3.

4.1. Probabilidade total de acesso ao ensino superior gratuito e por área de conhecimento

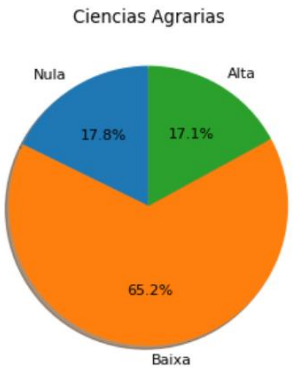
O primeiro passo da análise foi entender como a base estava dividida a partir dos resultados de probabilidade para cada área de conhecimento e para a probabilidade total de acesso ao ensino superior gratuito.

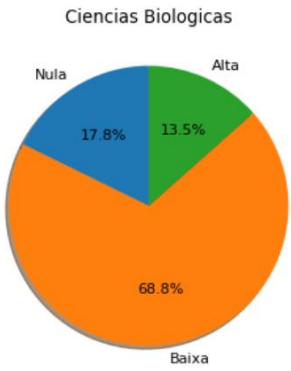
Nesta primeira análise, foi possível identificar, com as premissas e critérios utilizados no capítulo 3, que 29,7% dos inscritos teriam probabilidade ALTA, 52,6% teriam probabilidade BAIXA e 17,8% teriam probabilidade NULA, como mostra a figura abaixo:

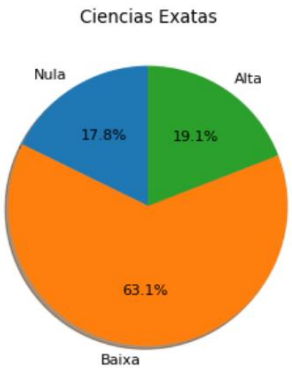


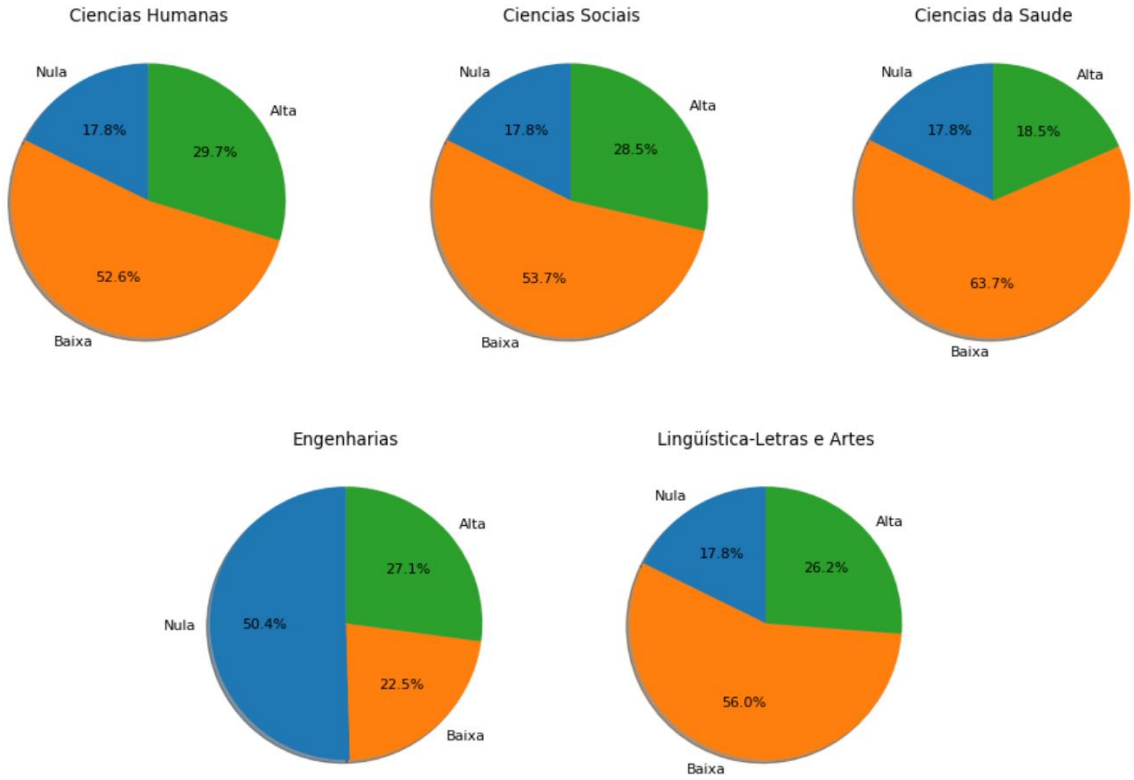
Também foram realizadas análises para evidenciar como os dados estão divididos por área de conhecimento, como demonstrado nas figuras abaixo:

4.1.1. Por área de conhecimento







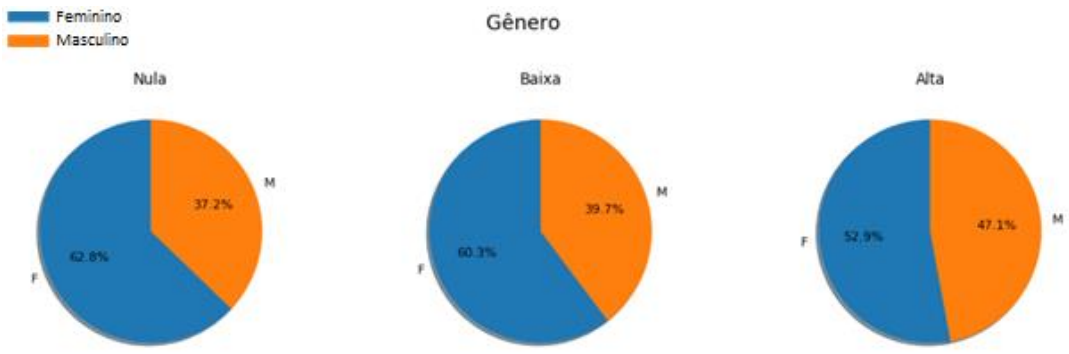


4.2. Análise dos dados a partir da probabilidade de acesso ao ensino superior gratuito

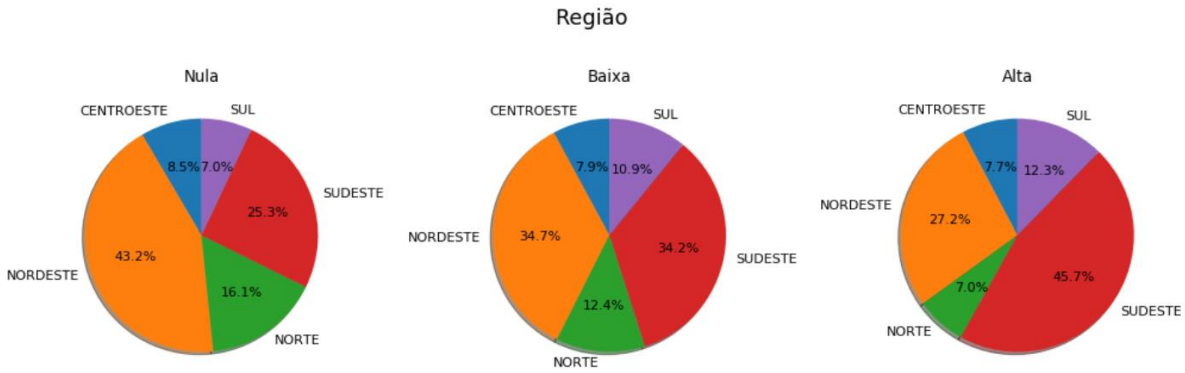
A partir da probabilidade (NULA, BAIXA e ALTA) de acesso ao ensino superior gratuito, foi realizada a análise para cada coluna, identificando a distribuição de cada valor (gênero, região, estado civil, etnia, ano de conclusão e tipo de escola). Para melhor entendimento, o objetivo desta análise é responder as seguintes perguntas, pegando um valor como exemplo: Entre os inscritos que tiveram probabilidade ALTA, quantos eram Homens? E Mulheres?

Abaixo, seguem figuras para evidenciar as análises:

4.2.1. Por gênero

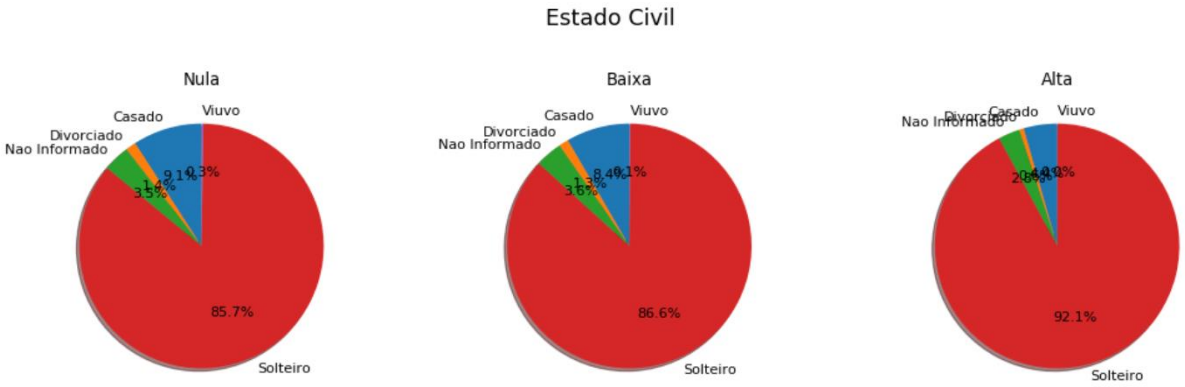


4.2.2. Por região

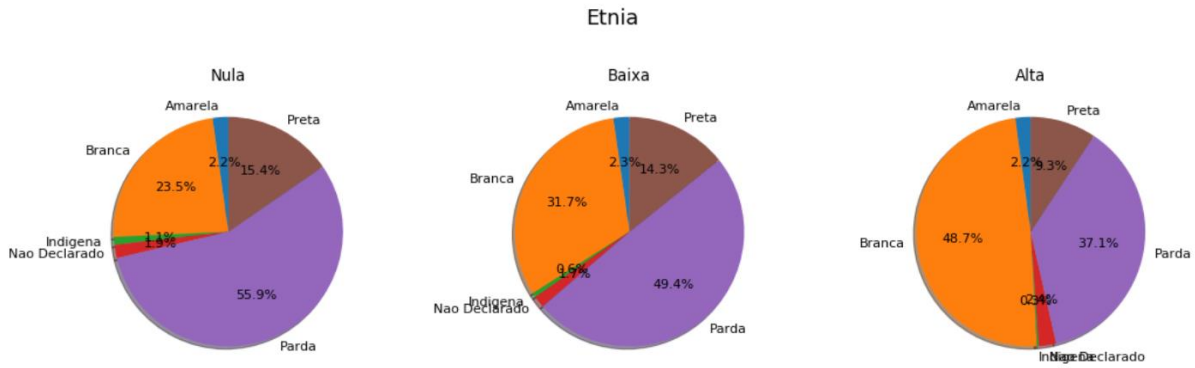


,

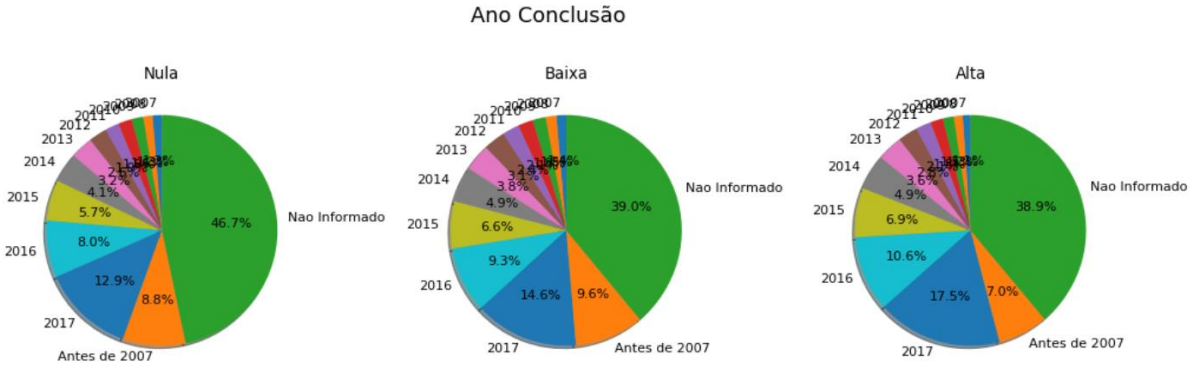
4.2.3. Por estado civil



4.2.4. Por etnia



4.2.5. Por ano de conclusão



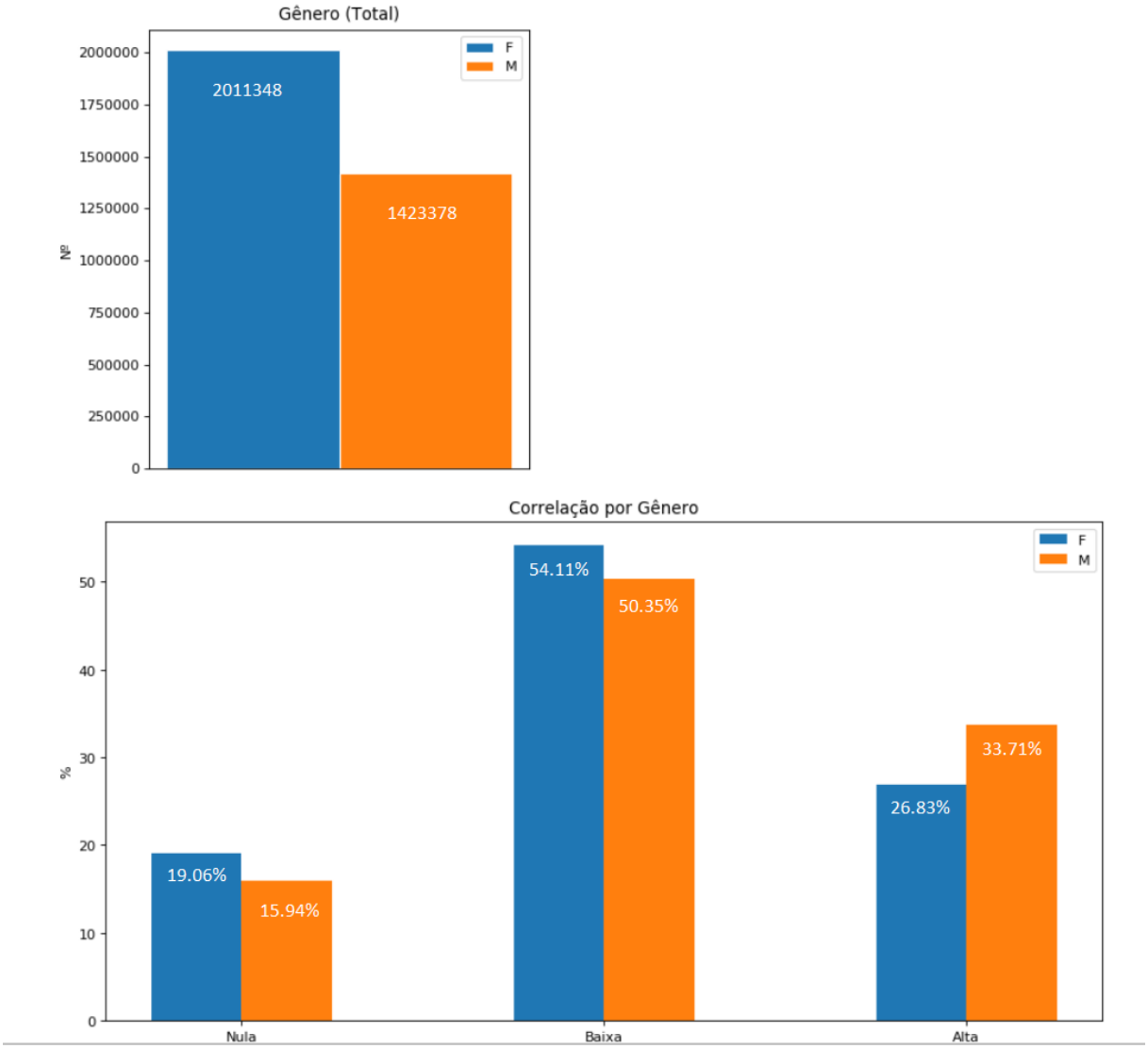
4.2.6. Por tipo de escola

4.3. Análise dos dados a partir dos valores de cada coluna

Na análise anterior, foi feita uma identificação de como os valores de cada coluna estavam divididos a partir da probabilidade de acesso ao ensino superior gratuito. Já nesta análise, o objetivo foi identificar qual a porcentagem que cada valor, em uma coluna específica, teria de acesso ao ensino superior gratuito. Para melhor entendimento, a análise busca responder as seguintes perguntas, como exemplo: De todos os solteiros que realizaram o ENEM, quantos tiveram probabilidade NULA? Quantos tiveram probabilidade BAIXA? E probabilidade ALTA?

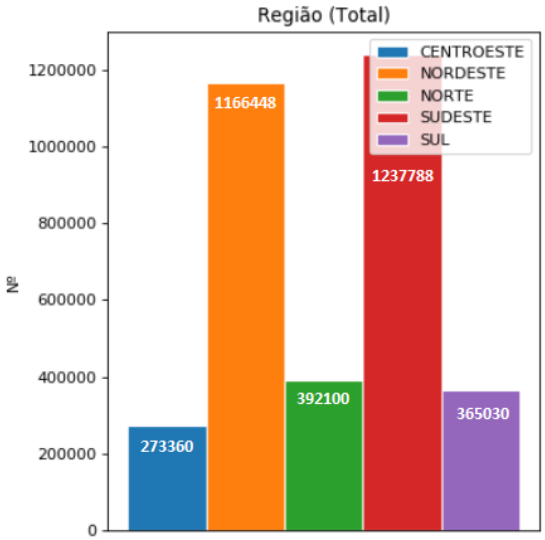
Então foi realizada as seguintes comparações:

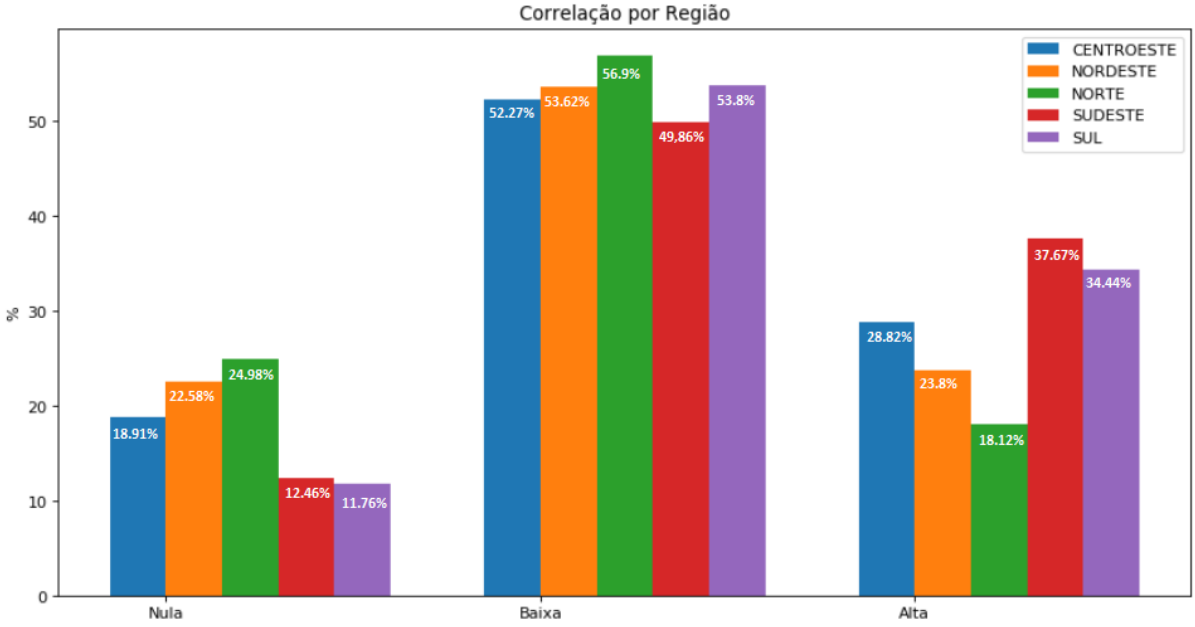
4.3.1. Por gênero



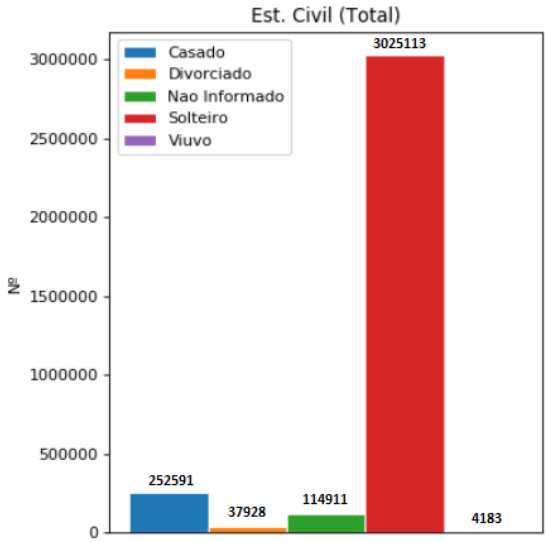
Onde “F” é Feminino e “M” corresponde a masculino.

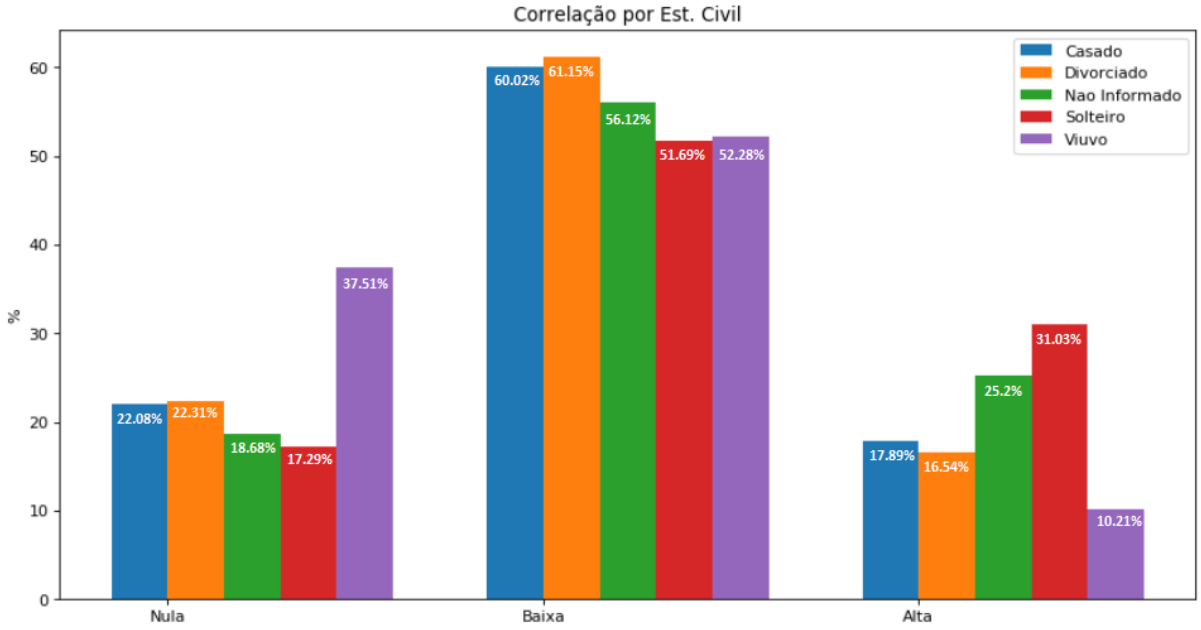
4.3.2. Por região



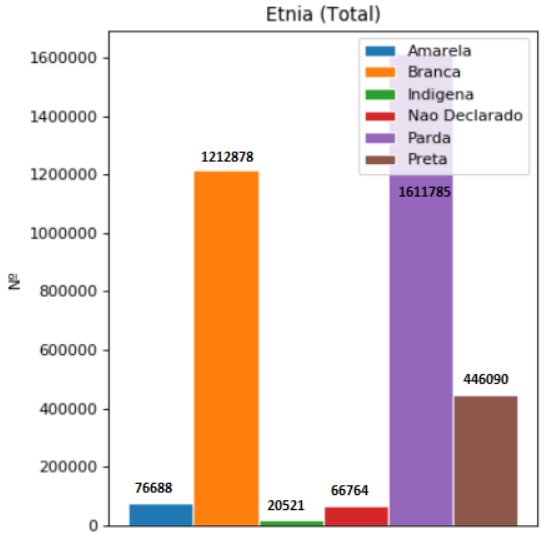


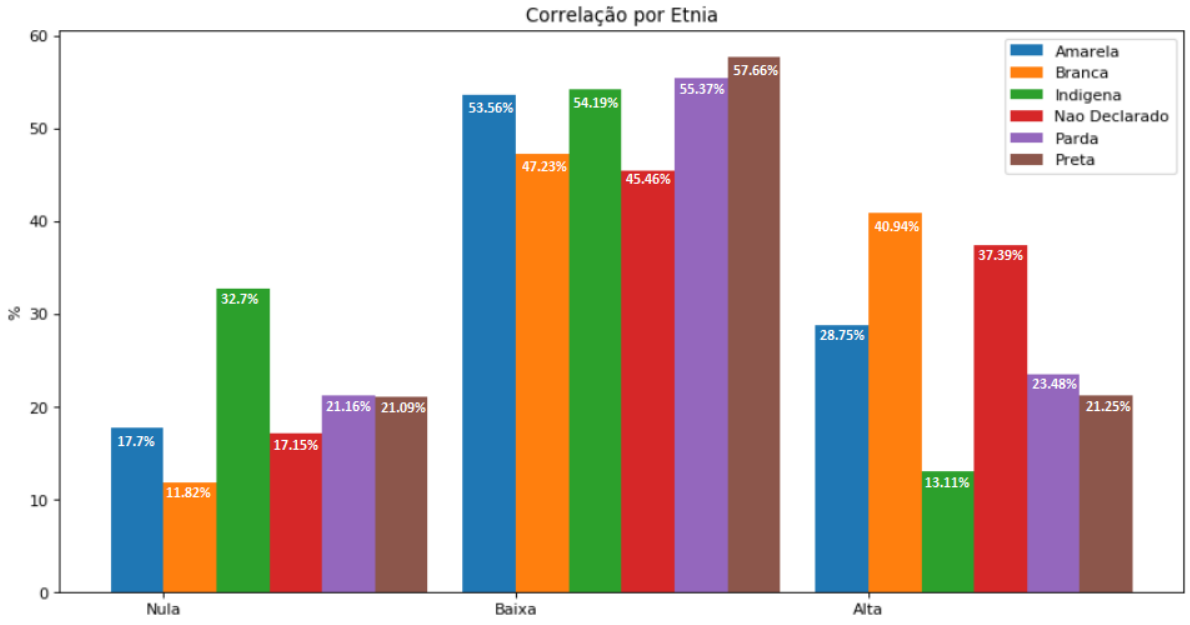
4.3.3. Por estado civil



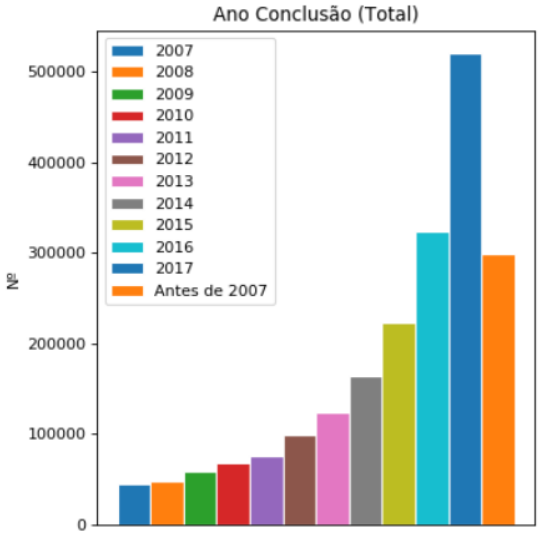


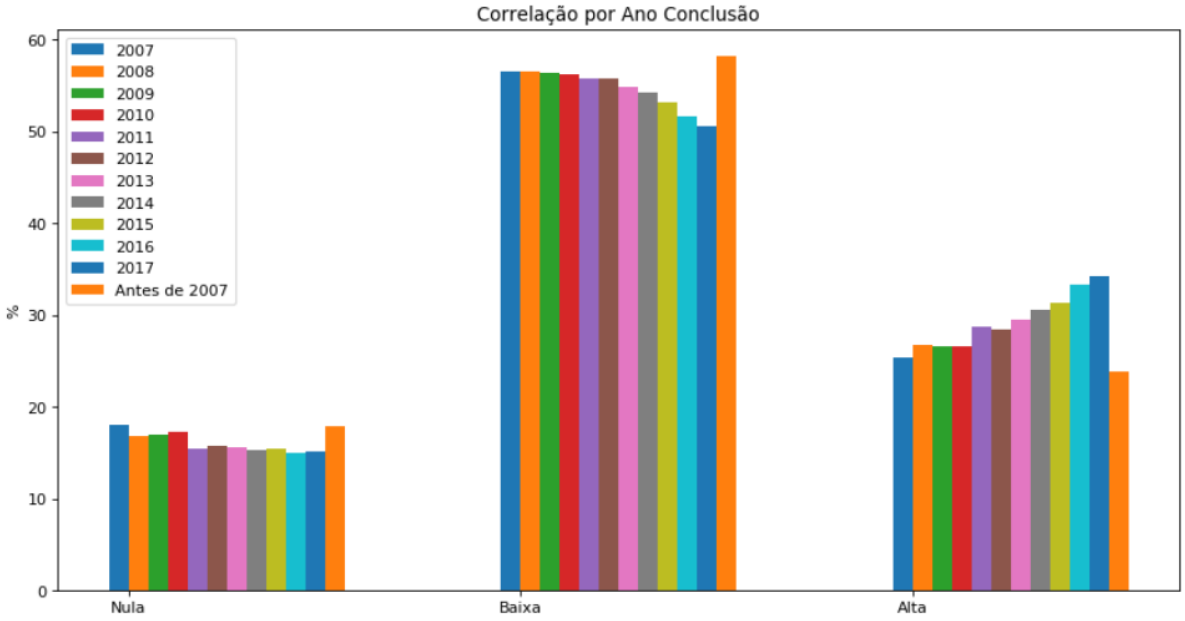
4.3.4. Por etnia



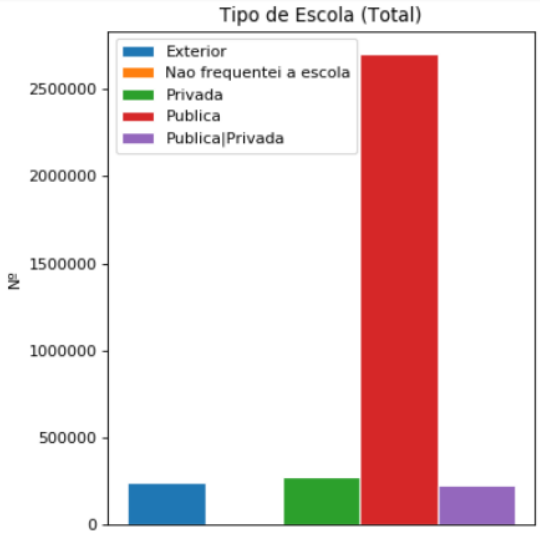


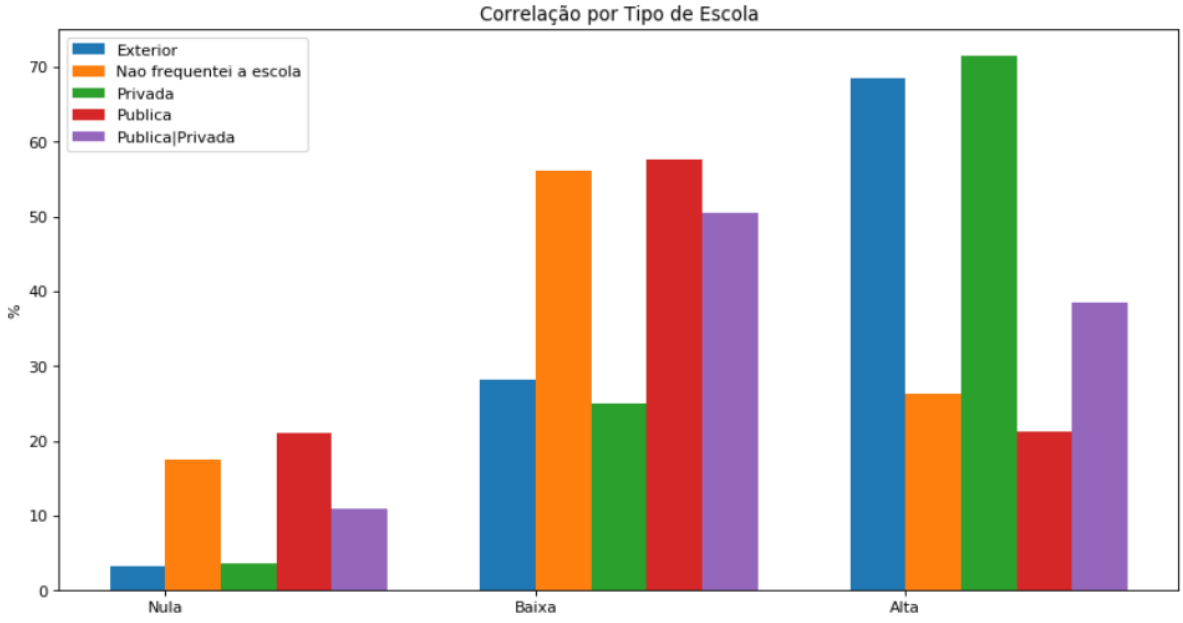
4.3.5. Por ano de conclusão



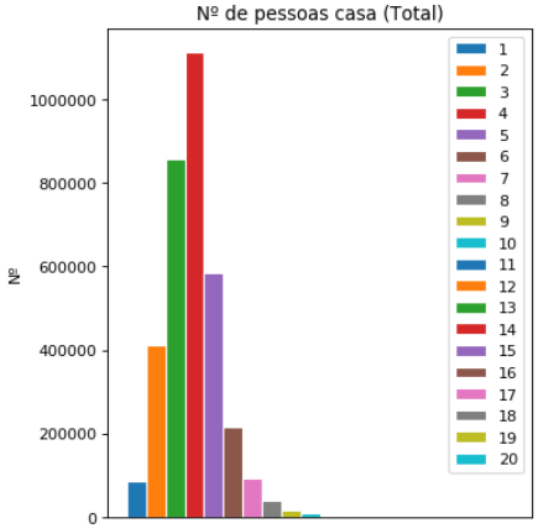


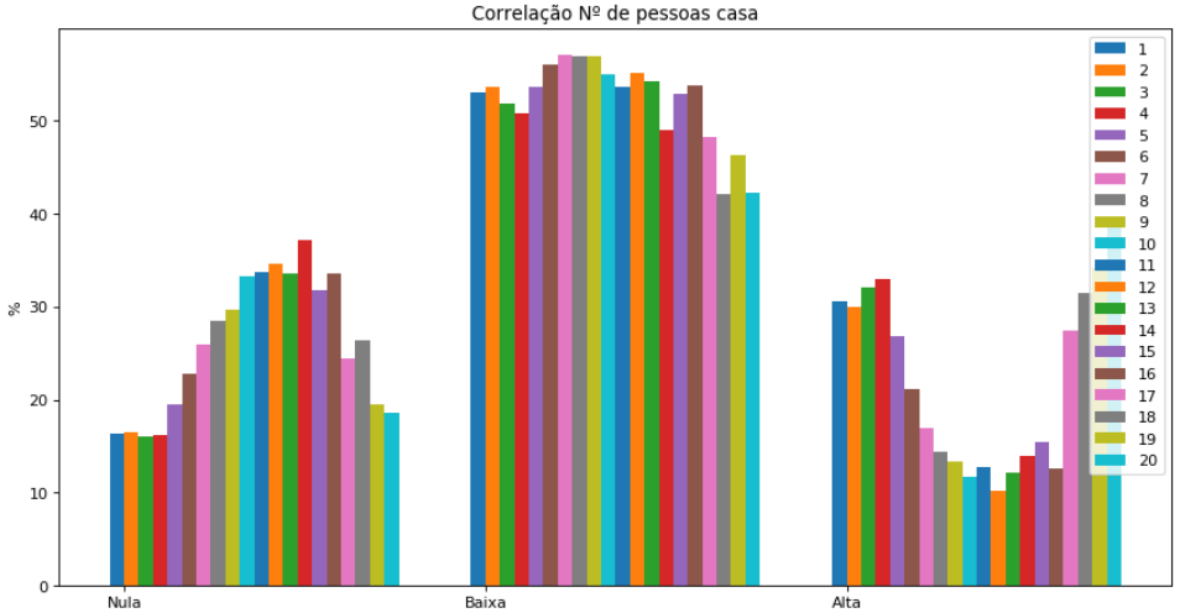
4.3.6. Por tipo de escola



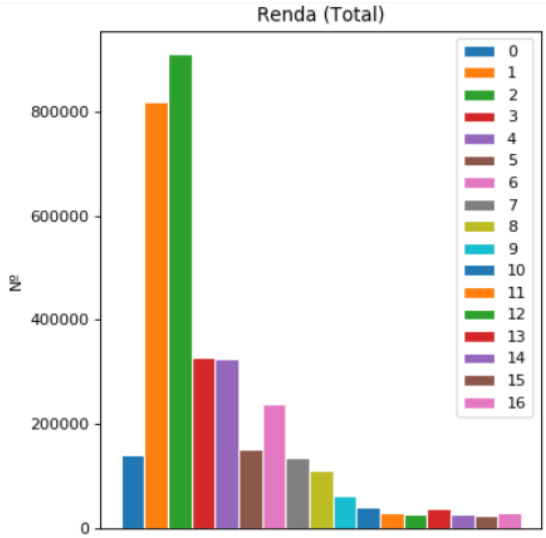


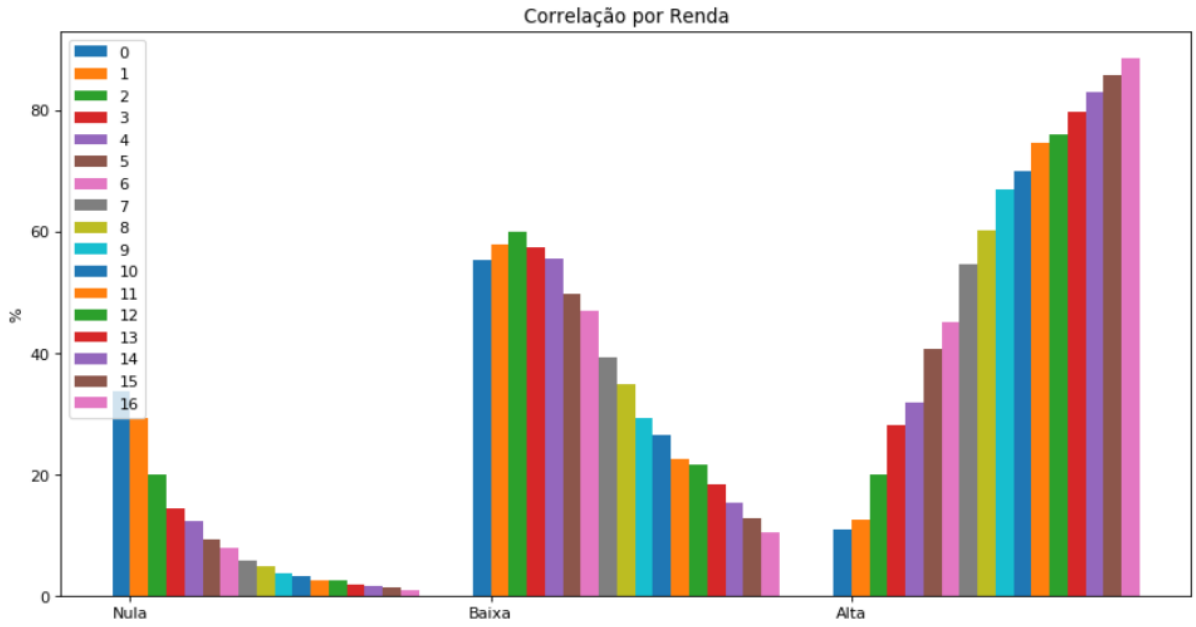
4.3.7. Por número de pessoas na residência





4.3.8. Por renda familiar





Onde está comparando o número de salários mínimos: “0” corresponde a menos que um salário mínimo, “1” corresponde a um salário mínimo e assim sucessivamente.

5. Criação de Modelos de *Machine Learning*

A partir da tabela resultante da camada 3 (*Specific View Data*) do capítulo 3, foi criado um modelo de predição, com o intuito de evidenciar a correlação entre o perfil socioeconômico dos inscritos.

Este estudo foi desenvolvido utilizando a linguagem *python* e a biblioteca *sklearn* para desenvolvimento dos modelos de predição.

No primeiro momento, para criar o modelo, foi utilizado o algoritmo de classificação *Random Forest, com n_estimators=100*, considerando como variável resposta a coluna "*Prob_Total*", que indica se o inscrito tem probabilidade ALTA, BAIXA ou NULA de acesso ao ensino público gratuito.

Também foi utilizado uma amostra de 20% da base total (409.772 registros) para treino do modelo, dividindo-a em 80% para base de treino e 20% para base de teste.

Rodando o algoritmo, foi obtido os seguintes valores de *F1_score* para a base de treino e para a base de teste:

F1_score	
Treino	0.88
Teste	0.56

Com o intuito de aprimorar o modelo, foi considerado que o inscrito com probabilidade BAIXA ou NULA, não teria acesso ao ensino superior gratuito, ou seja, apenas o inscrito com probabilidade ALTA teria acesso.

Com esta consideração, o modelo foi novamente treinado com duas possibilidades de classificação: "1" correspondendo ao inscrito que tem a probabilidade ALTA de ter acesso e "0" parra o inscrito com probabilidade NULA ou BAIXA.

Neste caso, o modelo obtido teve os seguintes resultados de *F1_score*:

F1_score	
Treino	0.92
Teste	0.72

Obtendo, dessa forma, um melhor resultado na predição dos inscritos que terão ou não acesso ao ensino superior gratuito.

6. Apresentação dos Resultados

Uma vez que obtivemos um modelo com mais de 0.9 de *f1_score*, é possível afirmar que temos um modelo com alto índice de acerto, mesmo para classes desproporcionais. Com isso, podemos fazer algumas inferências a partir dos padrões que o modelo encontrou:

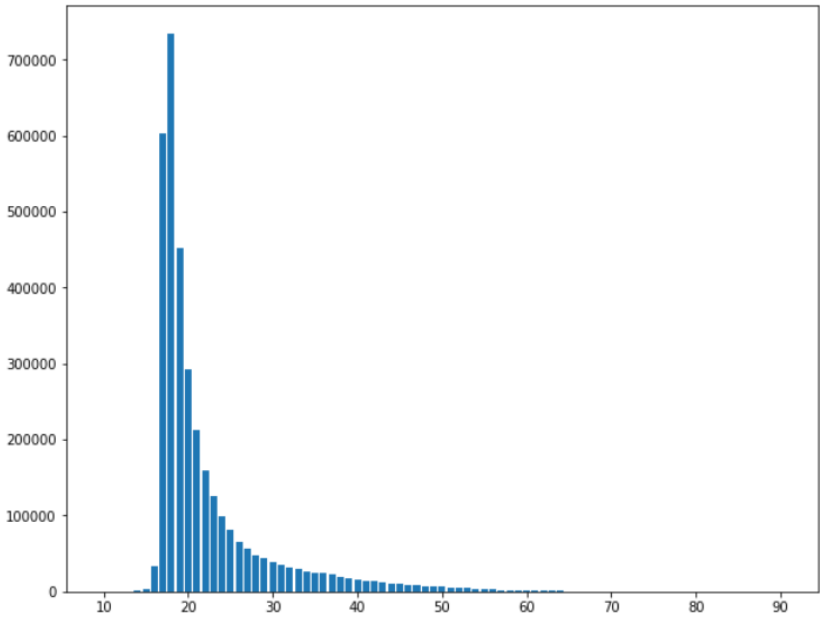
6.1. Correlação de Variáveis

Segue abaixo a correlação das variáveis de treino com a variável resposta, ordenadas de forma decrescente:

RENDA	0.415066
TP_ESCOLA	0.367819
TP_COR_RACA	0.187287
NU_IDADE	0.174051
REGIAO	0.171571
TP_ESTADO_CIVIL	0.102599
Q005	0.093108
TP_SEXO	0.083999
TP_ANO_CONCLUIU	0.081343

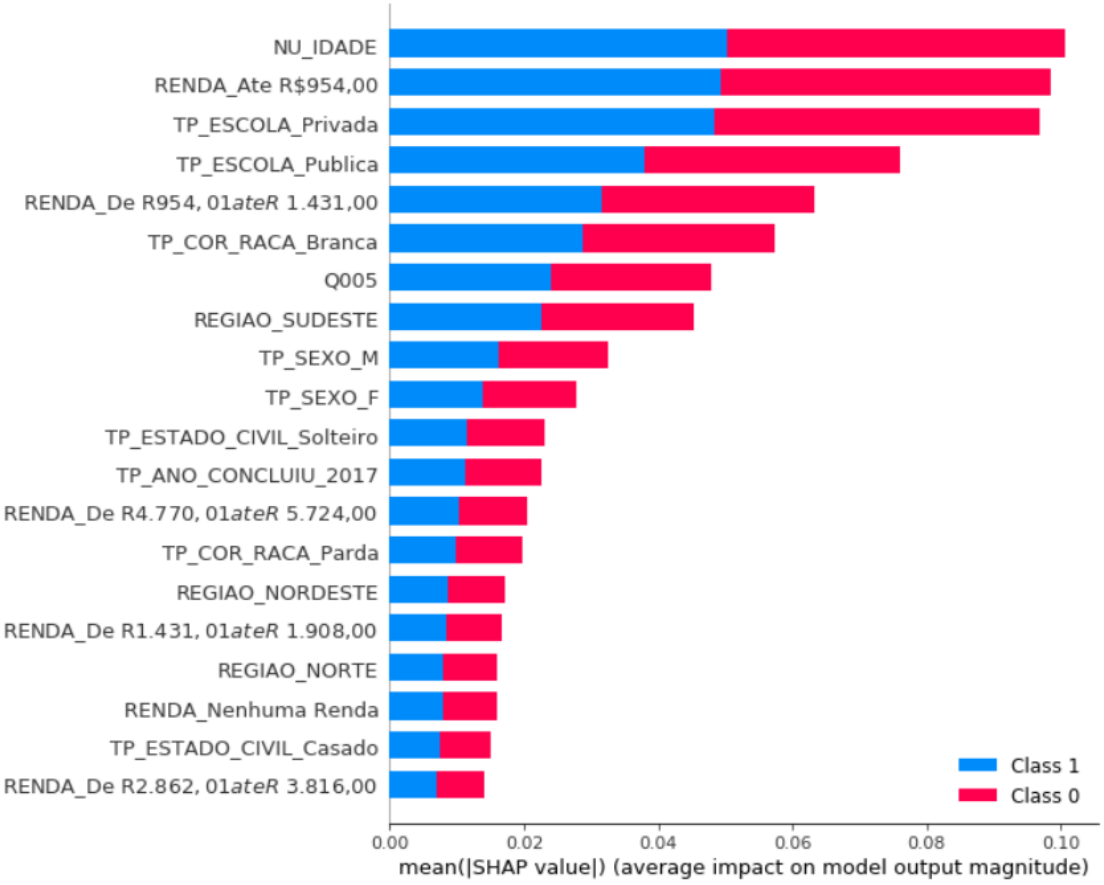
Analisando a correlação das variáveis acima, é possível afirmar que a renda familiar e o tipo de escola em que estudou são as variáveis que mais estão correlacionadas com o fato de o inscrito ter acesso ou não ao ensino superior. Também temos uma importante análise para a correlação da variável *TP_COR_RACA* e *REGIAO*, o que indica que a cor da pele e a região do Brasil onde o inscrito reside, também estão correlacionadas com a probabilidade de acesso ao ensino superior. No próximo tópico, irei mostrar uma análise mais detalhada, onde iremos observar que alguns valores destas variáveis importam mais para a predição do modelo.

Quanto a idade do inscrito, também temos uma importante correlação com a variável resposta. Porém ao analisarmos a base, temos a partir da função *describe* da biblioteca pandas (*python*), que 75% da base tem até 27 anos, enquanto nos outros 25% restante estão os inscritos com idade maior que 27 anos até 90 anos. Abaixo segue a distribuição das idades:



6.2. Importância do perfil do inscrito no resultado do modelo

Para obter o gráfico abaixo, foi utilizado a biblioteca *SHAP* para *python*, que tem como finalidade trazer alguns indicadores que temos em nosso modelo, como a funcionalidade utilizada neste estudo, por exemplo, que indica a influência de cada variáveis utilizadas no treino para a predição do resultado:



O gráfico acima nos diz as variáveis que mais são importantes para poder classificar um inscrito que terá (1) ou não (0) acesso ao ensino superior gratuito.

Como podemos observar e comparar com o tópico anterior, temos bastante relação com a correlação das variáveis.

Sabemos que a idade do inscrito tem algumas ressalvas, como citado anteriormente, por ter uma distribuição desbalanceada. Já as outras variáveis demonstradas no gráfico, nos diz a importância em que a renda familiar do inscrito tem na predição do modelo, uma vez que a partir desta variável podemos inferir o tipo de escola que o inscrito estuda, também correlacionando com a renda da família.

Outra variável que vale ressaltar é a *TP_COR_RACA* que aparece duas vezes no gráfico acima com o valor “Branca” e “Parda”.

7. Conclusão

Este estudo teve como objetivo realizar uma análise sobre como o perfil socioeconômico dos inscritos pode influenciar no acesso ao ensino superior gratuito. Entende-se que as análises realizadas aqui podem e devem ser aprimoradas e

estudadas com mais profundidade para que se tenha melhores resultados e que possam auxiliar outros estudos, inclusive de cunho social.

Outro ponto que vale ressaltar, são que as bases e os dados utilizados aqui foram tirados de uma fonte pública, onde qualquer pessoa pode inseri-las. Sendo assim, não podemos afirmar que os dados são oficiais, muito menos verdadeiros. Apesar disso, o objetivo também deste trabalho é demonstrar que, uma vez com uma base oficial, é possível replicar a metodologia aplicada e dessa forma, obter dados mais confiáveis para base de qualquer outro estudo. E por último, é também importante ressaltar que o modelo preditivo desenvolvido neste estudo não tem o objetivo de predizer ao inscrito de um próximo ENEM se ele irá ou não ter acesso ao ensino superior gratuito, mas sim, tem o objetivo de encontrar e evidenciar quais características de um perfil socioeconômico tem de importância no acesso ao ensino superior gratuito.

Mesmo assim, considerando as afirmações anteriores, foi possível desenvolver um estudo e também um modelo preditivo, evidenciando que o perfil socioeconômico do inscrito tem sim grande relação com o acesso ou não ao ensino superior gratuito.

Com isso, é entendido que este estudo foi uma análise superficial da relação do perfil socioeconômico com o acesso ao ensino superior gratuito, e com base nele, podemos citar possíveis trabalhos futuros que podem ser desenvolvidos, como por exemplo: uma análise mais detalhada de cada um dos tópicos da análise exploratória e o cruzamento dessas informações; a aplicação deste mesmo estudo utilizando o ENEM de outros anos, para identificar se este padrão também é identificado no todo; o uso de bases oficiais, a fim de evidenciar o próprio resultado que serviria de base para estudos sociais, com o intuito de melhorar o acesso ao ensino superior gratuito.

8. Links

Jupyter notebook com o trabalho desenvolvido:

https://github.com/tomazmatheus/pucminas-tcc/blob/master/notebooks/TCC_Enem_Prouni.ipynb

Base de dados ENEM 2018:

https://www.kaggle.com/ffmenezes/microdados-enem-2018 (Kaggle)

https://drive.google.com/file/d/1pNZKbv_3fh6HUJV28CWDGNVuZYMyfBQB/view?usp=sharing (Repositório pessoal)

Base de dados Notas de corte PROUNI 2018:

https://brasil.io/dataset/cursos-prouni/cursos/ (Brasil IO)

https://drive.google.com/file/d/1fiG443cxmsBBBiMFaJLWjTTWSu6lyL4l/view?usp=sharing (Repositório Pessoal)

Base de dados Cursos por área de conhecimento:

https://drive.google.com/file/d/1lCaLFtZpmzJAKp93Mfh23ayiHMjuK8qc/view?usp=sharing (Repositório Pessoal)

Metadados das bases:

https://drive.google.com/file/d/17HJPtZpqFEYj47-_puAaD39szL9pPXLD/view (Repositório pessoal)

Base de dados final para treino:

https://drive.google.com/file/d/1tnZdqrFVl-9u8OuDoKhREQk_QQ7GR1jF/view (Repositório pessoal)

Apresentação do Projeto no YouTube:

https://www.youtube.com/watch?v=1LCCw39u7Dc&feature

REFERÊNCIAS

ARRUDA, Daniel Péricles; VIDAL, Ricardo Flores. ProUni: sobre o direito de acesso e permanência e studantil. **Educação Online**, v. 15, n. 33, p. 1-25, 2020.

DE CASTRO, Maria Helena Guimarães; TIEZZI, Sergio. A reforma do ensino médio e a implantação d o Enem no Brasil. **Desafios**, v. 65, n. 11, p. 46-115, 2004.

SILVEIRA, Fernando Lang da; BARBOSA, Marcia Cristina Bernardes; SILVA, Roberto da. Exame Nac ional do Ensino Médio (ENEM): uma análise crítica. **Revista Brasileira de Ensino de Física**, v. 37, n. 1, p. 1101, 2015.

TRAVITZKI, Rodrigo; CALERO, Jorge; BOTO, Carlota. What does the National High School Exam (E NEM) tell Brazilian society?. **Cepal Review**, 2014.

TRAVITZKI, Rodrigo; FERRÃO, Maria Eugénia; COUTO, Alcino Pinto. Desigualdades educacionais e socioeconômicas na populaçao brasileira pré-universitária: Uma visão a partir da análise de dados do ENEM. **Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas**, v. 24, p . 1-32, 2016.