Modelo preditivo de classificação de porte de escola

ENEM por escola

02/06/2020

Pós-graduação em Análise em Big Data

Nome do Aluno:
Kaio Henrique Pedroza Silva

Coordenadores:
Profª Drª Alessandra de Álvila Montini
Profª Dr. Adolpho Walter Pimazoni Canton

Agenda

1. Objetivo do Trabalho
2. Contextualização do Problema
3. Base de Dados
    i. Bases originais
    ii. Filtros
    iii. Principais variáveis
4. Análise Exploratória de Dados

1. Objetivo do Trabalho

O objetivo do trabalho é classificar, por meio das variáveis qualitativas e quantitativas da base de dados, o porte das escolas da região Sudeste do Brasil que participaram da prova do ENEM de 2009 a 2015.

A predição será realizada utilizando dados históricos de desempenho das escolas no ENEM, modelos estatísticos e algoritmos de Inteligência Artificial, que selecionarão as características mais relevantes para explicar o porte de uma escola, seja pública ou privada.

Desta forma, o aluno poderá escolher o porte de uma escola que melhor se adequa às suas necessidades de maneira mais assertiva.

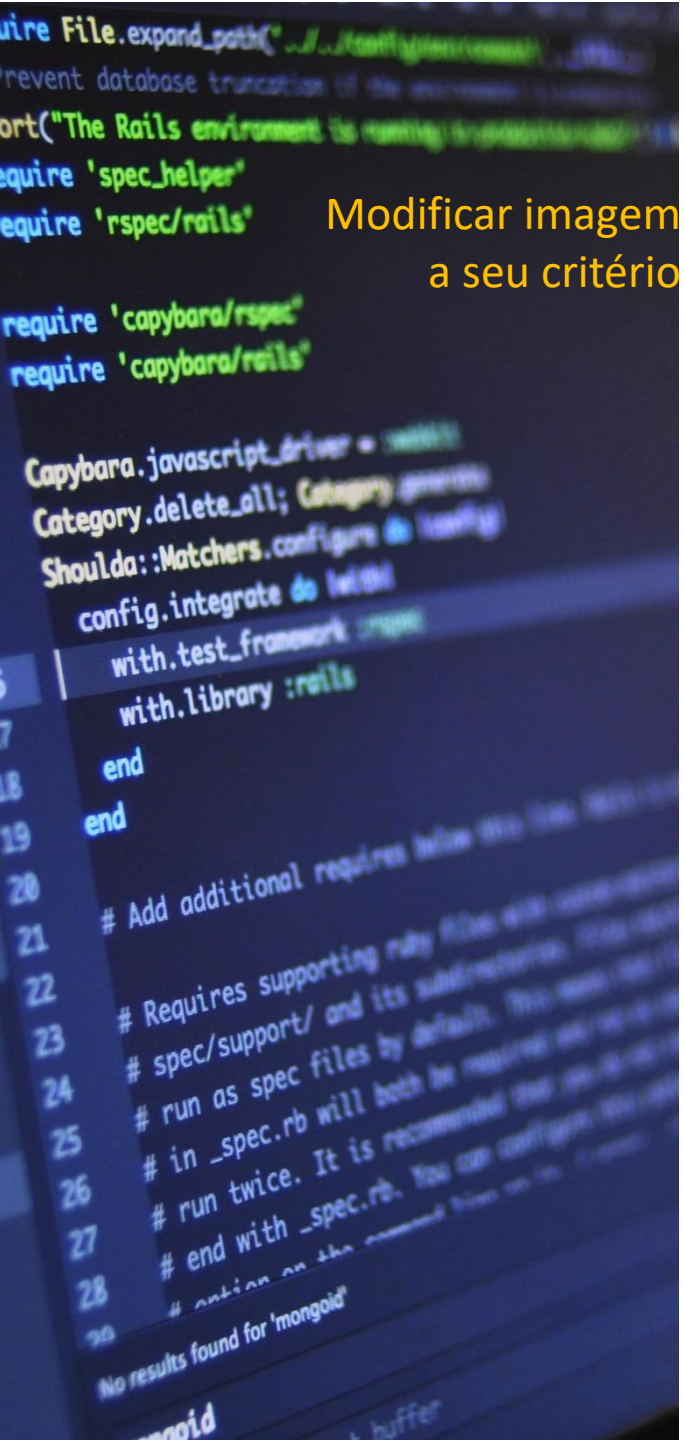
2. Contextualização do Problema

Escolher uma escola com um ensino de qualidade é fundamental para se ter um bom desempenho acadêmico, por meio de materiais eficientes e educadores que possam transmitir o conhecimento de forma que o aluno consiga absorvê-lo.

Através de variáveis como notas gerais no exame, taxa de participação e taxa de aprovação/reprovação, o intuito do modelo é classificar o porte das escolas em: mais de 90 alunos, de 61 a 90 alunos, de 31 a 60 alunos e de 1 a 30 alunos.

A classificação poderia gerar insights como: será que escolas maiores têm, de fato, maior desempenho no exame? Ou será que uma escola menor tem o melhor rendimento?

3. Bases de Dados



- A base de dados é encontrada no site http://inep.gov.br/microdados

- Única base disponível sobre o rendimento das escolas no ENEM

- O programa “ENEM por Escola” foi descontinuado a partir do ano de 2015, por isso não temos dados dos anos seguintes

3.i. Base Original

172035 registros

Visão da base
- Abrange os anos de 2005 até 2015
- Possui 27 variáveis (5 variáveis qualitativas e 22 variáveis quantitativas)

Filtros de inclusão
- Escolas do Sudeste do Brasil
- Ano de edição a partir de 2009

Preenchimento de NAs
- Preenchimento dos NA's das variáveis qualitativas com "Não informado"
- Preenchimento dos NA's das variáveis quantitativas com a mediana

Importante
- Nota média objetiva: calculada somente no ano de 2008
- Nota média total: calculada somente nos anos de 2005 até 2007

3.ii. Filtros

Base Original
172.035 observações

Base de partida

Base original com filtros de negócio com 172.035 observações de escolas, sendo que 45% são do Sudeste.

Filtro 1 – Ano de edição a partir de 2009
104.687 observações

Filtragem dos anos

Trazer as informações mais completas possíveis e redução de variáveis inutilizadas.

Filtro 2 – Escola do Sudeste brasileiro
77.533 observações

Filtragem de UFs

Diminuição do escopo do projeto, focando no Sudeste do Brasil.

Base Final
46.243 observações

Base final

Base final pronta para análise e futura modelagem.

3.iii. Principais variáveis

Variáveis da prova

- ano_edicao
- taxa_participacao
- nota_media_ciencias_natureza
- nota_media_ciencias_humanas
- nota_media_linguagens_codigos
- nota_media_matematica
- nota_media_redacao
- nota_media_objetiva
- nota_media_total
- taxa_aprovacao
- taxa_reprovacao
- taxa_abandono

Variáveis da escola

- cod_uf_escola
- sigla_uf_escola
- cod_municipio_escola
- nome_municipio_escola
- cod_escola
- nome_escola
- tipo_dependencia
- tipo_localizacao_escola
- numero_matriculas
- indicador_socio_economico_escola
- indicador_adequacao_escola
- indicador_permanencia_escola
- porte_escola

4. Análise Exploratória de Dados

A análise exploratória dos dados se baseou em:

1) Análise dos tipos de todas as variáveis

2) Contagem, média, desvio padrão, mínimo, máximo, mediana e quartis

3) Contagem de NA's

4) Análise da presença de outliers

4. Análise Exploratória de Dados

Tipos das variáveis

Object:
- sigla_uf_escola
- nome_municipio_escola
- nome_escola
- indicador_socio_economico_escola
- porte_escola

Int64:
- ano_edicao
- cod_uf_escola
- cod_municipio_escola
- cod_escola
- tipo_dependencia
- tipo_localizacao_escola
- numero_matriculas
- numero_participantes

Float64:
- numero_participantes_especiais
- taxa_participacao
- nota_media_ciencias_natureza
- nota_media_ciencias_humanas
- nota_media_linguagens_codigos
- nota_media_matematica
- nota_media_redacao
- indicador_adequacao_escola
- indicador_permanencia_escola
- taxa_aprovacao
- taxa_reprovacao
- taxa_abandono

4. Análise Exploratória de Dados

Describe das variáveis

	numero_matriculas	numero_participantes_especiais	numero_participantes	taxa_participacao	nota_media_ciencias_natureza	nota_media_ciencias_humanas	nota_media_linguagens_codigos	nota_media_matematica	nota_media_redacao	indicador_adequacao_escola	indicador_permanencia_escola	taxa_aprovacao	taxa_reprovacao	taxa_abandono
count	46243	19685	46243	46243	46243	46243	46243	46243	46080	19675	13430	45803	45803	45803
mean	86,86761	0.29398	51,88342	65,10767	504,90407	544,96059	525,70507	535,36639	573,70289	63,92313	77,97881	87,07303	9,29281	3,63415
std	84,99862	0.89626	53,4937	23,63627	53,58149	55,65068	45,91963	75,68077	75,73661	14,78753	52,49141	11,26883	8,30386	5,73012
min	0.00000	0.00000	10	3,8	372,31	368,1	363,34	370,48	194,55	0.00000	0.00000	0.00000	0.00000	0.00000
25%	29	0.00000	19	51,3	464,025	504,98	492,1	476,26	522,61	55,1	70,695	80,7	3	0.00000
50%	59	0.00000	34	66,7	490,31	540,78	519,14	514,65	574,44	65	81,02	90,1	7,1	0.50000
75%	115	0.00000	64	85,29	540,98	584,685	560,145	588,14	622,925	74,4	89,29	96	13,4	5,7
max	1038	27	670	100	755,16	758,04	712,35	873,65	930	100	5822	100	100	100

4. Análise Exploratória de Dados

NA's das variáveis

coluna	contagem	porcentagem
indicador_permanencia_escola	32813	71%
indicador_adequacao_escola	26568	57%
numero_participantes_especiais	26558	57%
taxa_abandono	440	0.95%
taxa_reprovacao	440	0.95%
taxa_reprovacao	440	0.95%
nota_media_redacao	163	0.3%

4. Análise Exploratória de Dados

Analisando a presença de outliers na base

coluna	contagem	porcentagem
tipo_localizacao_escola	1185	2.56%
numero_participantes	989	2.14%
numero_matriculas	917	1.98%
taxa_abandono	824	1.78%
taxa_reprovacao	629	1.36%
taxa_aprovacao	504	1.09%
numero_participantes_especiais	486	1.05%
nota_media_ciencias_natureza	227	0.49%
nota_media_redacao	215	0.46%
nota_media_matematica	201	0.43%
indicador_adequacao_escola	120	0.26%
nota_media_linguagens_codigos	45	0.1%
nota_media_ciencias_humanas	33	0.07%
indicador_permanencia_escola	1	0.002%