Modelo preditivo de classificação de ingresso em faculdade

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. Nota média geral e target
    iv. Principais variáveis
    v. Balanceamento
4. Análise Exploratória de Dados
5. Modelagem com Estatística Tradicional
6. Modelagem com Inteligência Artificial
7. Conclusões

1. Objetivo do Trabalho

O objetivo do trabalho é classificar, por meio das variáveis qualitativas e quantitativas da base de dados, se uma escola da região Sudeste do Brasil e participante da prova do ENEM entre 2009 e 2015 tem uma média suficiente para o ingresso de um aluno que deseja cursar Ciências da Computação.

A predição será realizada utilizando dados históricos de desempenho das escolas públicas e privadas no ENEM, modelos estatísticos e algoritmos de Inteligência Artificial, que selecionarão as características mais relevantes para explicar o ingresso de um aluno (que deseja cursar Ciências da Computação) em uma faculdade.

Desta forma, o aluno poderá escolher uma escola que ofereça mais respaldo para o ingresso na faculdade.

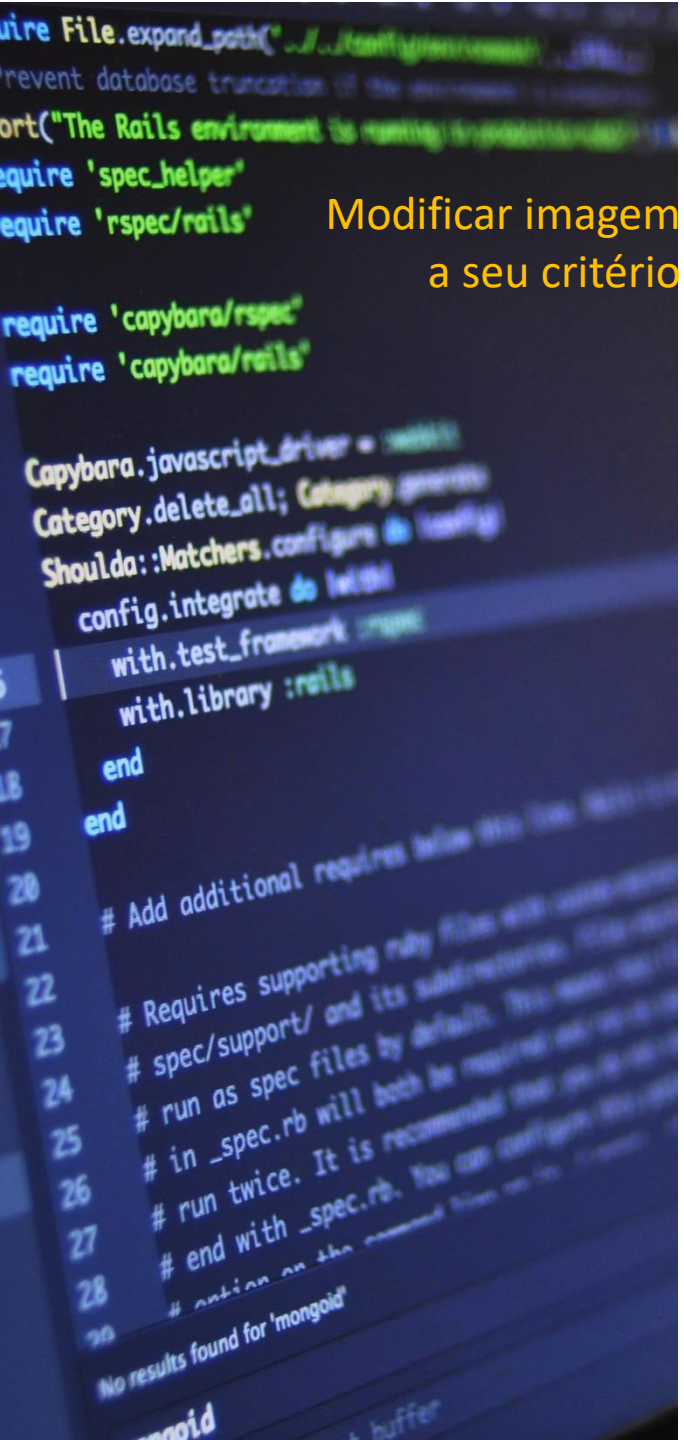
2. Contextualização do Problema

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. No caso do curso de Ciências da Computação pelo Sisu, é preciso ter uma nota mínima no ENEM de 576 pontos e uma nota superior a zero na redação. Porém, a nota de corte de redação para o modelo foi de 450.

Através de variáveis como notas gerais no exame, taxa de participação e taxa de aprovação/reprovação, o intuito do modelo é classificar se a nota média geral e da redação da escola possibilita o ingresso dos alunos em uma faculdade por meio Sisu.

A classificação poderia auxiliar na escolha de uma escola que ofereça mais respaldo para o ingresso na faculdade.

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

- Informação sobre a nota mínima para entrar em uma faculdade se encontra nos sites:
  - https://www.mundovestibular.com.br/enem/sisu/notas-de-corte-sisu-2020-2/#:~:text=Para%20ser%20selecionado%20na%20maioria,notas%20superiores%20a%20700%20pontos.
  - https://www.enemvirtual.com.br/notas-de-corte-sisu-2020-2/

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 de partida

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

Filtragem dos anos

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

Filtragem de UFs

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

Base final

Base final pronta para análise e futura modelagem.

3.iii. Criação da nota média geral e target

A criação da nota média geral, por se tratar de um curso de exatas, foi feita por meio de uma média ponderada.

media = (0.1*nota_media_ciencias_natureza) +
(0.1*nota_media_ciencias_humanas) +
(0.1*nota_media_linguagens_codigos) +
(0.7*nota_media_matematica)

target = 1 se media > 576 e nota_media_redação > 450 senão 0

3.iv. 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

3.v. Balanceamento

Como o resultado do target foi de 34072 como target=0 e 12171 como target=1, a base de dados foi balanceada com:

1) 13000 registros como target=0
2) 12171 registros como target=1

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, quartis e coeficiente de variação

3) Tabelas de frequências

4) Boxplots

5) Contagem de NA's

6) 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

Análise das variáveis quantitativas

COLUNA	MEDIA	MIN	25%	50%	75%	MAX	DESVIO PADRÃO	COEFICIENTE DE VARIAÇÃO
tipo_dependencia	2,85868	1	2	2	4	4	0,99889	0,34942
tipo_localizacao_escola	1,02563	1	1	1	1	2	0,15802	0,15407
numero_matriculas	86,86761	0	29	59	115	1038	84,99862	0,97848
numero_participantes_especiais	0,29398	0	0	0	0	27	0,89626	3,0487
numero_participantes	51,88342	10	19	34	64	670	53,4937	1,03104
taxa_participacao	65,10767	3,8	51,3	66,7	85,29	100	23,63627	0,36303
nota_media_ciencias_natureza	504,90407	372,31	464,025	490,31	540,98	755,16	53,58149	0,10612
nota_media_ciencias_humanas	544,96059	368,1	504,98	540,78	584,685	758,04	55,65068	0,10212
nota_media_linguagens_codigos	525,70507	363,34	492,1	519,14	560,145	712,35	45,91963	0,08735
nota_media_matematica	535,36639	370,48	476,26	514,65	588,14	873,65	75,68077	0,14136
nota_media_redacao	573,70289	194,55	522,61	574,44	622,925	930	75,73661	0,13201
indicador_adequacao_escola	63,92313	0	55,1	65	74,4	100	14,78753	0,23133
indicador_permanencia_escola	77,97881	0	70,695	81,02	89,29	5822	52,49141	0,67315
taxa_aprovacao	87,07303	0	80,7	90,1	96	100	11,26883	0,12942
taxa_reprovacao	9,29281	0	3	7,1	13,4	100	8,30386	0,89358
taxa_abandono	3,63415	0	0	0,5	5,7	100	5,73012	1,57674

4. Análise Exploratória de Dados

Tabela de frequência do indicador socio econômico de cada escola

INDICADOR_SOCIO_ECONOMICO_ESCOLA	QTTDE
Grupo 5	1957
Grupo 4	1935
Grupo 3	1857
Grupo 6	920
Grupo 2	148
Grupo 1	28

4. Análise Exploratória de Dados

Tabela de frequência nome do munícipio da escola (top 5)

NOME_MUNICIPIO_ESCOLA	QTTDE
São Paulo	4556
Rio de Janeiro	2882
Belo Horizonte	1289
Campinas	559
Guarulhos	520

4. Análise Exploratória de Dados

Tabela de frequência nome da escola (top 5)

NOME_ESCOLA	QTTDE
COLEGIO TIRADENTES PMMG	91
COLEGIO E CURSO PONTO DE ENSINO	72
SÃO JOSE COLEGIO	51
SISTEMA ELITE DE ENSINO	50
SEBRAE ESC TEC DE FORM GERENCIAL	46

4. Análise Exploratória de Dados

Tabela de frequência porte da escola

PORTE_ESCOLA	QTTDE
Maior que 90 alunos	15369
De 1 a 30 alunos	12204
De 31 a 60 alunos	11407
De 61 a 90 alunos	7263

4. Análise Exploratória de Dados

Tabela de frequência UF

SIGLA_UF_ESCOLA	QTTDE
SP	23004
MG	12202
RJ	8401
ES	2636

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%

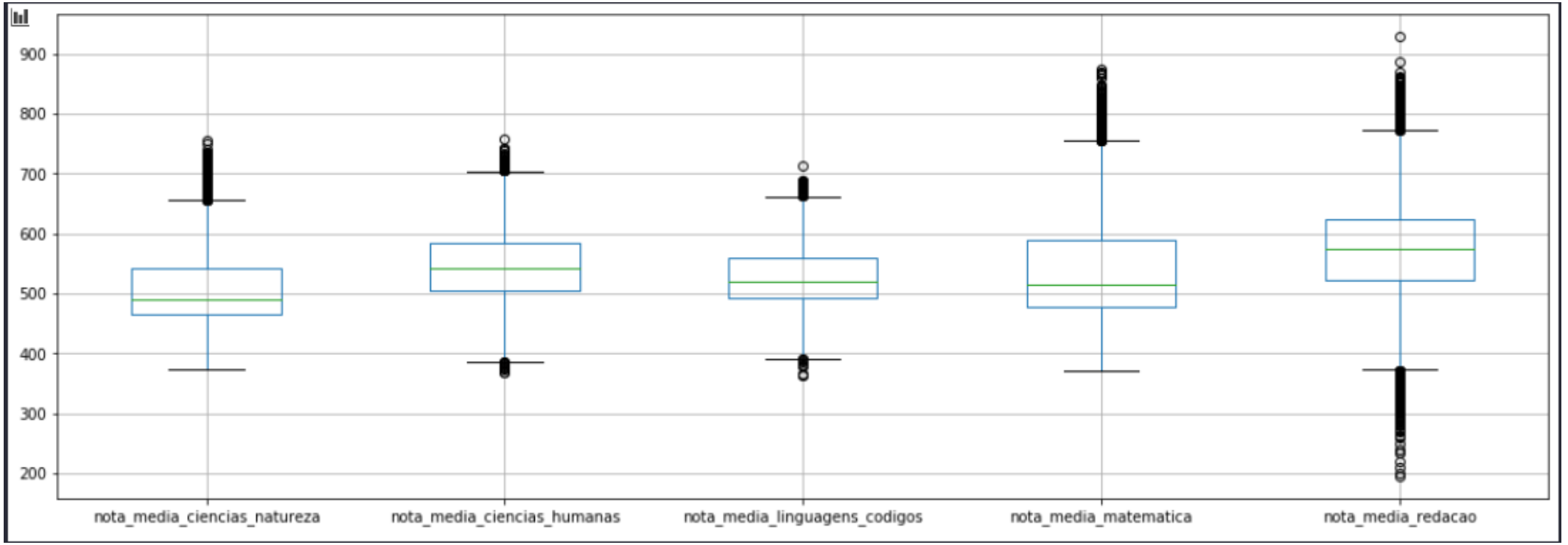
4. Análise Exploratória de Dados

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%

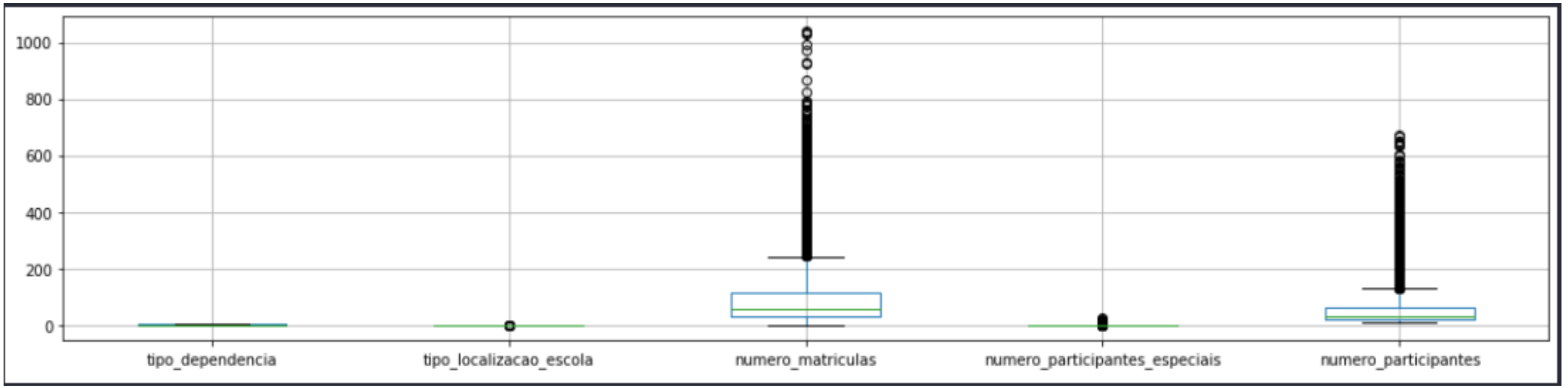
4. Análise Exploratória de Dados

Boxplot das notas de cada escola



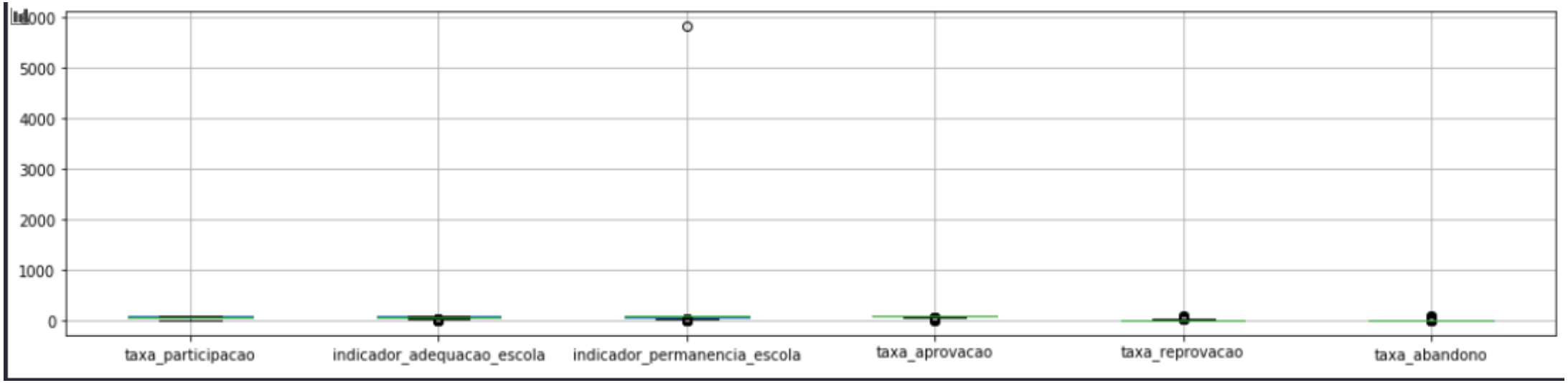
4. Análise Exploratória de Dados

Boxplot das informações da escola



4. Análise Exploratória de Dados

Boxplot das taxas



Foi escolhido como nota de corte média de 500 (exatamente a média de uma nota máxima), na parte da divisão de treino e teste, foi feito a divisão da seguinte forma:

1) Divisão da base em 2: acima da média e abaixo da média
2) 70% acima e abaixo da média foi reservada como treino
3) 30% acima e abaixo da média foi reservada como teste

Filtragem das features

Variáveis	Chi-2	RFE	RF	Total
nota_media_matematica	True	True	True	3
nota_media_geral	True	True	True	3
tipo_dependencia	False	True	True	2
nota_media_linguagens_codigos	True	False	True	2
nota_media_ciencias_natureza	True	False	True	2
nota_media_ciencias_humanas	True	False	True	2
cod_escola	True	False	True	2
taxa_abandono	True	False	False	1
numero_matriculas	True	False	False	1
nota_media_redacao	True	False	False	1
cod_uf_escola	False	True	False	1
cod_municipio_escola	True	False	False	1
ano_edicao	False	True	False	1

Observação: foram retiradas as colunas de nota por conta da correlação alta entre elas e o target.

Correlação entre as notas e o target

	nota_media_ciencias_natureza	nota_media_ciencias_humanas	nota_media_linguagens_codigos	nota_media_matematica	nota_media_redacao	target
nota_media_ciencias_natureza	1	0.84	0.89	0.87	0.76	0.77
nota_media_ciencias_humanas	0.84	1	0.84	0.74	0.66	0.66
nota_media_linguagens_codigos	0.89	0.84	1	0.85	0.74	0.73
nota_media_matematica	0.87	0.74	0.85	1	0.71	0.83
nota_media_redacao	0.76	0.66	0.74	0.71	1	0.6
target	0.77	0.66	0.73	0.83	0.6	1

5. Modelagem com Estatística Tradicional

Modelo	Acurácia de treino	Acurácia de teste
Regressão logística	64.53%	45.9%
Árvore de decisão	77.11%	87.89%

Regressão: random state=123 e parâmetros default

Árvore de decisão: random state=123, max_depth=2 e splitter="random"

6. Modelagem com Inteligência Artificial

Modelo	Score
Random Forest	74%
Gradient Boosting	73.9%

2328	11
0	5208

Matriz de confusão: RF

2323	16
0	5208

Matriz de confusão: GB

Mesmo sendo uma diferença mínima no score, o melhor modelo é o RF, pois tem um acerto maior de tanto dos targets quanto do não-target.

7. Conclusões

1) Algumas features tiveram muita influência na modelagem

2) Os modelos podem ser melhorados com o tunning dos parâmetros

3) Volume relativamente notório de outliers

4) Grande taxa de NA's em algumas variáveis

5) Houve uma pequena diferença entre o resultado dos modelos de Inteligência Artificial

LABDATA FIA – Laboratório de Análise de Dados

Unidade Pinheiros

Unidade Paulista