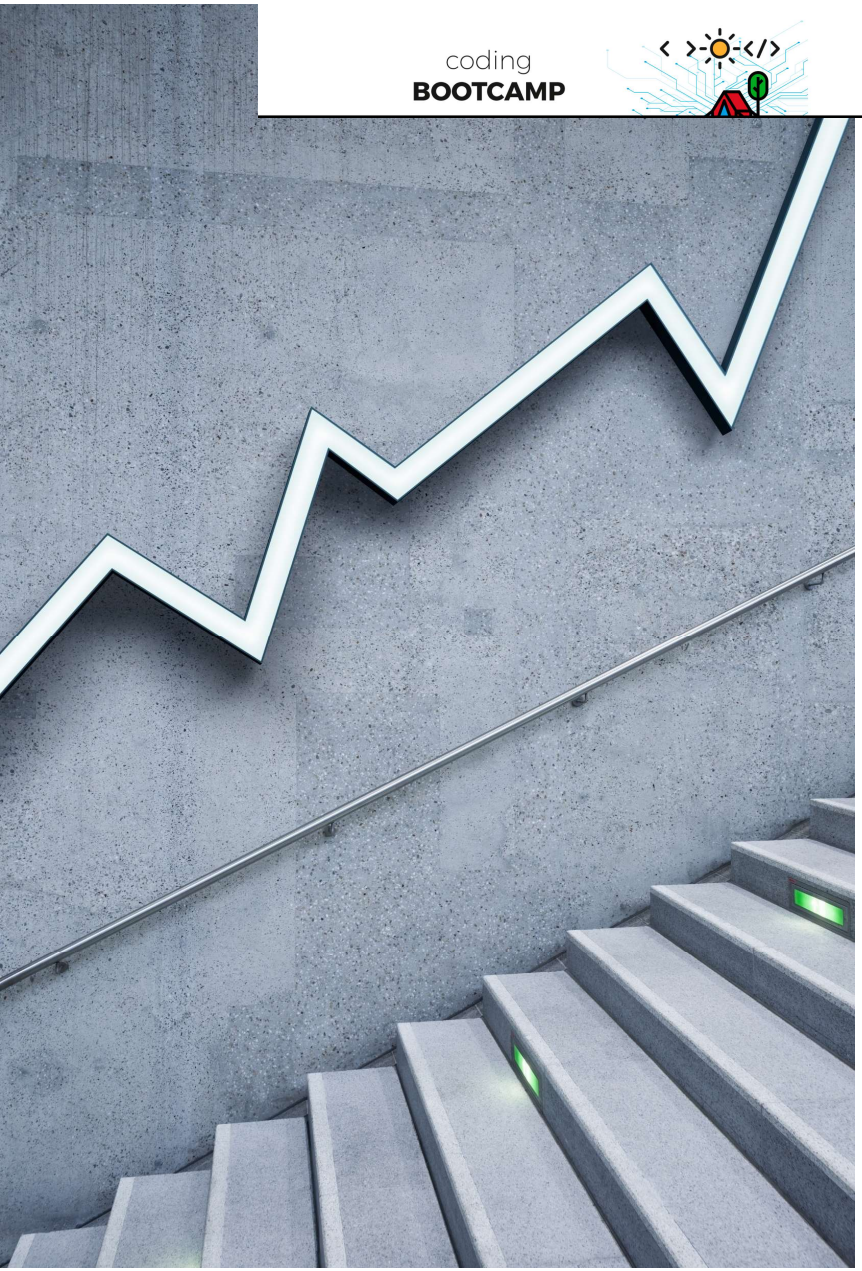


coding BOOTCAMP

Enap Escola Nacional de Administração Pública

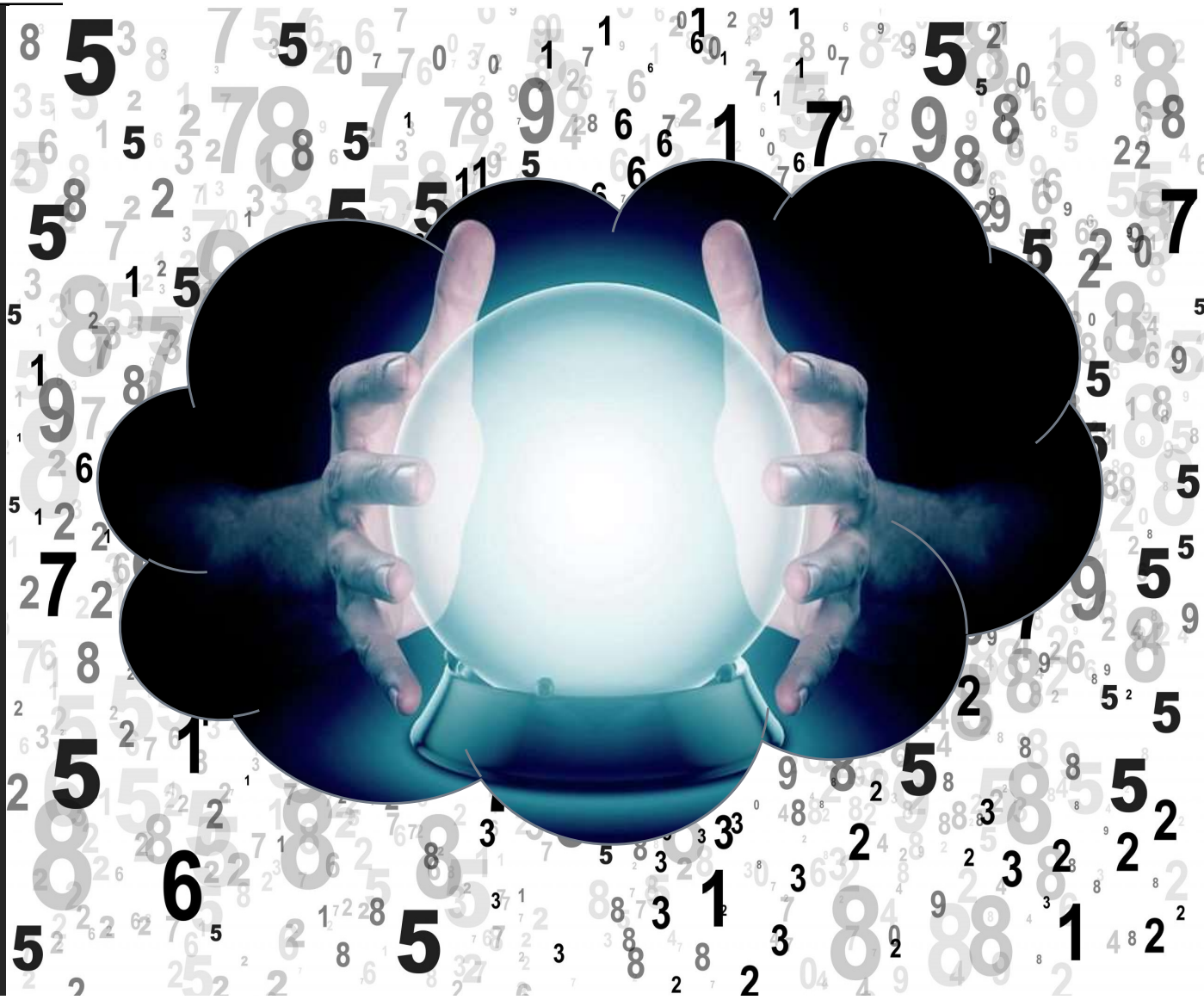
# Preditor de ingressantes em cursos superiores

MARCELO CARDOSO SILVA
COLÉGIO PEDRO II

ORIENTADORA
THACIANA CERQUEIRA

# problema

Prever a quantidade de ingressantes em um curso, a partir dos dados indicadores de fluxo da educação superior, com o objetivo de atingir a integralização com o maior número de concluintes possível.



coding BOOTCAMP

Machine Learning

Enap Escola Nacional de Administração Pública

# dataset

## Indicadores de Fluxo da Educação Superior

## Disponibilizados pelo Inep

https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-fluxo-da-educacao-superior

15/12/2021

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

**Indicadores de Trajetória de Curso de Graduação - Brasil - 2019**

Indicadores de Trajetória dos Alunos nos Cursos de Graduação da Educação Superior 2019 (*coorte* 2010), composto por: Taxa de Permanência, Taxa de Conclusão Acumulada, Taxa de Desistência Acumulada, Taxa de Conclusão Anual e Taxa de Desistência Anual, organizados por cursos e instituições de ensino.

| Código da Instituição | Categoria Administrativa | Código da Região Geográfica do Curso | Código da Unidade Federativa do Curso | Código do Município do Curso | Grau Acadêmico | Código da área do Curso segundo a classificação CINE BRASIL | Ano de Ingresso | Ano de Referência | Prazo de Integralização em Anos | Quantidade de Ingressantes no Curso | Indicadores de Trajetória | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | Taxa de Permanência - TAP | Taxa de Conclusão Acumulada - TCA | Taxa de Desistência Acumulada - TODA |
| 1 | 1 | 5 | 51 | 5103403 | 1 | 0421D01 | 2010 | 2010 | 6 | 107 | 91,6 | 0,9 | 7,5 |
| 1 | 1 | 5 | 51 | 5103403 | 1 | 0421D01 | 2010 | 2011 | 6 | 107 | 84,1 | 0,9 | 15,0 |
| 1 | 1 | 5 | 51 | 5107602 | 2 | 0115L13 | 2010 | 2018 | 5 | 47 | 0,0 | 38,3 | 61,7 |
| 1 | 1 | 5 | 51 | 5107602 | 2 | 0115L13 | 2010 | 2019 | 5 | 47 | 0,0 | 38,3 | 61,7 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2010 | 5 | 21 | 95,2 | 0,0 | 4,8 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2011 | 5 | 21 | 95,2 | 0,0 | 4,8 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2012 | 5 | 21 | 85,7 | 4,8 | 9,5 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2013 | 5 | 21 | 76,2 | 9,5 | 14,3 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2014 | 5 | 21 | 66,7 | 9,5 | 23,8 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2015 | 5 | 21 | 66,7 | 9,5 | 23,8 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2016 | 5 | 21 | 9,5 | 14,3 | 76,2 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2017 | 5 | 21 | 0,0 | 14,3 | 85,7 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2018 | 5 | 21 | 0,0 | 14,3 | 85,7 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2019 | 5 | 21 | 0,0 | 14,3 | 85,7 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L13 | 2010 | 2010 | 5 | 34 | 88,2 | 0,0 | 11,8 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L13 | 2010 | 2011 | 5 | 34 | 82,4 | 0,0 | 17,6 |
| 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L13 | 2010 | 2012 | 5 | 34 | 76,5 | 0,0 | 23,5 |

6 arquivos: 2010-2019, 2011-2019 ... 2015-2019

1 tabela no SQL Server com 1.199.470 registros

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

## Indicadores de Trajetória de Curso de Graduação - Brasil - 2019

Indicadores de Trajetória dos Alunos nos Cursos de Graduação da Educação Superior 2019 (*coorte* 2010), composto por: Taxa de Permanência, Taxa de Conclusão Acumulada, Taxa de Desistência Acumulada, Taxa de Conclusão Anual e Taxa de Desistência Anual, organizados por cursos e instituições de ensino.

| | A | C | G | H | I | J | L | P | Q | R | V | AA | AB | AC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | | | | | | | | | | | | Indicadores de Trajetória | | |
| 8 | Código da Instituição | Categoria Administrativa | Código da Região Geográfica do Curso | Código da Unidade Federativa do Curso | Código do Município do Curso | Grau Acadêmico | Código da área do Curso segundo a classificação CINE BRASIL | Ano de Ingresso | Ano de Referência | Prazo de Integralização em Anos | Quantidade de Ingressantes no Curso | Taxa de Permanência TAP | Taxa de Conclusão Acumulada - TCA | Taxa de Desistência Acumulada - TODA |
| 10 | 1 | 1 | 5 | 51 | 5103403 | 1 | 0421D01 | 2010 | 2010 | 6 | 107 | 91,6 | 0,9 | 7,5 |
| 11 | 1 | 1 | 5 | 51 | 5103403 | 1 | 0421D01 | 2010 | 2011 | 6 | 107 | 84,1 | 0,9 | 15,0 |
| 318 | 1 | 1 | 5 | 51 | 5107602 | 2 | 0115L13 | 2010 | 2018 | 5 | 47 | 0,0 | 38,3 | 61,7 |
| 319 | 1 | 1 | 5 | 51 | 5107602 | 2 | 0115L13 | 2010 | 2019 | 5 | 47 | 0,0 | 38,3 | 61,7 |
| 320 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2010 | 5 | 21 | 95,2 | 0,0 | 4,8 |
| 321 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2011 | 5 | 21 | 95,2 | 0,0 | 4,8 |
| 322 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2012 | 5 | 21 | 85,7 | 4,8 | 9,5 |
| 323 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2013 | 5 | 21 | 76,2 | 9,5 | 14,3 |
| 324 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2014 | 5 | 21 | 66,7 | 9,5 | 23,8 |
| 325 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2015 | 5 | 21 | 66,7 | 9,5 | 23,8 |
| 326 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2016 | 5 | 21 | 9,5 | 14,3 | 76,2 |
| 327 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2017 | 5 | 21 | 0,0 | 14,3 | 85,7 |
| 328 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2018 | 5 | 21 | 0,0 | 14,3 | 85,7 |
| 329 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L14 | 2010 | 2019 | 5 | 21 | 0,0 | 14,3 | 85,7 |
| 330 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L13 | 2010 | 2010 | 5 | 34 | 88,2 | 0,0 | 11,8 |
| 331 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L13 | 2010 | 2011 | 5 | 34 | 82,4 | 0,0 | 17,6 |
| 332 | 1 | 1 | 5 | 51 | 5103403 | 2 | 0115L13 | 2010 | 2012 | 5 | 34 | 76,5 | 0,0 | 23,5 |

coding BOOTCAMP

Machine Learning

Enap Escola Nacional de Administração Pública

# dataset

Filtro:

cursos presenciais: 76.010

**ores de Trajetória de Curso de Graduação - Brasil - 2019**

rior 2019 (*coorte* 2010), composto por: Taxa de Permanência, Taxa de Conclusão Acumulada, Taxa de Desistência Acumulada, Taxa de Conclusão Anual e Taxa de

| Ano de Ingresso | Ano de Referência | Prazo de Integralização em Anos | Quantidade de Ingressantes no Curso | Quantidade de Permanência no Curso no ano de referência | Quantidade de Concluintes no Curso no ano de referência | Quantidade de Desistência no Curso no ano de referência | Quantidade de Falecimentos no Curso no ano de referência | Indicadores de Trajetória | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Taxa de Permanência - TAP | Taxa de Conclusão Acumulada - TCA | Taxa de Desistência Acumulada - TODA |
| 2010 | 2010 | 6 | 107 | 98 | 1 | 8 | 0 | 91,6 | 0,9 | 7,5 |
| 2010 | 2011 | 6 | 107 | 90 | 0 | 8 | 0 | 84,1 | 0,9 | 15,0 |
| 2010 | 2018 | 5 | 47 | 0 | 0 | 0 | 0 | 0,0 | 38,3 | 61,7 |
| 2010 | 2019 | 5 | 47 | 0 | 0 | 0 | 0 | 0,0 | 38,3 | 61,7 |
| 2010 | 2010 | 5 | 21 | 20 | 0 | 1 | 0 | 95,2 | 0,0 | 4,8 |
| 2010 | 2011 | 5 | 21 | 20 | 0 | 0 | 0 | 95,2 | 0,0 | 4,8 |
| 2010 | 2012 | 5 | 21 | 18 | 1 | 1 | 0 | 85,7 | 4,8 | 9,5 |
| 2010 | 2013 | 5 | 21 | 16 | 1 | 1 | 0 | 76,2 | 9,5 | 14,3 |
| 2010 | 2014 | 5 | 21 | 14 | 0 | 2 | 0 | 66,7 | 9,5 | 23,8 |
| 2010 | 2015 | 5 | 21 | 14 | 0 | 0 | 0 | 66,7 | 9,5 | 23,8 |
| 2010 | 2016 | 5 | 21 | 2 | 1 | 11 | 0 | 9,5 | 14,3 | 76,2 |
| 2010 | 2017 | 5 | 21 | 0 | 0 | 2 | 0 | 0,0 | 14,3 | 85,7 |
| 2010 | 2018 | 5 | 21 | 0 | 0 | 0 | 0 | 0,0 | 14,3 | 85,7 |
| 2010 | 2019 | 5 | 21 | 0 | 0 | 0 | 0 | 0,0 | 14,3 | 85,7 |
| 2010 | 2010 | 5 | 34 | 30 | 0 | 4 | 0 | 88,2 | 0,0 | 11,8 |
| 2010 | 2011 | 5 | 34 | 28 | 0 | 2 | 0 | 82,4 | 0,0 | 17,6 |

coding BOOTCAMP

Machine Learning

Enap Escola Nacional de Administração Pública

# dataset

Filtro:

cursos presenciais: 76.010

**ores de Trajetória de Curso de Graduação - Brasil - 2019**

rior 2019 (*coorte* 2010), composto por: Taxa de Permanência, Taxa de Conclusão Acumulada, Taxa de Desistência Acumulada, Taxa de Conclusão Anual e Taxa de

| Ano de Ingresso | Ano de Referência | Prazo de Integralização em Anos | Quantidade de Ingressantes no Curso | Quantidade de Permanência no Curso no ano de referência | Quantidade de Concluintes no Curso no ano de referência | Quantidade de Desistência no Curso no ano de referência | Quantidade de Falecimentos no Curso no ano de referência | Indicadores de Trajetória | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Taxa de Permanência - TAP | Taxa de Conclusão Acumulada - TCA | Taxa de Desistência Acumulada - TODA |
| 2010 | 2010 | 6 | 107 | 98 | 1 | 8 | 0 | 91,6 | 0,9 | 7,5 |
| 2010 | 2011 | 6 | 107 | 90 | 0 | 8 | 0 | 84,1 | 0,9 | 15,0 |
| 2010 | 2018 | 5 | 47 | 0 | 0 | 0 | 0 | 0,0 | 38,3 | 61,7 |
| 2010 | 2019 | 5 | 47 | 0 | 0 | 0 | 0 | 0,0 | 38,3 | 61,7 |
| 2010 | 2010 | 5 | 21 | 20 | 0 | 1 | 0 | 95,2 | 0,0 | 4,8 |
| 2010 | 2011 | 5 | 21 | 20 | 0 | 0 | 0 | 95,2 | 0,0 | 4,8 |
| 2010 | 2012 | 5 | 21 | 18 | 1 | 1 | 0 | 85,7 | 4,8 | 9,5 |
| 2010 | 2013 | 5 | 21 | 16 | 1 | 1 | 0 | 76,2 | 9,5 | 14,3 |
| 2010 | 2014 | 5 | 21 | 14 | 0 | 2 | 0 | 66,7 | 9,5 | 23,8 |
| 2010 | 2015 | 5 | 21 | 14 | 0 | 0 | 0 | 66,7 | 9,5 | 23,8 |
| 2010 | 2016 | 5 | 21 | 2 | 1 | 11 | 0 | 9,5 | 14,3 | 76,2 |
| 2010 | 2017 | 5 | 21 | 0 | 0 | 2 | 0 | 0,0 | 14,3 | 85,7 |
| 2010 | 2018 | 5 | 21 | 0 | 0 | 0 | 0 | 0,0 | 14,3 | 85,7 |
| 2010 | 2019 | 5 | 21 | 0 | 0 | 0 | 0 | 0,0 | 14,3 | 85,7 |
| 2010 | 2010 | 5 | 34 | 30 | 0 | 4 | 0 | 88,2 | 0,0 | 11,8 |
| 2010 | 2011 | 5 | 34 | 28 | 0 | 2 | 0 | 82,4 | 0,0 | 17,6 |

[...] ; TX_CONCLUSAO
TX_DESISTENCIA

QT_INGRESSANTE

`LinearRegression()` RMSE: 72.7

*`OneHotEncoder`*

# dataset

Filtro:

cursos presenciais: 76.010

Experimentos:

Dados de treino: 60.808 (80%)

Dados de teste: 15.202 (20%)

42

`random_state`

| | verdade | predição |
|---|---|---|
| 0 | 152 | 77.703125 |
| 1 | 91 | 118.917969 |
| 2 | 24 | 37.365234 |
| 3 | 167 | 153.566406 |
| 4 | 24 | 49.099609 |
| ... | ... | ... |
| 15197 | 36 | 68.544922 |
| 15198 | 30 | 49.683594 |
| 15199 | 55 | 94.964844 |
| 15200 | 98 | 129.599609 |
| 15201 | 137 | 140.437500 |

Enap Escola Nacional de Administração Pública

# dataset

Filtro:

cursos presenciais: 76.010

Experimentos:

Dados de treino: 60.808 (80%)

Dados de teste: 15.202 (20%)

42
random_state

~~`LinearRegression()`~~ ~~*OneHotEncoder*~~ `RMSE: 72.7`

`CatBoostRegressor()` *CatBoostEncoder* `RMSE: 51.8`

| | verdade | predição |
|---|---|---|
| ~~0~~ | 152 | 81.079125 |
| ~~1~~ | 91 | 100.110729 |
| 2 | 24 | 34.019192 |
| 3 | 167 | 192.961188 |
| ~~4~~ | 24 | 19.702916 |
| ... | ... | ... |
| ~~15197~~ | 36 | 41.019172 |
| 15198 | 30 | 43.731528 |
| 15199 | 55 | 65.891742 |
| 15200 | 98 | 121.196599 |
| 15201 | 137 | 127.106463 |

Enap Escola Nacional de Administração Pública

# dataset

Filtro:

cursos presenciais: 76.010

Experimentos:

Dados de treino: 60.808 (80%)

Dados de teste: 15.202 (20%)

42
random_state

15/12/2021

~~`LinearRegression()`~~ RMSE: 72.7
~~*OneHotEncoder*~~

~~`CatBoostRegressor()`~~ RMSE: 51.8
~~*CatBoostEncoder*~~

`CatBoostRegressor()` MAE: 23.58
*CatBoostEncoder*

| | verdade | predição |
|---|---|---|
| ~~0~~ | 152 | 72.280098 |
| 1 | 91 | 83.117959 |
| ~~2~~ | 24 | 29.845508 |
| 3 | 167 | 174.956945 |
| ~~4~~ | 24 | 24.423530 |
| ... | ... | ... |
| ~~15197~~ | 36 | 36.610139 |
| 15198 | 30 | 39.392768 |
| 15199 | 55 | 46.413764 |
| ~~15200~~ | 98 | 98.329462 |
| ~~15201~~ | 137 | 95.568902 |

Enap Escola Nacional de Administração Pública

# dataset

Filtro:

cursos presenciais: 76.010

Experimentos:

Dados de treino: 60.808 (80%)

Dados de teste: 15.202 (20%)

42
random_state

`LinearRegression()` RMSE: 72.7
*OneHotEncoder*

`CatBoostRegressor()` RMSE: 51.8
*CatBoostEncoder*

`RandomForestRegressor()` RMSE: 55.2
*CatBoostEncoder*

| | verdade | predição |
|---|---|---|
| 0 | 170 | 168.41000 |
| 1 | 6 | 49.02000 |
| 2 | 51 | 53.79000 |
| 3 | 26 | 58.05000 |
| 4 | 120 | 109.86000 |
| ... | ... | ... |
| 15197 | 11 | 15.93000 |
| 15198 | 9 | 32.21848 |
| 15199 | 69 | 72.23000 |
| 15200 | 42 | 43.50000 |
| 15201 | 43 | 37.85000 |

Enap Escola Nacional de Administração Pública

# dataset

Filtro:

cursos presenciais: 76.010

Experimentos:

Dados de treino: 60.808 (80%)

Dados de teste: 15.202 (20%)

42
random_state

15/12/2021

~~`LinearRegression()`~~ RMSE: 72.7
~~*OneHotEncoder*~~

~~`CatBoostRegressor()`~~ RMSE: 51.8
~~*CatBoostEncoder*~~

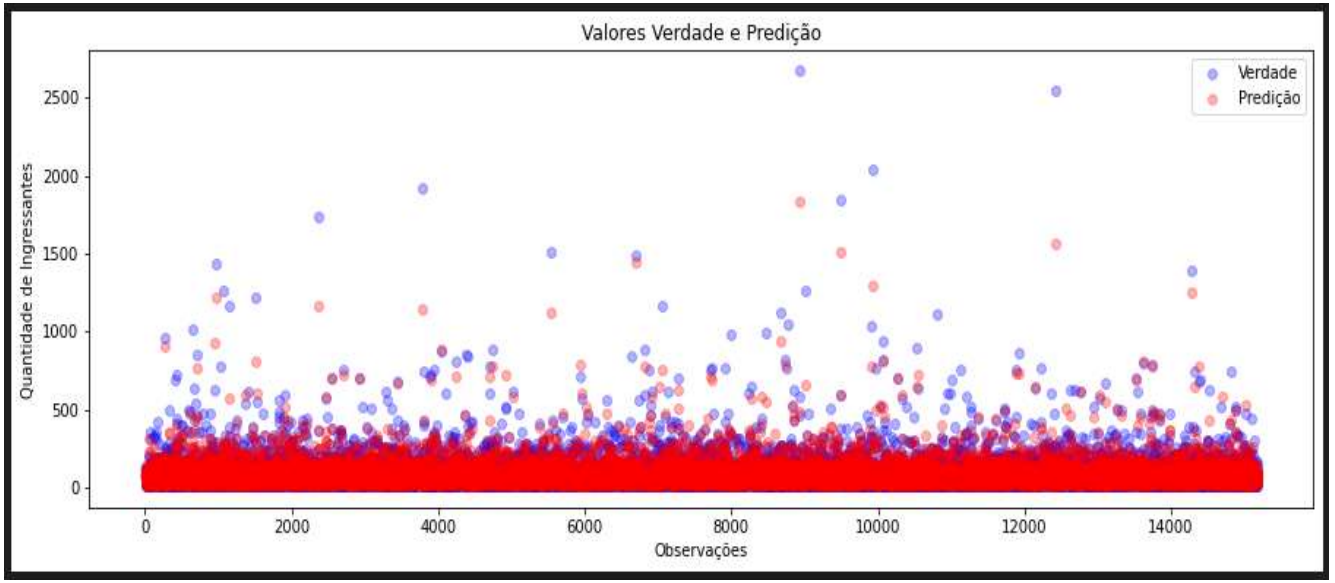
`CatBoostRegressor()` MAE: 23.58
*CatBoostEncoder*

~~`RandomForestRegressor()`~~ RMSE: 55.2
~~*CatBoostEncoder*~~

| | verdade | predição |
|---|---|---|
| ~~0~~ | 152 | 72.280098 |
| 1 | 91 | 83.117959 |
| ~~2~~ | 24 | 29.845508 |
| 3 | 167 | 174.956945 |
| ~~4~~ | 24 | 24.423530 |
| ... | ... | ... |
| ~~15197~~ | 36 | 36.610139 |
| 15198 | 30 | 39.392768 |
| 15199 | 55 | 46.413764 |
| ~~15200~~ | 98 | 98.329462 |
| ~~15201~~ | 137 | 95.568902 |

Enap Escola Nacional de Administração Pública

# futuro



# dataset

Filtro:

reduzir o dataset ( target <=25 )

cursos presenciais: 15.264

Experimentos:

Dados de treino: 12.211 (80%)

Dados de teste: 3.053 (20%)

42

random_state

CatBoostRegressor()  RMSE: 4.6

*CatBoostEncoder*

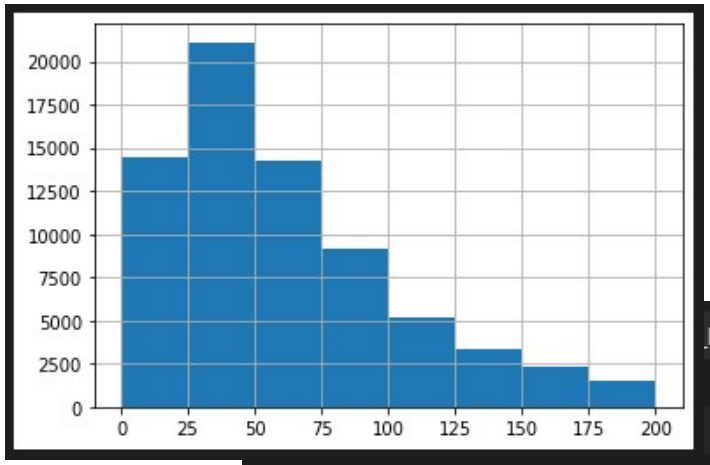

| | verdade | predição |
|---|---|---|
| 0 | 18 | 16.693773 |
| 1 | 22 | 18.402570 |
| 2 | 14 | 15.718735 |
| 3 | 5 | 8.550538 |
| 4 | 9 | 14.444949 |
| ... | ... | ... |
| 3048 | 25 | 12.226251 |
| 3049 | 4 | 5.085830 |
| 3050 | 4 | 5.016169 |
| 3051 | 23 | 17.902256 |
| 3052 | 6 | 11.851555 |

Enap Escola Nacional de Administração Pública

# futuro

Experimentos:

- Classificar os dados em intervalos
- Aplicar modelos de classificação




| | | | INGRESSANTE | CO_FX_QT_INGRESSANTE |
|---|---|---|---|---|
| | | | 118 | FX_101_125 |
| | | | 93 | FX_076_100 |
| 0.000000 | 3.703125 | 96.31250 | 27 | FX_026_050 |
| 6.667969 | 28.000000 | 65.31250 | 75 | FX_051_075 |
| 0.000000 | 50.000000 | 50.00000 | 12 | FX_<_025 |
| 0.000000 | 0.000000 | 100.00000 | 9 | FX_<_025 |
| 0.000000 | 80.562500 | 19.43750 | 36 | FX_026_050 |
| 0.000000 | 59.687500 | 40.31250 | 186 | FX_176_200 |
| 0.000000 | 34.718750 | 65.25000 | 167 | FX_151_175 |
| 0.000000 | 0.000000 | 100.00000 | 79 | FX_076_100 |





# fim ... ou início ?!

marcelocardoso@cp2.g12.br



# Obrigado !