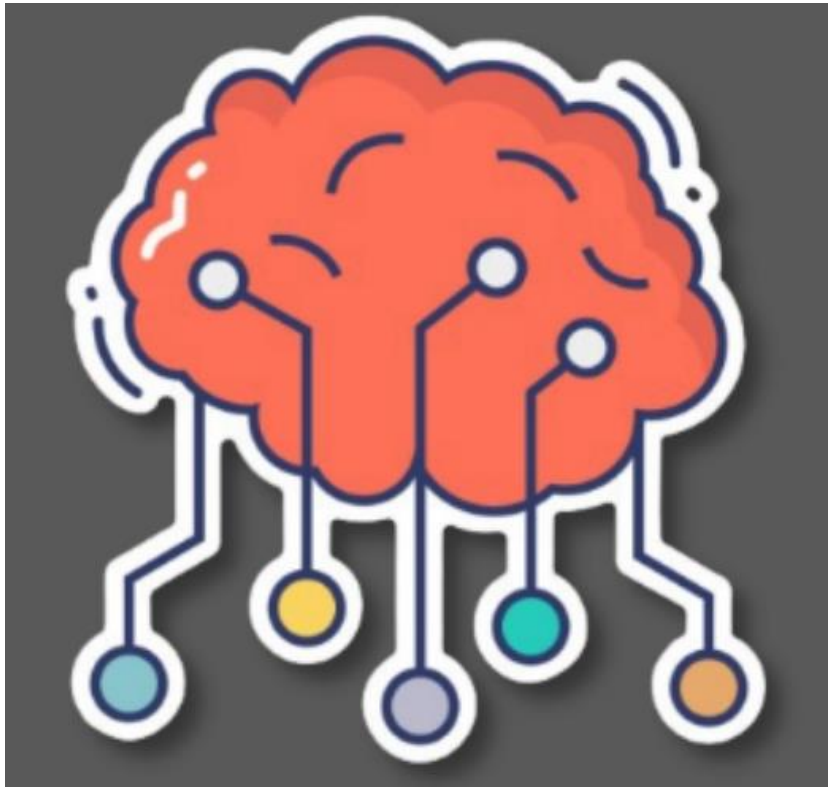
INSTITUTO FEDERAL DE
EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
São Paulo
Campus Campinas

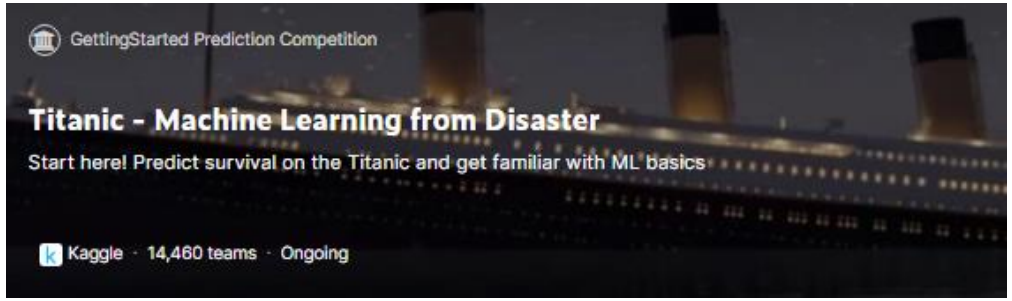
Projeto Interdisciplinar
Big Data + ML



- Pós Graduação em Ciência de Dados - 2022.2
- IFSP Campinas
- Profa. Bianca Pedrosa - bpedrosa@ifsp.edu.br
  Prof. Samuel Martins (Samuka) - @hisamuka
- Outubro de 2022
- aluno: Swift Yaguchi - CP301665X
- https://github.com/swiftyaguchi/IFSP-CMP-Projeto_Interdisciplinar-Pos-Ciencia_de_Dados-2022.2

Proposta do Projeto - “Frame the Problem”

- Inspiração :
  - "Titanic Problem" na plataforma do Kaggle



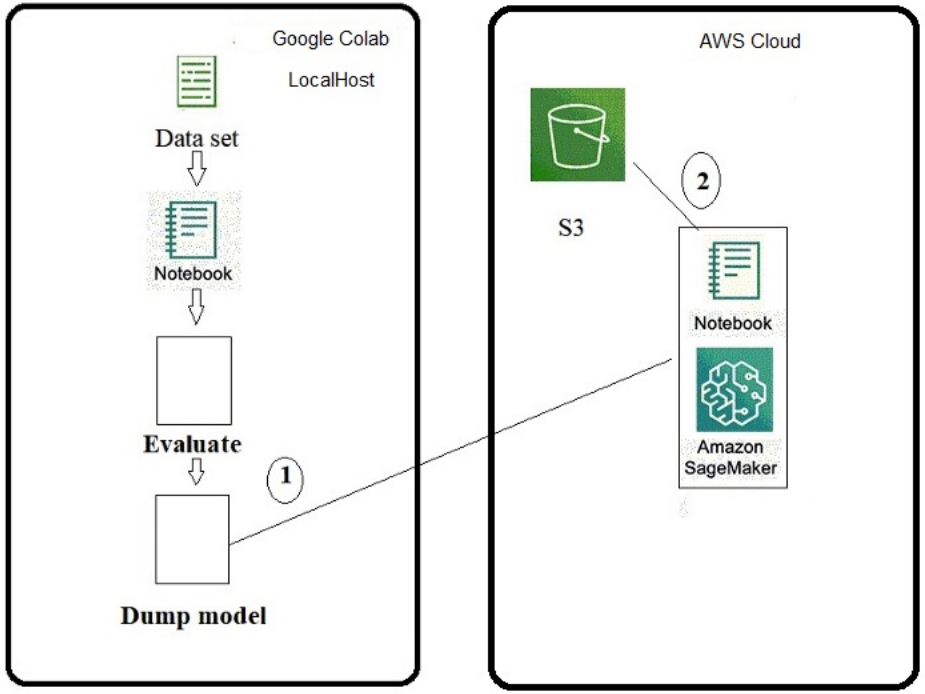
- Titanic survival prediction:
  - dados dos passageiros do navio
    >> *características dos sobreviventes do naufrágio*.

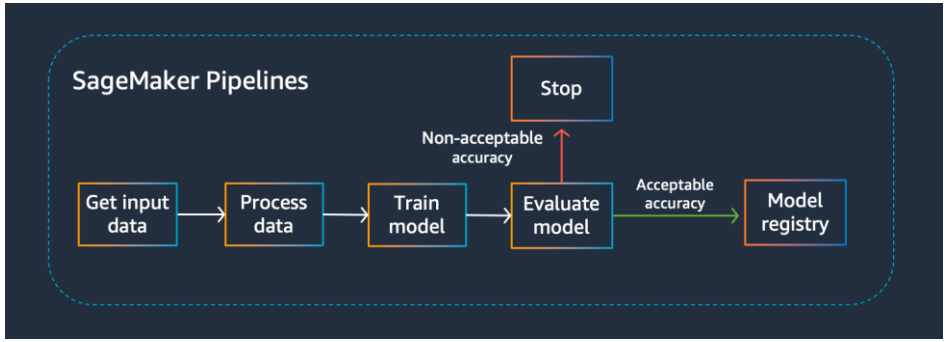
- Neste projeto:
  - dados de candidatos da eleição brasileira de 2022
    >> *características dos candidatos eleitos.*

TSE Portal de Dados Abertos do TSE

Arquitetura e Fluxo de Trabalho

- Desenvolvimento inicial do notebook Jupyterlab:
  - *ambiente do Google Colab*
  - *ambiente local do meu computador.*
- Carregado no AWS Cloud para finalização
  - *Fluxo de trabalho da figura do Sagemaker Pipeline*
  - *Dados carregados em bucket S3*
- Link do notebook Jupyterlab :
  - https://ifspcps-swiftyaguchi-proj-interdisciplinar-2022-teste1.notebook.us-east-1.sagemaker.aws/lab/tree/****%20%20Proj_Inter_2022.2%20%20****

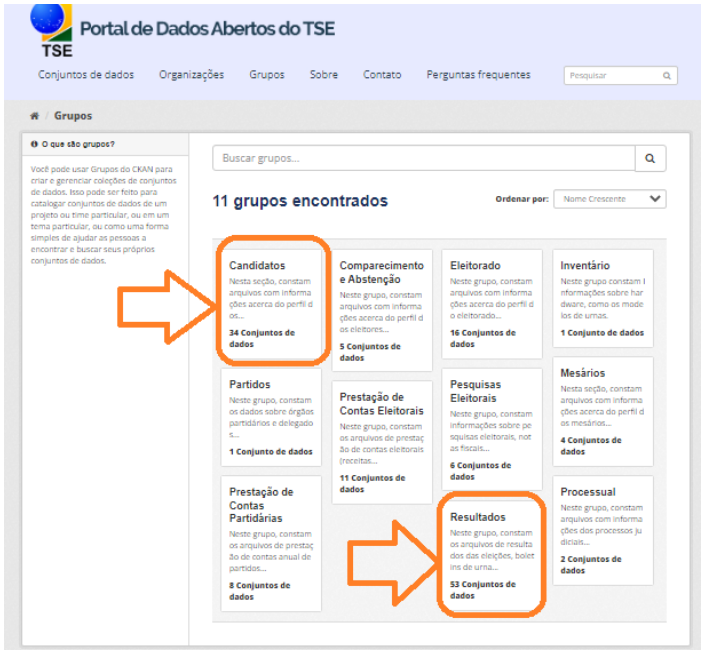




Get Data

Perfil dos candidatos (5.986KB zip)

Resultados de votação por municipio e por zona eleitoral (830.299KB zip)



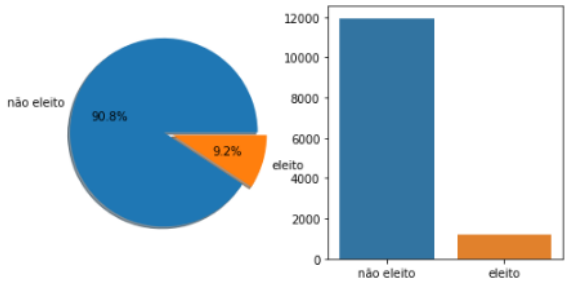
Limpeza e Pré-processamento

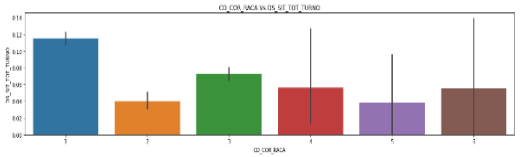
- 13165 – train
- 3292 - test

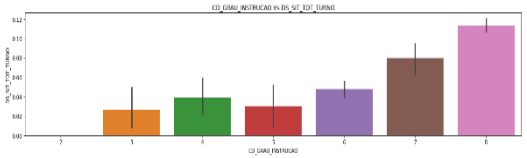
```
RangeIndex: 13165 entries, 0 to 13164
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CD_CARGO              13165 non-null  int64
 1   SQ_CANDIDATO          13165 non-null  int64
 2   NR_IDADE_DATA_POSSE_x 13165 non-null  float64
 3   CD_GENERO             13165 non-null  int64
 4   CD_GRAU_INSTRUCAO     13165 non-null  int64
 5   CD_ESTADO_CIVIL       13165 non-null  int64
 6   CD_COR_RACA           13165 non-null  int64
 7   CD_OCUPACAO           13165 non-null  int64
 8   ST_REELEICAO_x        13165 non-null  int64
 9   VR_BEM_CANDIDATO_x    13165 non-null  float64
 10  QT_VOTOS_NOMINAIS_x   13165 non-null  int64
 11  DS_SIT_TOT_TURNO      13165 non-null  int64
dtypes: float64(2), int64(10)
memory usage: 1.2 MB
```

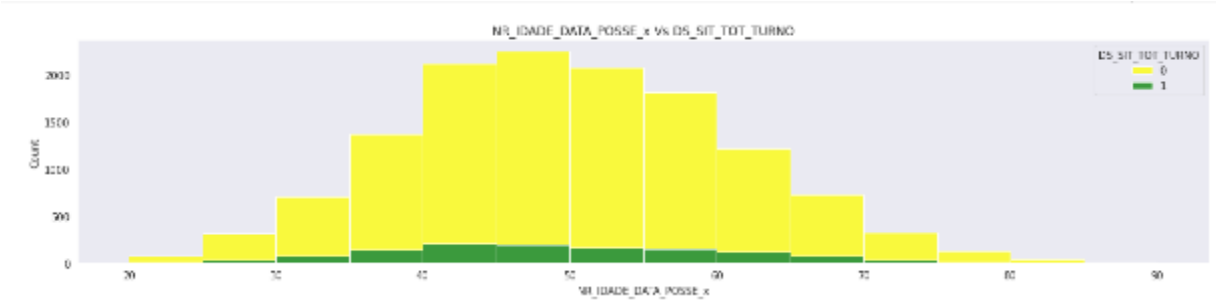
```
RangeIndex: 3292 entries, 0 to 3291
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CD_CARGO              3292 non-null   int64
 1   SQ_CANDIDATO          3292 non-null   int64
 2   NR_IDADE_DATA_POSSE_x 3292 non-null   float64
 3   CD_GENERO             3292 non-null   int64
 4   CD_GRAU_INSTRUCAO     3292 non-null   int64
 5   CD_ESTADO_CIVIL       3292 non-null   int64
 6   CD_COR_RACA           3292 non-null   int64
 7   CD_OCUPACAO           3292 non-null   int64
 8   ST_REELEICAO_x        3292 non-null   int64
 9   VR_BEM_CANDIDATO_x    3292 non-null   float64
 10  QT_VOTOS_NOMINAIS_x   3292 non-null   int64
 11  DS_SIT_TOT_TURNO      3292 non-null   int64
dtypes: float64(2), int64(10)
memory usage: 308.8 KB
```

Exploratory Data Analysis (EDA):







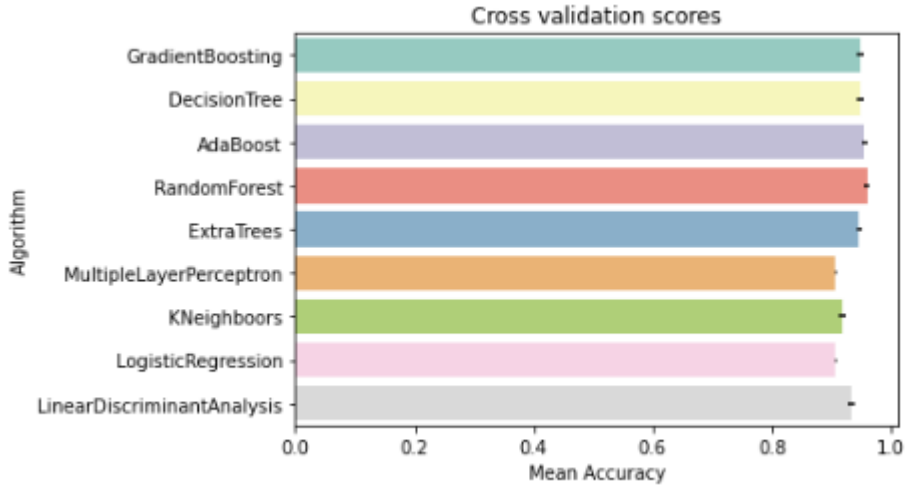


- Apenas 9,2% dos candidatos são eleitos
- Maior taxa de eleitos:
  - *Candidatos à reeleição*
  - *Faixa de idade entre 40 e 60 anos, e são os mais eleitos*
  - *Brancos*
  - *Casados*
  - *Grau de instrução superior*
  - *Homens*

Treinamento Machine Learning

Baseline

Cross Validation



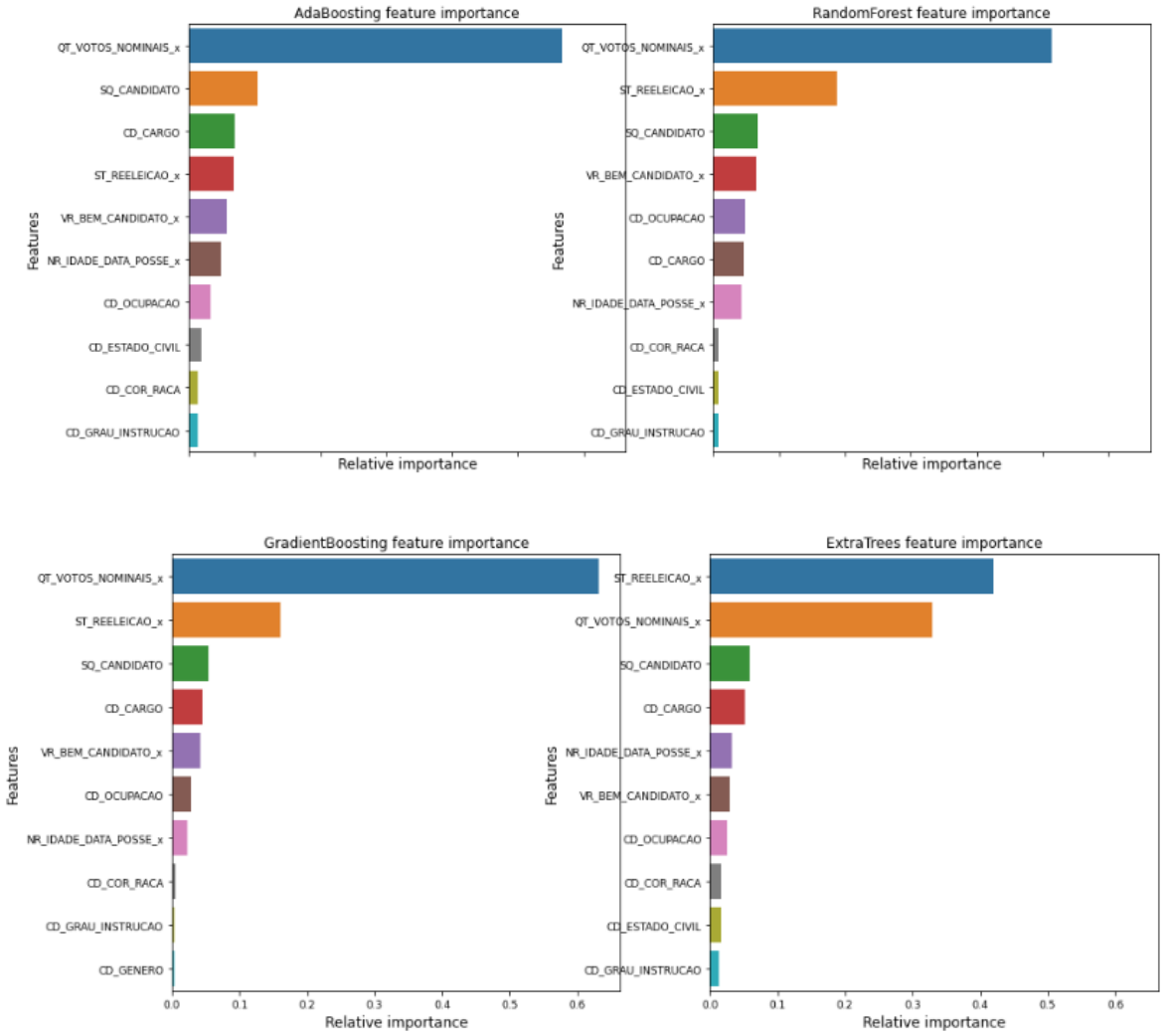
[15]:

	CrossValMeans	CrossValerrors	Algorithm
3	0.959742	0.004876	RandomForest
2	0.955260	0.005178	AdaBoost
1	0.948273	0.006755	DecisionTree
0	0.948273	0.007205	GradientBoosting
4	0.946373	0.004164	ExtraTrees
8	0.933840	0.006913	LinearDiscriminantAnalysis
6	0.918039	0.005036	KNeighboors
7	0.907634	0.000344	LogisticRegression
5	0.907330	0.001054	MultipleLayerPerceptron

Resultado Emsemble Modeling:

0.9577764277035237

Análise das variáveis

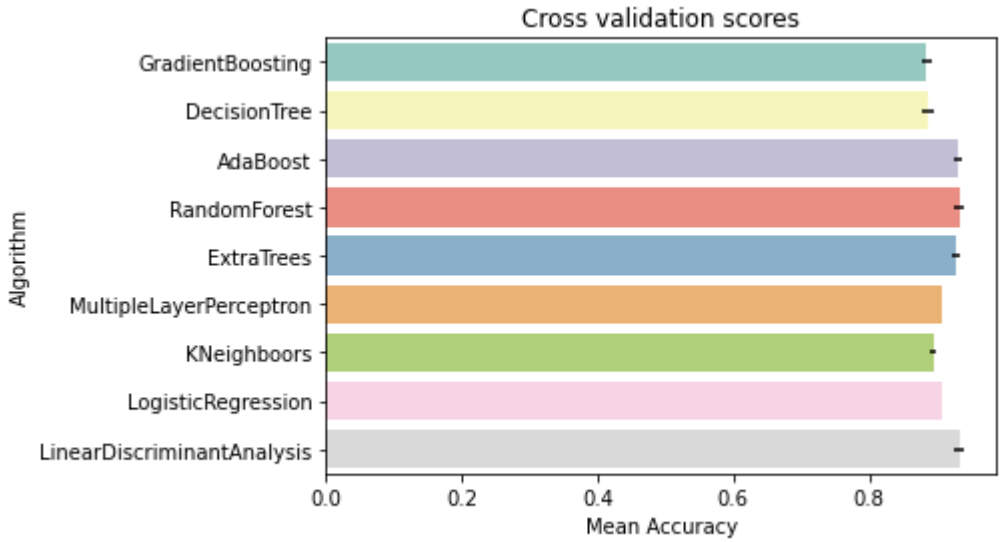


AdaBoosting, Random Forest e Gradient

- Quantidade de votos nominais tem maior importância

Treinamento Machine Learning

Retirando variável número de votos recebidos

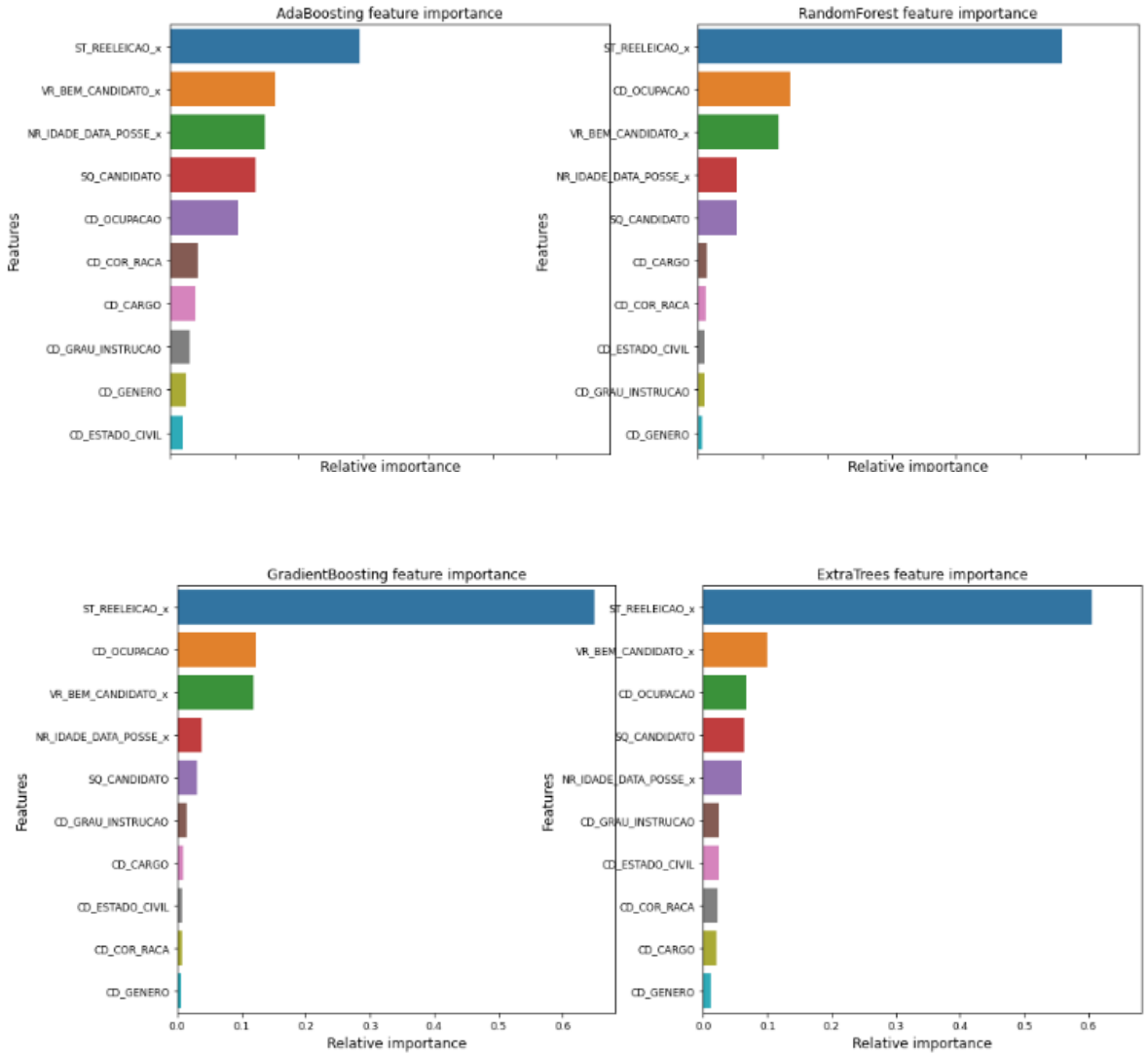


	CrossValMeans	CrossValerrors	Algorithm
8	0.933992	0.007103	LinearDiscriminantAnalysis
3	0.933385	0.006939	RandomForest
2	0.931865	0.005090	AdaBoost
4	0.928523	0.005194	ExtraTrees
7	0.907634	0.000344	LogisticRegression
5	0.907406	0.000963	MultipleLayerPerceptron
6	0.894493	0.003329	KNeighboors
1	0.886897	0.008041	DecisionTree
0	0.885378	0.007638	GradientBoosting

Resultado Emsemble Modeling:

0.925577156743621

Análise das variáveis



'ST_REELEICAO_x' tem maior importância para todos

Em seguida:

AdaBoosting e Extra Trees:

- VR_BEM_CANDIDATO com importância relativa maior que CD_OCUPACAO,

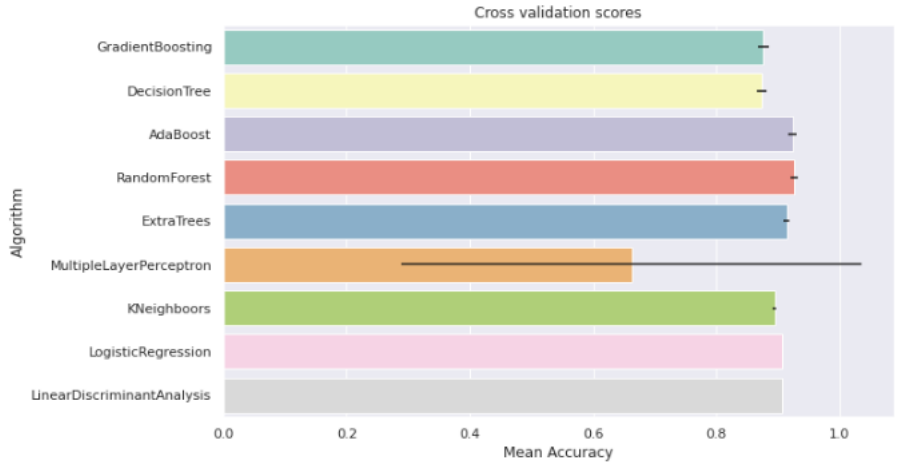
Random Forest e Gradient Boosting

- O contrário.

Treinamento Machine Learning

Estudo eliminando as variáveis:

- QT_VOTOS_NOMINAIS_x
- ST_REELEICAO_x



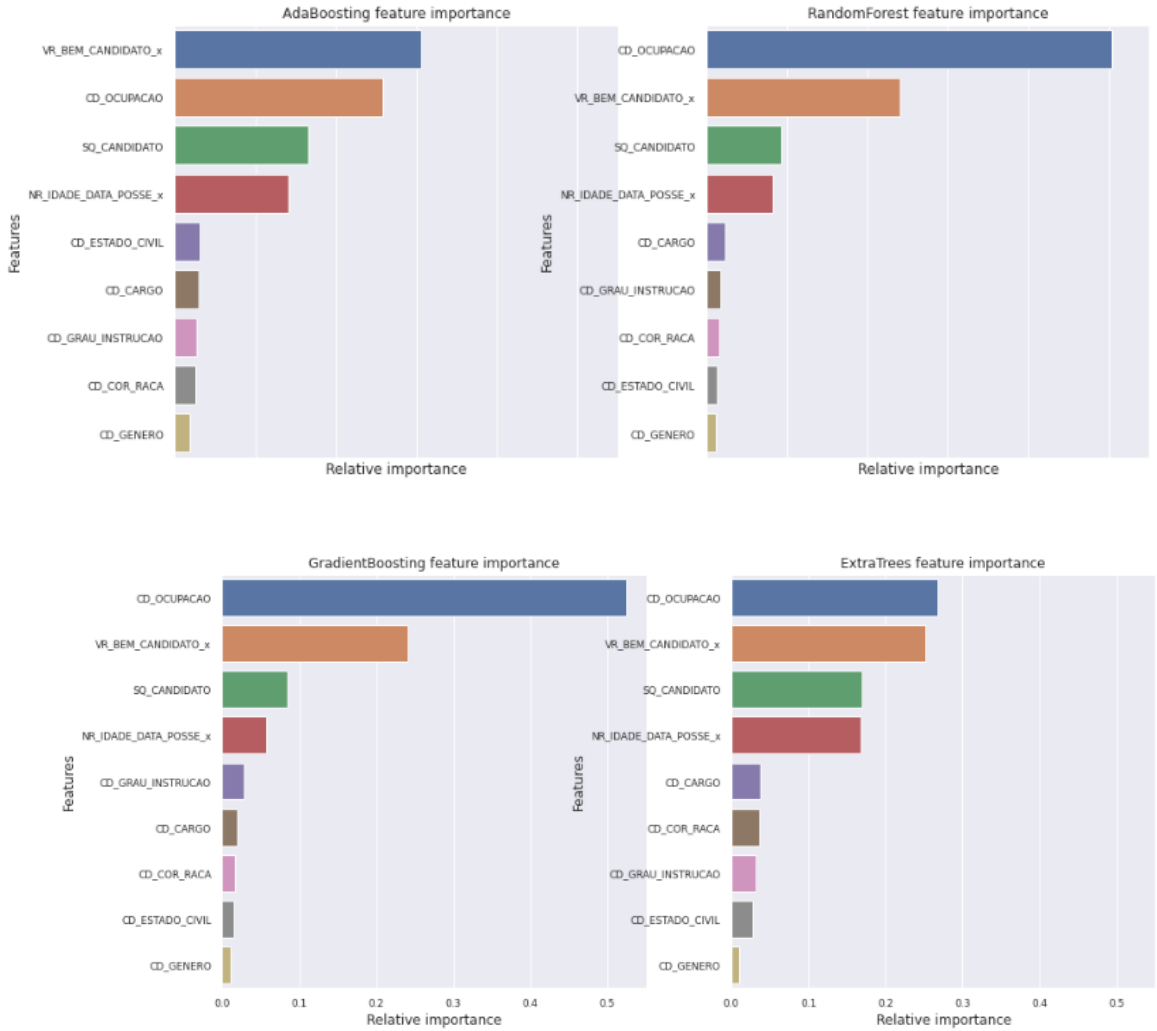
[18]:

	CrossValMeans	CrossValerrors	Algorithm
3	0.926244	0.005791	RandomForest
2	0.923890	0.006302	AdaBoost
4	0.914319	0.004322	ExtraTrees
7	0.907634	0.000344	LogisticRegression
8	0.907254	0.000706	LinearDiscriminantAnalysis
6	0.894493	0.003329	KNeighboors
0	0.876795	0.008635	GradientBoosting
1	0.873681	0.007695	DecisionTree
5	0.663001	0.373544	MultipleLayerPerceptron

Resultado Emsemble Modeling:

0.9170716889428918

Análise das variáveis



AdaBoosting :

- VR_BEM_CANDIDATO com importância relativa maior que CD_OCUPACAO,

Random Forest, e Extra Trees e Gradient Boosting :

- O contrário.

Conclusão

Estudos	Candidatos Eleitos	Score de Acurácia
Baseline	242	0.9577764277035237
Estudo eliminado variável número de votos recebidos	170	0.925577156743621
Estudo eliminando as variáveis: - QT_VOTOS_NOMINAIS - ST_REELEICAO	94	0.9170716889428918