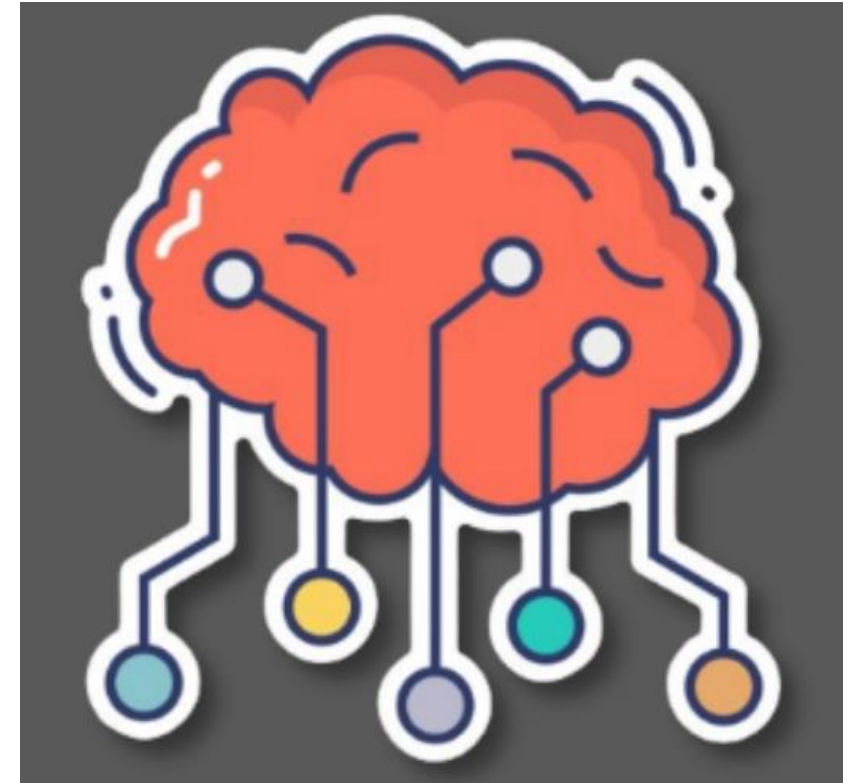


Projeto Interdisciplinar Big Data + ML

- 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



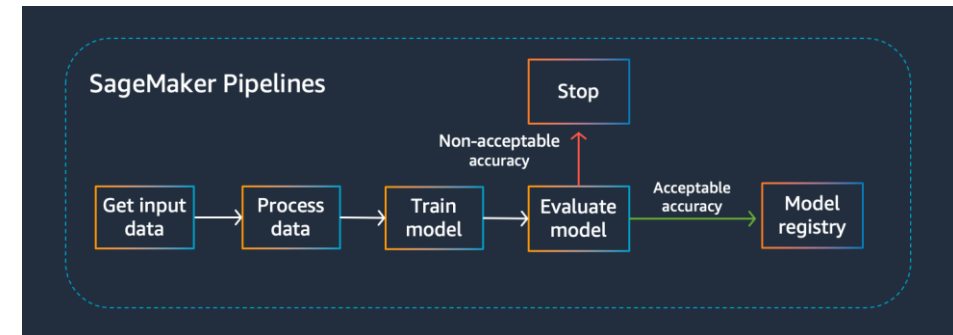
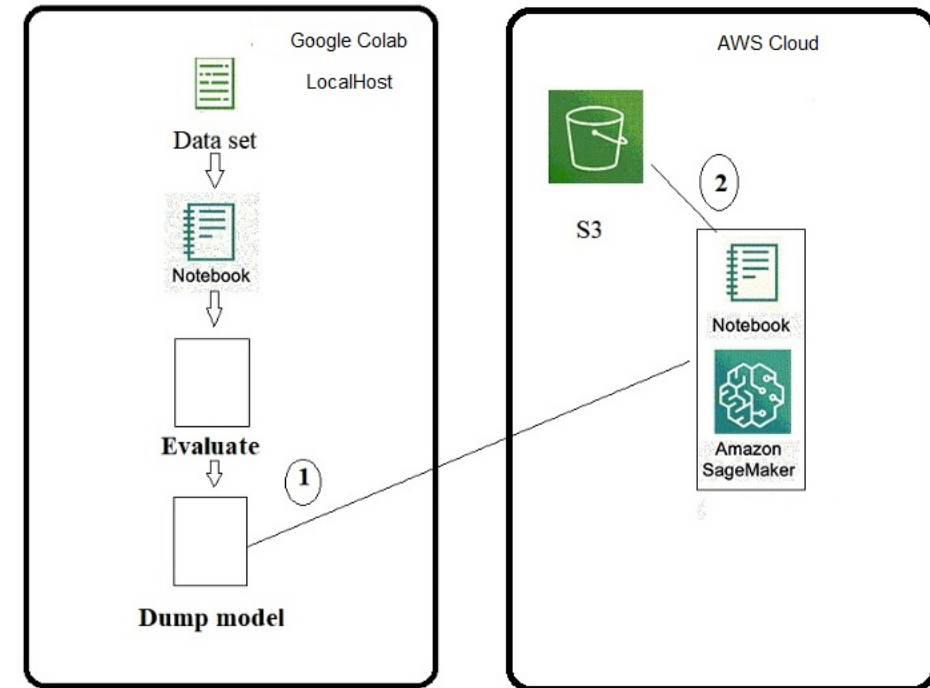
Proposta do Projeto - "Frame the Problem"

- 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.*



Arquitetura e Fluxo de Trabalho

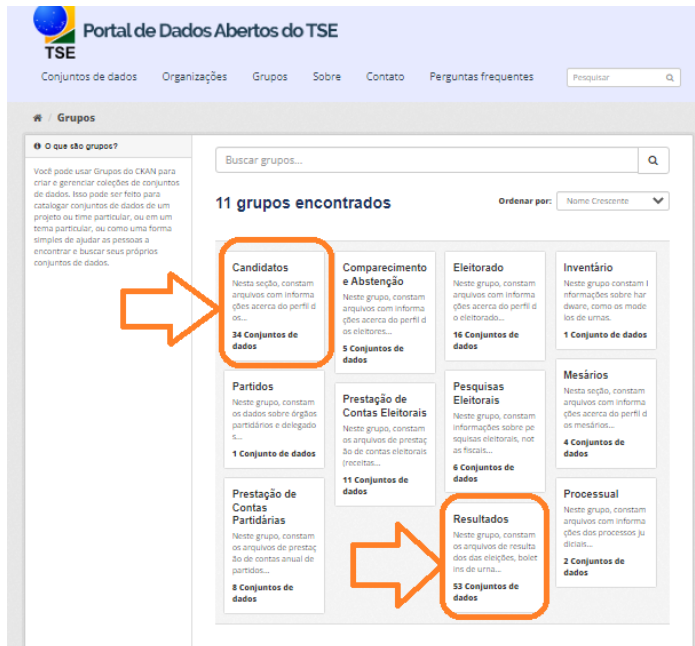
- Desenvolvimento inicial do notebook Jupyterlab:
 - *ambiente do Google Colab*
 - *ambiente local do meu computador.*
- Carregado no AWS Cloud para finalização
 - *Fluxo de trabalho da figura do Sagemaker Pipeline*
 - *Dados carregados em bucket S3*



Get Data

Perfil dos candidatos (5.986KB zip)

Resultados de votação por município 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_TURN0	13165 non-null	int64

dtypes: float64(2), int64(10)

memory usage: 1.2 MB

RangeIndex: 3292 entries, 0 to 3291

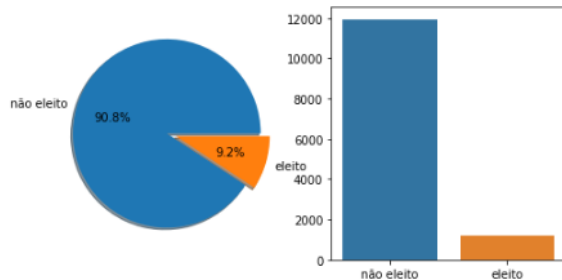
Data columns (total 12 columns):

#	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_TURN0	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*

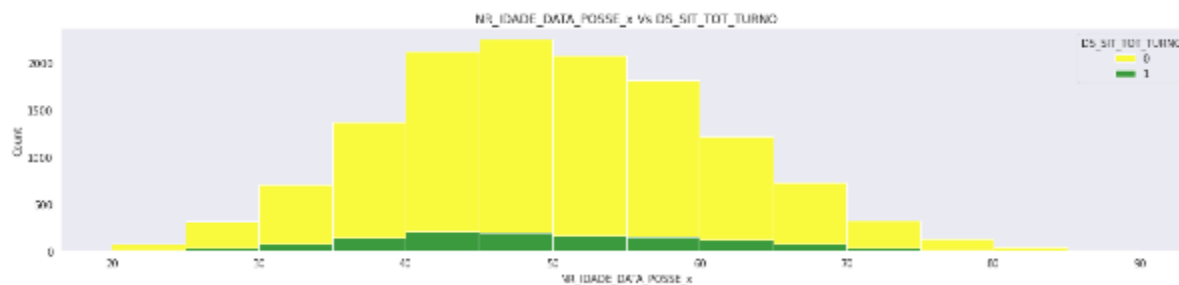
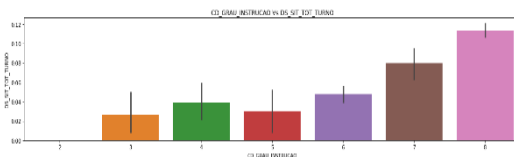
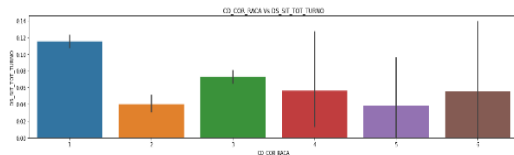
- *Faixa de idade entre 40 e 60 anos, e são os mais eleitos*

- *Brancos*

- *Casados*

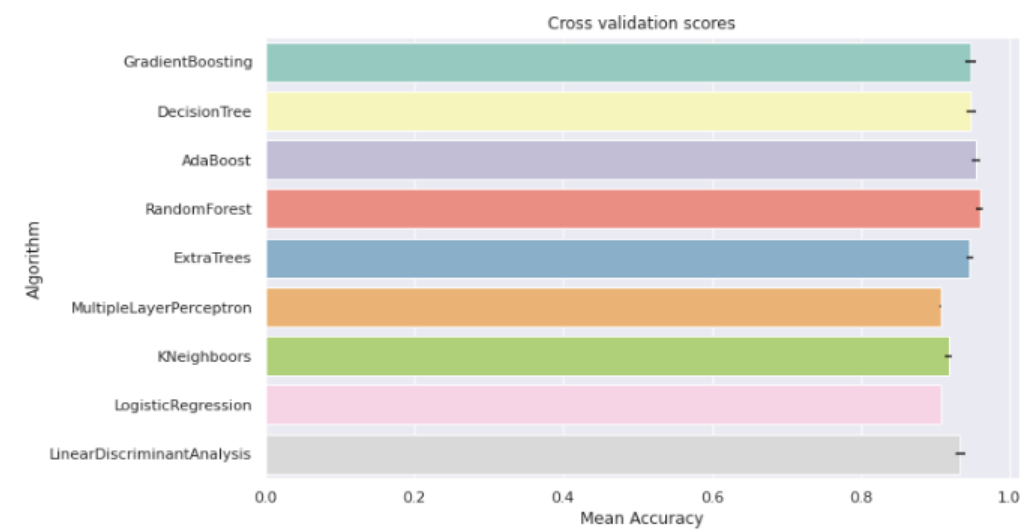
- *Grau de instrução superior*

- *Homens*



Treinamento Machine Learning

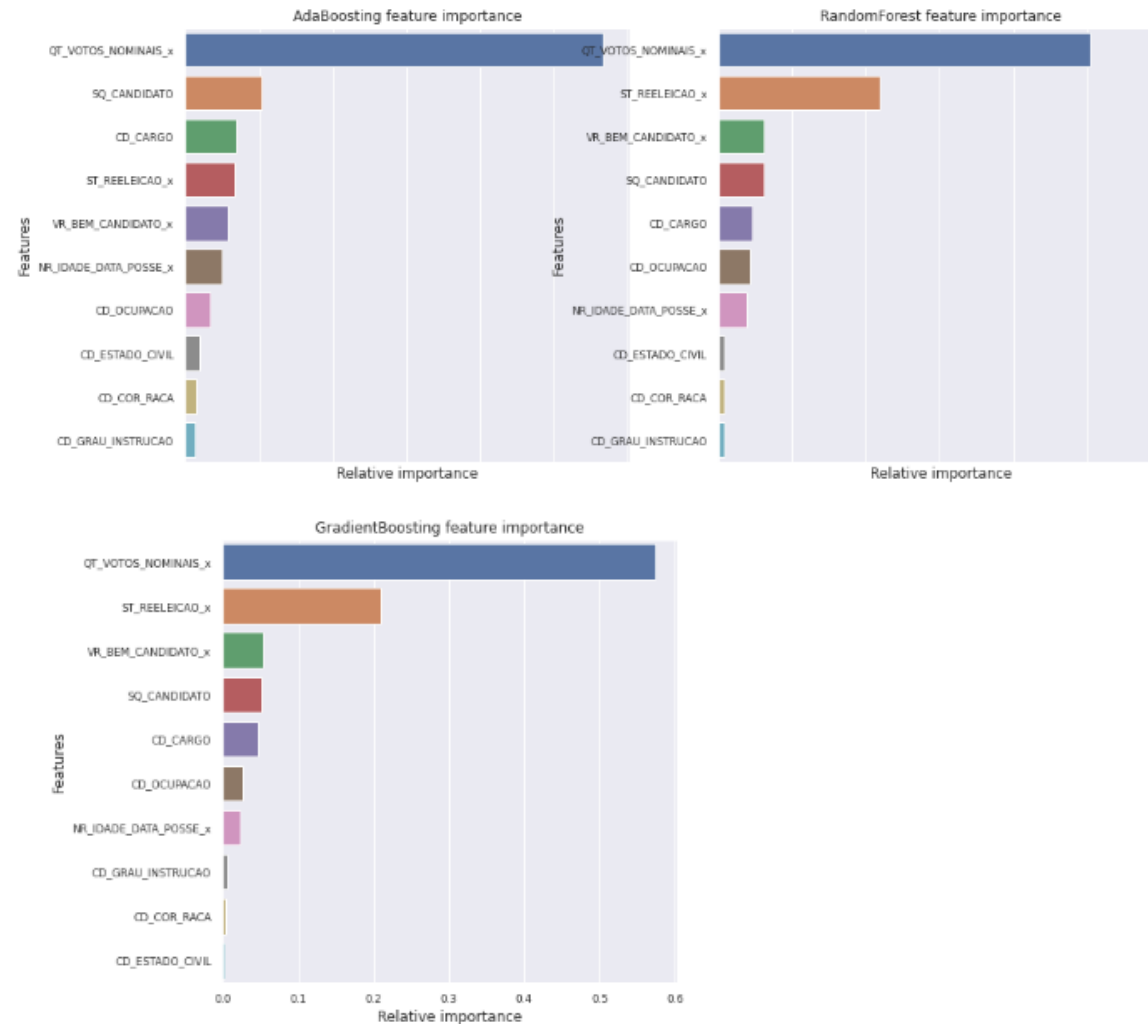
Cross Validation



	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	KNeighbors
1	0.886897	0.008041	DecisionTree
0	0.885378	0.007638	GradientBoosting

Resultado Emsemble Modeling:
0.925577156743621

Análise das variáveis

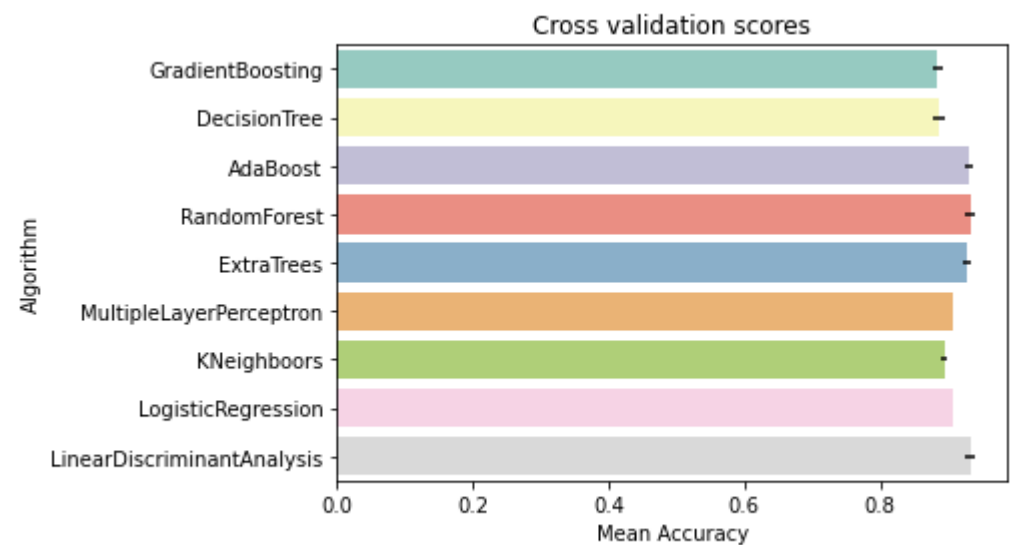


AdaBoosting, Random Forest e Gradient

- Qty de votos nominais tem maior importância

Treinamento Machine Learning

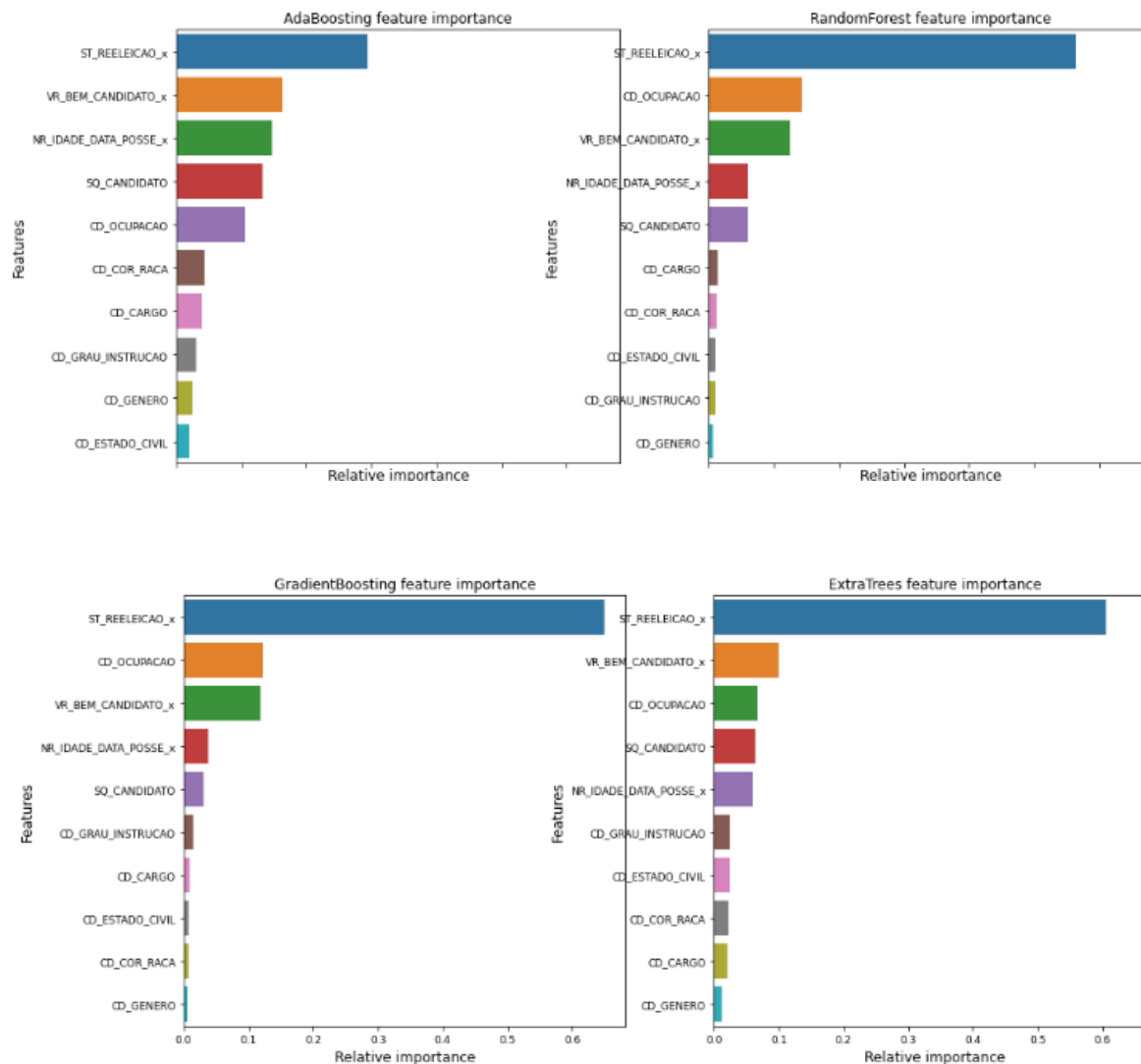
Retirando variável número de votos recebidos



Resultado Emsemble Modeling:
0.925577156743621

	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	KNeighbors
1	0.886897	0.008041	DecisionTree
0	0.885378	0.007638	GradientBoosting

Análise das variáveis



AdaBoosting e Extra Trees:

- VR_BEM_CANDIDATO com importância relativa maior que CD_OCUPACAO, Random Forest e Gradient Boosting
- O contrário.