Centro Brasileiro de Pesquisas Físicas
CBPF
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

Finding Stars

clustering analysis to separate star/galaxies in a survey

Thiago P. Carneiro, 2021

Introdução

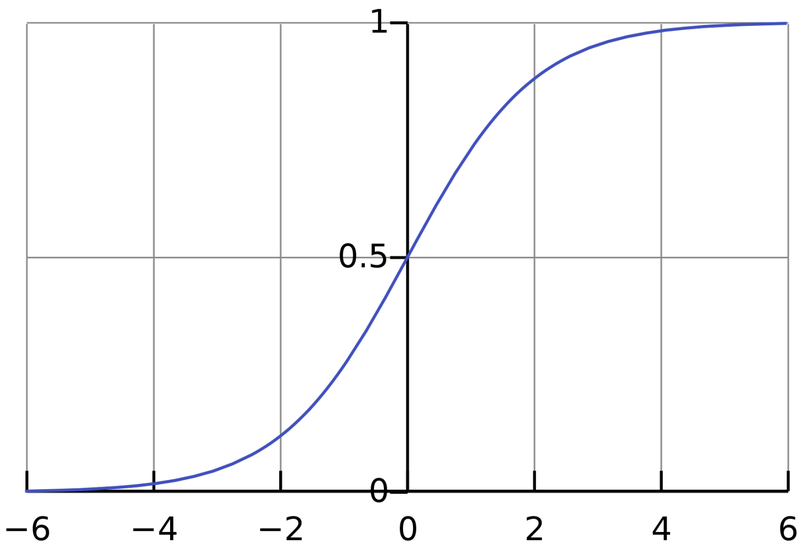
- Surveys mapeiam um número astronômico de estrelas e de galáxias em seus catálogos (além de outros corpos, como quasares, em menor grau).
- Como muitas vezes não há interesse em analisar estrelas, torna-se importante identificá-las para que sejam ignoradas.
- Visualmente, pode ser difícil distinguir uma estrela de uma galáxia elíptica.
- Considerando o aumento no volume de dados dos catálogos nos próximos anos, a automatização da identificação de estrelas torna-se imprescindível.

Planejamento

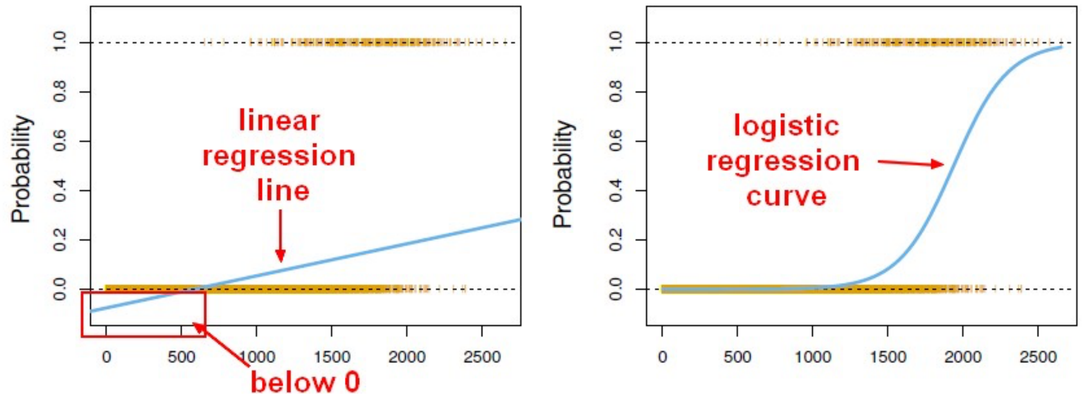
- Irei avaliar as capacidades de identificar estrelas e a velocidade de diferentes técnicas de aprendizado de máquina:
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - LightGBM
  - TabNet
  - Modelo Composto (Logistic Regression + Random Forest)

Logistic Regression

- Baseada no ajuste dos parâmetros de uma função logística para maximizar a verossimilhança do modelo aos dados de treinamento.
- A classificação se dá por simples projeção dos dados na curva ajustada.



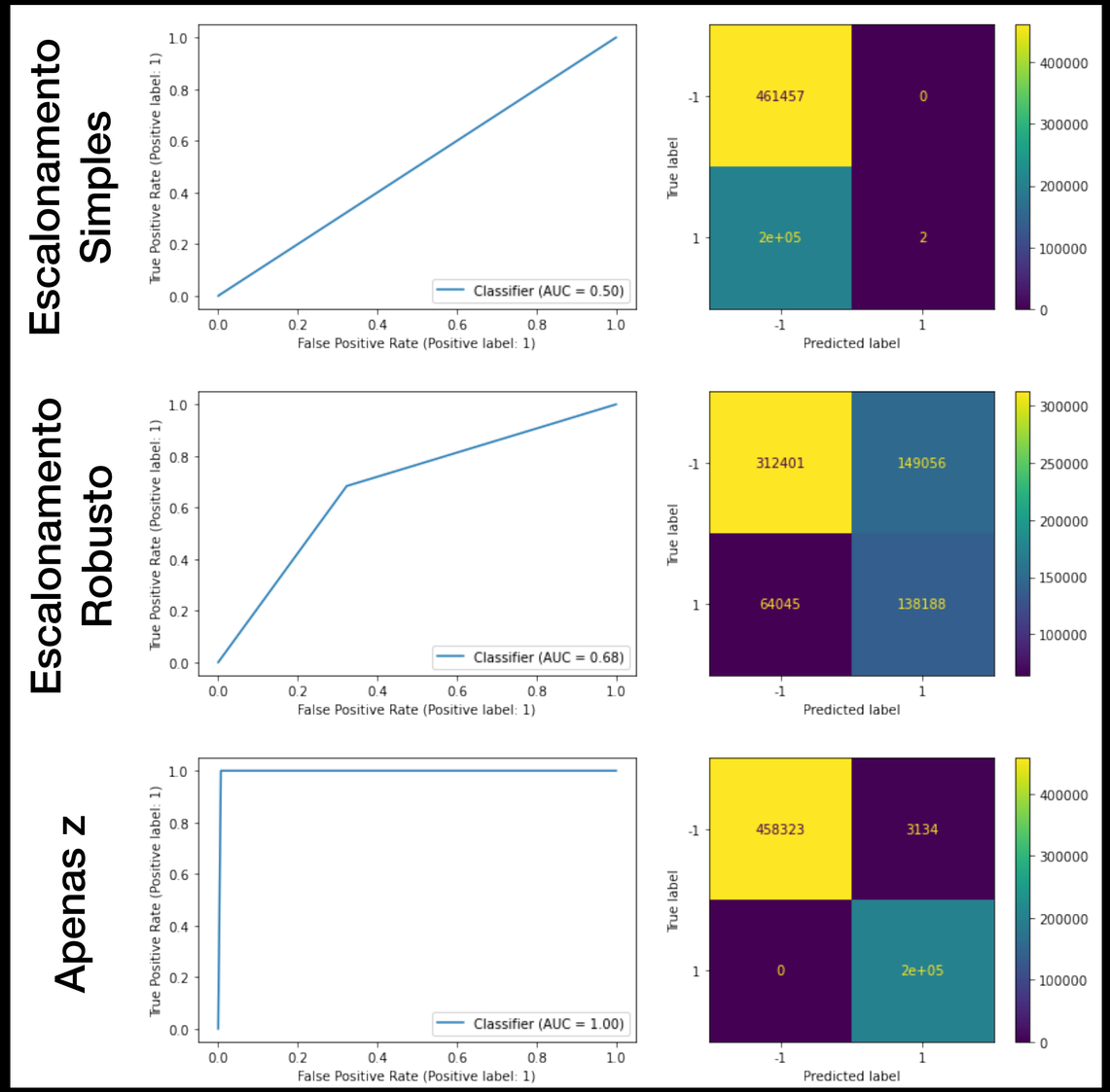
[https://commons.wikimedia.org/wiki/File:Logistic-curve.svg]



[https://datacadamia.com/data_mining/simple_logistic_regression]

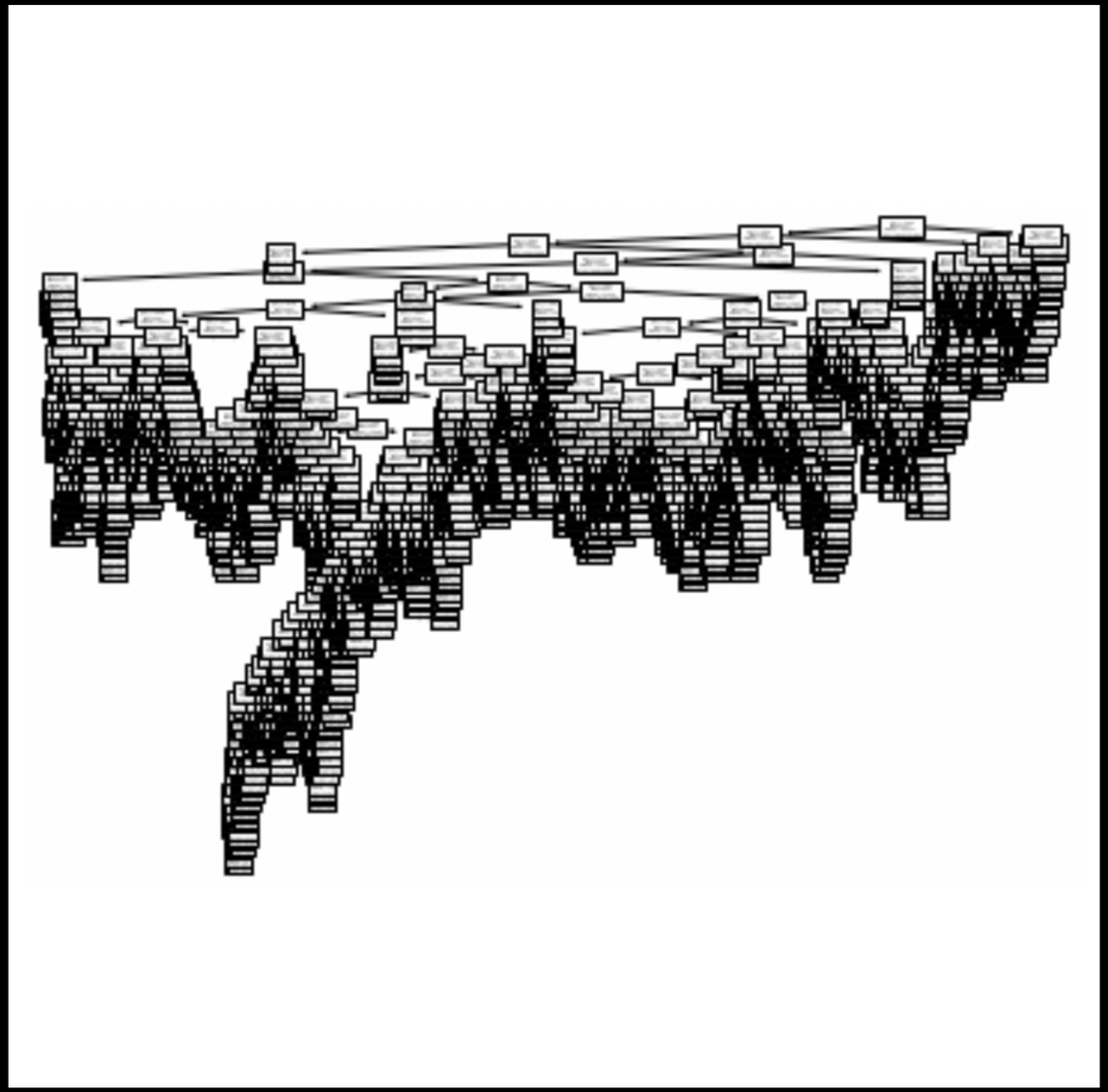
Logistic Regression

- Para este modelo, as técnicas usuais de escalonamento não deram bons resultados. Somente quando utilizamos apenas a informação z no treinamento o modelo pôde convergir para um resultado satisfatório.
- Nota-se que seu maior problema foi identificar muitas galáxias/quasares como estrelas.



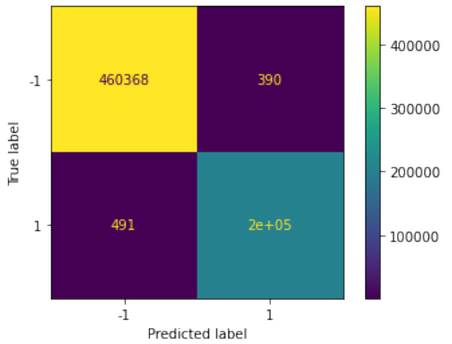
Decision Tree

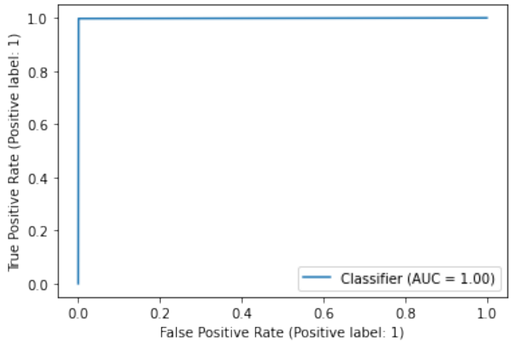
- Modelo em que sucessivos testes VERDADEIRO/FALSO avaliam os dados para definir a classificação.
- O resultado de cada teste define qual será o próximo teste (ou, eventualmente, qual a classificação do dado), formando a figura de uma árvore.



Decision Tree

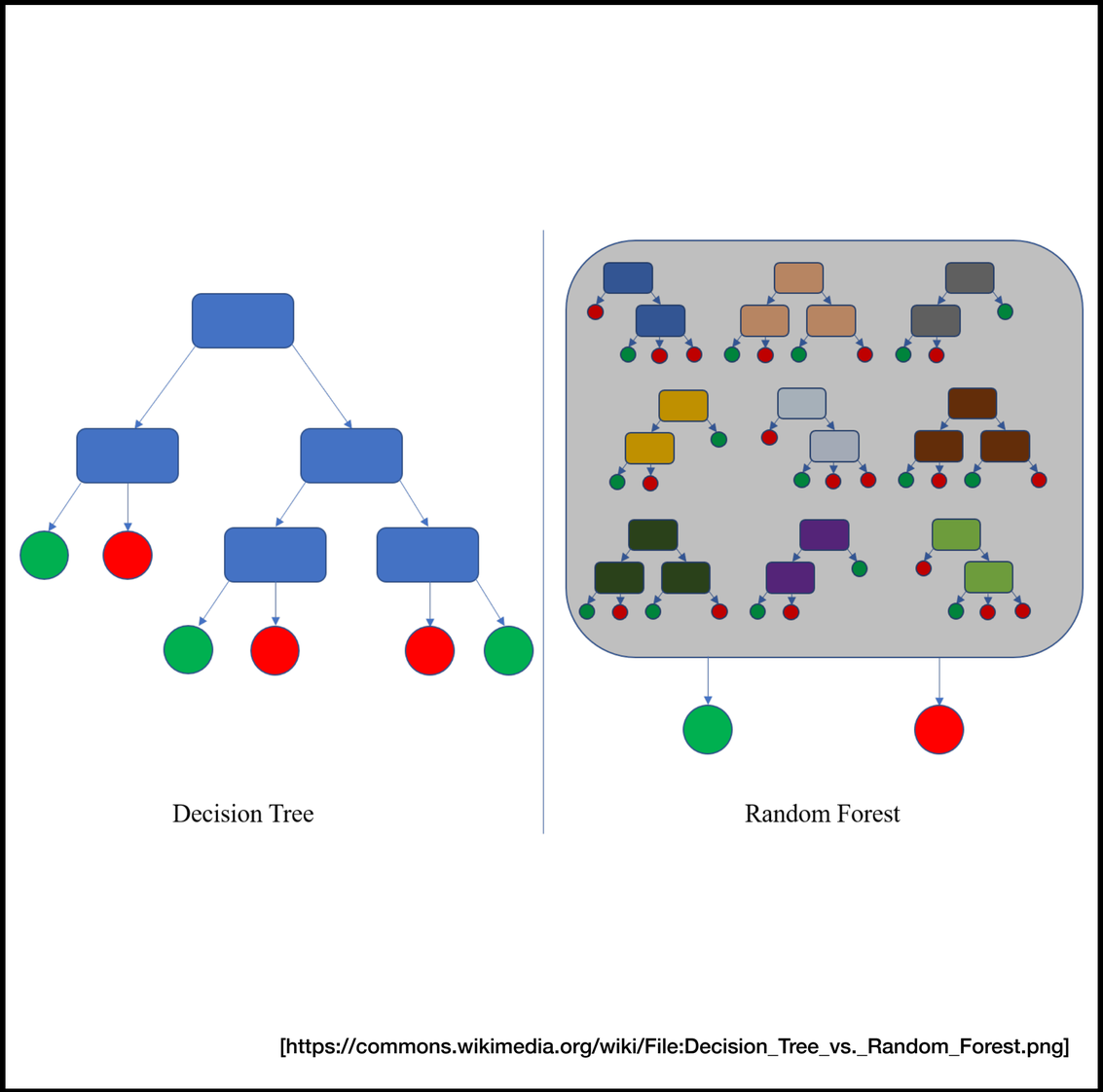
- O resultado foi bastante satisfatório em comparação com a Regressão Logística: houve uma grande diminuição do número de casos de Galáxias/QSOs tomados por Estrelas.





Random Forest

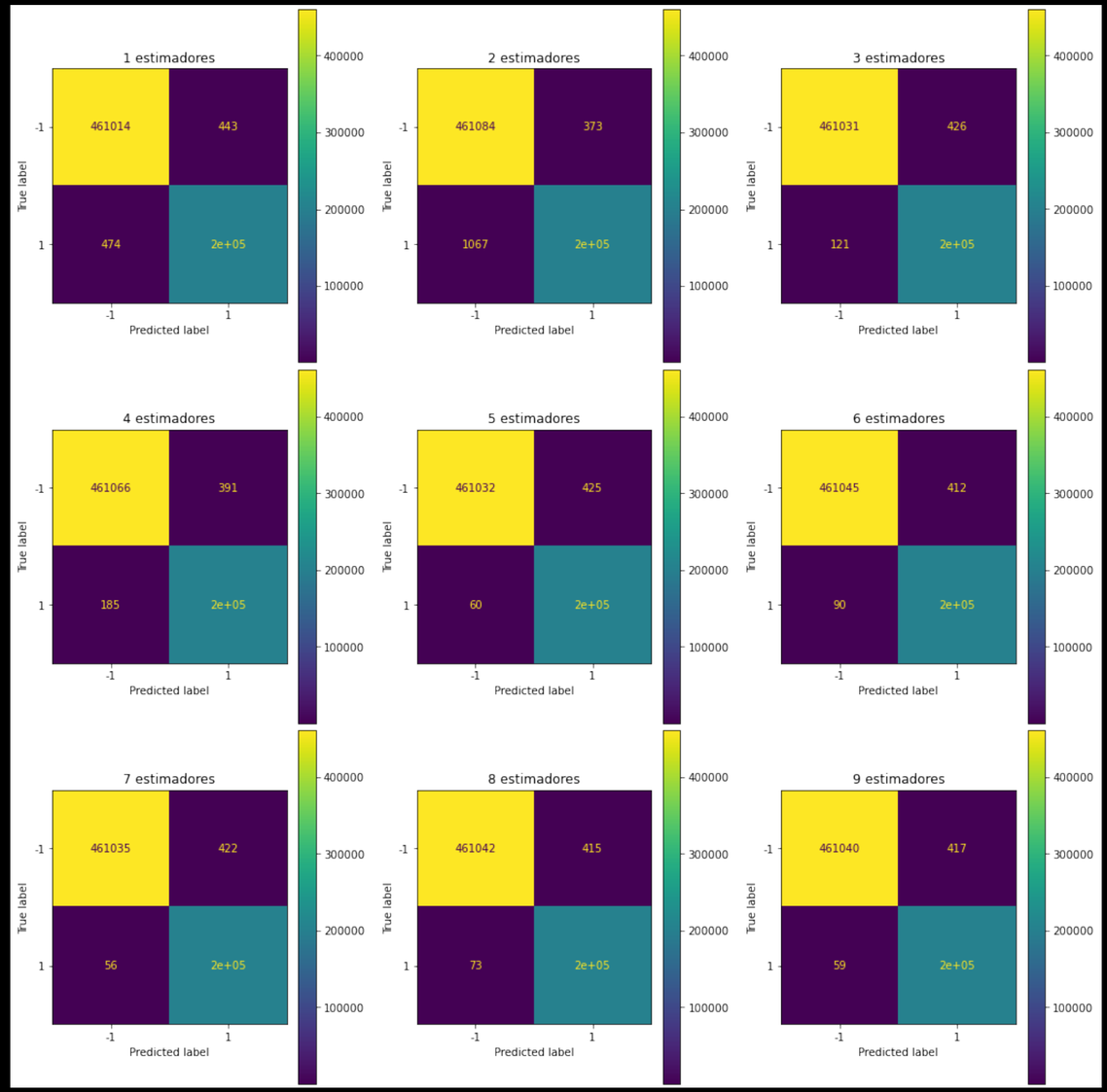
- São modelos que aplicam várias árvores de decisão para classificar o alvo.
- As árvores são criadas a partir de diferentes subconjuntos dos dados de treinamento, de modo a maximizar a variedade de árvores.
- No Scikit-Learn, a classificação final é obtida pela média das probabilidades de classificação dadas por todas as árvores.



[https://commons.wikimedia.org/wiki/File:Decision_Tree_vs._Random_Forest.png]

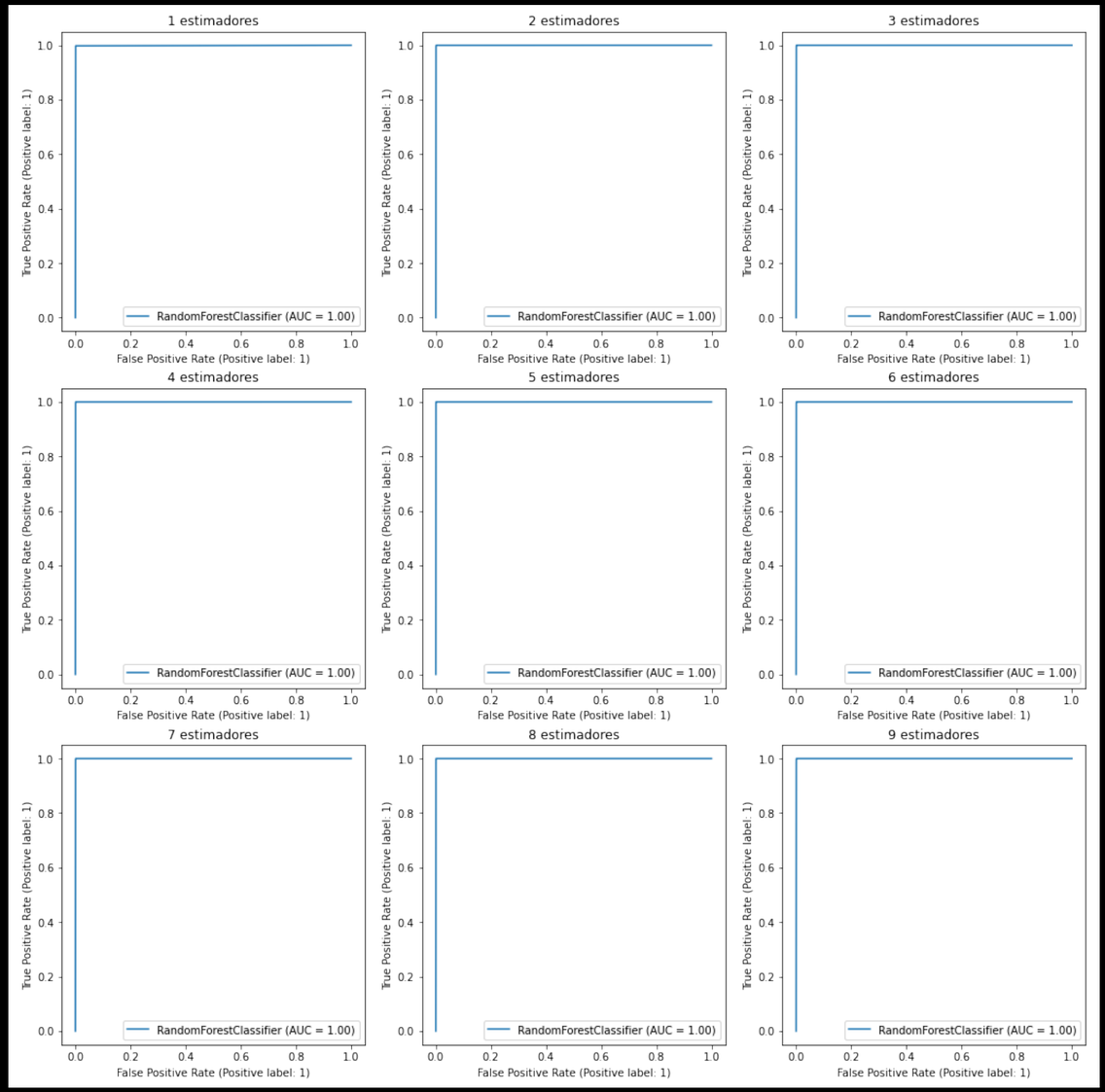
Random Forest

- Embora a taxa de galáxias/quasares (ID = -1) identificados como estrela (ID = 1) tenha se mantido aproximadamente constante, com o aumento do número de árvores há uma queda acentuada no número de estrelas identificadas como galáxias/quasares.



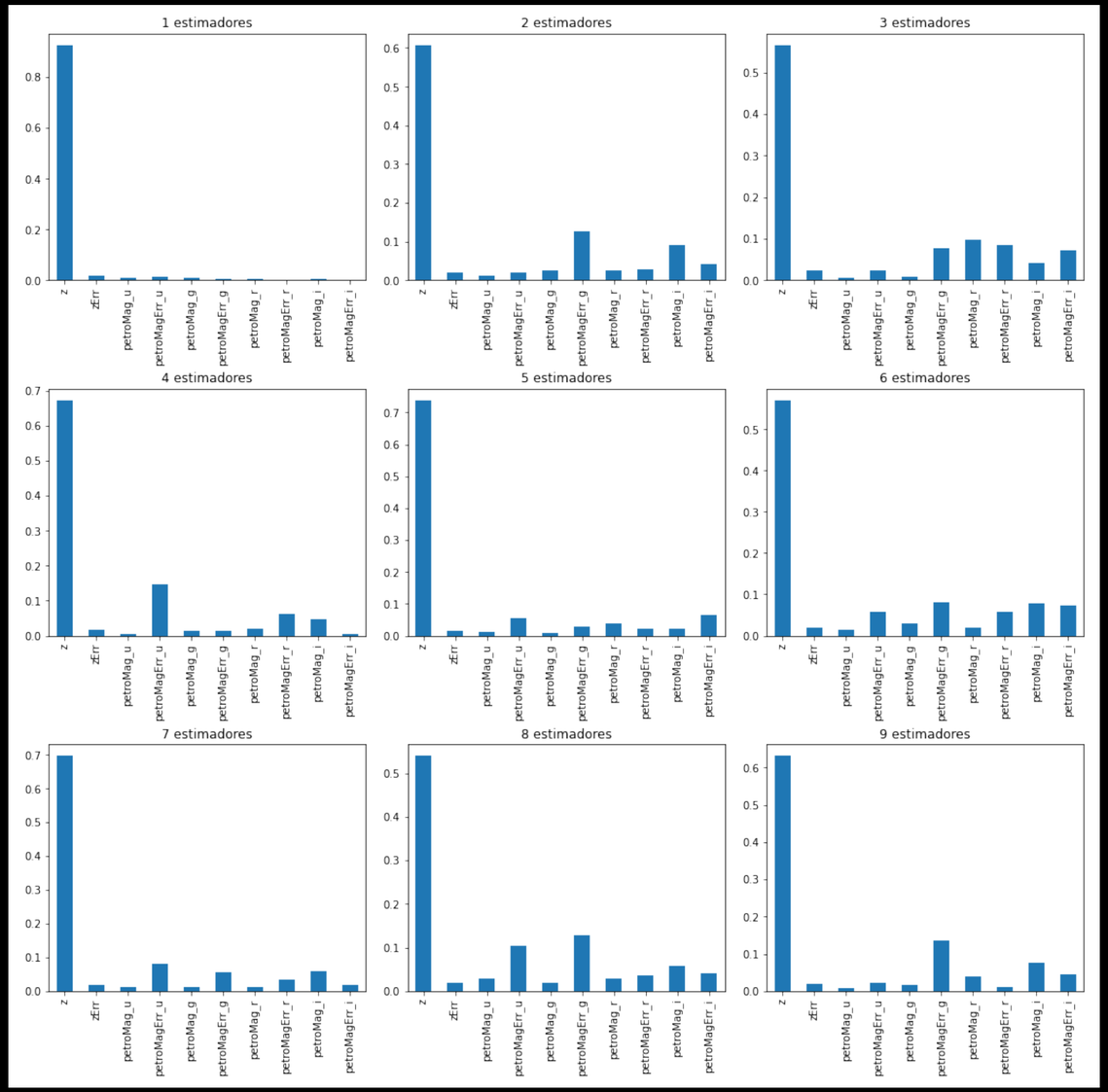
Random Forest

- A curva ROC ficou igual, perfeita, para todos os classificadores.

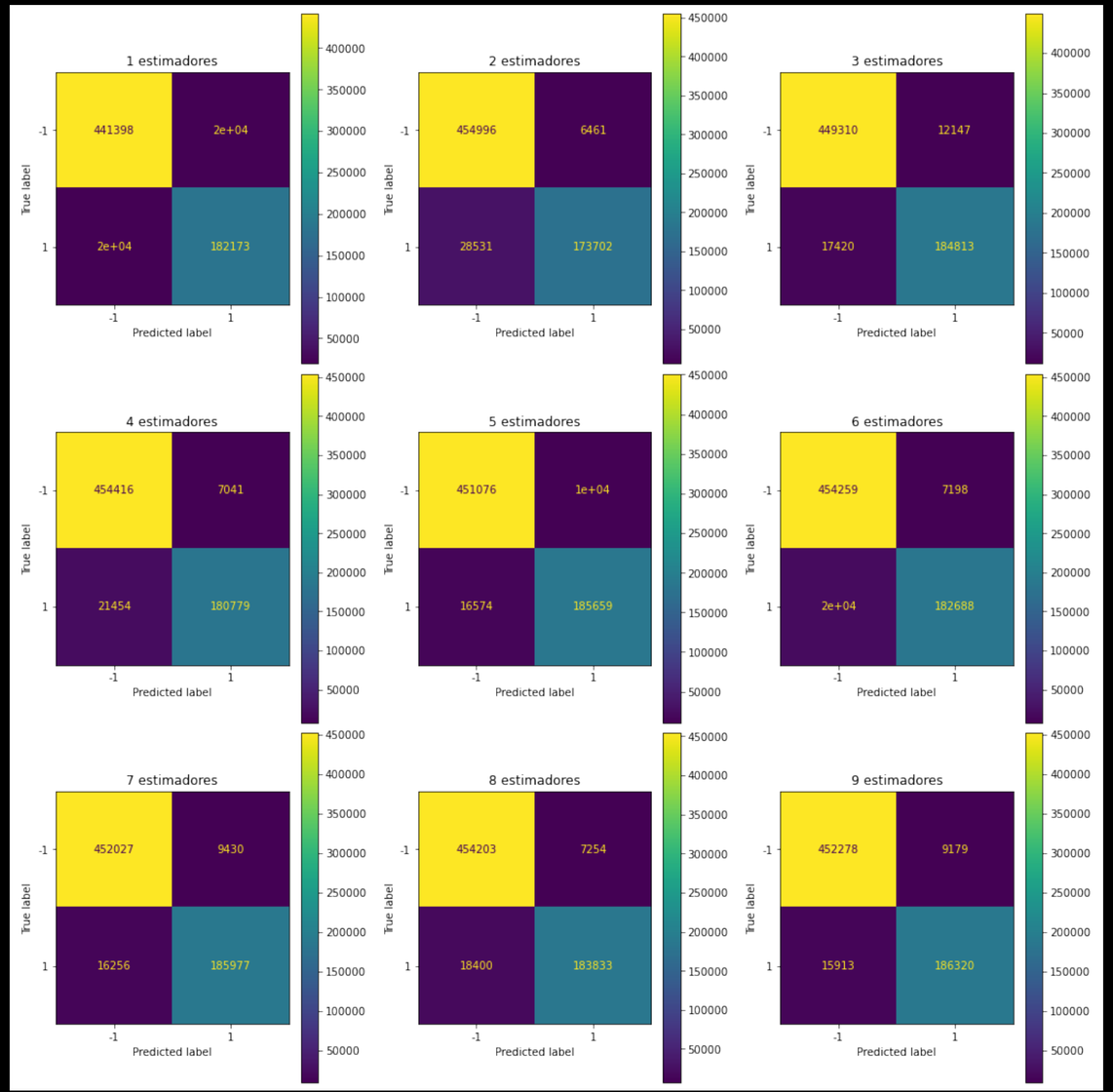


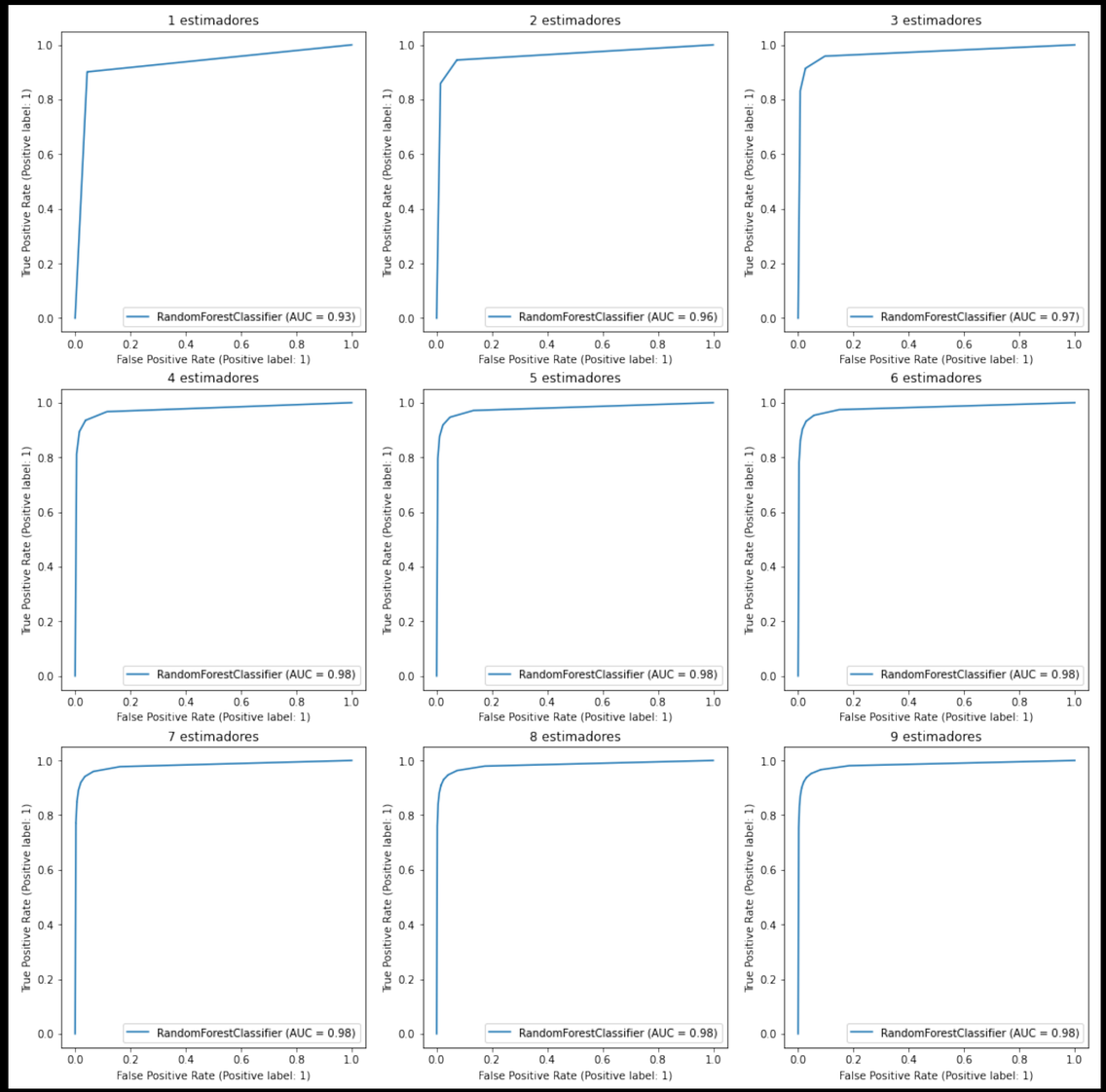
Random Forest

- Ao analisarmos o peso de cada informação na classificação, fica evidente a importância de z.
- Apenas por curiosidade, testamos treinar os classificadores sem a informação z…



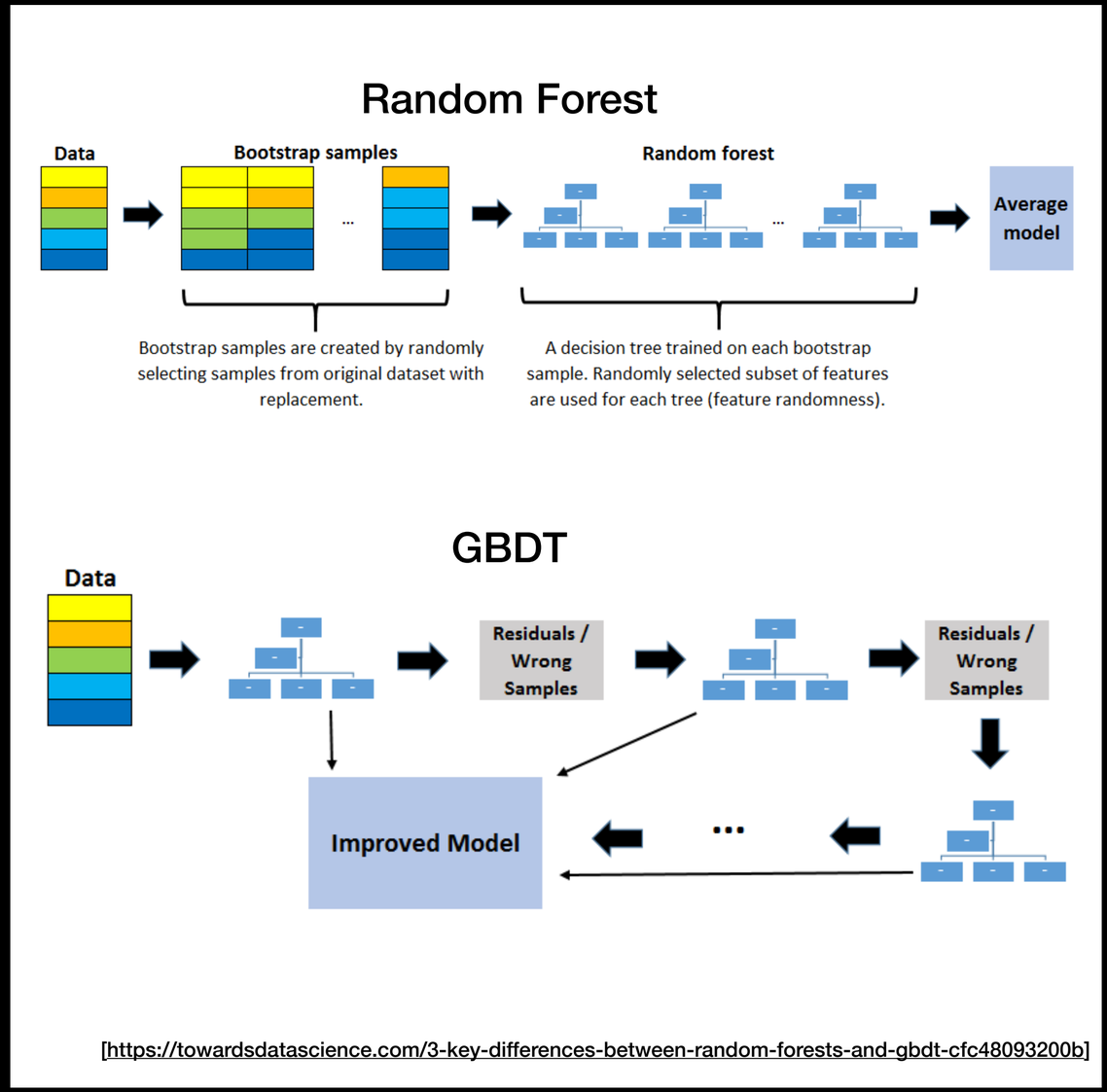
Random Forest (sem z)





LightGBM

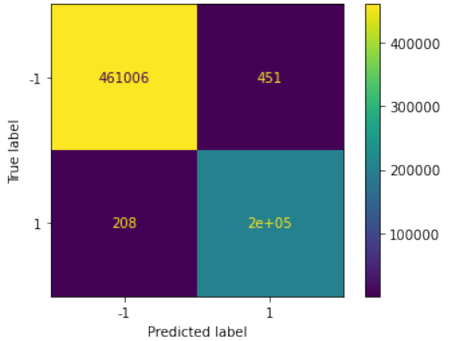
- Modelo do tipo Gradient Boosting Decision Tree: a cada ciclo de treinamento são criadas novas árvores para minimizar o erro (gradient) remanescente do ciclo anterior (boosting).

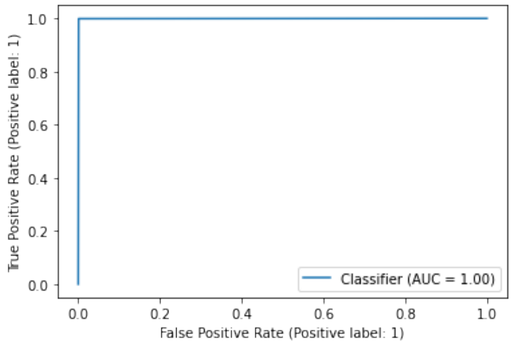


[https://towardsdatascience.com/3-key-differences-between-random-forests-and-gbdt-cfc48093200b]

LightGBM

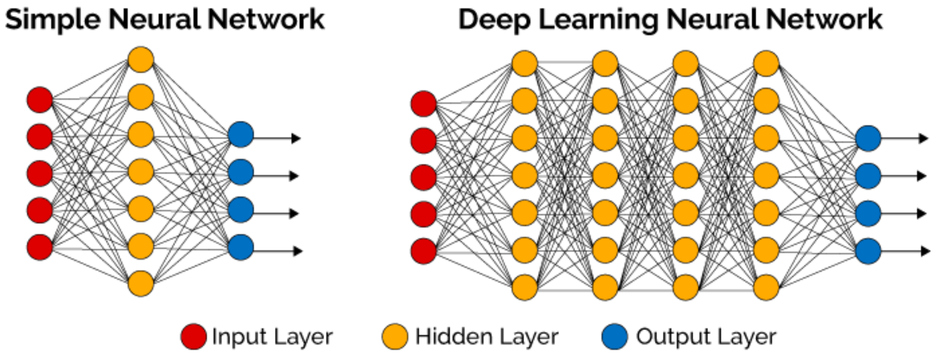
- O classificador LightGBM trouxe resultados semelhantes aos obtidos com Random Forest.





TabNet

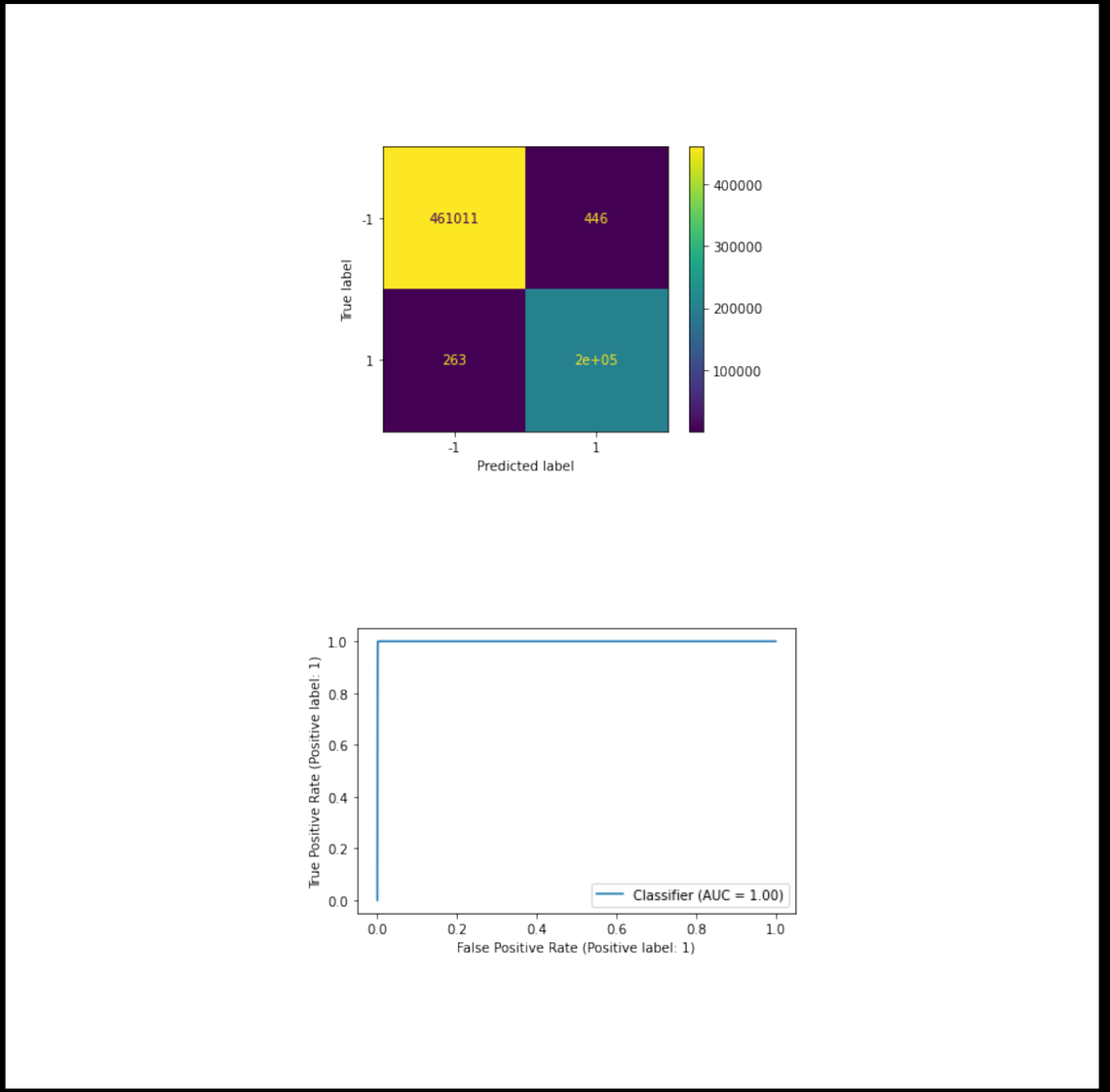
- Rede neural profunda para dados tabulares.
- Alguns testes indicam uma performance superior aos métodos baseados em árvores de decisão.



[https://www.deeplearningbook.com.br/o-que-sao-redes-neurais-artificiais-profundas/]

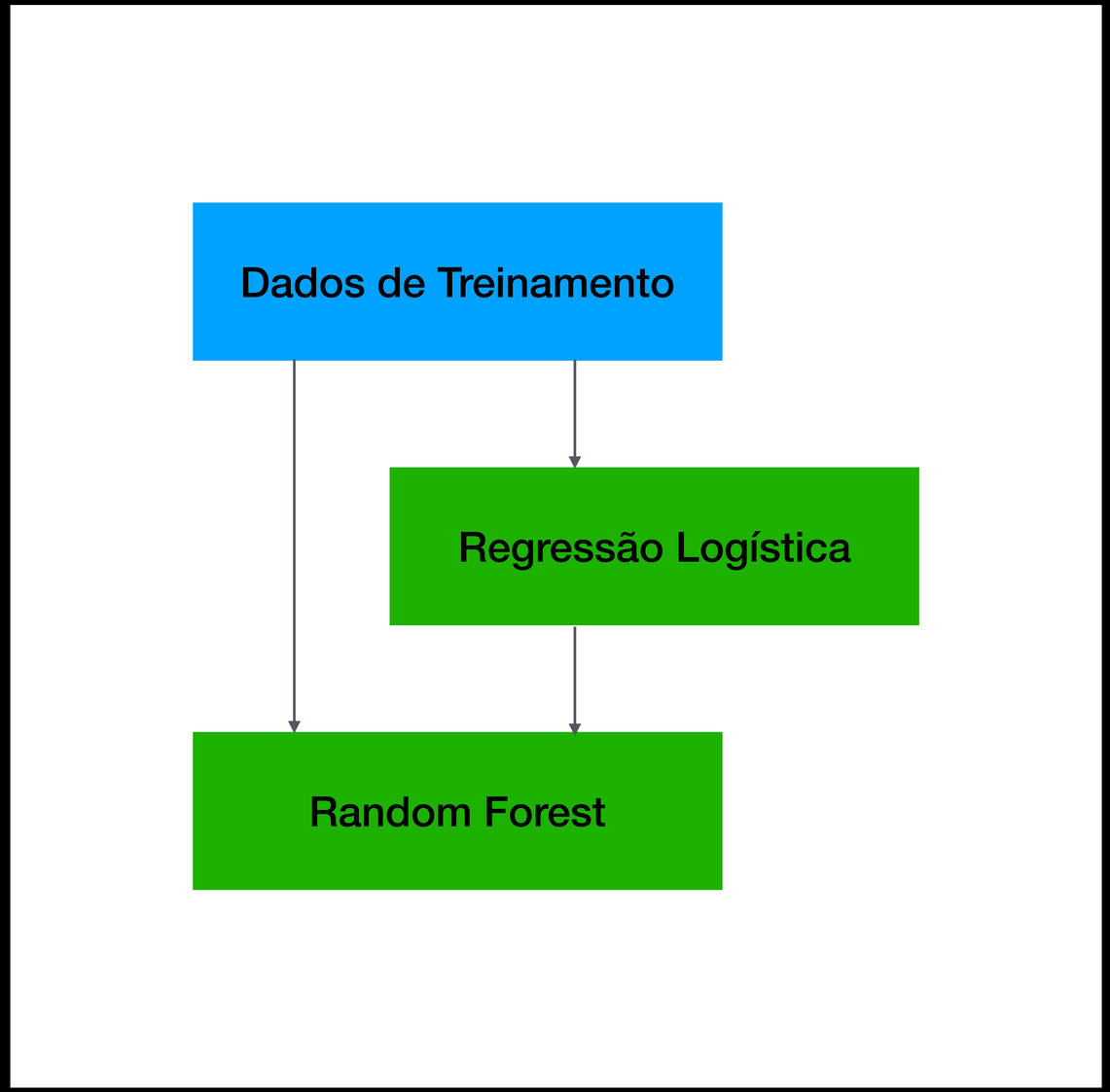
TabNet

- Os melhores resultados da rede neural TabNet foram bem semelhantes ao das Random Forests. Entretanto, a cada execução do notebook o resultado da TabNet variou bastante, mesmo mantendo o seed constante. Provavelmente por um novo sorteio do conjunto de treinamento.



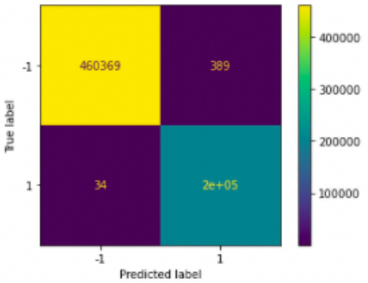
Modelo composto

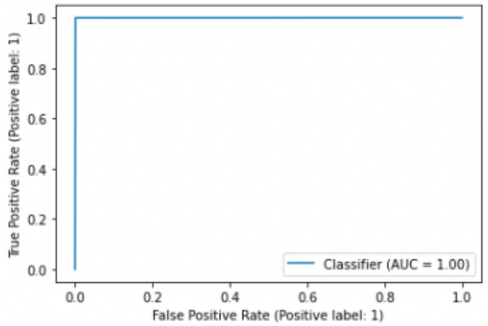
- Tentamos um modelo que usava a saída do modelo Logistic Regression como alimentação para uma Random Forest.



Modelo composto

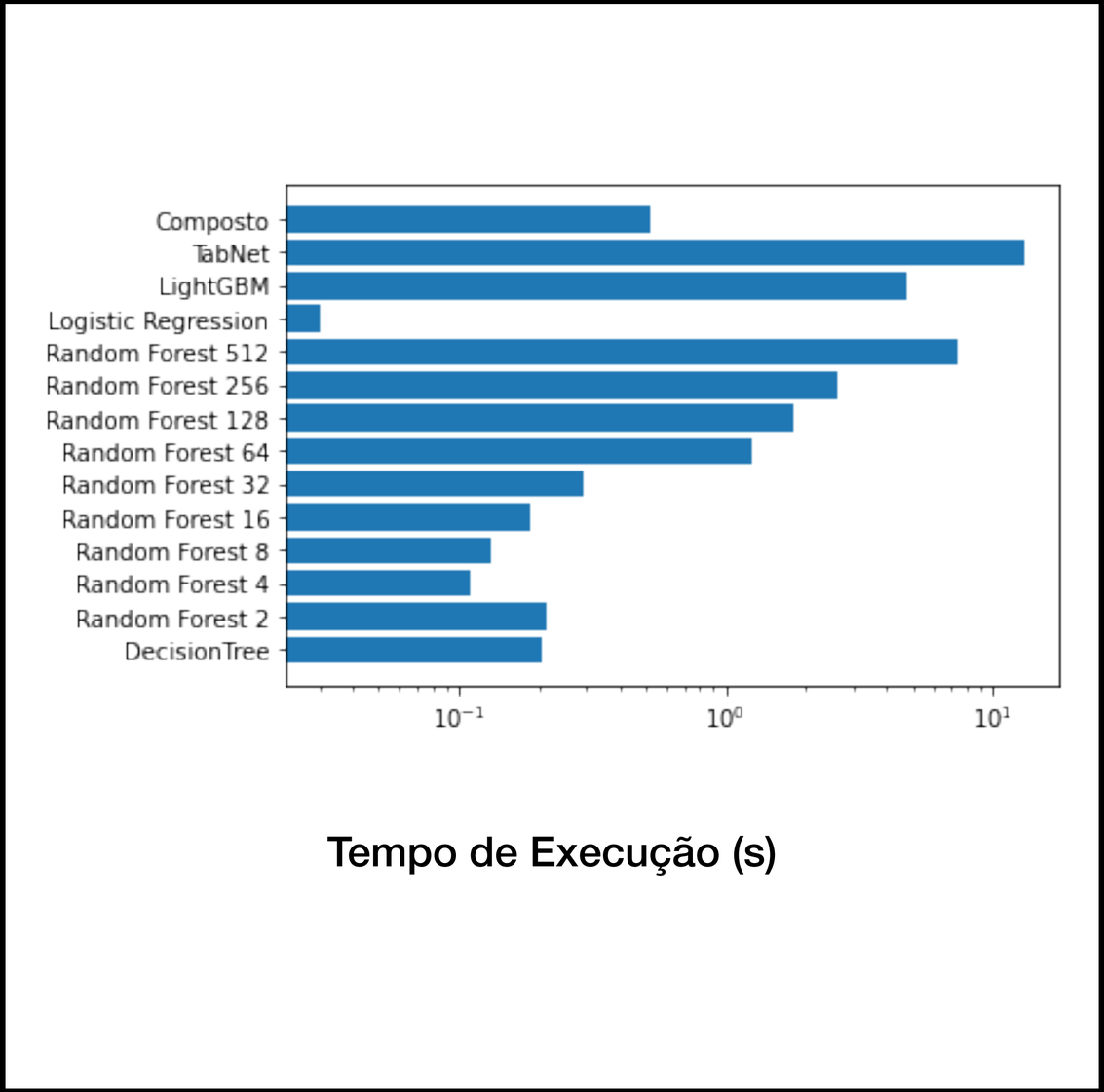
- Infelizmente o resultado foi equivalente ao da Random Forest pura.





Velocidade

- O tempo de execução dos diversos classificadores foi bastante distinto, com uma enorme diferença entre o método mais lento (TabNet) e o mais rápido (Logistic Regression).
- Note-se que não foi possível utilizar uma GPU no teste da TabNet, o que poderia ter melhorado substancialmente sua velocidade.



Dados Experimentais

- Podemos observar que o número de Falsos Positivos se manteve aproximadamente constate para diversos modelos.
- As Random Forest de 16 e 32 árvores parecem ter um melhor equilíbrio entre velocidade e recall.

	FP	FN	Tempo
DecisionTree	390	491	0.20383229199796915
Random Forest 2	350	991	0.21221791597781703
Random Forest 4	379	162	0.11020874997484498
Random Forest 8	380	74	0.1308629999984987
Random Forest 16	387	49	0.18594116700114682
Random Forest 32	391	41	0.29546383302658796
Random Forest 64	393	34	1.2564960420131683
Random Forest 128	396	35	1.7835329590016045
Random Forest 256	396	33	2.615519624989247
Random Forest 512	401	29	7.364277708024019
Logistic Regression	2977	0	0.030547125003067777
LightGBM	409	196	4.745788125001127
TabNet	355	0	13.067160374979721
Composto	389	34	0.5236535420117434

Conclusão

- Felizmente verificamos uma alta taxa de acertos em todos os métodos testados, bem como boa velocidade em quase todos.
- Em velocidade e precisão, podemos considerar que as Random Forest tiveram o melhor resultado.
- Mesmo que uma eventual execução em GPU melhorasse seu tempo, a biblioteca TabNet apresentou altíssima variância, produzindo resultados muito díspares a cada treinamento.
- Tentamos investigar um modelo composto, na esperança de que este aproveitasse o alto recall da Logistic Regression e a boa precisão das Random Forest. Infelizmente, o resultado não foi melhor do que o das Random Forest puras.

Bibliografia

1. Scikit-Learn User Guide - https://scikit-learn.org/stable/user_guide.html
2. LightGBM - https://lightgbm.readthedocs.io
3. Ke, G. et al. - LightGBM: A Highly Efficient Gradient Boosting Decision Tree - Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3149-3157 - https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
4. Shwartz-Ziv, R., Armon, A. - Tabular Data: Deep Learning is Not All You Need - https://arxiv.org/pdf/2106.03253.pdf
5. Ellis, C. - TabNet: simple binary classification example - https://www.kaggle.com/carlmcbrideellis/tabnet-simple-binary-classification-example
6. Carvalho, E. - TabNet : Attentive Interpretable Tabular Learning - https://github.com/dreamquark-ai/tabnet/blob/develop/README.md
7. Hjelle, G. - Python Timer Functions: Three Ways to Monitor Your Code - https://realpython.com/python-timer/