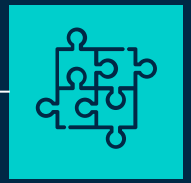
Projeto Final

CIÊNCIA DE DADOS PARA SEGURANÇA

Dante Aléo
Marcos Vinicius Pontarolo

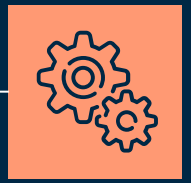
Índice



01

DATASET

- exploração



02

Machine Learning

KNN
Random Forest
MLP

03

Conclusão

DDoS Evaluation Dataset

- DDoS
- Exploração Dataset
- Tentar identificar um ataque DDoS no tráfego de rede.
- Tratamento do Dataset
- Rotulação do Dataset
- Classificação do Dataset

Repartição do Dataset

Dataset Completo:

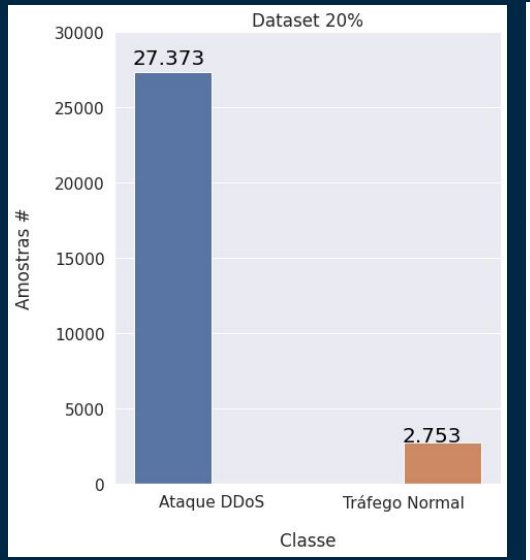
- 137.006 Tráfego de Rede DDoS.
- 13.623 Tráfego de Rede Normal.

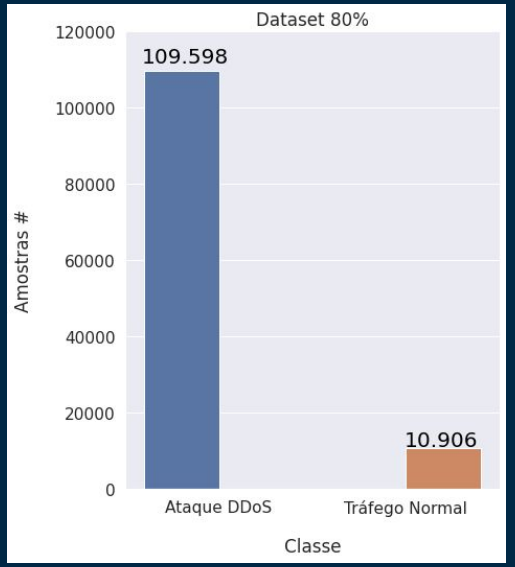
Porção de teste:

- 27.373 Tráfego de Rede DDoS.
- 2.753 Tráfego de rede Normal.

Porção de treino:

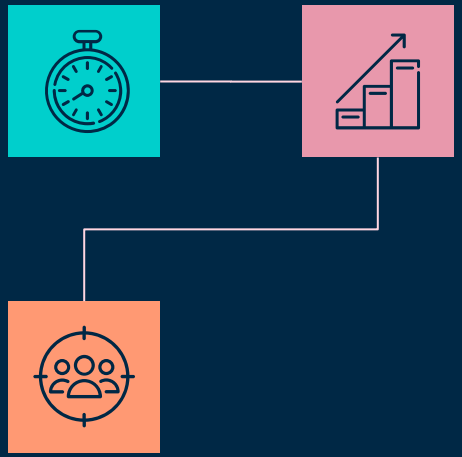
- 109.598 Tráfego de rede DDoS.
- 10.906 Tráfego de rede Normal





Modelos de Machine Learning

KNN
Algoritimo

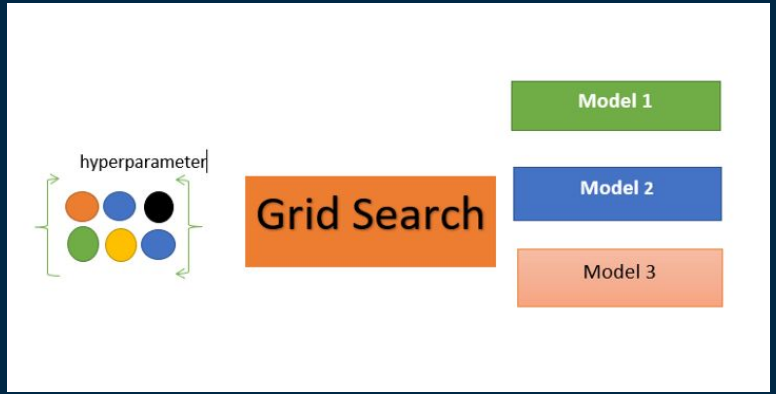


Random Forest
Algoritimo

MLP
Rede neural

Tuning, modificação de hiperparâmetros

- Testes individuais
- Filtro Grid Search
- Grid Search completo
- Latência
- Resultados incluindo os melhores parâmetros.

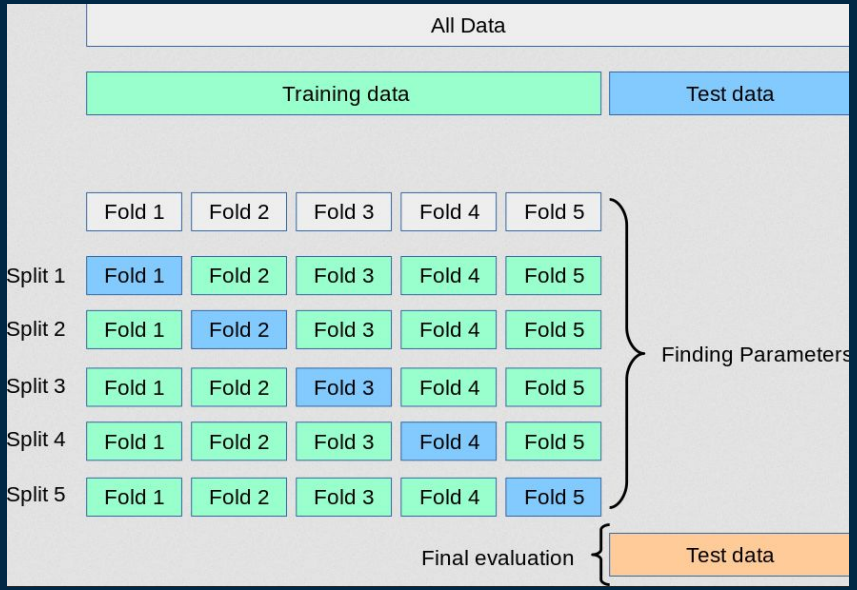


Dataset desbalanceado

- F1Score
- Precision Recall Curve - P/R ao invés de ROC

Validação

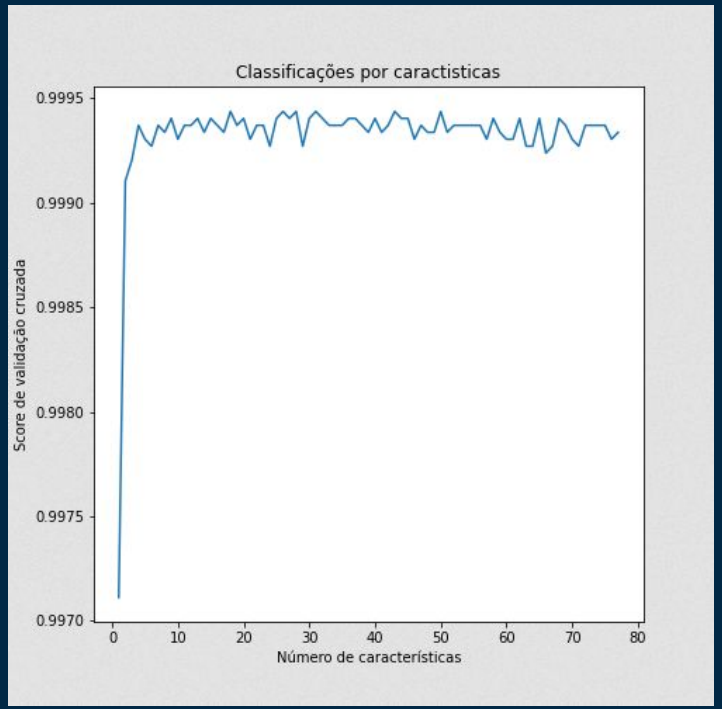
- K-fold Cross Validation



Resultados promissores

- Classificação binária
- Ataque DDoS ou tráfego normal
- Rotulação dos tipos de ataques
- Readequar dataset completo para balancear
- Agrupamento de tráfego de rede

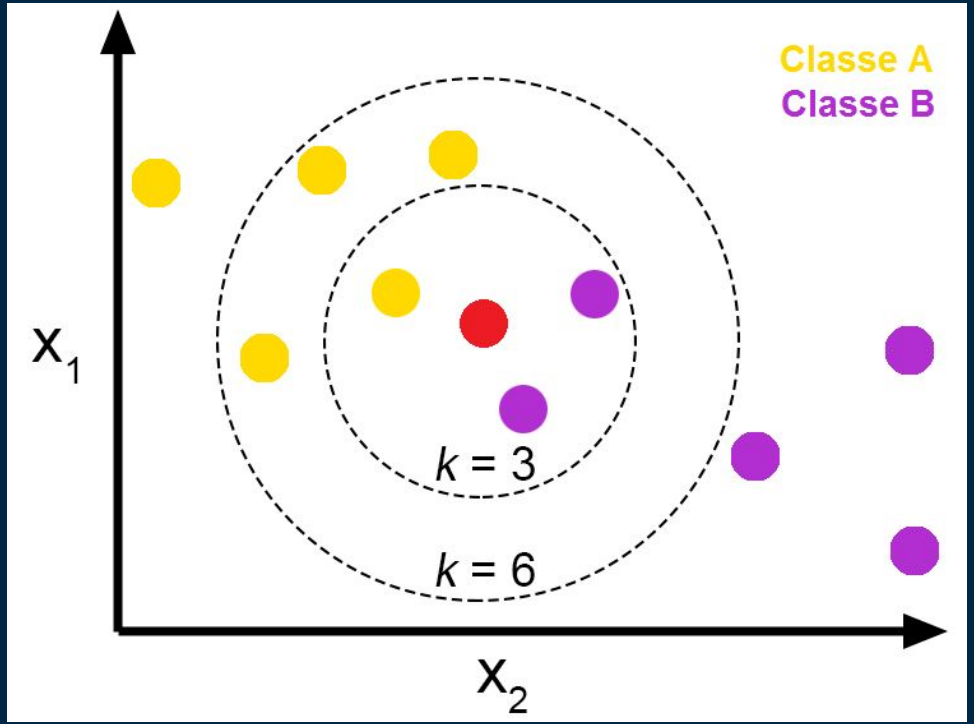
Decision Tree classificando características



- Premissa inicial
- Cerca de 20-30 características

KNN - K-nearest neighbors

- Espaço de características próximas umas das outros
- Mesma rotulação de ataque



Configurações e parâmetros - KNN

Parâmeros:

- K = quanto maior o K pior será a acurácia do nosso modelo de KNN
- Weights - Uniform e Distance
- Algorithm (baseado em outros 2 parâmetros)
- Metric - minkowski, euclidean e manhattan
- 5 K fold cross validation nos parâmetros do **Grid Search.**

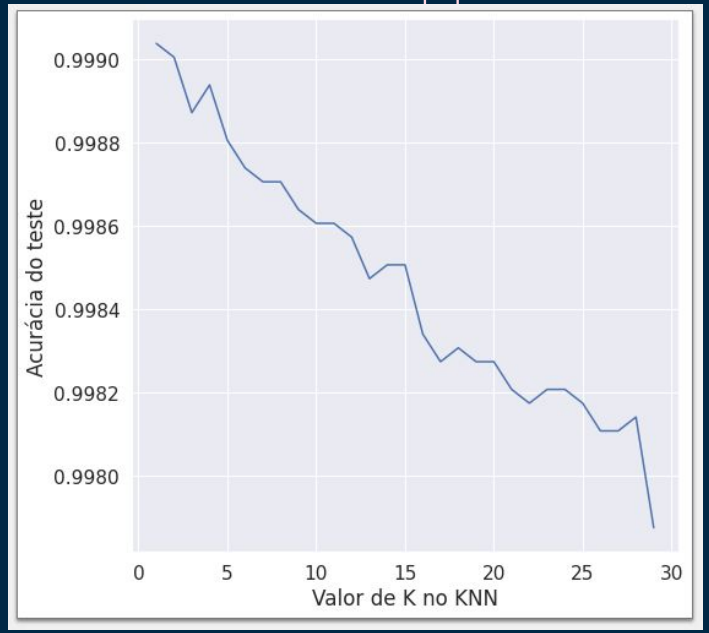
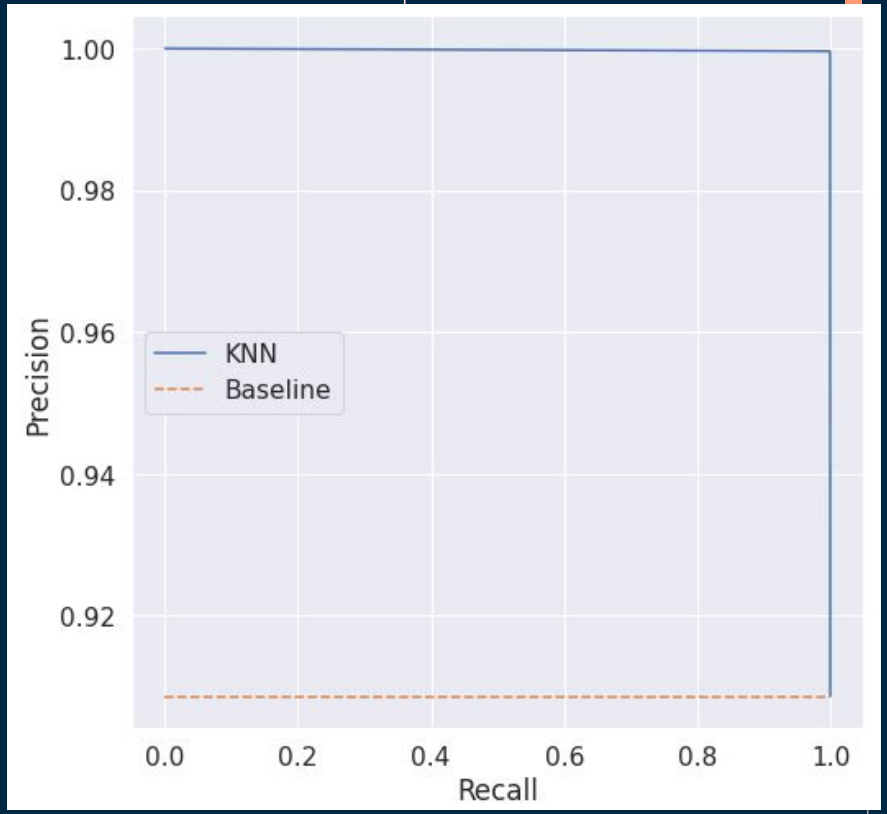


Gráfico P/R - KNN

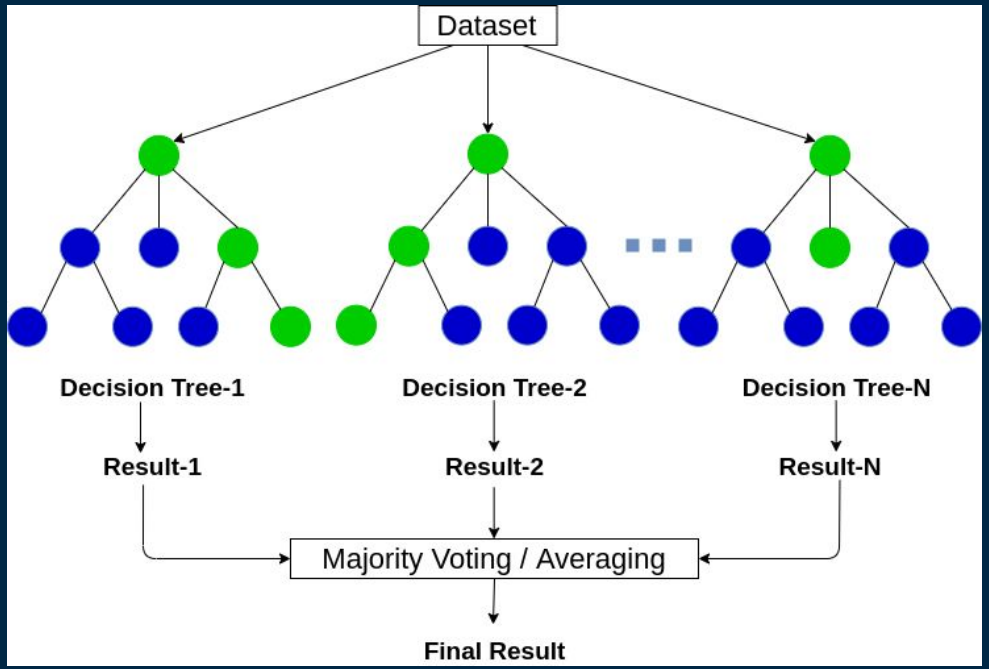


Conclusão - KNN

- KNN utilizando as configurações padrões já é bem preciso
- Com tuning obtivemos um desempenho um pouco melhor

Random Forest

- Problemas de classificação ou regressão
- Faz sentido com várias features



Configurações e parâmetros – Random Forest

- Inicialmente foram usadas configurações padrão
- N_estimators
- N final -> 650
- Max Features
- Min samples split y
- Max Depth z
- Grid Search resultados =
  - Max Depth = 25, Max Features = 50, Min Samples Split = 3, N Estimators = 200

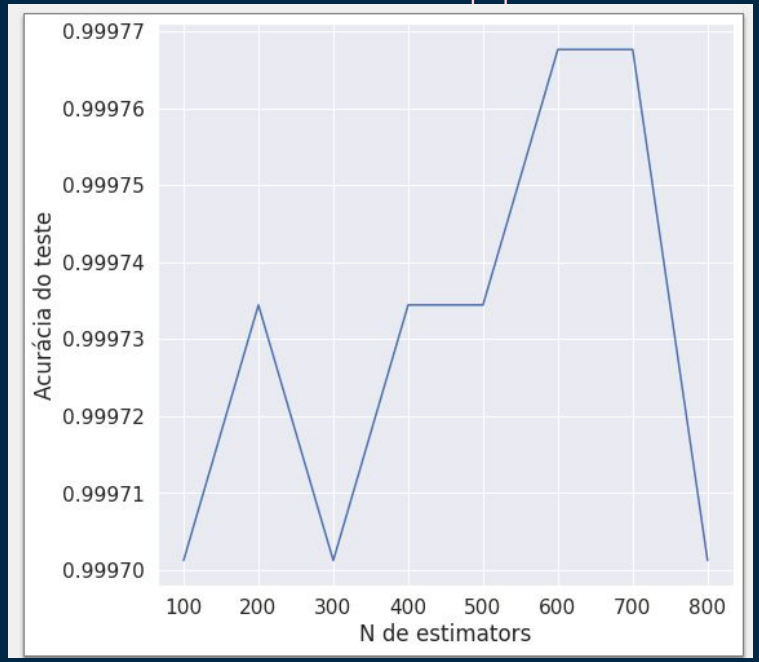
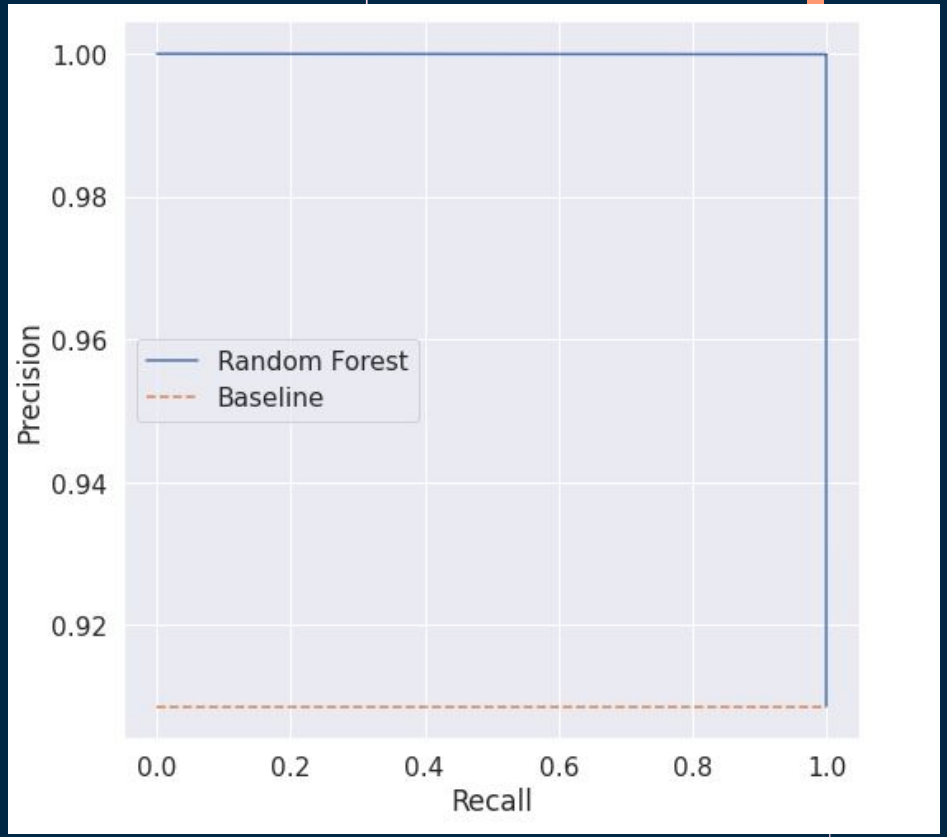


Gráfico P/R - Random Forest

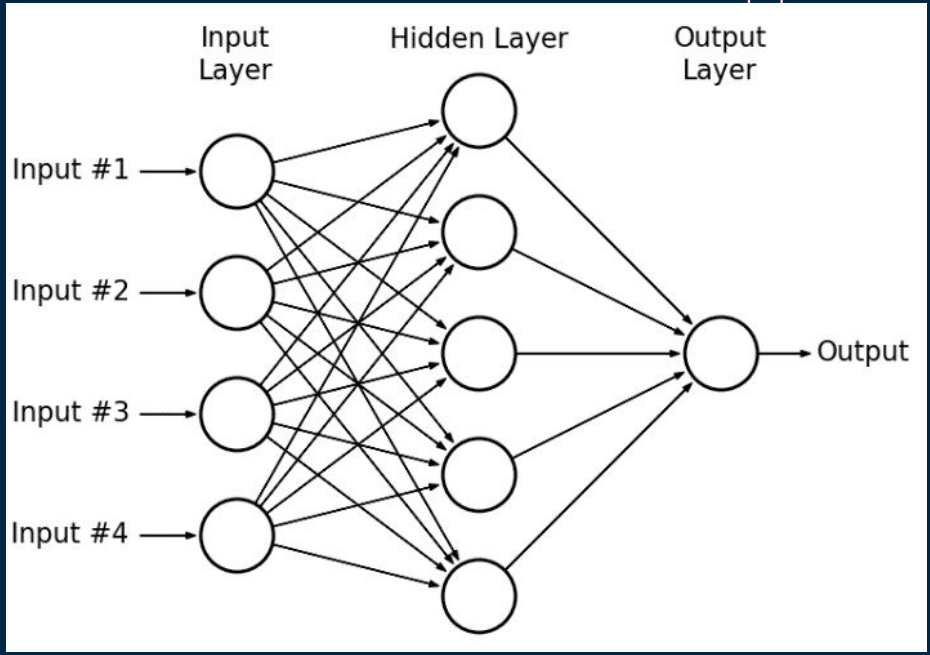


Conclusão - Randon Forest

- Modelo muito bom para nosso dataset
- Houve uma surpresa referente ao número de features

MPL – Multilayer perceptron

- Rede neural
- Ackpropagation/Backpropagation



Configurações e parâmetros- MPL

- Inicialmente foram usadas as configurações padrão
- Max iter -> mais iterações, maior acurácia
- hidden layer sizes
- Acivation (funções tahn e relu)
- Solver (sgd e adam para testes)
- Alpha -> testado com 0.001 e 0.05
- Learning Rate (Constant + Adaptative)

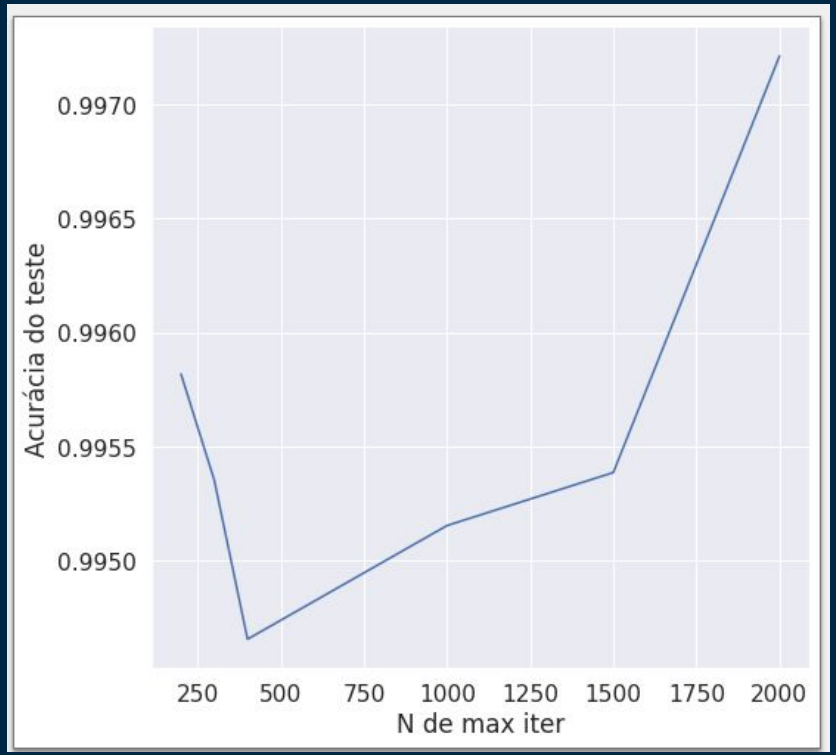
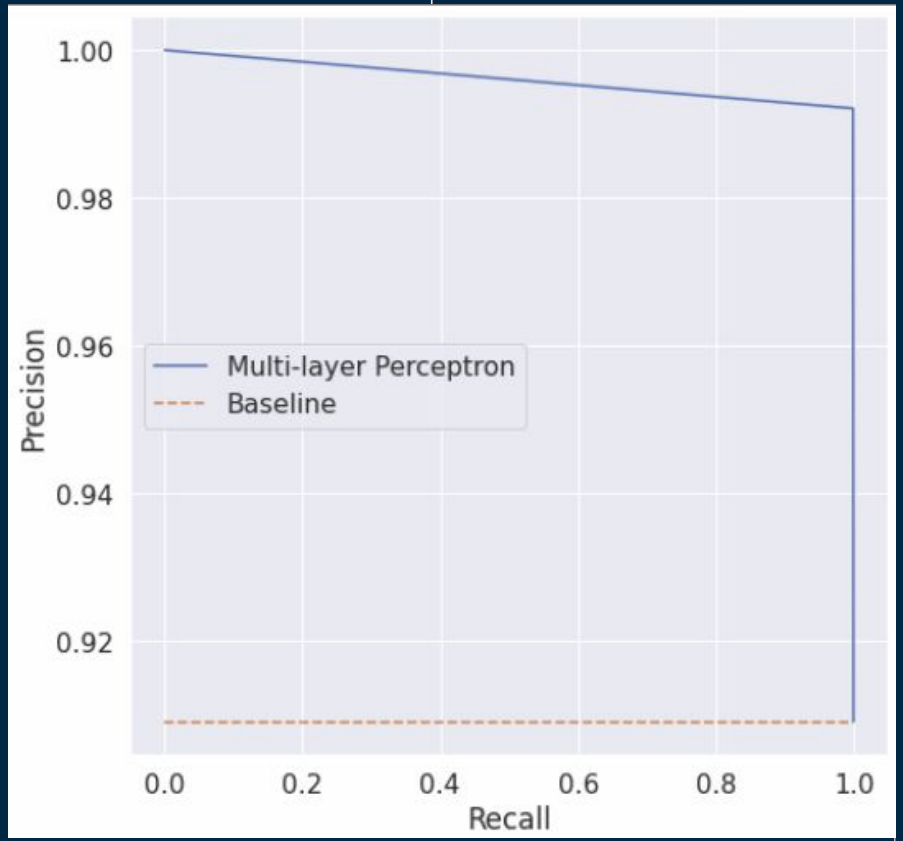


Gráfico P/R - MLP



Conclusão - MLP

- Um ótimo modelo para o nosso dataset
- Os parâmetros otimizados usando Grid Search foram promissores
- Um tempo maior para ajustes
- Curva P/R Diferente

Métricas

	PRECISÃO	RECALL	F1SCORE
KNN	0.999561	0.999487	0.999524
Random Forest	0.999963	0.999817	0.999890
MLP	0.992061	0.999306	0.995671

Métricas

	PRECISÃO	RECALL	F1SCORE
KNN	0.999561	0.999487	0.999524
Random Forest	0.999963	0.999817	0.999890
MLP	0.992061	0.999306	0.995671

Métricas k-fold cross validation com 5 pastas

	PRECISÃO	RECALL	F1SCORE
KNN	0.999438	0.999540	0.999489
Random Forest	0.999934	0.999796	0.999865
MLP	0.992133	0.998256	0.995169

Conclusão:

- Random Forest foi o melhor!
- Classificação binária
- Boa classificação
- Ideias futuras

Reprodutibilidade

- Dataset
- Exploração
- Projeto Final
- Gráficos
- Links
- Comentários sobre os códigos

https://github.com/marcospontarolo/CI1030

Obrigado!