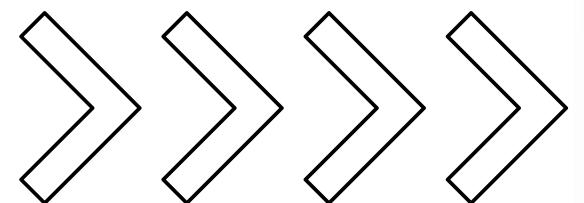


CLASSIFICAÇÃO E ANÁLISE

NOVO ENSINO SUPLEMENTAR

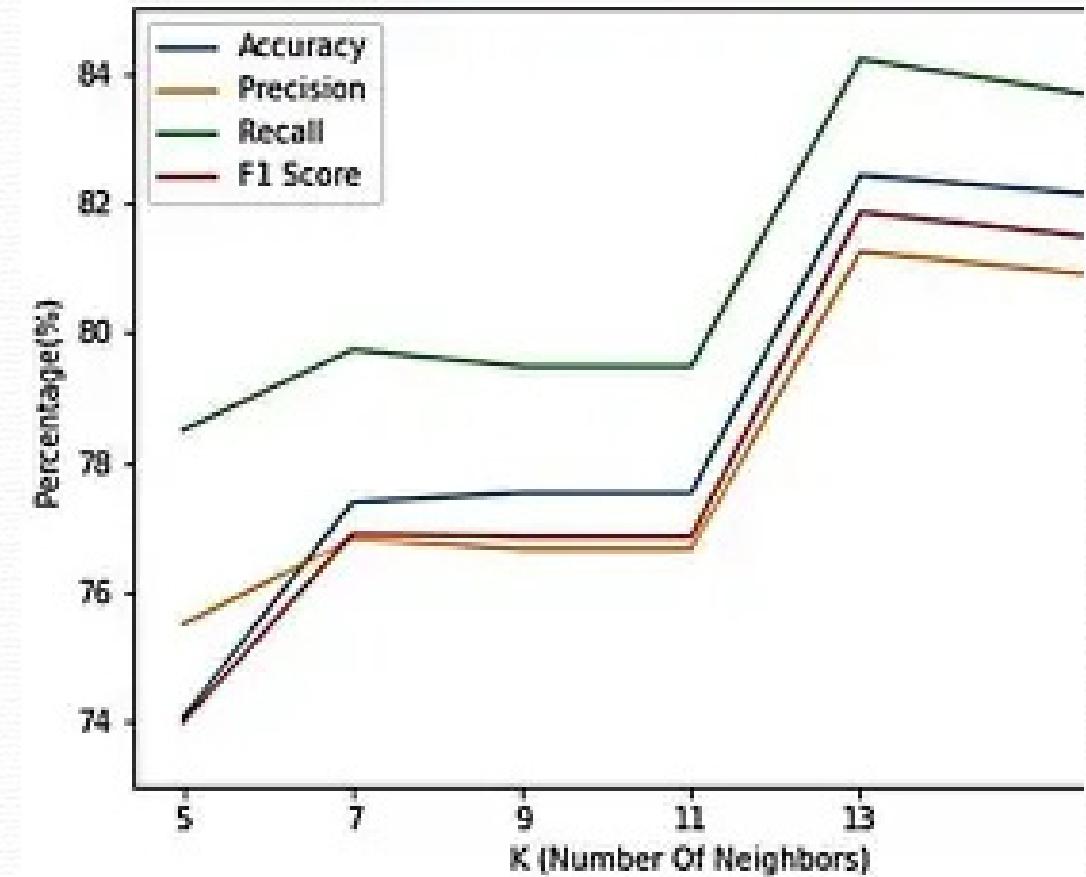
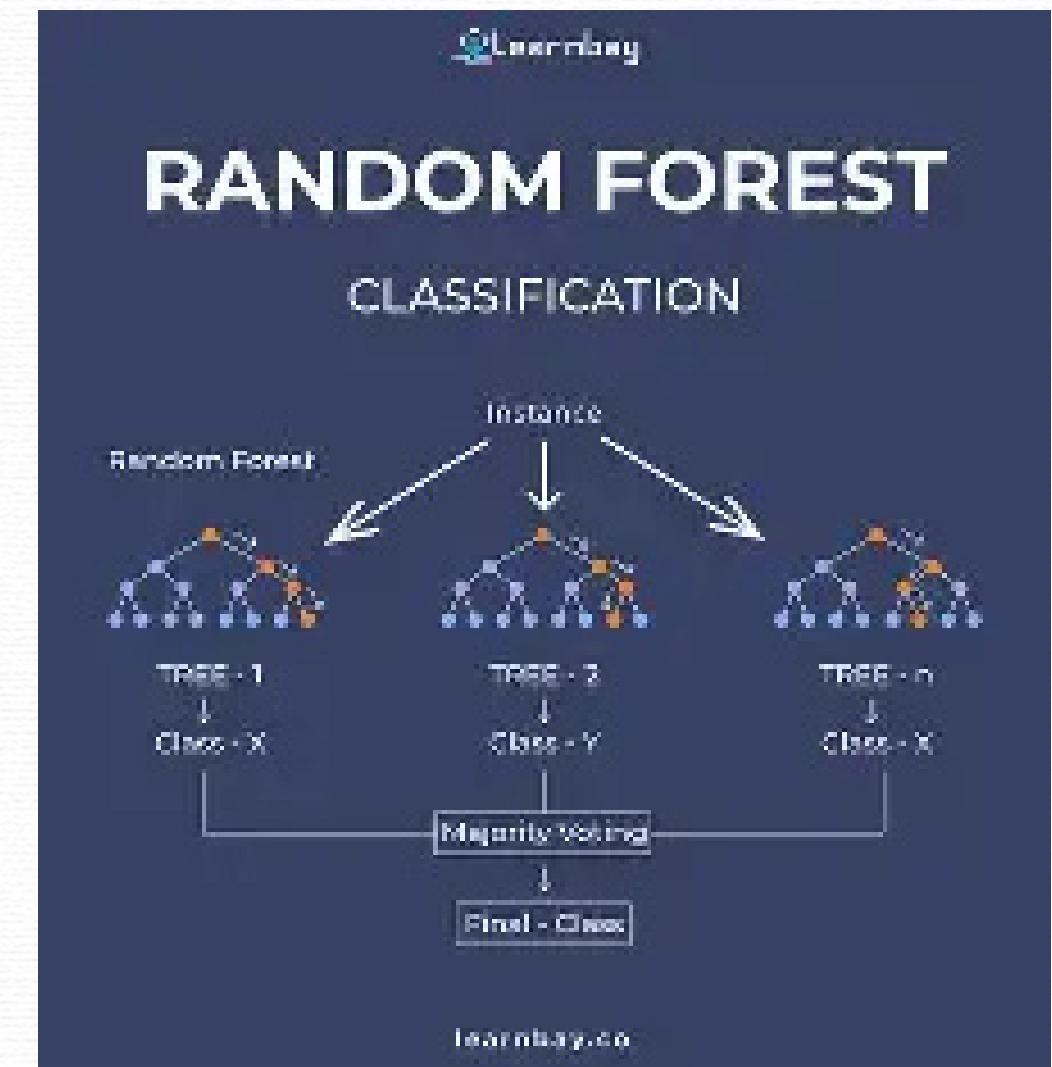


Aluno: João Pedro Pacheco Santos
Orientador: Lucas Lopes

O que vimos anteriormente:

Vimos que a base de dados utilizada já passou por tratamento e não contém dados nulos ou duplicados. A proporção de casos benignos e malignos é satisfatória para o aprendizado de máquina e contém muitos dados distintos e não-relacionados

Classificação



Por que RandomForest?

- ✓ Lida bem com muitas variáveis numéricas (como as 30 features do dataset).
- ✓ É robusto a overfitting, pois faz média de várias árvores (reduzindo erros individuais).
- ✓ Funciona bem mesmo sem ajuste fino, dando bons resultados iniciais.
- ✓ Mostra a importância das variáveis, ajudando a interpretar quais medidas influenciam mais no diagnóstico.
- ✓ Aguenta dados desbalanceados melhor que modelos mais simples.

Resultados:

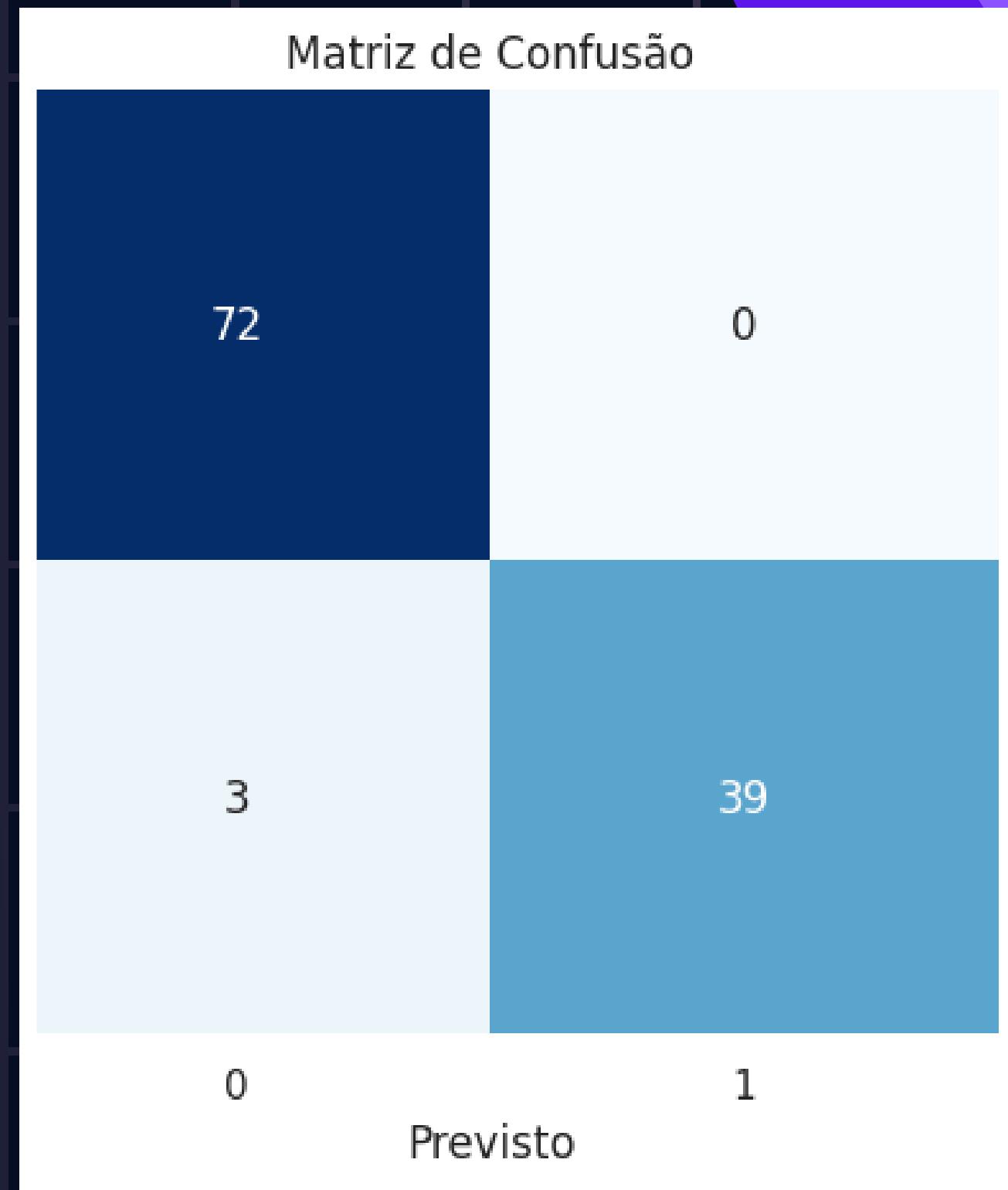
Acurácia: 0.9737

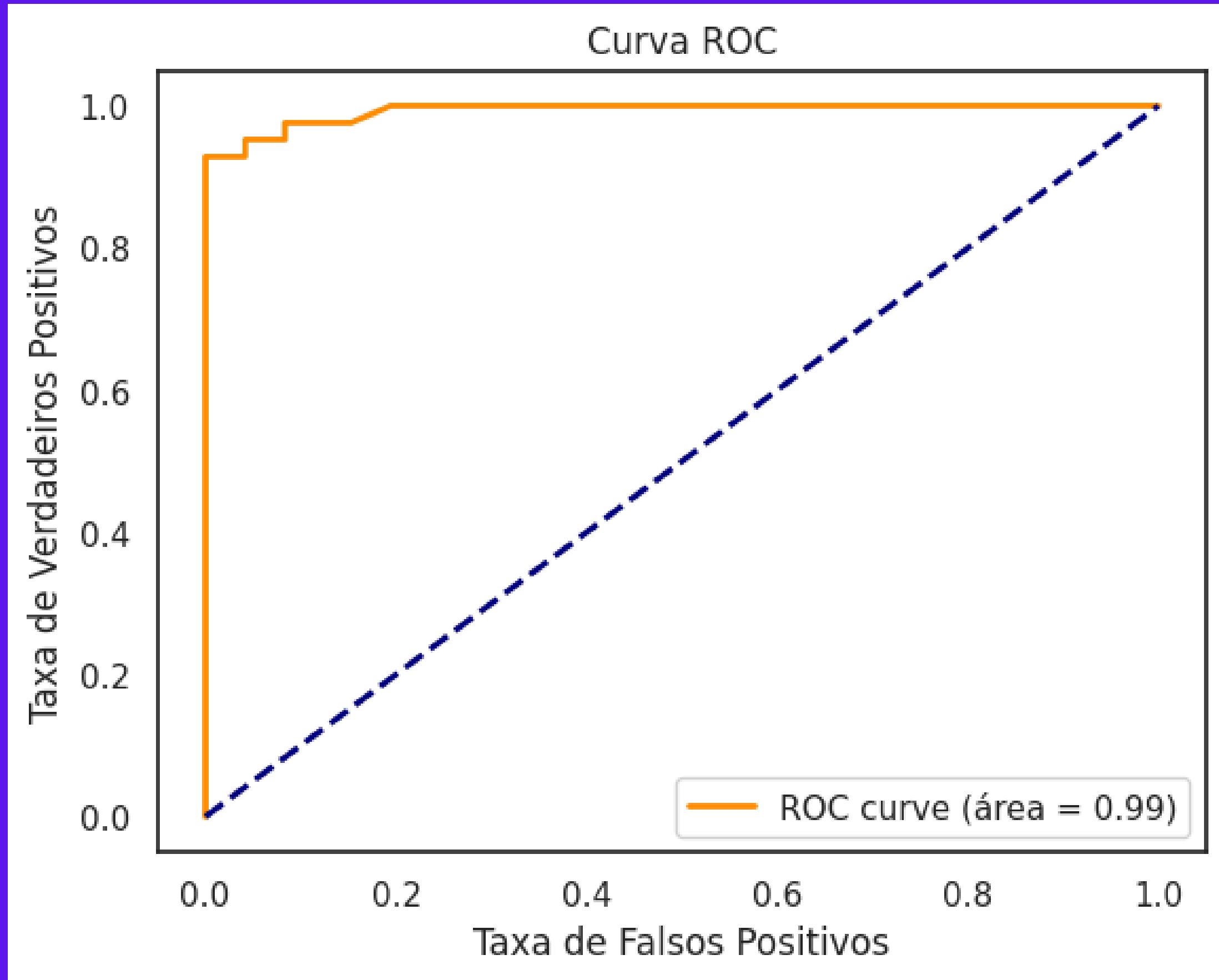
Recall: 0.9286

F1-score: 0.9630

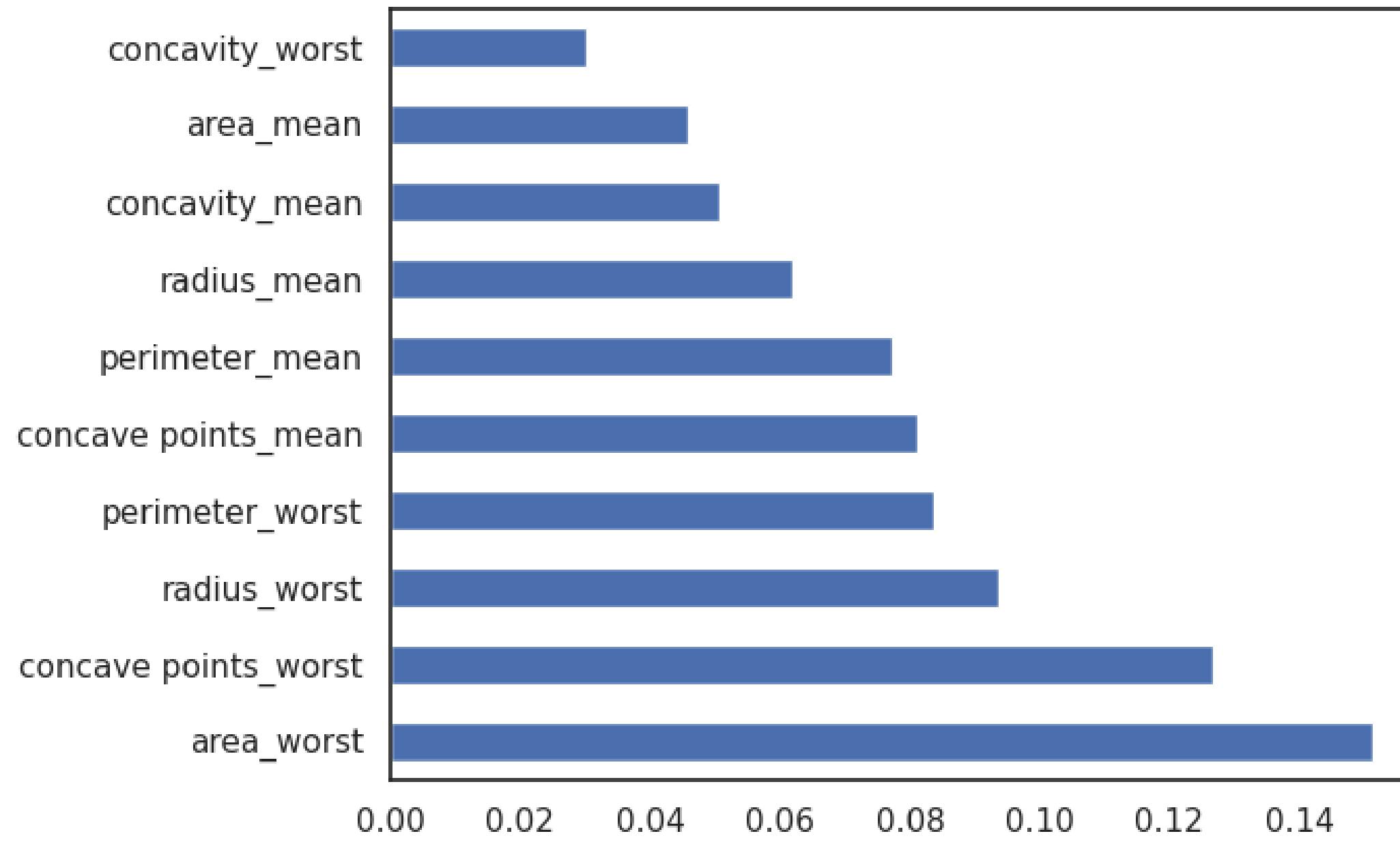
Precisão: 1.0000

Matriz de Confusão





Top 10 Features Importantes



Análise final

A base de dados é ótima para o aprendizado de máquina, pois tem taxas altas em todas as métricas, sem apresentar taxas muito distintas umas das outras.

Uma base de dados maior pode ser satisfatória para uma ampliação de casos.