OUTUBRO ROSA E CIÊNCIA DE DADOS: PREVENDO CÂNCER DE MAMA ATRAVÉS DE ALGORITIMOS DE MACHINE LEARNING

Internacional Comissária de Despachos Aduaneiros

Problema de negócio abordado

A base de dados trabalhada reúne informações de diversos exames de prevenção ao câncer de mama realizados por pacientes fictícios de um hospital igualmente hipotético[1].

Nela, distinguiram-se os resultados dos exames benignos dos malignos, sendo os primeiros o evento que monitoraremos através de modelagem em Data Science.

A necessidade da criação de um modelo de machine learning que auxilie os profissionais de saúde a detectarem o câncer é de suma importância pois o tempo médio para diagnóstico do câncer no Brasil é de 270 dias na rede pública de saúde. Essa demora faz com que cerca de 80% dos pacientes com algum tipo da doença comecem o tratamento em estágios mais avançados – portanto, com menores chances de cura.

O objetivo é tornar o processo de diagnóstico de câncer mais célere.

Base utilizada: https://www.dropbox.com/s/z8nw6pfumdw3bb9/breast-cancer-wisconsin.csv?raw=1

Características da base de dados

A base de dados possui 570 entradas e 11 features; serão aplicadas técnicas de amostragem e feature selection para fins de otimização do modelo.

As features da base são:

- **Diagnosis**: Diagnóstico se é um tumor maligno (M) ou benigno (B);
- **Radius**: Média da distância do centro até os pontos do contorno;
- **Texture**: A textura do núcleo da célula é medida pelo desvio padrão da intensidade da escala de cinza nos pixels componentes;
- **Perimeter**: Perímetro do núcleo da célula;
- **Area**: Área da superfície do núcleo da célula;
- **Smoothness**: A diferença entre o tamanho da linha do raio e a média das linhas ao redor do núcleo do tumor. Ou também a Variação local no comprimento de raio;

- **Compactness**: O perímetro e a área são combinados para dar uma medida compactness dos núcleos das células usando a fórmula **perímetro^2 / área**;
- **Concavity**: Mede a magnitude das concavidades do contorno;
- **Concave points**: Número de porções côncavas do contorno;
- **Symmetry:** Para medir a simetria, o eixo principal, ou corda mais longa através do centro, é encontrado. Em seguida, medimos a diferença de comprimento entre as linhas perpendiculares ao eixo principal até o limite nuclear em ambas as direções;
- **Fractal_dimension:** Traçar o log do perímetro observado contra o log do tamanho da régua e medir a inclinação descendente para encontrar uma aproximação da dimensão fractal;

Principais técnicas e métricas a serem utilizadas

Inicialmente, por se tratar de uma modelagem supervisionada com objetivo de detecção de resultado câncer mamário, vamos utilizar a RandomForestClassifier que é utilizado para tarefas de classificação e regressão.

Com relação ao tratamento das variáveis da base, utilizamos somente ‘drop’ para excluir colunas de valores nulos

Para monitorar o desempenho do modelo, utilizaremos como métricas o *F1-score*, o *ROC*, o *AUC*, *precision*, *recall*, e a acurácia; a escolha dessas métricas de monitoramento complementares à acurácia se deve ao fato desta última não ser métrica de desempenho suficientemente satisfatória para análise de modelos cujas bases sejam altamente desbalanceadas, como a do presente estudo.

Principais hipóteses a serem exploradas

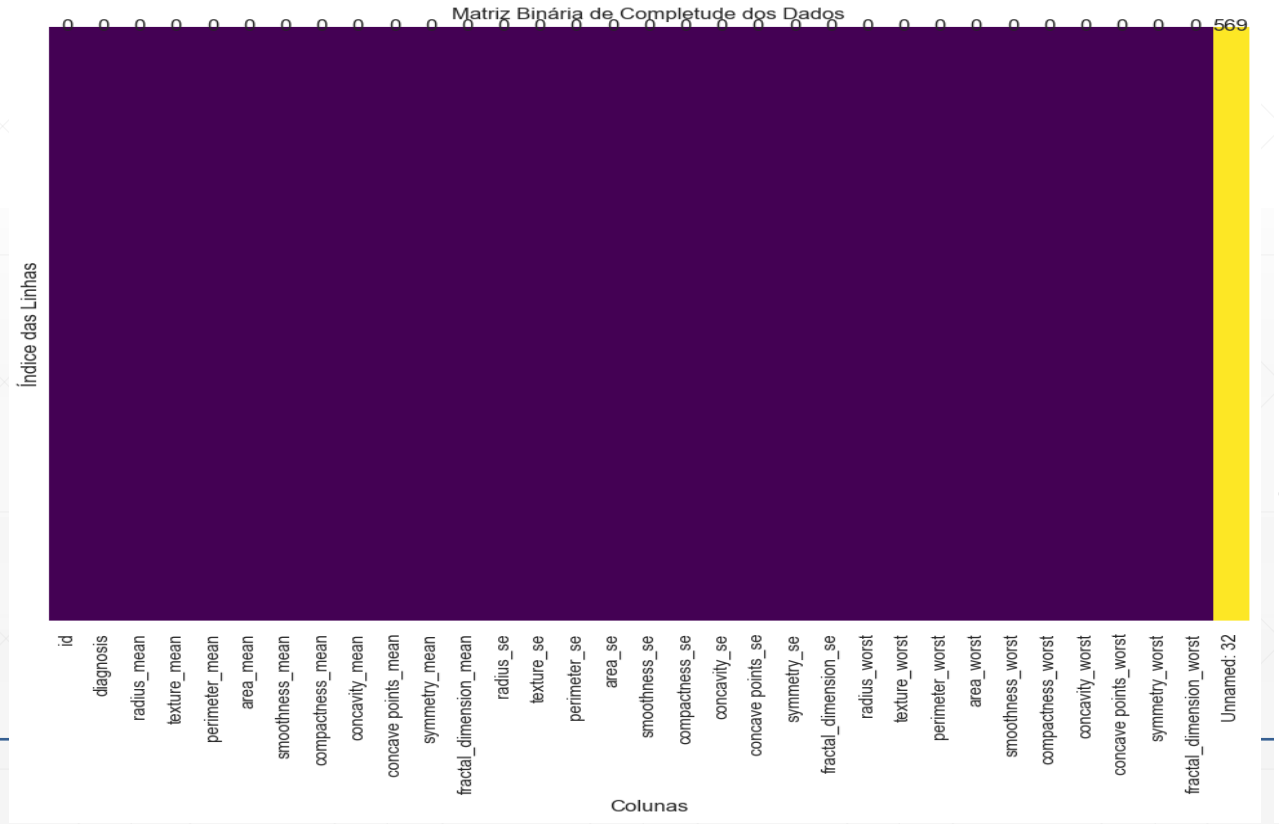
O objetivo máximo do presente estudo é identificar as principais características e padrões que envolvem o resultado e detecção de câncer de mama em resultados de exames do dataset. Para tal, tentaremos buscar meios de extrair insghts dos seguintes questionamentos:

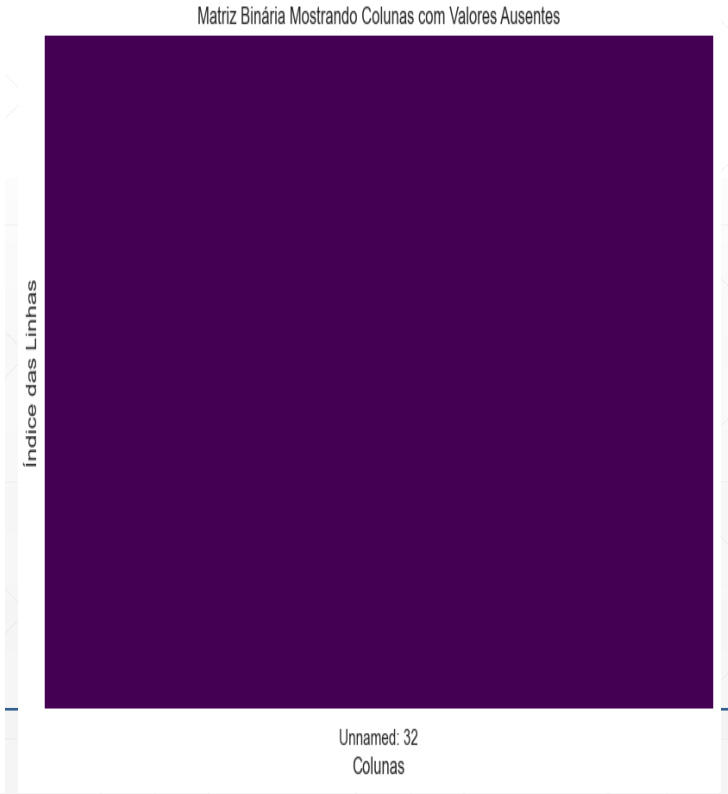
- Quão assertivo pode ser a detecção de câncer de mama por algoritmos de Machine Learning;
- Relevância de certas variáveis ou características clínicas e patológicas na previsão do câncer de mama;
- Qual percentual de diagnósticos benignos e malignos;

Tratamento inicial dos dados

Inicialmente, verificamos se haviam valores nulos em nossa base (Colunas nulas e células nulas). Constatamos que a última coluna denominada “Unnamed:32” continha todos os valores nulos. Excluímos a coluna citada e também a coluna de identificação do paciente que não é relevante para nosso modelo.

Somatória dos possíveis missing values pela função ‘isna’.



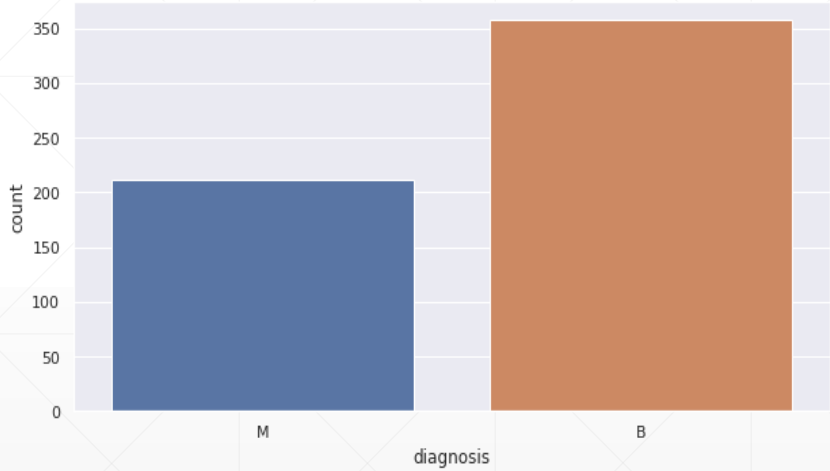


Análise exploratória dos dados

Determinaremos as distribuições de cada atributo, começando pelo diagnóstico:
Diagnóstico (maligno x benigno)

```
M = df.diagnosis.value_counts()[1]
B = df.diagnosis.value_counts()[0]
TOTAL = B + M
print(f'Maligno: {M}\nBenigno: {B}\nMaligno(%): {M/TOTAL:.2f}%\nBenigno(%): {B/TOTAL:.2f}%')
```

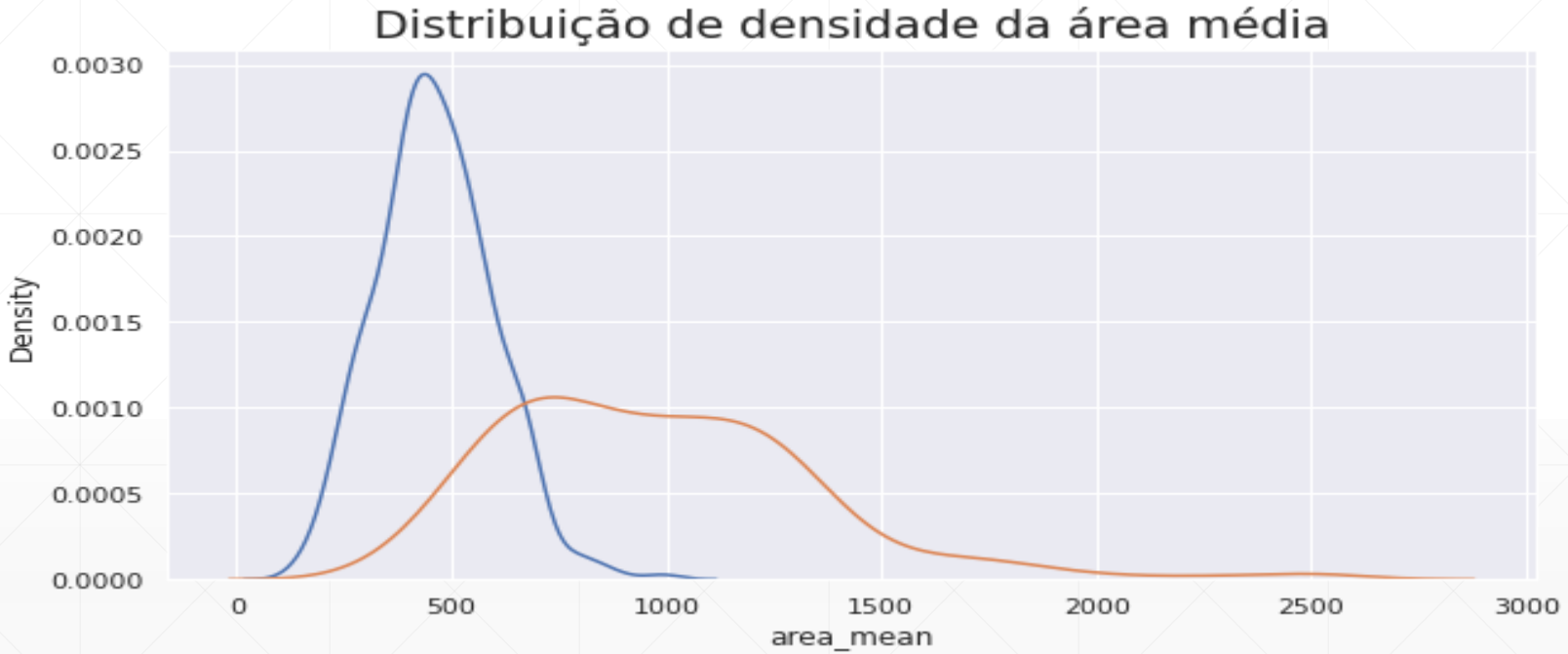
```
Maligno: 212
Benigno: 357
Maligno(%): 0.37%
Benigno(%): 0.63%
```



Análise exploratória dos dados

- **Distribuição de densidade da área média**

  Façamos agora uma análise da distribuição para os diagnósticos malignos e benignos. O pico da distribuição azul (benigna) é mais alto e mais estreito, indicando que a maioria dos tumores benignos tem uma área média relativamente pequena e que os valores estão mais concentrados em torno da mediana.


Distribuição de densidade da área média
Density
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
500
1000
1500
2000
2500
3000
area_mean

Análise exploratória dos dados

- **Distribuição média dos pontos côncavos**

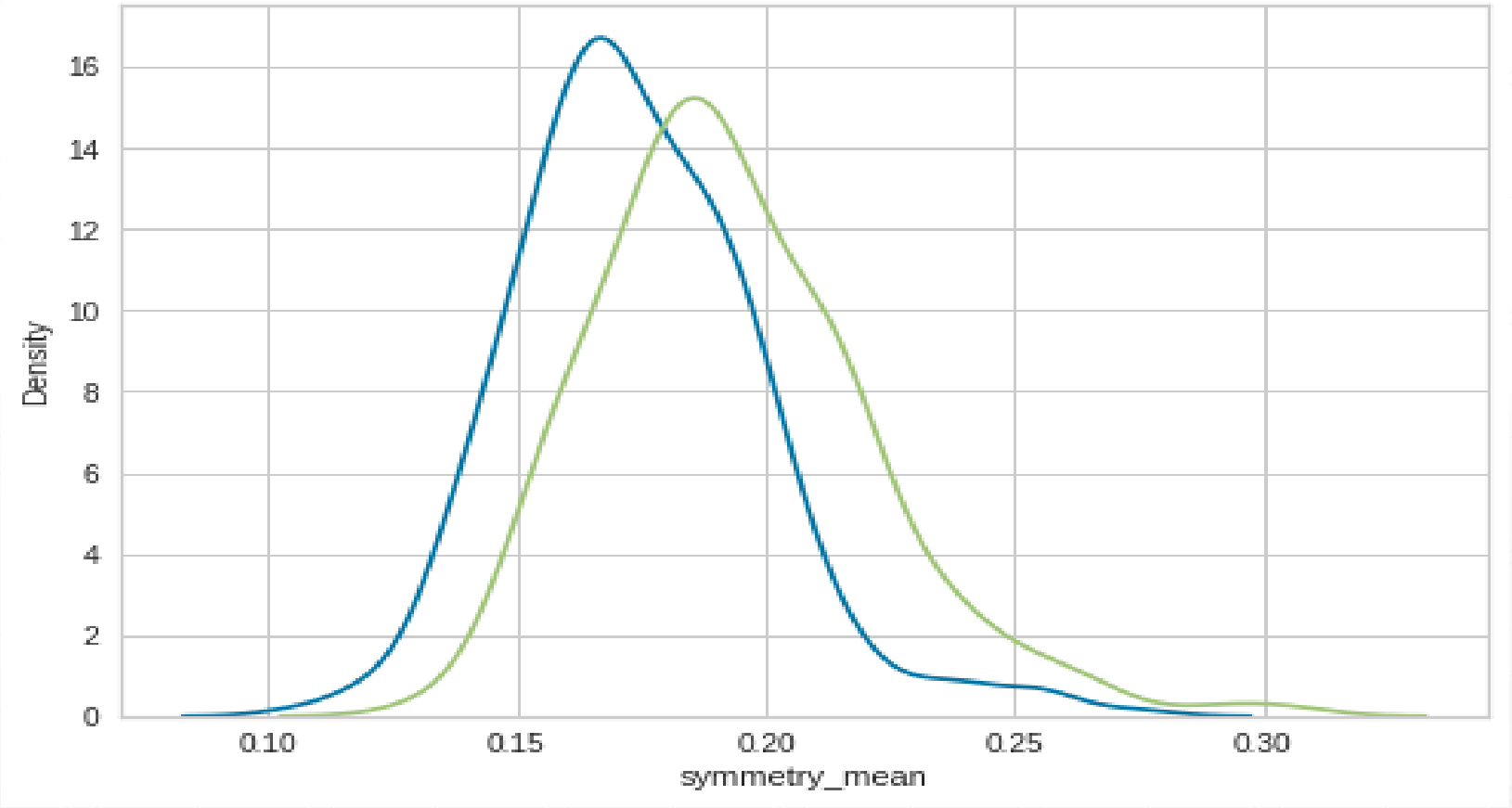
Agora verifiquemos se há algum padrão que pode ser notado quando analisamos a target (concave points_mean) através de um boxplot e um gráfico KDE (estimativa de densidade kernel).

Nota-se que os resultados benignos tendem a ter menos pontos côncavos em média, com uma concentração mais alta de exames em torno de um valor baixo. Existem outliers tanto em exames benignos quanto malignos, mas eles são mais prevalentes e extremos nos exames malignos, o que pode indicar uma variabilidade particular nos casos de câncer mais agressivos ou avançados

Análise exploratória dos dados

- **Simetria média dos resultados**

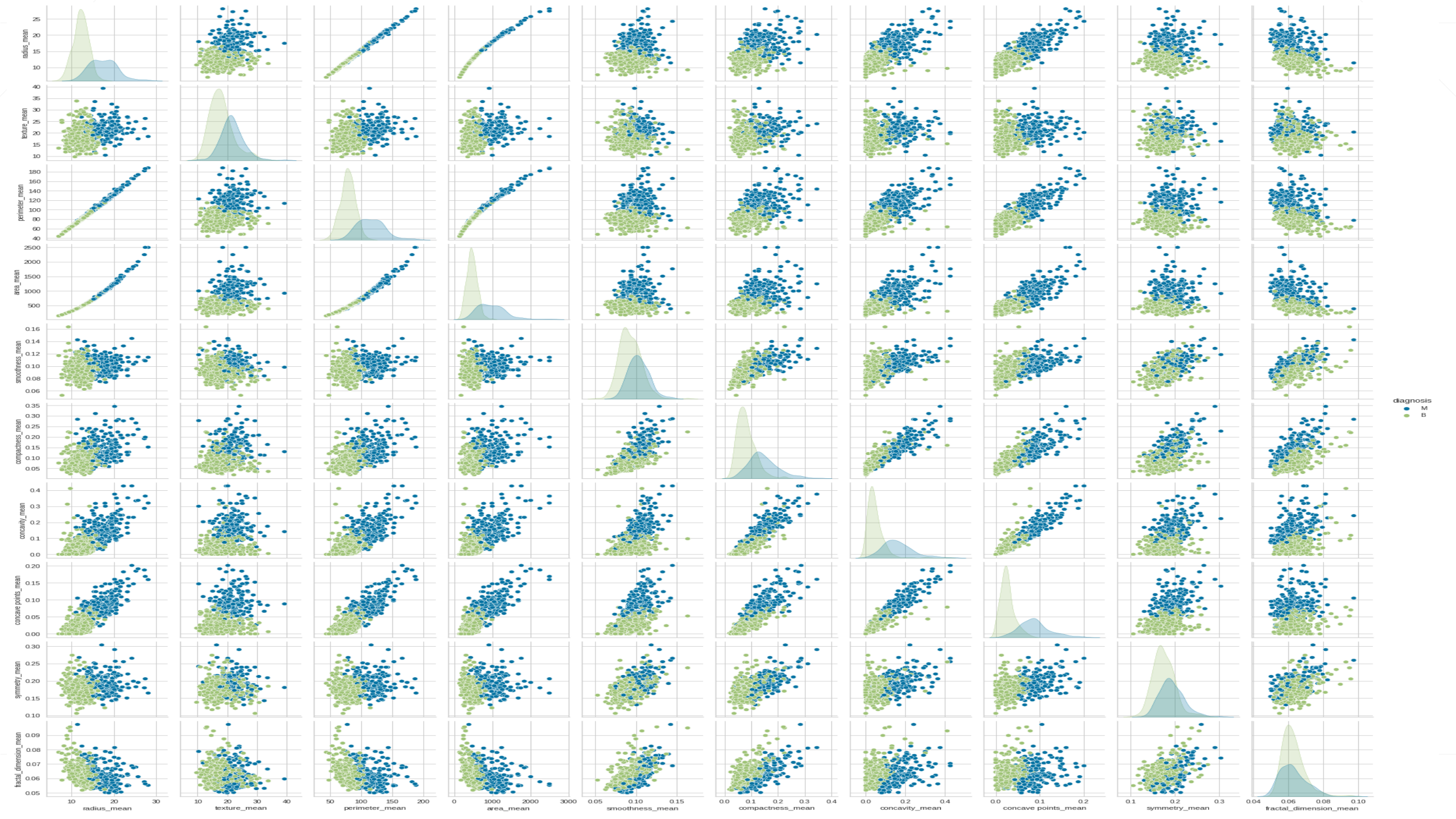
Existe uma sobreposição significativa entre as duas distribuições, o que significa que, embora haja uma tendência para os tumores malignos terem maior simetria média, há muitos casos em que a simetria sozinha não seria suficiente para fazer uma distinção clara entre benigno e maligno A simetria é um dos fatores considerados na classificação dos tumores. No entanto, este gráfico sugere que, enquanto a simetria pode contribuir para o diagnóstico, ela não é conclusiva por si só e deve ser combinada com outras medidas para uma avaliação mais precisa.


16
14
12
10
8
6
4
2
0
Density
0.10
0.15
0.20
0.25
0.30
symmetry_mean

Análise exploratória dos dados

- **Correlação do conjunto de dados**

  Esse gráfico irá correlacionar todos os atributos, para termos um panorama geral de como estão separados e correlacionados.


radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave points_mean
symmetry_mean
fractal_dimension_mean
diagnosis
M
B

Construção do Modelo

Construção do modelo usando PyCaret

Passada a etapa de análise exploratória, agora construiremos nosso modelo de Machine Learning.

Pycaret é uma biblioteca de auto - Machine Learning, capaz de construir e comparar diversos modelos de modo muito simples e rápido, fazendo com que possamos focar nas explicações acerca do assunto. E como é um tema bem importante e complexo de ser abordado, a melhor maneira de manter a produtividade é construindo modelos de modo mais rápido. Portanto, PyCaret é a escolha perfeita para o momento
Primeiro construiremos nosso modelo, usando 80% dos dados como treino e 20% como teste (leia sobre Princípio de Pareto)

- **Métricas de avaliação**

Durante o processo de criação de um modelo de machine learning nós precisamos medir a qualidade dele de acordo com o objetivo da tarefa. Existem funções matemáticas que nos ajudam a avaliar a capacidade de erro e acerto dos nossos modelos, e agora você conhecerá algumas das mais utilizadas. No artigo, usarei a palavra métrica para me referir a essas funções.

Tão importante quanto saber escolher um bom modelo, é saber escolher a métrica correta para decidir qual é o melhor entre eles.

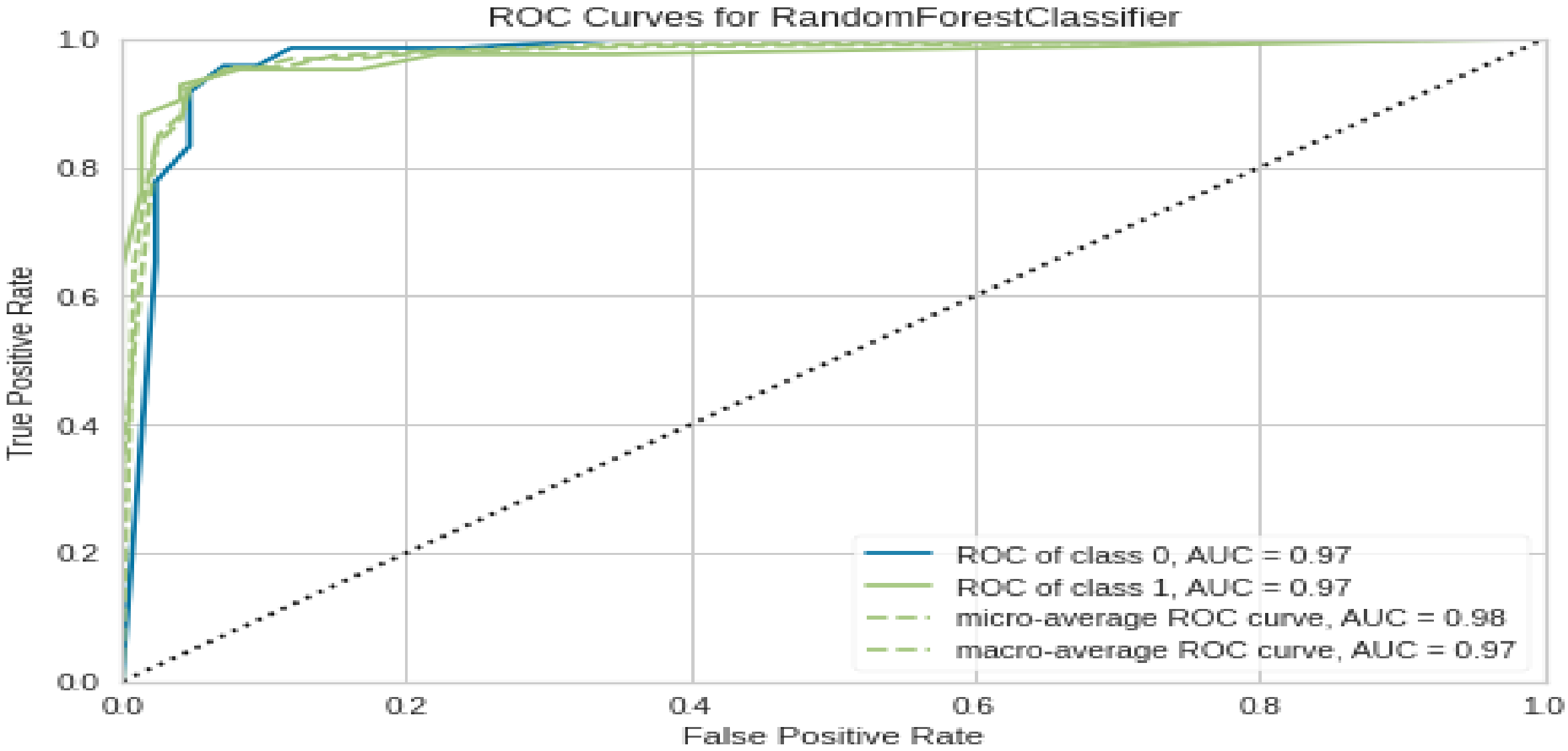
Existem métricas mais simples, outras mais complexas, algumas que funcionam melhor para datasets com determinadas características, ou outras personalizadas de acordo com o objetivo final do modelo.

Ao escolher uma métrica deve-se levar em consideração fatores como a proporção de dados de cada classe no dataset e o objetivo da previsão (probabilidade, binário, ranking, etc). Por isso é importante conhecer bem a métrica que será utilizada, já que isso pode fazer a diferença na prática.

Construção do Modelo

Nenhuma destas funções é melhor do que as outras em todos os casos. É sempre importante levar em consideração a aplicação prática do modelo. O objetivo deste artigo não é ir a fundo em cada uma delas, mas apresentá-las para que você possa pesquisar mais sobre as que achar interessante.
A curva **ROC** mostra o quão bom o modelo criado pode distinguir entre duas coisas (já que é utilizado para classificação). Essas duas coisas podem ser 0 ou 1, ou positivo e negativo. Os melhores modelos conseguem distinguir com precisão o binômio.

O valor do **AUC** varia de 0,0 até 1,0 e o limiar entre a classe é 0,5. Ou seja, acima desse limite, o algoritmo classifica em uma classe e abaixo na outra classe. Quanto maior o AUC, melhor.


ROC Curves for RandomForestClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0, AUC = 0.97
ROC of class 1, AUC = 0.97
micro-average ROC curve, AUC = 0.98
macro-average ROC curve, AUC = 0.97

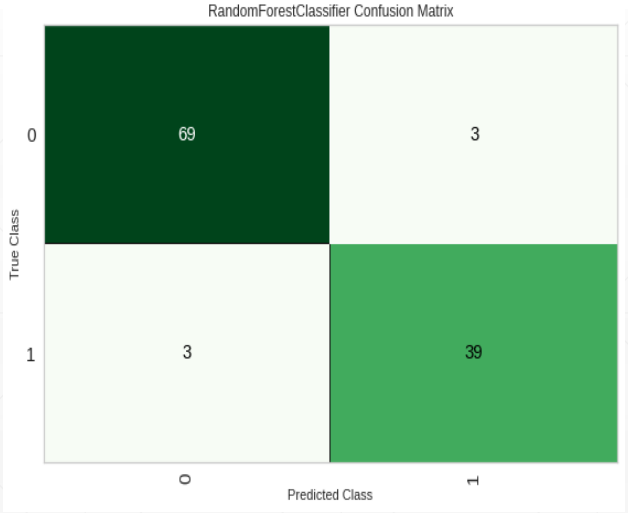
Construção do Modelo

- **Matriz de Confusão**

É uma tabela que mostra as frequências de classificação para cada classe do modelo.

- **Verdadeiro positivo** (true positive — TP): ocorre quando no conjunto real, a classe que estamos buscando foi prevista corretamente. Por exemplo, quando a mulher está grávida e o modelo previu corretamente que ela está grávida.
- **Falso positivo** (false positive — FP): ocorre quando no conjunto real, a classe que estamos buscando prever foi prevista incorretamente. Exemplo: a mulher não está grávida, mas o modelo disse que ela está.
- **Falso verdadeiro** (true negative — TN): ocorre quando no conjunto real, a classe que não estamos buscando prever foi prevista corretamente. Exemplo: a mulher não estava grávida, e o modelo previu corretamente que ela não está.
- **Falso negativo** (false negative — FN): ocorre quando no conjunto real, a classe que não estamos buscando prever foi prevista incorretamente. Por exemplo, quando a mulher está grávida e o modelo previu incorretamente que ela não está grávida.

Sobre matriz de confusão

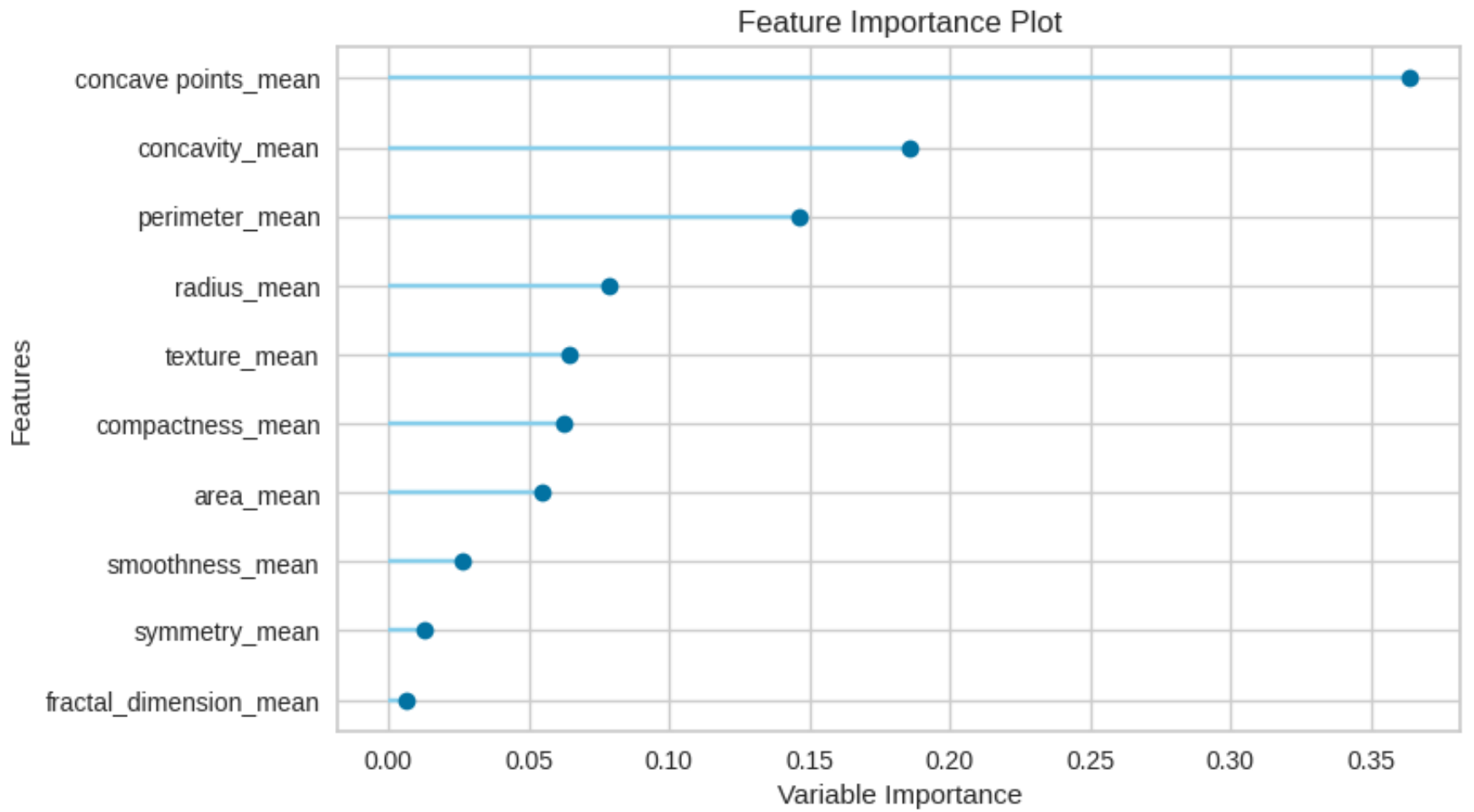


- ***Feature importance***

As pontuações relativas podem destacar quais características podem ser mais relevantes para o alvo e, ao contrário, quais características são menos relevantes. Isso pode ser interpretado por um especialista no domínio e pode ser usado como base para a coleta de mais ou diferentes dados.

A maioria das pontuações de importância são calculadas por um modelo preditivo que foi ajustado ao conjunto de dados. A inspeção da pontuação de importância fornece uma visão sobre esse modelo específico e quais recursos são os mais e menos importantes para o modelo ao fazer uma previsão. Este é um tipo de interpretação de modelo que pode ser executado para os modelos que o suportam.

A importância do recurso pode ser usada para melhorar um modelo preditivo. Isso pode ser obtido usando as pontuações de importância para selecionar os recursos a serem excluídos (pontuações mais baixas) ou os recursos a serem mantidos (pontuações mais altas). Este é um tipo de seleção de recurso e pode simplificar o problema que está sendo modelado, acelerar o processo de modelagem (a exclusão de recursos é chamada de redução de dimensionalidade) e, em alguns casos, melhorar o desempenho do modelo.
(Leia.)

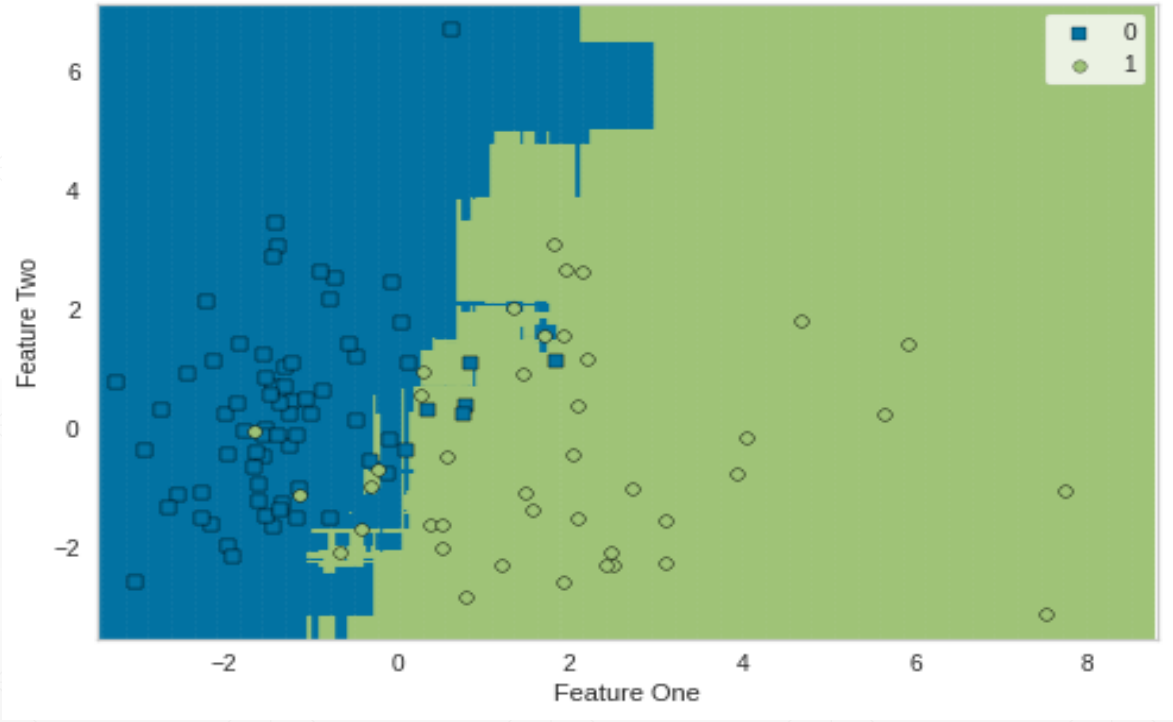

Feature Importance Plot
concave points_mean
concavity_mean
perimeter_mean
radius_mean
texture_mean
compactness_mean
area_mean
smoothness_mean
symmetry_mean
fractal_dimension_mean
Features
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Variable Importance

Superfície de decisão

Em um problema de classificação estatística com duas classes, um limite de decisão ou superfície de decisão é uma hipersuperfície que divide o espaço vetorial subjacente em dois conjuntos, um para cada classe. O classificador classificará todos os pontos de um lado da fronteira de decisão como pertencentes a uma classe e todos os do outro lado como pertencentes à outra classe. Um limite de decisão é a região de um espaço de problema em que o rótulo de saída de um classificador é ambíguo.

(https://en.wikipedia.org/wiki/Decision_boundary#:~:text=In%20a%20statistical%2Dclassification%20problem,sets%2C%20one%20for%20each%20class.&text=A%20decision%20boundary%20is%20the,of%20a%20classifier%20is%20ambiguous.)

Boundary



Conclusão

Nosso modelo de classificação random forest obteve sucesso em **75%** dos dados de teste (20% dos dados totais).

A todo momento vem ocorrendo mudanças e avanços na utilização de técnicas de inteligência artificial, principalmente na área da saúde. Detecção de câncer por meio de algoritmos de Machine Learning tem ajudado cada vez mais os médicos a diagnosticarem seus pacientes, assim aumentando as chances dos tratamentos convencionais.

No entanto, mesmo a construção de modelos altamente confiáveis para a detecção de câncer de mama não substitui uma variável, a principal feature, o **diagnóstico precoce**.