Algoritmos Supervisionados – Árvore de Decisão

ROCKETSEAT

Plataforma completa de aprendizado contínuo em programação.

#BoostingPeople

rocketseat.com.br

Algoritmos Supervisionados

Árvore de Decisão

O objetivo deste módulo é apresentar conceitualmente os principais algoritmos de classificação para que possamos desenvolver projetos de machine learning que fazem previsões de categorias ou classes. E faremos um projeto explorando o primeiro destes algoritmos, que é o de árvore de decisão, onde faremos o processo completo, desde o EDA até a entrega do modelo através de uma aplicação para inferência batch.

Agenda

- O que é um algoritmo de classificação?
- Um passeio pelos algoritmos de classificação
- O que é Árvore de Decisão
- Projeto – Árvore de Decisão

O que é um algoritmo de classificação ?

Os algoritmos de classificação têm como propósito organizar dados de maneira eficaz, sendo empregados em diversas situações. Diferentemente da análise de regressão, que explora relações numéricas, a classificação se destina a categorizar instâncias em grupos distintos. Isso pode ocorrer de diversas formas, desde a classificação binária, que divide os dados em duas categorias, até a classificação multi-classe e multi-rótulo, que lidam com múltiplas categorias. Essa diversidade de abordagens permite a aplicação dos algoritmos em uma ampla gama de casos, como a identificação de spam (classificação binária), o diagnóstico médico de diferentes doenças (classificação multi-classe), e a categorização de documentos com múltiplos temas (classificação multi-label).

A escolha criteriosa do método de classificação depende da natureza específica do problema e da estrutura dos dados envolvidos. Seja na identificação de padrões em imagens ou na segmentação de clientes para estratégias de marketing personalizado, os algoritmos de classificação desempenham um papel vital na análise de dados, possibilitando a tomada de decisões fundamentadas e a compreensão de padrões no mundo real.

Um passeio pelos algoritmos de classificação

Árvore de Decisão

As árvores de decisão dividem o conjunto de dados em subconjuntos com base em características, criando uma estrutura de árvore onde as folhas representam as classes. Cada nó interno representa uma decisão baseada em uma característica.

Exemplo de Uso: Diagnóstico médico, como a previsão de doenças com base em sintomas.

Regressão Logística

A regressão logística é um método de classificação que estima a probabilidade de uma instância pertencer a uma classe específica. Utiliza uma função para mapear a saída linear para o intervalo 0 e 1, interpretada como a probabilidade da classe positiva.

Exemplo de Uso: Identificação de e-mails como spam ou não spam com base em características como palavras-chave e comprimento do texto.

Naive Bayes

Baseado no teorema de Bayes, o Naive Bayes assume independência condicional entre as características. É eficaz e rápido, sendo especialmente útil em conjuntos de dados com muitas características.

Exemplo de Uso: Classificação de opiniões de clientes entre positivas, negativas ou neutras (Análise de Sentimentos).

Um passeio pelos algoritmos de classificação

K Nearest Neighbors

O KNN classifica uma instância com base nas classes das instâncias vizinhas mais próximas, utilizando uma métrica de distância (como a euclidiana). A instância é atribuída à classe mais comum entre seus k vizinhos mais próximos.

Exemplo de Uso: Sistemas de Recomendação para empresas de varejo digital

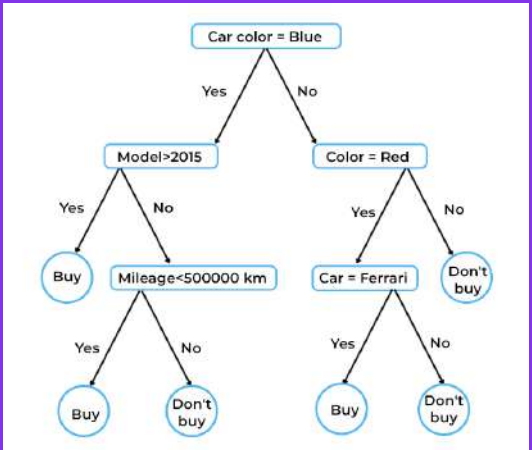
SVM

Support Vector Machines procuram encontrar o hiperplano que melhor separa as instâncias de diferentes classes, maximizando a margem entre elas. Podem ser estendidas para problemas de classificação binária e multi-classe.

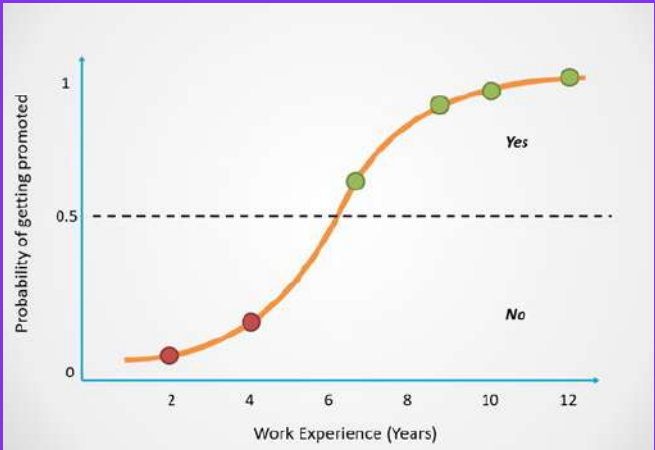
Exemplo de Uso: Classificação de documentos em categorias, como notícias, esportes e entretenimento.

Um passeio pelos algoritmos de classificação

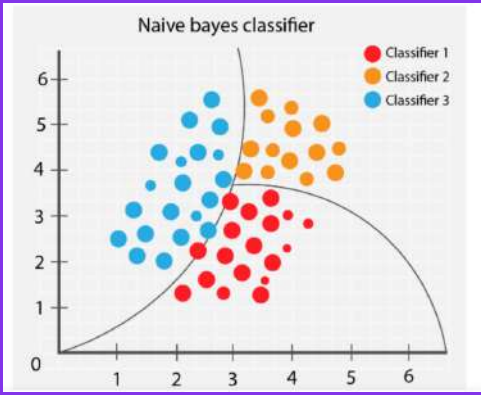
Árvore de Decisão



Regressão Logística

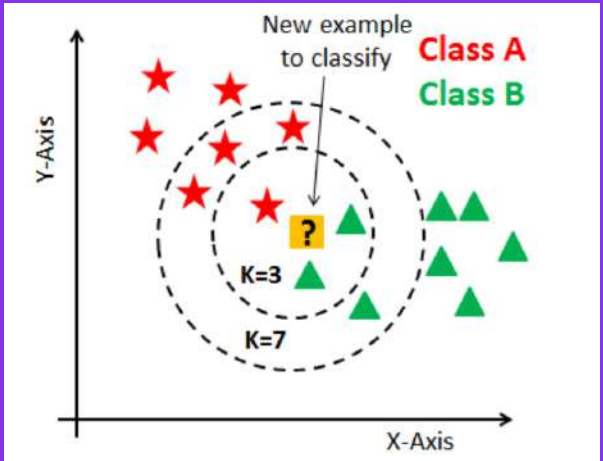


Naive Bayes

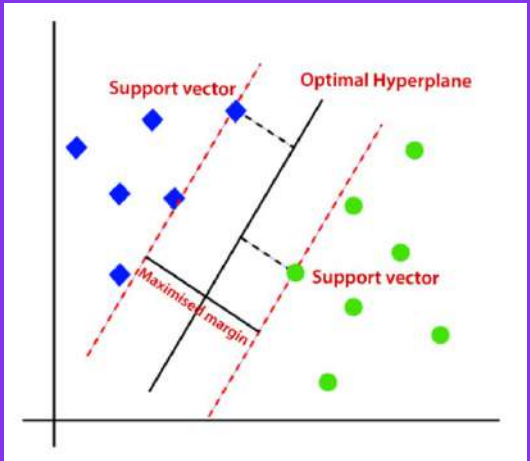


Um passeio pelos algoritmos de classificação

K Nearest Neighbors

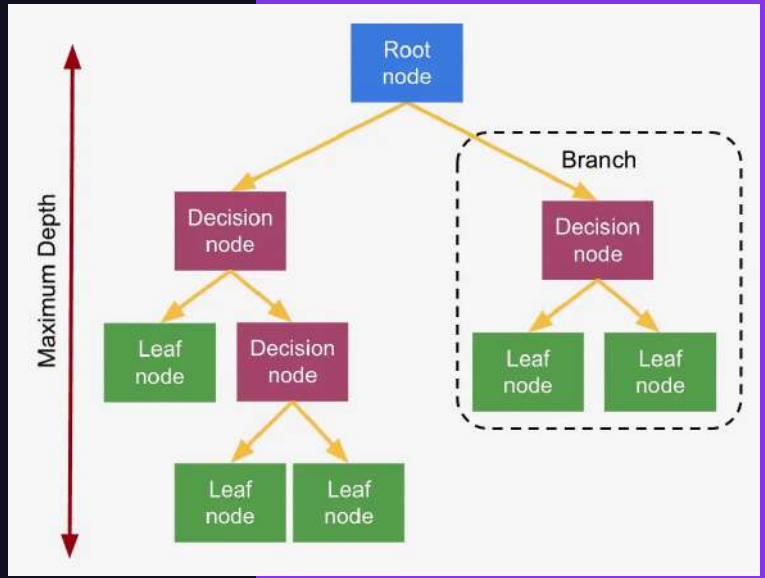


SVM



O que é uma árvore de decisão?

Uma árvore de decisão é um modelo de aprendizado de máquina que representa uma estrutura hierárquica de decisões baseadas em características dos dados. Inicia-se com a escolha da variável que melhor separa os dados, utilizando métricas como ganho de informação ou índice de Gini para avaliar a pureza das divisões. Em cada nó da árvore, a variável escolhida gera ramos que continuam o processo de subdivisão até atingir folhas que representam as decisões finais ou classes. O processo de construção da árvore visa minimizar a entropia ou impureza, resultando em uma estrutura que captura padrões complexos nos dados.



O que é uma árvore de decisão?

Entropia e Ganho de Informação:

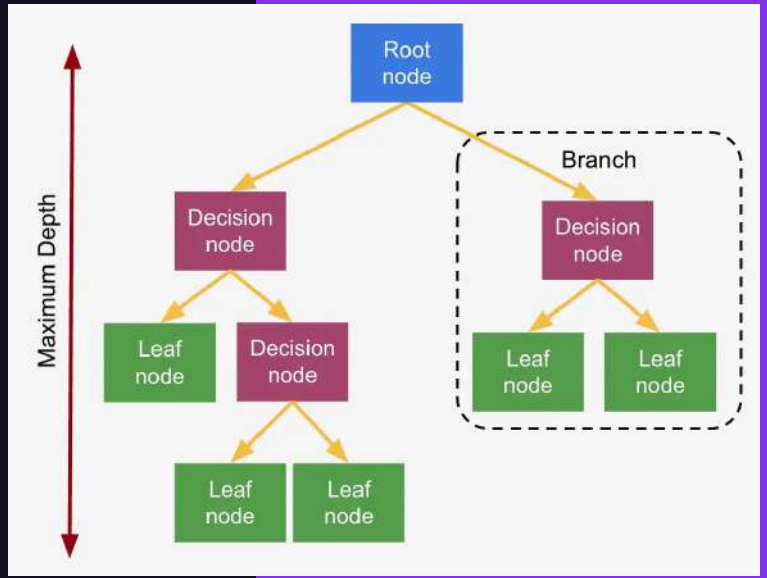
A entropia é uma medida de impureza em um conjunto de dados. O ganho de informação é usado para decidir como dividir os dados em cada nó da árvore, visando reduzir a entropia. Em termos simples, o objetivo é criar divisões que tornem os grupos resultantes mais homogêneos.

Corte (Pruning):

O corte é uma técnica utilizada para evitar o sobreajuste (overfitting) da árvore. Às vezes, as árvores podem se tornar muito complexas e se ajustar demais aos dados de treinamento. O corte envolve a poda de ramos da árvore para melhorar a generalização para novos dados.

Profundidade da Árvore:

A profundidade de uma árvore refere-se à quantidade de níveis ou perguntas feitas antes de chegar a uma decisão. Árvores mais profundas podem se ajustar demais aos dados de treinamento, enquanto árvores rasas podem não capturar padrões complexos.



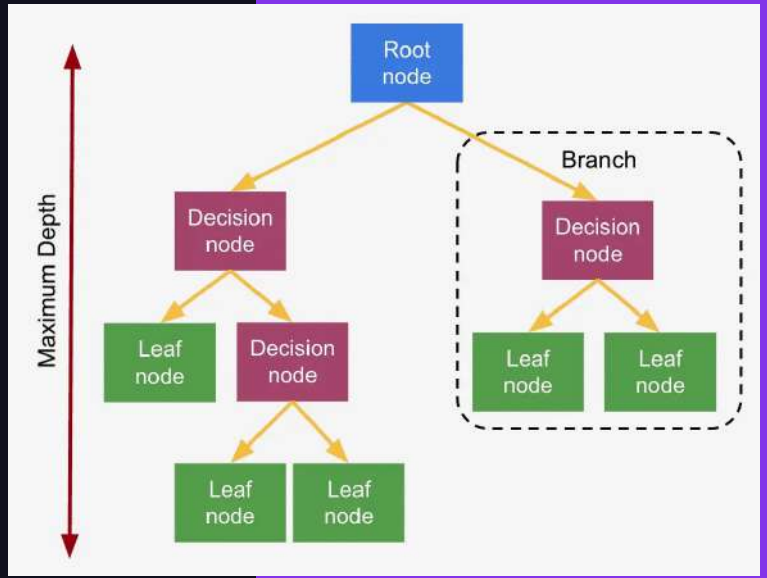
O que é uma árvore de decisão?

Variáveis Categóricas e Numéricas:

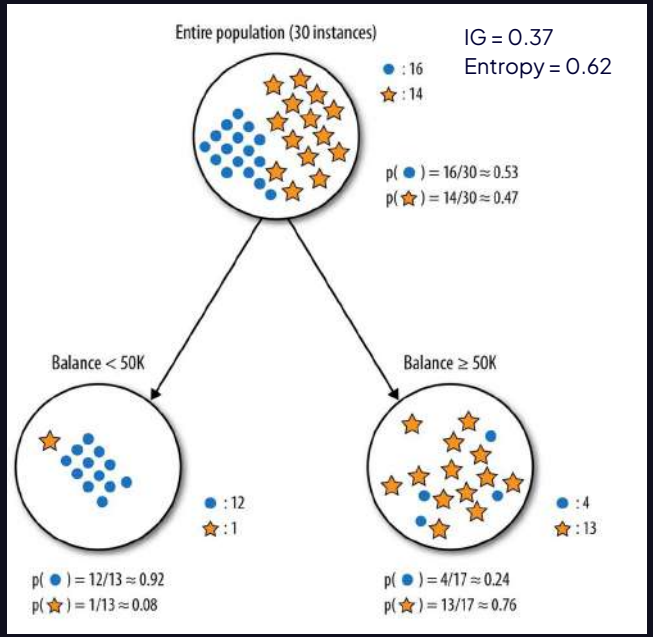
Árvores de decisão lidam com diferentes tipos de variáveis de maneira distinta. Para variáveis categóricas, a árvore faz perguntas do tipo "é igual a X?" ou "pertence à categoria Y?". Para variáveis numéricas, as árvores fazem perguntas do tipo "é maior que X?" ou "está no intervalo Y-Z?".

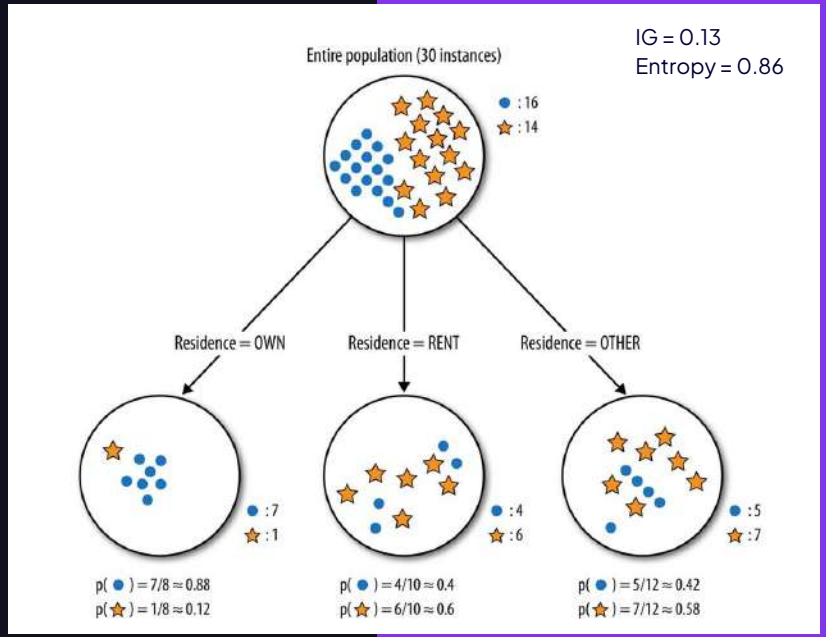
Grid Search / Random Search

Grid Search e Random Search são técnicas usadas para encontrar os melhores hiperparâmetros para o modelo, ajustando diferentes combinações e escolhendo aquelas que resultam no melhor desempenho, muitas vezes usando a validação cruzada para avaliação.



O que é uma árvore de decisão?



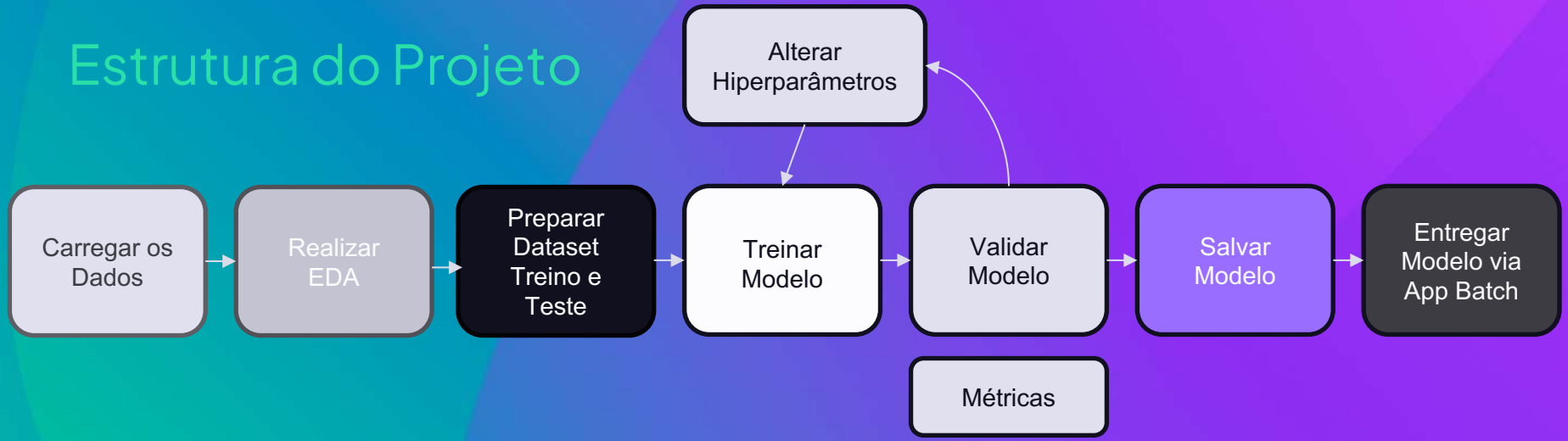


Projeto – Árvore de Decisão

Uma empresa de concessão de crédito para pequenas empresas possui um catálogo de seus clientes (PJ) com informações como idade da empresa, faturamento mensal, nível de inovação, entre outras. E para que esta empresa de concessão possa dar um atendimento mais apropriado para cada tipo de cliente, eles classificam os clientes em segmentos: Starter, Bronze, Silver e Gold.

Desta forma, para que seja possível classificar novos clientes, iremos construir um classificador que determina o segmento do cliente, com base nas informações disponíveis sobre o mesmo.

Estrutura do Projeto



Code Time ...

rocketseat.com.br