Aluno: Anderson Mendes dos Santos
Bootcamp: Machine Learning

01. Explique com suas palavras, o que é machine learning?

R: Machine Learning (ML) ou Aprendizado de máquina é um subcampo da Inteligência Artificial (IA) que se concentra no desenvolvimento de algoritmos e modelos que permitem aos computadores aprenderem padrões a partir de dados e fazerem previsões e fazerem ou tomar decisões sem serem explicitamente programados para cada tarefa específica.
O Machine Learning envolve o uso de dados para treinar modelos. Esses modelos são capazes de identificar padrões, relações ou tendências nos dados e, com base nisso, realizar tarefas como: classificação, regressão, clusterização e recomendação, por exemplo.
Existem três tipos de Aprendizado de Máquina, a saber por:

- **Aprendizado Supervisionado**: O modelo é treinado com dados que incluem entradas e saídas corretas (rótulos). O objetivo é aprender a mapear entradas para saídas;
- **Aprendizado Não-Supervisionado**: O modelo é treinado com dados sem rótulos e o objetivo é encontrar padrões ou estruturas ocultas;
- **Aprendizado por Reforço:** O modelo aprende por tentativa e erro, recebendo recompensas ou penalidades com base em suas ações;

02. Explique o conceito de conjunto de treinamento, conjunto de validação e conjunto de teste em machine learning.

R: **Conjunto de treinamento**: É o conjunto de dados usado para treinar o modelo. O modelo de Machine Learning aprende os padrões e relações presente nesses dados. Suas principais funções são o ajuste dos parâmetros modelo (por exemplo, os pesos de uma rede neural) e permitir que o modelo aprenda a mapear as entradas (features) para as saídas (labels).
Conjunto de Validação: É o conjunto de dados utilizado para ajustar hiperparâmetros e avaliar o desempenho do modelo durante o treinamento. Ele não é utilizado para treinar o modelo diretamente. Suas principais funções são ajudar a escolher o melhor modelo ou configuração e evitar o overfitting (quando o modelo se ajusta demais aos dados de treinamento e não generaliza bem para novos dados).
Conjunto de Teste: É o conjunto de dados não visto pelo modelo durante o treinamento ou validação. Ele é utilizado para avaliar o desempenho final do modelo em dados desconhecidos. Suas principais funções são simular como o modelo se comportará em situações reais e fornecer uma métrica imparcial de desempenho (por exemplo, acurácia, precisão, recall).

03. Explique como você lidaria com dados ausentes em um conjunto de dados de treinamento.

R: A presença de dados ausentes podem prejudicar o desempenho do modelo ou até mesmo inviabilizar seu treinamento. Para tratar dados ausentes poderíamos adotar algumas estratégias como:

- Identificar dados ausentes utilizando funções como **isnull()** ou **info()** da biblioteca Pandas para identificar quais colunas possuem dados ausentes e quantos valores estão faltando;
- Remover dados ausentes (de linhas ou colunas) caso a quantidade seja consideravelmente pequena, levando em consideração que o tamanho do conjunto de dados poderá ser reduzido significativamente;
- Preencher dados ausentes (imputação) com valores constantes, com medidas estatísticas (por exemplo, média, mediana, moda) ou até mesmo utilizando algoritmos de Machine Learning, por exemplo.
- Avaliar o impacto da imputação é importante para verificar se houve a introdução de viés ou distorções no conjunto de dados;

- Comparar as estatísticas descritivas (por exemplo, média, desvio padrão) antes e depois da imputação de valores;

Tratar dados ausentes é uma etapa essencial para para garantir a qualidade dos dados e o sucesso do modelo de Machine Learning. Então, é de suma importância o entendimento da natureza do problema e se a remoção de dados ausentes pode impactar nos resultados finais.

04. O que é uma matriz de confusão e como ela é usada para avaliar o desempenho de um modelo preditivo?

R: Uma matriz de confusão (Confusion Matrix) é utilizada para avaliar o desempenho de modelos de classificação em Machine Learning. Ela fornece uma visão detalhada de como o modelo está performando, mostrando os acertos e os erros das previsões em relações aos valores reais.
Para um problema de classificação binária, por exemplo, a matriz de confusão possui a seguinte estrutura:

	Previsto Positivo	**Previsto Negativo**
Real Positivo	Verdadeiro Positivo (VP)	Falso Negativo (VN)
Real Negativo	Falso Positivo (FP)	Verdadeiro Negativo (FN)

-Verdadeiro Positivo (VP): O Modelo prevê corretamente a classe positiva;
-Falso Positivo (FP): O modelo prevê erroneamente a classe positiva;
-Falso Negativo (FN): O modelo prevê erroneamente a classe negativa;
-Verdadeiro Negativo (VN): O modelo prevê corretamente a classe negativa;

A matriz de confusão calcula métricas de desempenho que ajudam a entender a qualidade do modelo. Algumas métricas são conhecidas como:

- Acurácia: Mede a proporção de previsões corretas em relação ao total de previsões;
  - *Acurácia = VP + VN / (VP + FP + VN + FN)*
- Precisão: Mede a proporção de previsões positivas corretas em relação ao total de previsões positivas;
  - *Previsão = VP / (VP + FP)*
- Recall: Mede a proporção de positivos reais que foram corretamente identificados pelo modelo;
  - *Recall = VP / (VP + FN)*
- F1-score: É a média harmônica entre precisão e recall. Útil quando há um desequilíbrio entre as classes;
  - *F1-score = 2 * (precisão * recall/(precisão + recall))*

O uso dessas métricas pode variar a depender das especificações do problema.

05. Em quais áreas (tais como construção civil, agricultura, saúde, manufatura, entre outras) você acha mais interessante aplicar algoritmos de machine learning?

R: Machine Learning tem aplicações vastas e transformadoras em praticamente todas as áreas. Mas, particularmente, a aplicação de ML se torna interessante e promissora nas áreas como saúde (para diagnóstico médico), finanças (para detecção de fraudes e análise de crédito) e transporte e logística (para veículos autônomos) .