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

R = É uma subárea da inteligência artificial, faz com que sistemas aprendam e melhorem automaticamente com base em dados, analisando padrões nos dados para tomar decisões, fazer previsões ou classificações. Tendo como exemplo o machine learning o sistema de identificação de spam do e-mail, onde ele aprende com base em e-mails anteriores que foram classificados como spam ou legítimos.

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

R = **Conjunto de treino** - É um conjunto de dados usados para treinar o modelo, aprendendo padrões a partir desses dados

Conjunto de validação - É utilizado para avaliar o desempenho do modelo enquanto ele está sendo treinado

Conjunto de teste - É um conjunto de dados usado após o treinamento e a validação do modelo, para avaliar seu desempenho final. O conjunto de teste simula o comportamento do modelo em dados novos.

Tendo como exemplo uma base de dados de questões do enem, onde o objetivo é classificar as questões com base no texto e enunciado da questão para que assim possa dizer a que área a questão pertence (Biologia, Química, Física).

Essas questões têm seus enunciados e rótulos já conhecidos (como Biologia, Química ou Física). O modelo usa essas questões para aprender os padrões no texto e enunciado que caracterizam cada área. Esse seria o conjunto de treino.

Um conjunto de questões são usadas para verificar como o modelo está se saindo e ajustar os parâmetros, garantindo que ele não esteja apenas se dando bem com os dados do conjunto de treinamento. Esse seria o conjunto de validação.

Depois um conjunto de dados usado após o treinamento para avaliar como o modelo classifica questões nunca vistas, vendo sua capacidade de generalização para classificar corretamente questões de Biologia, Química e Física com base no texto e enunciado. Esse seria o conjunto de teste.

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

R = Caso seja uma quantidade pequena e com baixa interferência no conjunto de dados, esses dados podem ser apagados e desconsiderados

Caso a remoção cause uma perda de dados considerável, pode ser usado a média ou moda ou mediana da coluna para substituir os valores ausentes

Também podendo usar até modelos de machine learning para prever os dados que estão faltando, como algoritmos de regressão.

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

R = Fornece uma visualização clara das previsões feitas pelo modelo em comparação com os rótulos reais das classes.

Ela é composta por:

Verdadeiros Positivos (TP): Casos em que o modelo previu corretamente a classe positiva (o modelo acertou).

Falsos Positivos (FP): Casos em que o modelo previu a classe positiva, mas a classe real era negativa (o modelo errou).

Falsos Negativos (FN): Casos em que o modelo previu a classe negativa, mas a classe real era positiva (o modelo errou).

Verdadeiros Negativos (TN): Casos em que o modelo previu corretamente a classe negativa (o modelo acertou).

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

R = Na área da saúde os modelos podem prever doenças com base em dados clínicos, assim podendo auxiliar no diagnóstico precoce da doença. Também podendo classificar e identificar imagens de radiografia e ressonância detectar anomalias.