

**Aluno: José Lisiomar de Souza**  
**Professor: Renê Ripardo Calixto**  
**Monitora: Angélica Viana**

### ATIVIDADE 01

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

**Resposta:**

Machine Learning é uma subárea da Inteligência Artificial, formada através de algoritmos computacionais capazes de criar um processo de dedução pelo aprendizado baseado em exemplos. Isso envolve alimentar o algoritmo com um conjunto de dados para que o sistema reconheça padrões nesses dados. Nesse sentido, quanto maior o tempo de exposição do algoritmo aos dados e exemplos apresentados maior será a sua assertividade quanto ao que é proposto.

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

**Resposta:**

**Conjunto de treinamento:** É a base para o aprendizado da máquina, já que contém todos os exemplos que o modelo utilizará para aprender a realizar a tarefa desejada. No treinamento, o modelo utiliza esse conjunto de dados para aprender os padrões e efetuar os ajustes dos parâmetros necessários. Quanto mais diversificado e representativo for o conjunto de dados melhor será o resultado do modelo.

**Conjunto de validação:** É usado para avaliar o desempenho do modelo durante o treinamento. Depois de treinar o modelo com o conjunto de exemplos é necessário verificar se ele está generalizando bem para os dados não vistos. Com base nesse resultado, é possível ajustar todas as variáveis externas (hiperparâmetros) e escolher a melhor configuração que otimiza o desempenho do modelo.

**Conjunto de testes:** É usado para avaliar o desempenho final do modelo após o treinamento e fazer ajustes com o conjunto de treinamento e validação. Fornece uma estimativa de como o modelo atuará com novos dados do mundo real. É essencial que o conjunto de teste não seja usado durante o treinamento ou ajuste do modelo, para evitar vies nos resultados de avaliação.

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

**Resposta:**

Primeiro é importante identificar quais variáveis possuem dados ausentes e verificar sua proporção, calculando a quantidade de valores ausentes em cada variável. Depois, tentar entender se os motivos da ausência dos dados relacionam-se com erros de coleta, corrupção dos dados ou se realmente estão ausentes. Ao compreender a causa pode-se influenciar a abordagem escolhida para lidar com os erros. Se a proporção de dados ausentes em uma variável ou amostra específica for muito alta e não puder ser estimada com confiança, pode-se considerar a remoção da amostra inteira ou da variável. Isso deve ser feito com cautela, pois a remoção indiscriminada de dados pode reduzir a quantidade de informações disponíveis para o modelo.

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

**Resposta:**

Uma matriz de confusão é uma tabela que é usada para descrever o desempenho de um modelo de classificação em um conjunto de dados para os quais os valores verdadeiros são conhecidos. Ela é chamada de matriz de "confusão" porque permite visualizar facilmente se o sistema está confundindo duas classes. As linhas da tabela representam as classes verdadeiras e as colunas representam as classes previstas pelo modelo. Para um problema de classificação binária, a matriz de confusão terá a seguinte estrutura:

	<b>Previsão Positiva</b>	<b>Previsão Negativa</b>
<b>Classe Positiva</b>	Verdadeiro Positivo (VP)	Falso Negativo (FN)
<b>Classe Negativa</b>	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Para entender melhor vamos considerar os conceitos básicos relacionados à classificação:

Verdadeiros positivos (VP): São os casos em que o modelo previu corretamente a classe positiva (por exemplo, a classe "sim" em um problema de detecção de spam).

Falsos positivos (FP): São os casos em que o modelo previu incorretamente a classe positiva quando na verdade a classe verdadeira era negativa (por exemplo, o modelo classificou erroneamente um e-mail legítimo como spam).

Verdadeiros negativos (VN): São os casos em que o modelo previu corretamente a classe negativa (por exemplo, a classe "não" em um problema de detecção de spam).

Falsos negativos (FN): São os casos em que o modelo previu incorretamente a classe negativa quando na verdade a classe verdadeira era positiva (por exemplo, o modelo não identificou corretamente um e-mail de spam).

A matriz de confusão é usada para calcular várias métricas de desempenho, como precisão, recall, F1-score e AUC-ROC. Cada uma dessas métricas fornece uma perspectiva diferente sobre o desempenho do modelo, e juntas, elas podem fornecer uma visão abrangente de sua eficácia.

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

**Resposta:** Por ordem de interesse, segue:

1. Diagnóstico Médico e personalização de tratamentos: Algoritmos de machine learning podem ser utilizados para análise de imagens médicas (como tomografias e ressonâncias magnéticas) e dados clínicos para ajudar no diagnóstico precoce de doenças. Além de prever respostas individuais a tratamentos médicos com base no perfil genético ou clínico do paciente.

2. Segurança Cibernética: O aprendizado de máquina pode ser usado para detectar atividades suspeitas, prever ataques cibernéticos e responder a incidentes de segurança de maneira mais eficaz.