1. Machine Learning é uma área da Inteligência Artificial na qual se usa a capacidade dos computadores em reconhecer padrões para tomada de decisões, de grande utilidade quando se tem um problema que envolve grande complexidade matemática entre as entradas e os resultados dos quais queremos estabelecer uma relação.

2. Os conjuntos citados são subconjuntos do conjunto de dados total, que possui elementos com as classificações já conhecidas, que são utilizadas para comparar com os resultados do modelo e avaliar se o mesmo é capaz de fazer uma boa predição . O conjunto de treinamento será utilizado no primeiro treino do algoritmo. Ao fim do treino, é feita a comparação entre as classificações feitas pelo algoritmo e as classificações reais.. Um dos problemas que se pode enfrentar ao realizar o treino com o conjunto de treinamento é o overfitting, onde o modelo se comporta especificamente para o conjunto de treino, sendo um resultado ruim, pois o modelo deve generalizar os resultados para qualquer conjunto no qual seja aplicado. Para evitar isso, utiliza-se o conjunto de validação, que traz outros elementos diferentes dos anteriores para treinar o algoritmo. Nesta etapa, é feito o ajuste dos hiperparâmetros para garantir que o modelo seja robusto. O conjunto de teste utiliza os dados que não estavam nos conjuntos anteriores para validar o modelo gerado após o treino e a validação como um modelo robusto e imparcial que está pronto para ser usado em dados que não possuem resultados conhecidos.

3. Primeiramente, é importante analisar a distribuição dos dados e que tipo de dados faltantes se tem para decidir como tratá-los. Mas, de modo geral, pode-se fazer a retirada dos dados faltantes do conjunto de dados ou substituí-los por algum valor. Alguns exemplos de como fazer essa substituição são: substituição pela mediana; uso de regressão para previsão dos dados; substituição por um valor arbitrário; uso de distribuição de probabilidades; etc.

4. A matriz de confusão é uma métrica utilizada para atestar matematicamente o desempenho do modelo criado. É representada por uma tabela que traz as frequências de cada classificação feita pelo modelo em comparação com as classificações reais, fazendo, então, as relações de Positivo e Negativo, indicando os caminhos da classificação, e Verdadeiro ou Falso, indicando se foi uma classificação coerente com a classificação real. Diante disso, são retornados alguns valores que permitem estimar a performance do algoritmo, como precisão e acurácia

5. Sou entusiasta da aplicação de machine learning em qualquer área que seja possível, acho muito interessante a pluralidade de casos que podemos facilitar com o uso dessa ferramenta. Mas gostaria de citar especificamente uma, que é a minha área de interesse, que é o uso de machine learning na Astronomia. Temos casos interessantes de classificação e regressão que envolvem relações não lineares bem complexas e sem modelo definido. Existe uma grande motivação do uso de machine learning na determinação de parâmetros essenciais para estudos acerca do Universo, incluindo a

criação de algoritmos específicos para Astronomia. A Astronomia é uma área que utiliza muito a ciência de dados, uma vez que é uma ciência essencialmente observacional.