A seguir, você encontra as principais etapas de um projeto de ML e a descrição do que é feito em cada etapa, desde o entendimento do problema até a utilização efetiva do sistema. Confira.

* 1. **Coleta de dados:** etapa em que o cientista de dados usa a linguagem SQL para coletar dados de um ou vários bancos de dados para construir sua tabela inicial de análise. Em muitos casos, o Python é usado para coletar dados via Application Programming Interface (API).
* 2. **Limpeza dos dados:** após a coleta, o cientista de dados avalia o conjunto de dados, procurando, principalmente, por outliers através de métodos matemáticos, devido a erros no sistema e inconsistência nos dados, por exemplo, a idade de um ser humano acima de 200 anos.
* 3. **Exploração de dados:** momento de analisar os dados em busca de respostas para as hipóteses levantadas pelo time de negócio, utilizando estatística para encontrar insights. A maioria dos projetos de Ciência de Dados acaba nessa fase, pois um modelo de ML nem sempre é a solução necessária para o negócio.
* 4. **Modelagem de dados:** preparação dos dados para "ensinar" os algoritmos de ML. As principais atividades nessa fase são transformações de variáveis categóricas e seleção de features.
* 5. **Algoritmos de ML:** fase da aplicação dos algoritmos de ML para criar um modelo do fenômeno observado. Seja um fenômeno de venda, churn entre outros. Lembrando que algoritmos de ML são, em sua essência, funções matemáticas, certo?
* 6. **Avaliação dos algoritmos de ML:** etapa para avaliar a performance dos algoritmos de ML em termos de suas principais métricas. As métricas, puramente estatísticas, variam de acordo com o problema e com o que o time de negócio espera do modelo de ML.
* 7. **Publicação do modelo em produção:** última etapa do processo, em que o cientista de dados torna seu modelo público e com acesso aos seus resultados. Geralmente, feito por meio de um banco de dados ou de um API.

Vários algoritmos de ML utilizam distância para prever ou classificar algo, por exemplo, k-means, KNN etc. em que o objetivo é determinar a qual grupo uma determinada amostra vai pertencer com base nas amostras vizinhas. Ao contrário dos outros algoritmos, ele não constrói um modelo, ele faz somente o cálculo de distância.

A seguir, há duas imagens, uma de um triângulo retângulo com dois de seus lados medindo 3 e 4 centímetros cada, e a outra imagem contém um plano cartesiano com dois pontos, posicionados nas coordenadas (1,3) e (3,2).

Figura 3 – Triângulo pitagórico

Figura 4 – Plano cartesiano com dois pontos especificados

