Aula 06 - Conjunto de dados (Datasets)

Olá, bem vindo a mais uma aula do curso de inteligência artificial. Nas aulas anteriores nós vimos várias definições e aplicações do aprendizado de máquina. Nesta aula, vamos aprender sobre conjunto de dados conhecidos também como datasets.
Mas o que é um conjunto de dados?
Quando estamos trabalhando com IA e precisamos treiná-la, precisamos passar a ela dados os quais ela usará como parâmetros para o aprendizado. Um conjunto de dados são arquivos que contém registros sobre um determinado assunto. São organizados em linha e colunas e são usados para análise de dados e estatística. Como já foi citado antes, para que uma IA seja treinada ela necessita de uma base de dados.

Essa base de dados também é conhecida como datasets, conjunto de dados, ou ainda entrada de dados. É o principal elemento de um processo de análise de dados.

vamos ver agora a estrutura de um conjunto de dados. Suponha que tenhamos uma tabela que representa uma pesquisa de refrigerante preferido de um determinado grupo de pessoas.

ID	Nome	Idade	Refrigerante
1	Matheus	18	Coca-cola
2	Iago	21	Guaraná
3	Sabrina	24	Pepsi
4	Lucas	23	Coca-cola
5	Luana	15	Fanta

Um conjunto de dados é composto por colunas que fazem o registro da informação. As linhas fazem o registro do que foi definido pela coluna. E existe também o ID, que é um conjunto de números para atribuir a cada registro uma identificação única. É útil se queremos facilitar e até mesmo preservar a identidade dos nossos indivíduos. E existem ainda os formatos de arquivos das bases de dados. Os mais comuns e utilizados são os seguintes: Planilhas do excel no formato .XLS, documentos no formato .CSV (c*omma separated value* em português valores separados por vírgula) e o formato de arquivo .TSV que é semelhante ao .CSV mas os valores são separados por TAB. Nós podemos tanto criar uma base de dados quanto utilizar uma base de dados que já existe.

Podemos usar uma base de dados chamada **Kaggle** que é uma das principais comunidades de para aprendizado e até mesmo competições para ciência de dados. Existe também o **UCL** que é um repositório de dados para aprendizado de máquina. Existe também o portal **brasileiro de dados abertos** onde é possível encontrar dados sobre várias e setores do Brasil. Mas de que maneira criamos um conjunto de dados?

veremos isso mais adiante, quando tivermos conhecimento em python para criar nossos primeiros projetos. Por essa aula é isso até a próxima aula!