# A base de dados escolhida para o projeto A3 é a **“Amazon Products Dataset 2023”** referente ao mês de setembro, estando disponível no Kaggle: (<https://www.kaggle.com/datasets/asaniczka/amazon-products-dataset-2023-1-4m-products?select=amazon_products.csv>).

Utilizaremos esta base de dados para a UC de “Análise de Dados e Big Data” e “Inteligência Artificial”.

As colunas da base de dados estão divididas da seguinte forma:

**Arquivo “amazon\_products”:**

* “asin”: Coluna referente aos ID dos produtos (Tipo String)
* “title”: Coluna referente aos nomes dos produtos (Tipo String)
* “imgUrl”: Coluna referente ao Url da imagem dos produtos (Tipo String)
* “productUrl”: Coluna referente ao Url dos produtos (Tipo String)
* “stars”: Coluna referente a avaliação dos clientes sobre os produtos, variando de 0 a 5 (Tipo Float)
* “reviews”: Coluna referente ao número de comentários sobre os produtos (Tipo Int)
* “price”: Coluna referente ao preço dos produtos (Tipo float)
* “listPrice”: Coluna referente ao preço inicial dos produtos, caso o produto não tenha sofrido alteração de preço, o valor será 0 (Tipo float)
* “category\_id”: Coluna referente a categoria dos produtos, se baseando no arquivo “amazon\_categories”(Tipo Int)
* “isBestSeller”: Coluna referente ao produto estar ou não na categoria de “mais vendido” (Tipo Boolean)
* “boughtInLastMonth”: Coluna referente ao número de vendas no último mês (Tipo Int)

**Arquivo “amazon\_categories”**

* “id”: Coluna referente ao ID das categorias dos produtos (Tipo Int)
* “category\_name”: Coluna referente ao nome das categorias dos produtos (Tipo String)

**Variável Target**

A variável target escolhida foi a coluna “IsBestSeller” para prever os produtos presentes na categoria dos mais vendidos do próximo mês.  
  
Escolhemos essa variável target , pois é possível traçar o desempenho de um produto, facilitando uma tomada de decisão para um possível fornecedor.

Para aplicação das técnicas de IA será necessário a transformação da variável target(“IsBestSeller”).

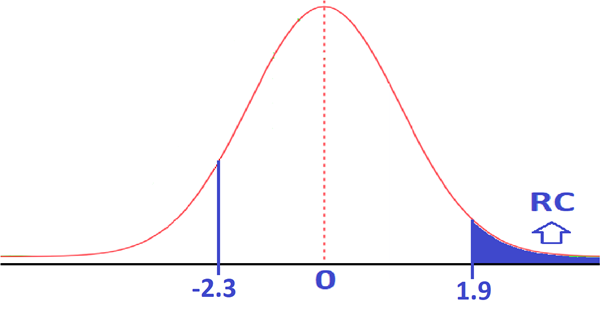
**Levantamento de hipótese**

A hipótese que será levantada e testada a partir da base de dados escolhida terá como base a seguinte indagação: " O preço dos produtos da categoria “Video Game” best sellers é maior do que o do que os produtos da categoria “Video Game” não são?"

Para esse caso, será observada uma amostra A da base de dados, apenas formada por linhas onde a coluna "IsBestSeller" = 1, e outra amostra B onde "IsBestSeller" = 0, com essas amostras seria possível calcular a média e o desvio padrão do preço de cada uma dessas amostras.

A hipótese alternativa seria a seguinte: "A média de preço da amostra A é maior que da amostra B?" Já a hipótese nula seria o inverso.

Para calcular o teste de hipótese, foi utilizado o PowerBi para levantamento dos dados da base. Após o cálculo do teste de hipótese, conclui-se que a hipótese alternativa levantada anteriormente foi descartada e, estatisticamente, a hipótese nula foi comprovada como verdadeira.

****

**1º Método de aprendizagem de máquina**

Como primeiro método de aprendizagem de máquina utilizamos a Árvore de Decisão com os seguintes parâmetros:

test\_size=0.35

criterion='entropy'

max\_depth= 5

Após o treinamento, foram obtidos os seguintes resultados:

precision recall f1-score support

False 0.99 1.00 1.00 496229

True 0.73 0.00 0.01 2989

accuracy 0.99 499218

macro avg 0.86 0.50 0.50 499218

weighted avg 0.99 0.99 0.99 499218

**2º Método de aprendizagem de máquina**

Como segundo método de aprendizagem de máquina utilizamos o **Naive Bayes** com os seguintes parâmetros:

test\_size= 0.2

modelo1 = GaussianNB()

Após o treinamento, foram obtidos os seguintes resultados:

A acurácia do modelo foi 63.72%

Matriz de confusão:

Predito     0 1  Todos

Real

0        1568   30   1598

1        1131  471   1602

Todos 2699  501   3200

**3º Método de aprendizagem de máquina**

Como terceiro método de aprendizagem de máquina utilizamos a **Regressão Linear** com os seguintes parâmetros:

test\_size=0.12

random\_state = 102

solver='lbfgs'

max\_iter=1000

Após o treinamento, foram obtidos os seguintes resultados:

Acuracia: 76.2 %

Matriz de confusão:

Predito     0 1  Todos

Real

0         873   84 957

1         373  590 963

Todos 1246  674   1920

Com base nos resultados dos 3 métodos acima de aprendizagem de máquina, podemos observar que o método que obteve o melhor desempenho foi o de **Regressão Linear**.