CLASSIFICAÇÃO DE SPAM

André Estevam ra166348
Karen Malzoni ra177493
Pedro Artico ra185545
Kaulitz Guimarães ra188530

PROBLEMA

Vídeos do Youtube recebem muitos comentários diariamente:

-Fãs - Haters - Pessoas mal intencionadas/ SpamBots

Pessoas mal intencionadas/ SpamBots:

Disseminam comentarios contendo:

- Spyware
- Vírus
- Malware

CONJUNTO DE DADOS

- **O conjunto de dados foi retirado do site : https://archive.ics.uci.edu/ml/datasets.html**
- **Os dados vieram divididos em 5 conjuntos diferentes.**
- **Cada conjunto refere-se à um artista diferente :**

Shakira

PSY

LMFAO

Eminem

Katy Perry

ATRIBUTOS

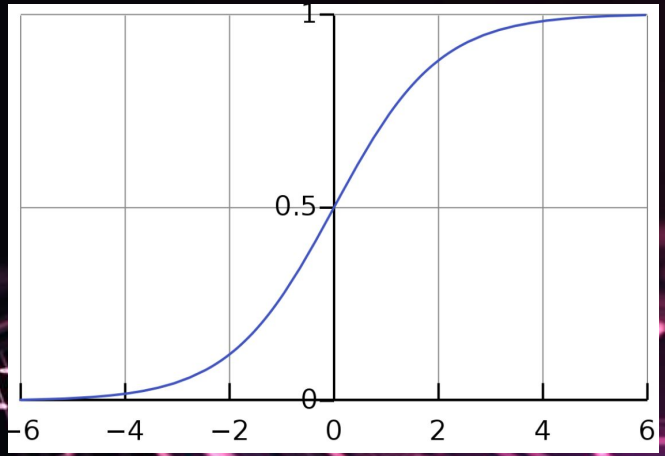
- **ANTES DA LIMPEZA :**
  - **Atributos presentes - author (textual), date (data/textual), content (textual), class (binário).**
- **DEPOIS DA LIMPEZA :**
  - **Limpeza de ruídos no atributo date.**
  - **Concatenação dos conjuntos.**
  - **Atributos presentes - author (textual), date (data/textual), content (textual), class (binário), artist (textual/descritivo).**

MINERAÇÃO DE TEXTOS

- **ATRIBUTO CONTENT**
- **TOKENIZING**
- **STOPWORDS**
- **BAG OF WORDS**

TAREFA

- **CLASSIFICAÇÃO :**
  - Logistic Regression :
    - Modelo linear.
    - Sigmoid function.
    - Probabilidade.



$$h_\theta(x) = g(\theta^T x)$$

$$z = \theta^T x$$

$$g(z) = \frac{1}{1 + e^{-z}}$$