**Guia 1**

Análise e Predição de Vendas Globais de Jogos

Aluno: Bruno Felipe Soares Santos

[bfss@cesar.school](mailto:bfss@cesar.school)

# Introdução

Este projeto tem como objetivo a análise e previsão das vendas globais de jogos ao longo dos anos, de 1980 a 2015. Para isso, foi utilizado um dataset do Kaggle (https://www.kaggle.com/datasets/thedevastator/global-video-game-sales), contendo informações sobre títulos, plataformas, gêneros, anos de lançamento e vendas globais. Vale notar que primeiro foi definido o dataset e após uma análise dos dados é que foi definido que seria a quantidade de vendas globais. Além disso, observa-se que o dataset contém informações precisas até o ano de 2015, a partir de 2016 existem muitos dados sem informação do ano, o que compromete a utilização destas informações para uma previsão com mais acurácia.

A abordagem utilizada incluiu uma análise descritiva dos dados, modelagem com métodos clássicos e redes neurais, e uma avaliação comparativa entre os modelos.

# Análise Descritiva dos Dados

Inicialmente, os dados foram carregados e inspecionados para entender sua estrutura e consistência. Algumas etapas importantes foram:

* Remoção de valores nulos na coluna 'Year' e conversão dos valores para inteiro.
* Exclusão de dados após 2015, devido às limitações do dataset para anos mais recentes.
* Substituição de valores nulos em 'Publisher' por 'Desconhecido'.
* Geração de visualizações exploratórias, como distribuição de vendas por ano, frequência de vendas globais, distribuição de jogos por plataforma e gênero.
* Conversão da coluna 'Year' para o formato de data e agregação das vendas globais por ano para análise temporal.

Após o tratamento de dados inicial, foram analisadas as informações existentes no dataset a fim de encontrar relações interessantes para definição do que seria trabalhado, escolhendo então a quantidade de vendas globais.

# Modelagem com ARIMA

Para a previsão utilizando métodos clássicos, foi utilizado o modelo ARIMA. As etapas incluíram:

* Análise das funções de autocorrelação (ACF) e autocorrelação parcial (PACF) para definir a ordem do modelo, escolhendo então os valores p = 1, d = 1, e q = 2.
* Treinamento do modelo ARIMA e previsão das vendas globais para os próximos cinco anos.
* Cálculo de métricas de avaliação: MSE (Erro Quadrático Médio), RMSE (Raiz do Erro Quadrático Médio) e MAE (Erro Absoluto Médio).

# Modelagem com LSTM

Além do modelo ARIMA, foi utilizado o modelo de redes neurais recorrentes (LSTM) para previsão temporal. Os principais passos foram:

* Normalização dos dados utilizando MinMaxScaler.
* Criação de sequências temporais para treinamento e teste do modelo.
* Construção de uma rede neural LSTM com duas camadas LSTM e uma camada densa de saída.
* Treinamento do modelo utilizando otimizador Adam e Early Stopping para evitar overfitting.
* Geração das previsões e inversão da normalização.
* Cálculo das mesmas métricas de avaliação utilizadas no ARIMA.

Vale notar que foi testado diversos valores de time\_step (1, **2**, 3, 4, 5, 10, 15), diferentes valores de units (**50**,**100**,150) para return\_sequences sendo True e sendo False, com e sem dropout e com diferentes valores (0.1, **0.2**, 0.3), diferentes valores de patience (5, 10 e 15), quantidades de épocas distintas (20, 30, 50, 70, **100**) e diferentes tamanhos de batch size (16, 32, **50** e 64), além da inclusão do callback de EarlyStopping. Sendo então os valores encontrados no notebook (e destacados aqui) os utilizados.

# Análise Comparativa

Os resultados das previsões foram comparados utilizando as métricas obtidas:

* **ARIMA:**
  + MSE: 28983.12
  + RMSE: 170.24
  + MAE: 156.24
* **LSTM:**
  + MSE: 13546.70
  + RMSE: 116.39
  + MAE: 98.79

A LSTM apresentou melhor desempenho que o modelo ARIMA, com menor erro absoluto e quadrático médio, indicando maior precisão nas previsões.

# Conclusão

O projeto demonstrou a importância da análise temporal para compreender padrões de vendas no mercado de jogos. Modelos de redes neurais, como LSTM, mostraram-se mais eficazes que métodos clássicos como ARIMA, ao capturar padrões mais complexos nos dados. Os resultados indicam que o uso de abordagens mais sofisticadas pode contribuir para previsões mais acuradas no setor.