**INSTITUTO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DA PARAÍBA**
**COORDENAÇÃO DO CURSO BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO**

**JOÃO PEDRO DE LIMA E SILVA**

# Projeto Final
# Mineração de Dados

CAMPINA GRANDE - PB
2025

## 1. Resumo da Proposta da Análise

Este projeto de mineração de dados teve como foco principal o desafio da manutenção preditiva, uma área crítica na engenharia e na indústria aeronáutica. O objetivo era desenvolver um modelo preditivo capaz de estimar a Vida Útil Restante (RUL - Remaining Useful Life) de motores a jato. Para isso, foi utilizado o conjunto de dados C-MAPSS (Commercial Modular Aero-Propulsion System Simulation), disponibilizado pela NASA, que simula o comportamento de motores turbofan desde o início de sua operação até o ponto de falha.

O projeto foi segmentado em etapas, começando pela Análise Exploratória de Dados (EDA), para entender as características das variáveis e os padrões de degradação. Em seguida, a modelagem preditiva foi realizada com o objetivo de construir e avaliar modelos de regressão que pudessem prever a RUL com a maior precisão possível. O foco inicial foi no subconjunto FD001, que apresentava condições operacionais e um único modo de falha, servindo como um cenário controlado para o desenvolvimento e validação da metodologia.

## 2. O que deu certo

A execução do projeto alcançou resultados promissores, destacando-se os seguintes pontos:

### 2.1. Análise Exploratória e Pré-processamento de Dados

A etapa de EDA foi fundamental para transformar os dados brutos em uma base utilizável. Conseguimos estruturar o conjunto de dados, atribuindo nomes significativos às colunas, o que facilitou a interpretação. A análise descritiva revelou que a vida útil média dos motores no conjunto de treino era de aproximadamente **206 ciclos**. Além disso, a inspeção das variáveis operacionais e dos sensores permitiu identificar que alguns sensores, como o s_18 e o s_19, não apresentavam variabilidade, indicando que poderiam ser descartados em etapas futuras.

### 2.2. Criação da Variável-Alvo

Um dos passos mais críticos foi a criação da variável-alvo RUL, que não estava presente no conjunto de treino original. Partindo do pressuposto de que a degradação dos motores é um processo linear e cumulativo, conseguimos calcular a RUL para cada ciclo de operação, transformando o problema em um de regressão supervisionada. Essa etapa foi essencial para a aplicação dos modelos de Machine Learning.

### 2.3. Modelagem Preditiva e Avaliação de Desempenho

A fase de modelagem foi um sucesso, com a implementação de três modelos distintos para comparação. Os resultados finais (conforme o notebook Projeto_Final.ipynb) demonstram o desempenho superior da **Rede Neural (MLP Regressor)**.

- **Rede Neural (MLP Regressor):** Obteve um desempenho significativamente superior, com **de 0.96** no conjunto de treino e R2 **de 0.95** no conjunto de teste. O **RMSE** no conjunto de teste foi de **52.83 ciclos**. A pequena diferença entre o desempenho nos conjuntos de treino e teste é uma forte indicação de que o modelo generalizou bem e não sofreu de sobreajuste (overfitting).

- **Comparação com Outros Modelos:** Conforme a análise inicial presente no README, a Rede Neural superou modelos como a Regressão Linear e o Random Forest, cujos R2 **foram de 47.28% e 48.85%**, respectivamente. Este contraste ressalta a capacidade da rede neural em capturar as relações não-lineares e complexas presentes nos dados de degradação.

## 3. O que deu errado

Apesar dos bons resultados, o projeto apresentou algumas limitações que poderiam ser melhoradas em um futuro trabalho:

### 3.1. Análise e Tratamento de Outliers

A proposta inicial da Sprint 03 incluía a detecção de outliers por meio de boxplots, mas essa análise não foi aprofundada. Não foi feito um tratamento sistemático desses valores discrepantes, o que poderia ter impactado a robustez dos modelos. Outliers podem introduzir ruído e distorcer os resultados, especialmente em modelos de regressão, e uma investigação mais detalhada poderia ter levado a um desempenho ainda melhor.

### 3.2. Foco Exclusivo em um Subconjunto de Dados

O projeto limitou-se ao subconjunto FD001, que apresenta um cenário simplificado. A base de dados da NASA inclui outros subconjuntos (FD002, FD003 e FD004) com diferentes condições de operação e múltiplos modos de falha. A falta de análise e treinamento do modelo com esses dados restringe a aplicabilidade e a generalização do modelo a cenários mais complexos e realistas.

## 4. O que faria diferente?

Se houvesse a oportunidade de reiniciar o projeto, as seguintes melhorias seriam implementadas:

1. **Pré-processamento Mais Agressivo:** Eu dedicaria mais tempo à engenharia de features, removendo não apenas os sensores com variabilidade nula, mas também explorando a criação de novas variáveis a partir das existentes.
2. **Análise de Outliers e Valores Ausentes:** Integraria uma etapa de tratamento de outliers e valores ausentes mais robusta, testando diferentes estratégias (remoção, imputação) e avaliando o impacto de cada uma no desempenho final do modelo.
3. **Expansão para Múltiplos Subconjuntos:** O projeto seria expandido para incluir a análise dos subconjuntos FD002, FD003 e FD004. Isso exigiria uma etapa adicional de normalização dos dados para lidar com as diferentes condições operacionais, mas resultaria em um modelo muito mais generalizável e aplicável a uma variedade maior de situações do mundo real.
4. **Ajuste de Hiperparâmetros da Rede Neural:** Embora a Rede Neural tenha apresentado o melhor desempenho, a busca por hiperparâmetros (número de camadas, neurônios, taxa de aprendizado) não foi exaustiva. Um ajuste fino poderia otimizar ainda mais o modelo.

## 5. Aprendizados sobre Mineração de Dados

Este projeto foi uma experiência valiosa que consolidou diversos conceitos da mineração de dados. Os principais aprendizados foram:

A Abordagem Interativa: Um projeto de mineração de dados não é linear. A EDA e a modelagem se complementam, e os insights obtidos em uma etapa podem levar a ajustes nas anteriores.

- A Importância da Preparação dos Dados: A qualidade e a preparação dos dados são tão, ou até mais, importantes que a escolha do modelo. A criação da variável RUL e a identificação de variáveis irrelevantes na EDA foram os pilares para o sucesso da modelagem.
- Compreensão das Métricas de Avaliação: A análise do R2 e do RMSE foi crucial para entender não apenas quão bem o modelo se ajustava aos dados, mas também a sua capacidade de generalização.
- O Poder dos Modelos Avançados: O desempenho superior da Rede Neural em um problema com dados complexos e não-lineares demonstrou o potencial desses modelos em relação a abordagens mais tradicionais, como a Regressão Linear.