**Universidade Federal de Pernambuco**
**Centro de Ciências Exatas e da Natureza**
**Departamento de Estatística**

### TÓPICOS ESPECIAIS EM ESTATÍSTICA COMPUTACIONAL

**Data:** 21 de março de 2025

# Introdução

Prever possíveis falhas em máquinas e equipamentos industriais é uma ótima forma de combater um dos problemas mais recorrentes em empresas. Quando um equipamento falha ele pode acarretar vários problemas que podem gerar problemas de produção, acidentes e grandes percas financeiras, e a melhor forma de evitar isso é prevenindo esses erros. Com isso, esse trabalho tem como objetivo buscar um forma de prever essas falhas através de modelos de predição com aprendizado de máquina utilizando dados sintéticos reflete a manutenção preditiva real encontrada no setor, onde são apresentados variáveis que possam nos ajudar a chegar o mais próximo do resultado desejado e aumentar o máximo possível a possibilidade de evitar as falhas nas maquinas.

# Fundamentos Teóricos e Metodológicos

O conjunto de dados utilizado nesse projeto se chama "Explainable Artificial Intelligence for Predictive Maintenance Applications" e é composto por 14 variáveis e 10.000 observações. Foi utilizada a linguagem Python para obter as análises e a aplicação dos modelos de Árvore de Decisão e Floresta Aleatória, a variável de classificação é a variável binária "Target" que apresenta 0 para as observações que a maquina apresentou falha e 1 para as que não apresentaram falha.

**Variáveis utilizadas:**

- **UID**: identificador único que varia de 1 a 10000.
- **productID**: consistindo em uma letra L, M ou H para baixo (50% de todos os produtos), médio (30%) e alto (20%) como variantes de qualidade do produto e um número de série específico da variante.
- **Type**: Apresenta somente as variantes de qualidade de produto (L, M e H).
- **air temperature [K]**: gerada usando um processo de caminhada aleatória posteriormente normalizada para um desvio padrão de 2 K em torno de 300 K.
- **process temperature [K]**: gerada usando um processo de caminhada aleatória normalizado para um desvio padrão de 1 K, adicionado à temperatura do ar mais 10 K.
- **rotational speed [rpm]**: calculada a partir da potência de 2860 W, sobreposta com um ruído normalmente distribuído.
- **torque [Nm]**: os valores de torque são normalmente distribuídos em torno de 40 Nm com um Ïf = 10 Nm e sem valores negativos.
- **tool wear [min]**: As variantes de qualidade H/M/L adicionam 5/3/2 minutos de desgaste da ferramenta à ferramenta usada no processo.
- **Failure Type**: Indica o tipo de falha da máquina. Não utilizamos essa variável pois também é uma variável de classificação.

# Aplicação

De início vamos apresentar a analise exploratória dos dados para que possamos entender como os dados a distribuição dos dados de cada variável presente no conjunto de dados.

### Análise Exploratória

Table 1: Medidas descritivas das variáveis númericas

| Medida | Air temperature [K] | Process temperature [K] | Rotational speed [rpm] | Torque [Nm] | Tool wear [min] |
|---|---|---|---|---|---|
| Min. | 295.3 | 305.7 | 1168 | 3.80 | 0 |
| 1st Qu. | 298.3 | 308.8 | 1423 | 33.20 | 53 |
| Median | 300.1 | 310.1 | 1503 | 40.10 | 108 |
| Mean | 300.0 | 310.0 | 1539 | 39.99 | 108 |
| 3rd Qu. | 301.5 | 311.1 | 1612 | 46.80 | 162 |
| Max. | 304.5 | 313.8 | 2886 | 76.60 | 253 |

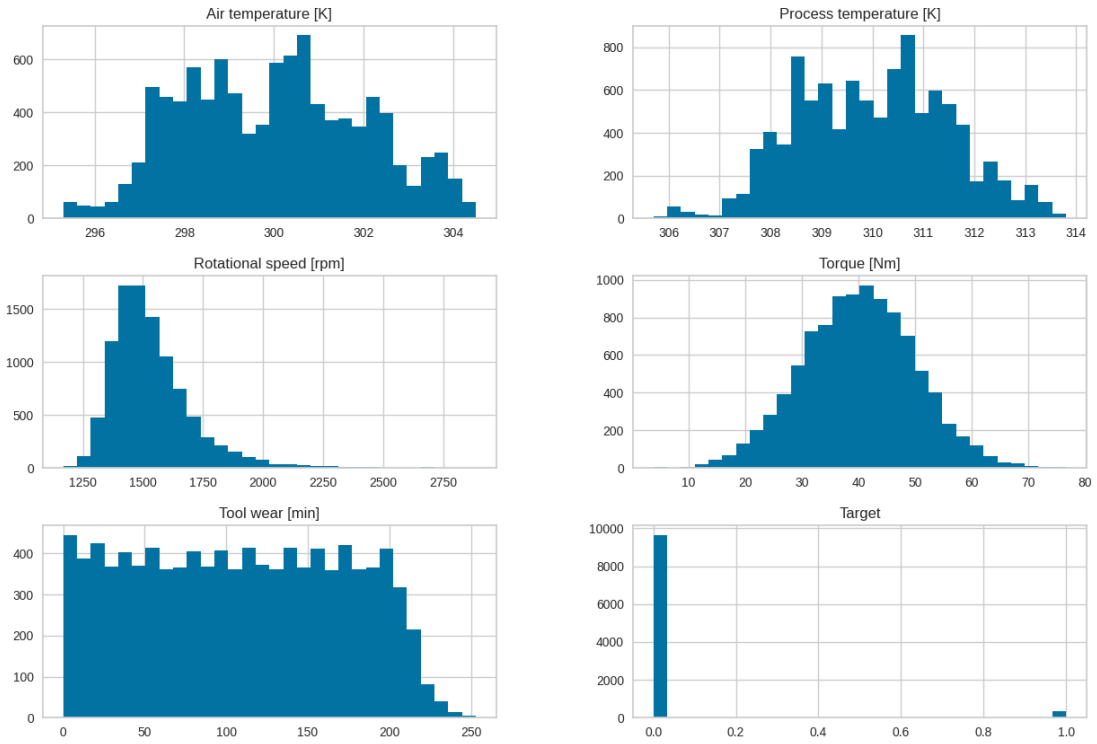



Figure 1: Gráficos com de distribuição das variáveis númericas

Na **Table 1** podemos observar que as variáveis *"Air temperature [K]"* e *"Process temperature [K]"* apresentam baixa variação na distribuição dos dados quando observamos os quartis, já as demais variáveis possuem grande variação, todas essas afirmações podem ser observadas na **Figure 1**.

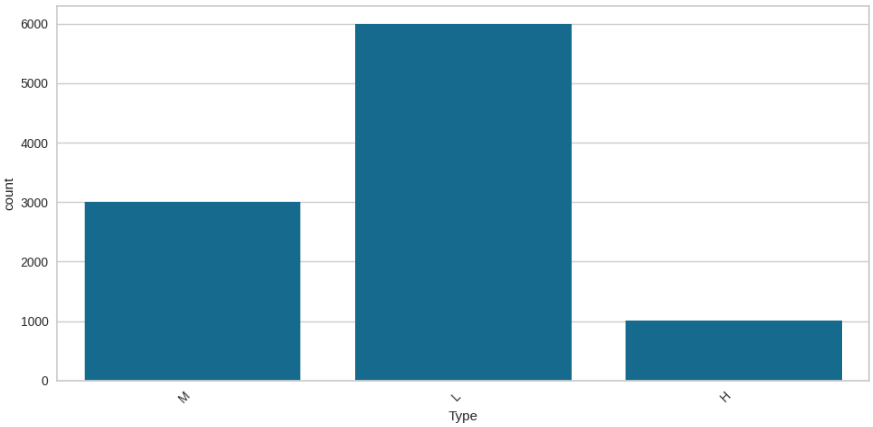



Figure 2: Gráfico com a distribuição da váriavel “Type”

A única variável categórica que temos no conjunto de dados é a várivavel *”Type”* que classifica os produtos por qualidade, sendo L para baixa qualidade, M para média qualidade e H para Alta qualidade. Na **Figure 2** podemos ver que as maquinas de baixa qualidade são as mais presentes nos dados.

### Aplicação do modelo

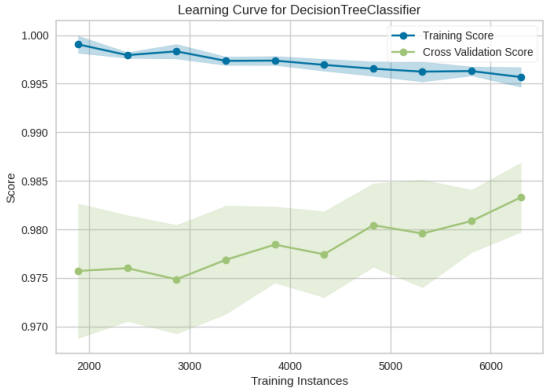
Para essa estudo escolhemos os modelos de **Árvore de decisão** e **Floresta aleatória** para comparar qual modelo apresenta melhor desempenho resolver nosso problema.

| Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| DT Classifier | 0.9850 | 0.7785 | 0.4872 | 0.6552 | 0.5588 | 0.5514 | 0.5576 |

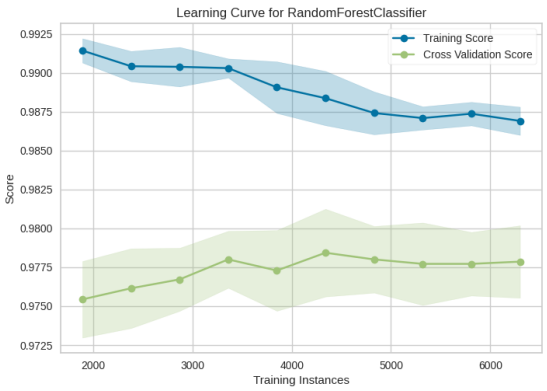
(a) Métricas do modelo de Árvore de decisão

| Model | Accuracy | AUC | Recall | Prec. | F1 | Kappa | MCC |
|---|---|---|---|---|---|---|---|
| RF Classifier | 0.9915 | 0.9884 | 0.5641 | 1.0000 | 0.7213 | 0.7173 | 0.7478 |

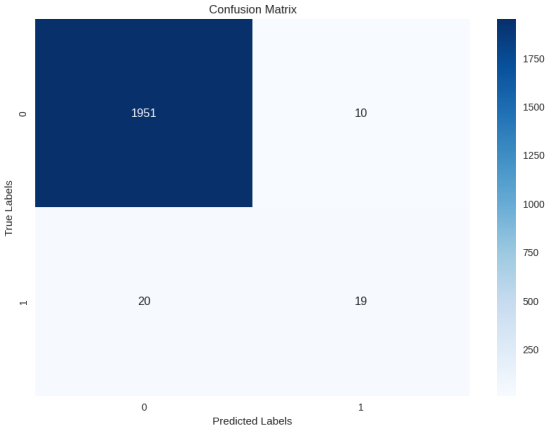
(b) Métricas do modelo de Floresta Aleatória



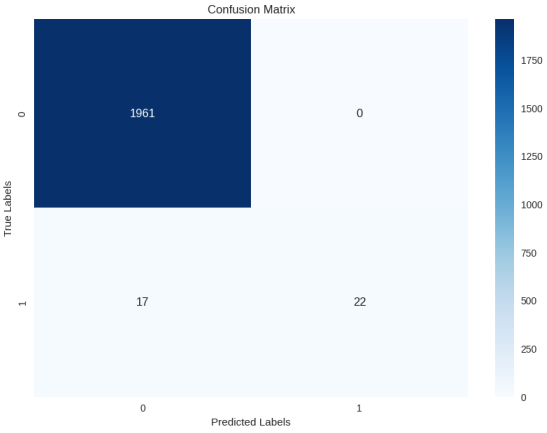

(c) Curva de aprendizado da Árvore de decisão




(d) Curva de aprendizado da Floresta Aleatória




(e) Matriz de confusão da Árvore de decisão




(f) Matriz de confusão da Floresta Aleatória

Nas imagens acima podemos ver que o modelo que apresentou melhor desempenho foi o modelo de árvore aleatória, isso pode ser influenciado pela dimensão dos dados já que o modelo tem melhor desempenho para grandes quantidades de observações se comparado com o modelo de árvores de decisão. Nas matrizes de confusão podemos observar que apesar do desempenho da floresta aleatória se sobressair em relação a árvore de decisão, podemos que o modelo obteve um resultado muito baixo para classificar as máquinas com falha, identificando apensa 56% das máquinas.

# Conclusão

Nesse estudo utilizamos os dados em modelos de classificação e tentamos chegar ao melhor resultado possível comparando esses dois modelos e observando o desempenho de ambos para encontrarmos uma forma de classificar maquinas com possíveis futuras falhas com objetivo de diminuir problemas nas empresas.

Vimos que modelo de floresta aleatória obteve um resultado melhor se comparado com o resultado do modelo de árvore de decisão, contudo, não podemos considerar esse resultado o melhor possível podemos pensar: Qual foi o problema?

Mesmo com a grande quantidade de observações e variáveis que utilizamos, podemos perceber que os resultados não apresentaram a melhor performance possível dos modelos, contudo tivemos a oportunidade de comparar o desempenho dos dois modelos se testados com o mesmo conjunto de dados e observar suas performances.

# Contribuição da equipe

- *Edgar Carlos Rodrigue de Oliveira:* Fez a ànalise exploratória e aplicou o modelo de árvore aleatória (50% de contribuição).
- *Vitória Karolinny dos Santos Gonçalves:* Aplicou o modelo de Floresta Aleatória e criou o relatório e slide (50% de contribuição).

# Referências


- Link para ter acesso aos dados: https://archive.ics.uci.edu/dataset/601/ai4i+2020+predictive+maintenance+dat