# \*\*Situação problema\*\*

Um problema persistente na área da manutenção é a identificação e redução de falhas nos equipamentos em serviço. Uma das técnicas mais utilizadas para a redução de falhas em equipamentos é a aplicação da manutenção preditiva, onde é realizado o acompanhamento, monitoramento e inspeção de máquinas com o objetivo de coletar dados que informam o desgaste dos equipamentos e seu processo natural de degradação. Assim, com esses dados, é possivel realizar manutenções em períodos anteriores a falha, aumentando a vida útil do equipamento.

Porém, como definir qual é o momento ideal para realizar a parada do equipamento?

Visando definir este tempo ideal, será desenvolvido um algoritmo de machine learning capaz de prever com antecedência quando ocorrerá a falha do equipamento, sendo possível executar a manutenção preditiva com maior eficiência.

Para a avaliação do modelo será utilizado a métrica Recall e AOC, visando assim penalisar mais os erros de falsos negativos para que não ocorra falhas indesejadas.

Para a elaboração deste projeto será realizado as seguintes etapas:

- Introdução: Visando esclarecer o objetivo do projeto, as métricas e as etapas a serem realizadas no desenvolvimento do modelo de machine learning.

- Limpeza dos dados e pré-processamento: Será feito uma pré-visualização dos dados, visando identificar e tratar dados ausentes, verificar quantas variáveis existem e o tipo de cada uma, além da criação de features simples com base nas features iniciais.

- Análise Exploratória dos Dados (EDA): Será realizado visualizações com o objetivo encontrar insights relevantes para o desenvolvimento do algoritmo, juntamente realizando o tratamento de algumas features com base em insights levantados.

- Correlação e processamento final: Será analisado a correlação entre os dados visando evitar possíveis problemas de multicolinearidade, realizando a remoção das variáveis indesejáveis.

- Seleção e Implementação de modelo: É a etapa onde será elaborado o modelo de machine learning para a solução do problema. Nesta etapa também será realizado o tunning de hiperparâmetros do modelo, visando melhorar o resultado obtido inicialmente.

- Conclusão: Etapa onde será realizado comentários em relação ao projeto realizado e também alguns levantamentos de melhorias futuras.

# \*\*Limpeza dos dados e pré-processamento\*\*

<img src="/Predictive Maintenance/prints project/data\_info.png ">

No dataset tem-se o total de 11 variáveis, excluindo a variável target, sendo composto por 9 sensores em formato numérico, uma variável data em formato object necessitando conversão para datetime, e uma variável device em formato categórico.

<img src="/Predictive Maintenance/prints project/data\_desc.png ">

Observa-se que as variáveis 'metric1', 'metric5' e 'metric6' são as variáveis que possuem uma oscilação de valores.

As variáveis 'metric2', 'metric3', 'metric4', 'metric7', 'metric8' e 'metric9' possuem uma maior presença de valores iguais à 0 e alguns valores máximos elevados que podem estar relacionados com a presença de falhas.

<img src="/Predictive Maintenance/prints project/data\_date.png ">

Após a conversão da variável ‘date’ nota-se que os dados são coletados diariamente e englobam datas entre o dia 01/01/15 até 02/11/15.

<img src="/Predictive Maintenance/prints project/data\_unique.png ">

Analisando a tabela, vemos que no dataset temos 1169 categorias diferentes em 'device'.

Nota-se que as 'metric' 2, 3, 4, 5, 7, 8 e 9 possuem poucos valores únicos medidos.

<img src="/Predictive Maintenance/prints project/data\_unique.png ">

Nota-se inicialmente que não existe dados ausentes em nenhuma variável e somente 1 linha duplicada. Baseado nisso será feito a remoção desta linha duplicada.

Como manipulação inicial será realizado a criação de variáveis auxiliares para data, separando o dia, mês, ano, dia da semana e semana do mês. Também será realizado a separação da variável 'devices' em dois, sendo uma relacionada aos quatro primeiros valores da string, que indica o setor onde o equipamento está instalado, e a segunda feature relacionada aos outros caracteres da string, indicando o tipo de equipamento.

O correto para fazer a divisão da TAG do 'device' é saber o padrão da empresa trabalhada, para assim poder identificar corretamente o setor, qual equipamento está sendo abordado e o identificador único do equipamento. Como não temos o acesso a esta informação será feito a divisão simplificada inicialmente citada.

<img src="/Predictive Maintenance/prints project/data\_nf.png ">

Nota-se que subdividindo a variável 'device' encontramos que ela opera em 7 setores diferentes e com 1169 de equipamentos diferentes, igual a quantidade de devices inicialmente observada.

# \*\*Análise Exploratória dos Dados (EDA)\*\*

## Distribuição de falhas

<img src="/Predictive Maintenance/prints project/img\_01.png ">

Nota-se pelo gráfico uma proporção de dados normais muito maior do que os dados de falha, tendo somente 106 dispositivos que apresentaram falha.

Posteriormente será realizado uma operação visando balancear essas classes desta variável objetivo.

Como poucos equipamentos sofreram falhas, talvez eles tenham sido substituídos ou trocados por outro equipamento ao longo do tempo, portanto será feito uma verificação da quantidade de 'devices' ao longo do tempo neste dataset e se os dispositivos que apresentaram falha estão presentes no dataset desde o início.

<img src="/Predictive Maintenance/prints project/img\_02.png ">

Nota-se que ocorreu uma queda expressiva dos equipamentos em funcionamento com o passar dos meses, principalmente do mês de janeiro para fevereiro.

Observando os dias em que ocorreram falhas nos dispositivos, nota-se que esses dias não acarretam na diminuição dos equipamentos em funcionamento.

Também é possível observar que só houve aumento de quantidade significativa de dispositivos em dois períodos, no começo do mês de maio e no começo do mês de setembro.

Como a quantidade de dispositivos apresentado no início do mês de janeiro é igual a quantidade de dispositivos apresentados no dataset não foram adicionados novos equipamentos com o passar do tempo, mas pode ter ocorrido a parada de funcionamento desses equipamentos ou estabelecido uma rotatividade de funcionamento entre os dispositivos.

## Observação

\*\*Como as métricas são coletadas diariamente será considerado que as métricas coletadas para cada dispositivo foram coletadas ao mesmo tempo do mesmo dia.\*\*

\*\*Como não é informado em que momento a métrica foi colhida, seja no dia em que aconteceu a falha, um dia anterior a falha, seja com as métricas coletadas com o equipamento operando na iminência da falha ou em qualquer outra situação que possa ser decisiva para a construção do modelo de previsão, será suposto que 'a falha ocorrerá no dia posterior a medição realizada’.\*\*

Com essas considerações o modelo a ser planejado e construído será com base neste um dia de antecedência, para que assim possa ser feito a parada e a manutenção do equipamento com pelo menos um dia de antecedência a falha.

<img src="/Predictive Maintenance/prints project/img\_03.png ">

Nota-se pelo gráfico que não foram todos os equipamentos que tiveram falha que operaram durante o ano inteiro, tendo o mesmo perfil de diminuição de operação com o passar do tempo que o gráfico anterior.

É possível observar que ocorrem oscilações neste grupo de equipamentos em funcionamento, portanto neste caso é possível inferir que pode ocorrer uma alternância de funcionamento dos equipamentos ao longo dos meses, visto que o equipamento quando falha ou ele é substituido ou é realizado a sua manutenção e colocado em operação novamente.

Como existe esta alternância de funcionamento será criado uma variável para contabilizar a quantidade de dias trabalhado por cada dispositivo e uma outra variável para contabilizar quantas vezes o equipamento foi religado.

<img src="/Predictive Maintenance/prints project/img\_04.png ">

Pelo histograma de todo os dispositivos é possível inferir que mais de 350 dispositivos possui menos que 10 dias de trabalho. Além disso, é notável mais duas concentrações, sendo uma delas com tempo de trabalho entre 50 a 100 dias e outro com tempo de trabalho entre 200 e 250 dias.

Aproximadamente 27 dispositivos possuem mais de 300 dias em operação no dataset.

Analisando o histograma com somente os dispositivos que falharam, ocorre duas concentrações, sendo uma delas entre 10 e 50 dias de funcionamento e outra entre 100 e 150 dias.

<img src="/Predictive Maintenance/prints project/img\_05.png ">

Nota-se que a maioria dos equipamentos só foi ligado uma vez e que no máximo 5 equipamentos foram religados até 3 vezes. Isso pode indicar que na maioria dos casos foram feitas substituições dos equipamentos ao invés de um revezamento entre eles.

Analisando somente os equipamentos que apresentaram falha, a maioria só foi ligada uma vez e no máximo religado uma vez, indicando que na maioria dos casos foi feito a substituição do dispositivos e que em poucos casos ocorreu o reparo do equipamento e o mesmo foi posto para funcionar novamente.

## Avaliando os sensores

<img src="/Predictive Maintenance/prints project/img\_06.png ">

Na 'metric1' aparentemente todos os valores são divisíveis por 8, mas como a visualização dos valores únicos não é completa será feito uma verificação da multiplicidade para depois realizar a divisão.

<img src="/Predictive Maintenance/prints project/img\_07.png ">

Analisando os valores únicos da 'metric2' nota-se que todos são múltiplos de 8, exceto o valor 55.

Examinando os valores de ‘metric2’ só existe um único valor de 55, sendo todos os outros são múltiplos de 8. Portanto ele será modificado para o valor mais próximo, 56, e será feito a divisão dos dados por 8, criando uma nova métrica.

<img src="/Predictive Maintenance/prints project/img\_08.png ">

Para 'metric3' não é possível perceber nenhum padrão, porém é visível a presença de alguns saltos de valores.

<img src="/Predictive Maintenance/prints project/img\_09.png ">

Para 'metric4' não é possível perceber nenhum padrão.

<img src="/Predictive Maintenance/prints project/img\_09\_1.png ">

Para 'metric5' não é possível perceber nenhum padrão entre seus valores.

<img src="/Predictive Maintenance/prints project/img\_10.png ">

Aparentemente, observando 'metric6' não é possível identificar nenhum padrão, mas é possível observar um range elevado de valores nesta variável.

<img src="/Predictive Maintenance/prints project/img\_11.png ">

Para 'metric7' nota-se que seus valores são múltiplos de 2. Portando da mesma forma do que foi realizado para a métrica 2 será feito a divisão dos seus valores criando uma nova feature.

<img src="/Predictive Maintenance/prints project/img\_12.png ">

Para 'metric8' nota-se que seus valores são múltiplos de 2. Portando da mesma forma do que foi realizado para a métrica 2 será feito a divisão dos seus valores criando uma nova feature.

<img src="/Predictive Maintenance/prints project/img\_13.png ">

Para 'metric9' não é possível perceber nenhum padrão, porém é visível a presença de alguns valores muito elevados

Como algumas métricas foram alteradas, será feito a remoção das métricas anteriores, deixando somente as novas features.

<img src="/Predictive Maintenance/prints project/img\_14.png ">

Analisando os gráficos acima temos que:

- Apesar da variável 'mnw1' possuir valores elevados, eles são bem distribuidos.

- As variáveis 'mnw2','metric3', 'metric4', 'mnw7','mnw8' e 'metric9' possuem uma concentração de dados maior em torno do valor 0 e uma pequena quantidade de dados em valores elevados, possuindo uma skewness elevada.

- As variáveis 'metric5' e 'metric6' apresentam pequenas concentrações em valores distintos e espaçados.

- Nenhuma das variáveis apresentou uma diferença significativa entre as classes 0 e 1 de 'failure'.

Com base nesses insights será feita as seguintes manipulações:

- Como as variáveis 'metric5' e 'metric6' apresentam concentrações ditintas elas podem servir como uma medição contínua de uma característica do equipamento. portanto será criado uma variável auxiliar que representa a diferença entre o valor inicial da metrica com o seu valor ao longo dos dias para o dispositivo analisado.

- Como as variáveis 'mnw2','metric3', 'metric4', 'mnw7','mnw8' e 'metric9' possuem uma skewness elevada, será criado uma nova variável em formato logarítmico delas.

<img src="/Predictive Maintenance/prints project/img\_15.png ">

<img src="/Predictive Maintenance/prints project/img\_16.png ">

Observa-se que tanto para a métrica 'dif\_m5' quanto para a 'dif\_m6' não ocorre a presença de valores negativos, indicando que as variáveis 'metric5' e 'metric6' são variáveis contínuas crescentes.

# \*\*Correlação\*\*

Visando evitar possíveis problemas de multicolinearidade na resolução do problema será utilizado o método de correlação de Pearson, visando identificar a correlação linear entre as variáveis.

Como valor de corte será utilizado o valor de 0.75 e de -0.75, sendo esses valores definidos empiricamente em projetos anteriores.

<img src="/Predictive Maintenance/prints project/img\_17.png ">

Como as features 'log\_m7' e 'log\_m8' possuem correlação de 1, portanto a 'log\_m8' será removida.

A feature 'weekmonth' possui alta correlação com a feature 'day', portanto ela será removida.

Como a feature 'month' possui alta correlação com a feature 'op\_period', será feito a remoção de 'month'

Como visto anteriormente os dados foram coletados em um mesmo ano, logo a feature 'year' será removida.

Como a feature 'device' foi separada em 'sector' e 'equipment' ela será removida.

Como a feature 'equipment' representa um ID único do equipamento, em muitos casos ele pode enviesar muito o modelo, portanto ele será removido.

# \*\*Implementação do modelo\*\*

Para o desenvolvimento do modelo será elaborado uma \*\*'Rede Neural Direta (FNN)'\*\* levando em consideração a métrica \*\*'recall'\*\* para a verificação de desempenho do modelo. Também será realizado o tunning dos hiperparâmetros do modelo com base no pacote \*\*'OPTUNA'\*\* para otimizar o seu desempenho.

## Treinamento do modelo

Para o treinamento do algoritmo o dataset inicialmente será dividido em um dataset de treino e um dataset de teste. Após isso será aplicado a técnica de 'cross-validation' para 10 folds visando um algoritmo final mais generalizado.

Como existe um desbalanceamento das classes será aplicado o \*\*'StratifiedKFold'\*\*, que tem como objetivo a separação dos folds levando em consideração a proporção de cada variável, ou seja, se a classe minoritária existir em 1% do dataset, ele irá dividir esse 1% em X partes iguais e unirá com os outros 99% divididos também em X partes.

Durante o treinamento do algoritmo será realizado o método de \*\*upsample\*\* da classe minoritária por meio do pacote \*\*SMOTE\*\*, tornando a classe minoritária com a mesma quantidade da classe majoritária.

Para a variável categórica 'sector' será aplicado o \*\*'OneHotEncoder'\*\*, dividindo os setores em novas colunas.

Após isso será aplicado o \*\*'StandardScaler'\*\* nas variáveis para melhorar o desempenho e precisão do modelo.

<img src="/Predictive Maintenance/prints project/img\_18.png ">

Nota-se pelo gráfico do histórico do tunning do modelo que ocorreu uma melhora na performance do modelo com o decorrer das iterações, obtendo um \*\*'recall'\*\* médio de 80.38% na validação cruzada.

<img src="/Predictive Maintenance/prints project/img\_19.png ">

Observando o gráfico de importância dos hiperparâmetros, os que tiveram mais impacto no modelo é a quantidade de épocas no treinamento, o valor do learning rate e a números de neurônios da camada inicial.

<img src="/Predictive Maintenance/prints project/img\_20.png ">

Treinando novamente o modelo sem a validação cruzada, temos um resuldado com um recall de aproximadamente 62%, porém um valor de precisão próximo de 1%.

Nota-se que a quantidade de Falsos Positivos é equivalente a aproximadamente 10% da quantidade de Verdadeiros Positivos, que na prática em uma situação de manutenção, não seriam paradas/verificações muito recorrentes para garantir a confiabilidade e a seguraça do equipamento.

## Avaliando o dataset de teste

<img src="/Predictive Maintenance/prints project/img\_21.png ">

Aplicando o modelo para os dados de teste, obtivemos um resultado similar aos dados de treino.

# \*\*Conclusão\*\*

O resultado alcançado no modelo final foi aceitável possuindo um recall aproximado de 0.66, porém obteve um valor abaixo para a precisão.

A fim de melhorar o projeto em trabalhos futuros, as seguintes atividades podem ser realizadas:

- Aplicação de um modelo de clusterização para a criação de novas features, visando incrementar a capacidade preditiva do modelo.

- Teste de outros modelos de aprendizado de máquina, tentando aperfeiçoar o recall.

- Avaliação de importância das features na entrada do modelo, visando a remoção das features com pouca capacidade preditiva.