AGÊNCIA BRASILEIRA DE DESENVOLVIMENTO INDUSTRIAL (ABDI)

INSTITUTO DE INTELIGÊNCIA ARTIFICIAL APLICADA (I2A2)

AI PARA INDUSTRIA

JOÃO PAULO DA SILVA CARDOSO

RELATÓRIO:

ANÁLISE EXPLORATÓRIA DE DADOS DE SENSORES DE BOMBA D’ÁGUA PARA PREDIÇÃO DE FALHAS

27/02/2023

**1. INTRODUÇÃO**

Este relatório tem como objetivo apresentar uma análise exploratória de dados referente a um conjunto de sensores utilizados em uma bomba d'água. Em virtude de questões de anonimato, informações restritas foram fornecidas sobre as grandezas monitoradas pelos sensores, apenas sabe-se que uma equipe reduzida é responsável pela manutenção da bomba d'água em uma área distante da cidade grande e que ocorreram sete falhas no sistema no ano anterior, gerando graves consequências para diversas pessoas, bem como para algumas famílias. A equipe encarregada da manutenção não conseguiu identificar nenhum padrão nos dados quando ocorrem as falhas, o que dificulta a identificação da área na qual se deve concentrar mais atenção para evitar esses eventos. Dessa forma, o objetivo da análise é compreender de maneira mais aprofundada os padrões de falhas identificados pelo conjunto de dados dos sensores.

**2. DESENVOLVIMENTO**

**2.1 Conjunto de dados**

A base de dados utilizada encontra-se disponível em uma plataforma online de ciência de dados chamada Kaggle, a base chama-se “pump\_sensor\_data” e encontra-se no seguinte endereço <https://www.kaggle.com/datasets/nphantawee/pump-sensor-data>. Este arquivo contém 3 principais grupos de dados:

* Dados de timestamp: medido a cada 1 minuto.
* Dados dos sensores (52 séries): Valores brutos sem especificar a grandeza medida.
* Status da máquina: 3 classes (Broken, Normal, Recovering)

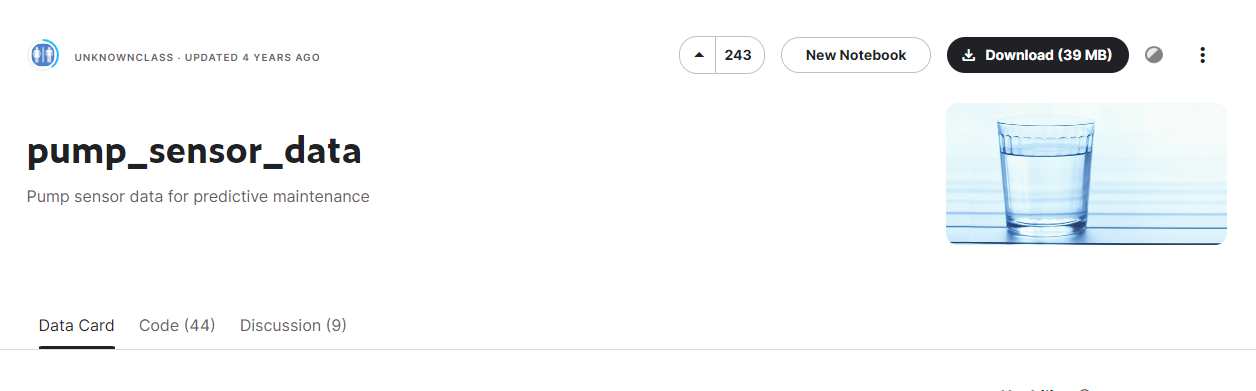


Figura Imagem base de dados

**2.2 Análise exploratória dos dados**

**2.2.1 Análise inicial**

Os dados iniciais da base de dados mostram que existem 52 sensores, cada um representado por uma coluna, que registraram informações a cada 1 minuto. O conjunto de dados contém 220320 registros, representando um período de 153 dias. É importante notar que as medições foram realizadas em escalas diferentes. A coluna machine\_status possui 3 classes com as seguintes quantidades de instancias: (NORMAL) com 205836, (RECOVERING) tendo 14477 e (BROKEN) com apenas 7.

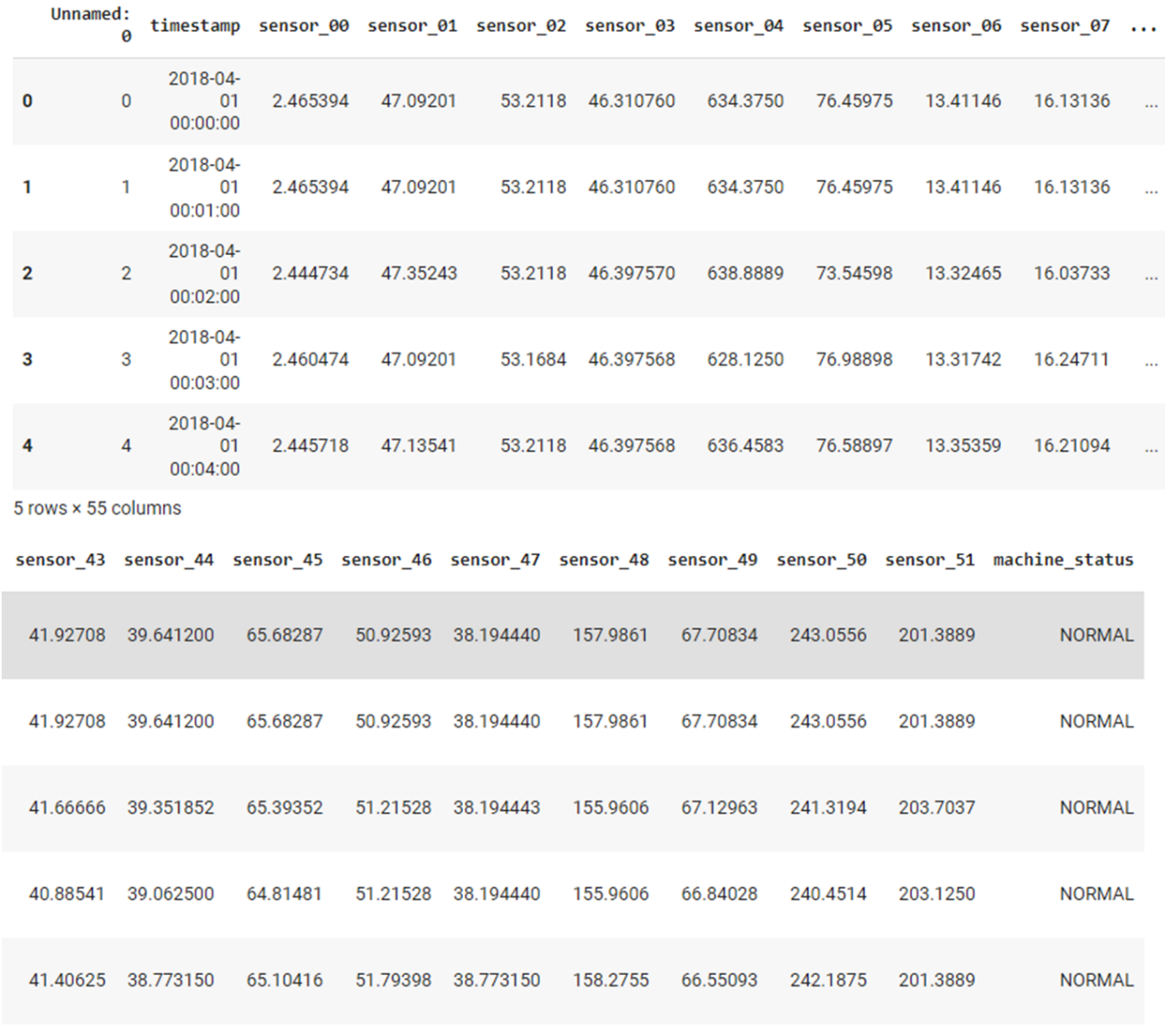


Figura Visão geral do dataset.

Quanto a presença de dados faltantes, na tabela a seguir, parece mostrar o número e a porcentagem de valores ausentes para cada coluna. O conjunto de dados possui um total de 220.320 registros, sendo que a coluna "sensor\_15" não possui valores. A maioria das outras colunas possui uma pequena porcentagem de valores ausentes, variando de 0,0% a 4,6%, exceto por "sensor\_00" (4,6%), "sensor\_50" (35,0%) e "sensor\_51" (7,0%). A coluna "machine\_status", variável alvo, também não possui valores ausentes.

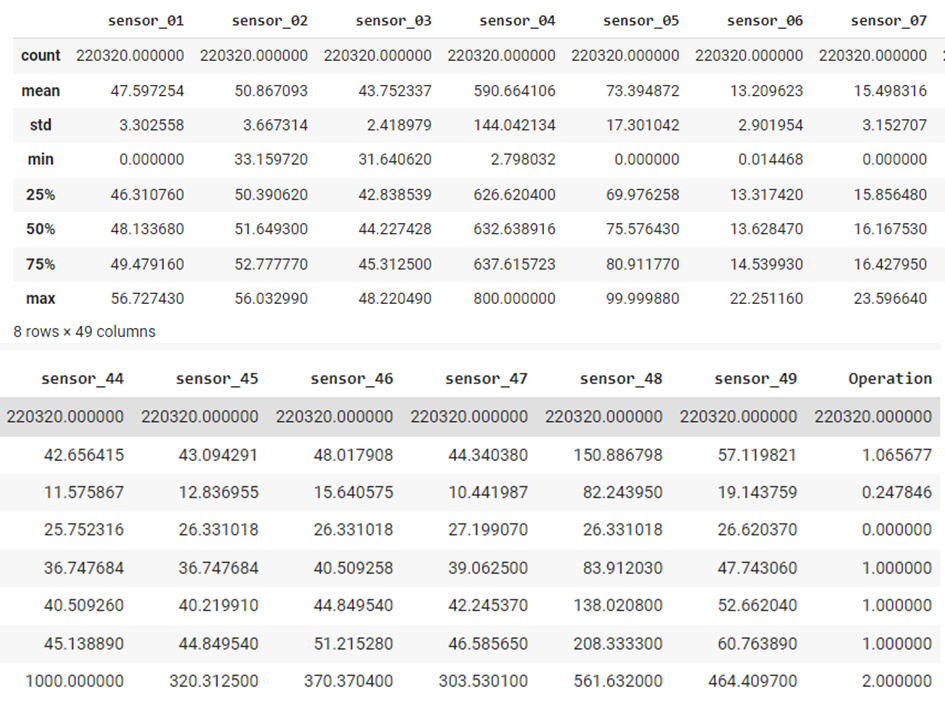
|  |
| --- |
| Quantidade Percentual  Unnamed: 0 0 0.0  timestamp 0 0.0  sensor\_00 10208 4.6  sensor\_01 369 0.2  sensor\_02 19 0.0  sensor\_03 19 0.0  sensor\_04 19 0.0  sensor\_05 19 0.0  sensor\_06 4798 2.2  sensor\_07 5451 2.5  sensor\_08 5107 2.3  sensor\_09 4595 2.1  sensor\_10 19 0.0  sensor\_11 19 0.0  sensor\_12 19 0.0  sensor\_13 19 0.0  sensor\_14 21 0.0  sensor\_15 220320 100.0  sensor\_16 31 0.0  sensor\_17 46 0.0  sensor\_18 46 0.0  sensor\_19 16 0.0  sensor\_20 16 0.0  sensor\_21 16 0.0  sensor\_22 41 0.0  sensor\_23 16 0.0  sensor\_24 16 0.0  sensor\_25 36 0.0  sensor\_26 20 0.0  sensor\_27 16 0.0  sensor\_28 16 0.0  sensor\_29 72 0.0  sensor\_30 261 0.1  sensor\_31 16 0.0  sensor\_32 68 0.0  sensor\_33 16 0.0  sensor\_34 16 0.0  sensor\_35 16 0.0  sensor\_36 16 0.0  sensor\_37 16 0.0  sensor\_38 27 0.0  sensor\_39 27 0.0  sensor\_40 27 0.0  sensor\_41 27 0.0  sensor\_42 27 0.0  sensor\_43 27 0.0  sensor\_44 27 0.0  sensor\_45 27 0.0  sensor\_46 27 0.0  sensor\_47 27 0.0  sensor\_48 27 0.0  sensor\_49 27 0.0  sensor\_50 77017 35.0  sensor\_51 15383 7.0  machine\_status 0 0.0 |

**2.2.2 Limpeza e tratamento**

Para melhorar a qualidade dos dados, foi realizada uma limpeza na tabela inicial, eliminando o índice inicial "Unnamed: 0" e sensores que possuíam um percentual de dados faltantes ou NaN (not a number) acima de 2%. Os sensores que foram removidos são: sensor\_00, sensor\_15, sensor\_50 e sensor\_51.

Além disso uma série de conversões foram adotadas, como: Converter timestamp para um objeto DatetimeIndex, que é necessário para análise de séries temporais em pandas. Converter coluna machine\_status para dado do tipo categórico, isso ajudará a reduzir o tamanho do dataframe e a acelerar as operações de análise. Transformação de timestamp para o índice, essa transformação é necessária quando você deseja realizar uma análise de séries temporais, pois ela fornece uma maneira fácil de acessar e manipular dados com base em uma escala de tempo.

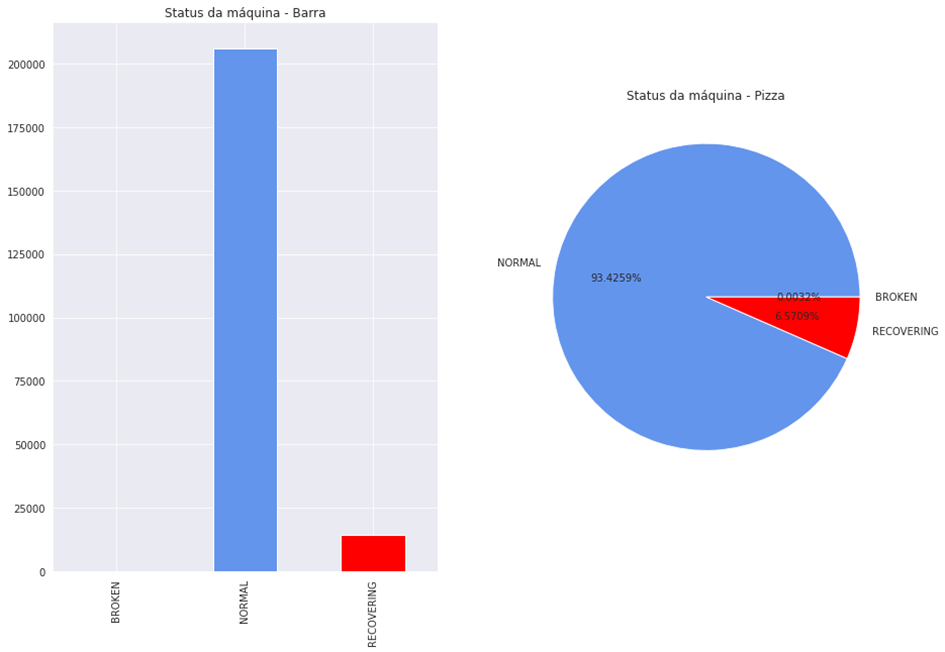
A criação de uma nova coluna de inteiros “Operation”, para equiparar classes categóricas de machine\_status, para fins estatísticos, sendo 0 para BROKEN, 1 para NORMAL e 2 para RECOVERING. Por fim, a verificação se não há duplicação no índice, essa verificação é importante porque, em uma análise de séries temporais, é essencial que os índices não contenham duplicatas, pois isso pode levar a problemas como perda de dados e imprecisão em cálculos de estatísticas descritivas e outras métricas.



Figura

**2.3 Visualização**

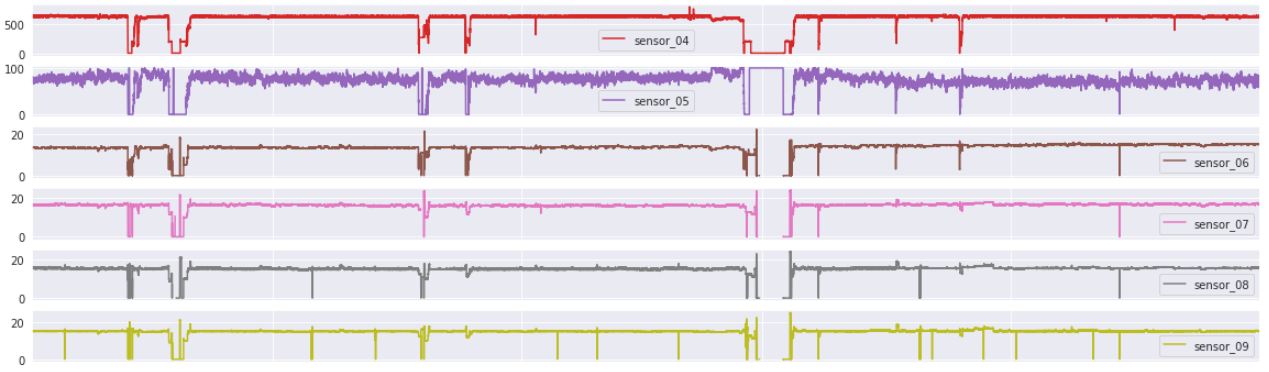
A seguir, foi realizada uma seleção de colunas para plotagem, para gráfico de barras, pizza.

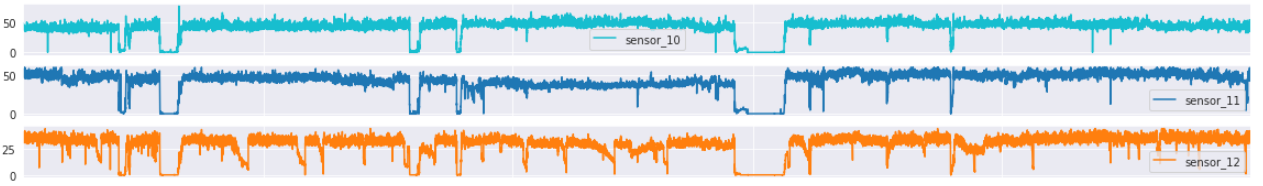


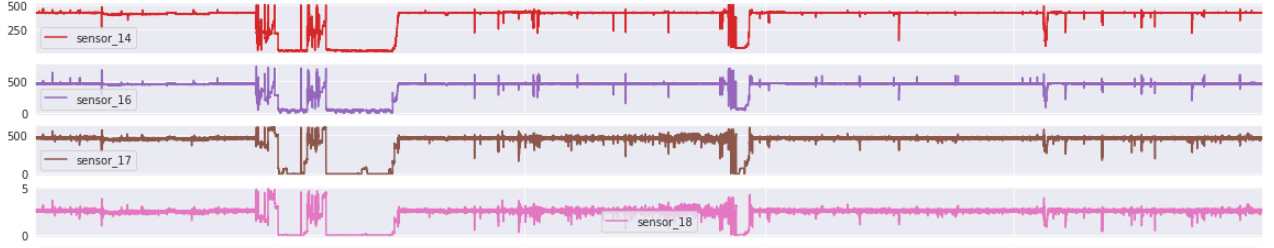
Figura

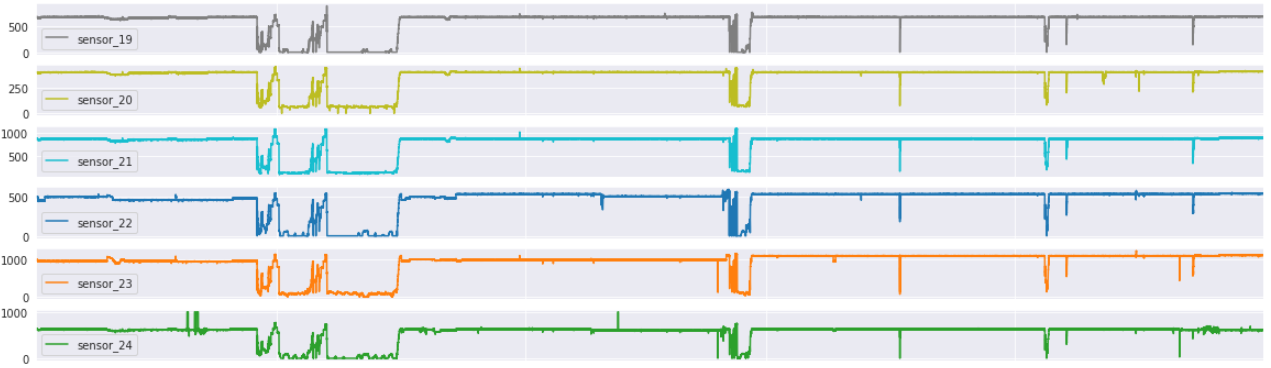
Na sequência de figuras a seguir, observa-se as séries ao longo do tempo, pode ser visto que existem padrões de sinais sendo capturados pelos os respectivos grupos de sensores: (1,2,3), (4,5,6,7,8,9), (10,11,12), (14,16,17,18), (19,20,21,22,23,24), (25,26,28,29,30,31,32,33), (34,35), (38,39,40,41,42,43,45,46,47). Por sua vez, há sinais que são muito ruidosos e parecem não seguir nenhuma tendência em particular.

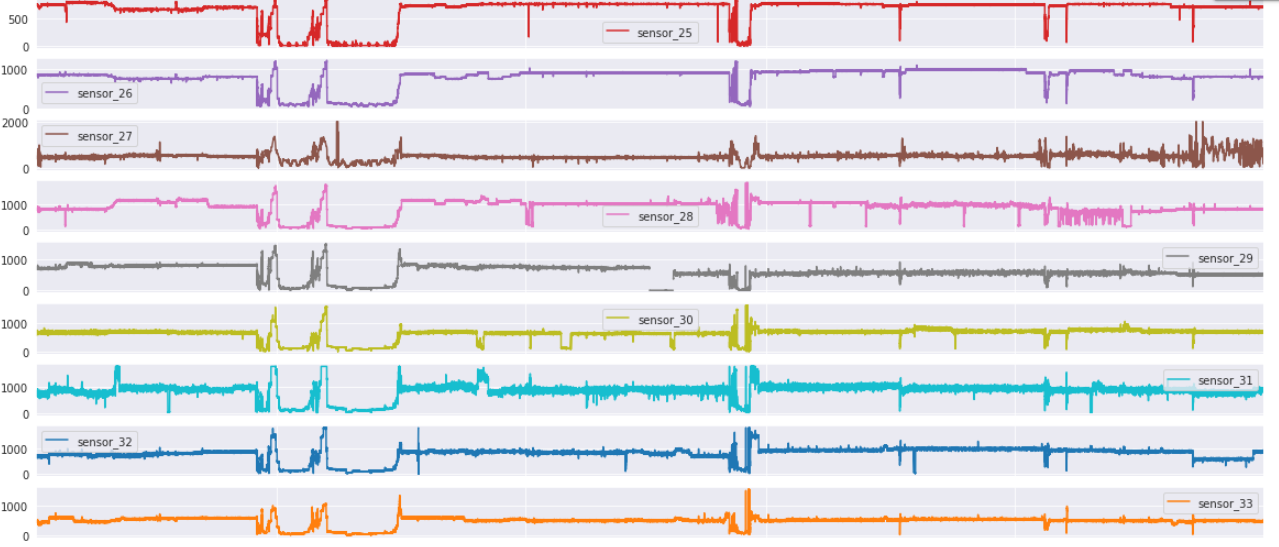
Sensores (1,2,3) 

Sensores (4,5,6,7,8,9) 

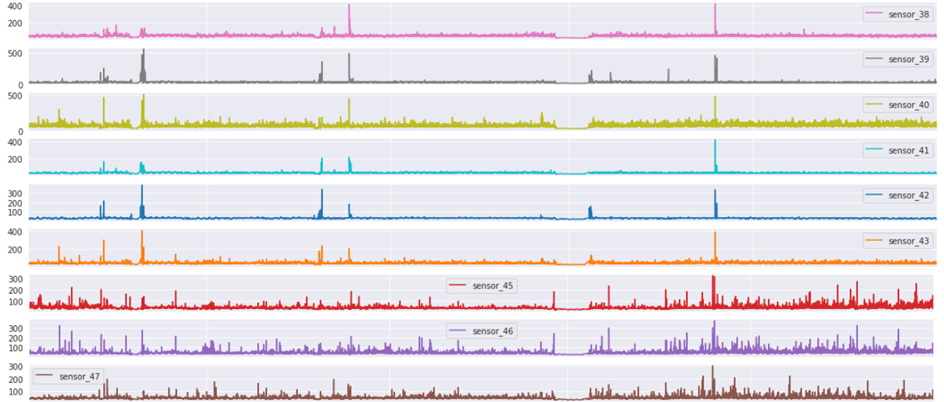
Sensores (10,11,12) 

Sensores (14,16,17,18) 

Sensores (19,20,21,22,23,24) 

Sensores (25,26,28,29,30,31,32,33) 

Sensores (34,35) 

Sensores (38,39,40,41,42,43,45,46,47) 

Após esta etapa, uma nova avaliação do dataframe foi realizada, para eliminar todos os valores nulos existentes, preenchendo os valores com a função “ffill”, sendo está uma abreviação para "forward fill", que em português significa "preenchimento para a frente". Ela é utilizada em pandas, uma biblioteca de análise de dados em Python, para preencher valores ausentes (NaN - "Not a Number") em uma série ou em um DataFrame. é útil na limpeza e preparação de dados, especialmente quando se lida com conjuntos de dados que possuem muitos valores ausentes. Ela permite que os dados sejam preenchidos de forma eficiente com valores que fazem sentido, preservando assim a integridade dos dados.

**2.3.1 Relatório com Profiling**

Um relatório utilizando o Pandas Profiling foi gerado, no entanto, a versão completa não funcionou devidamente no google colab, por consequência, a versão “minimal” foi gerada em uma página html para melhor interação e encontra-se anexa a este relatório. A seguir uma breve amostra das informações obtidas com o Profiling.

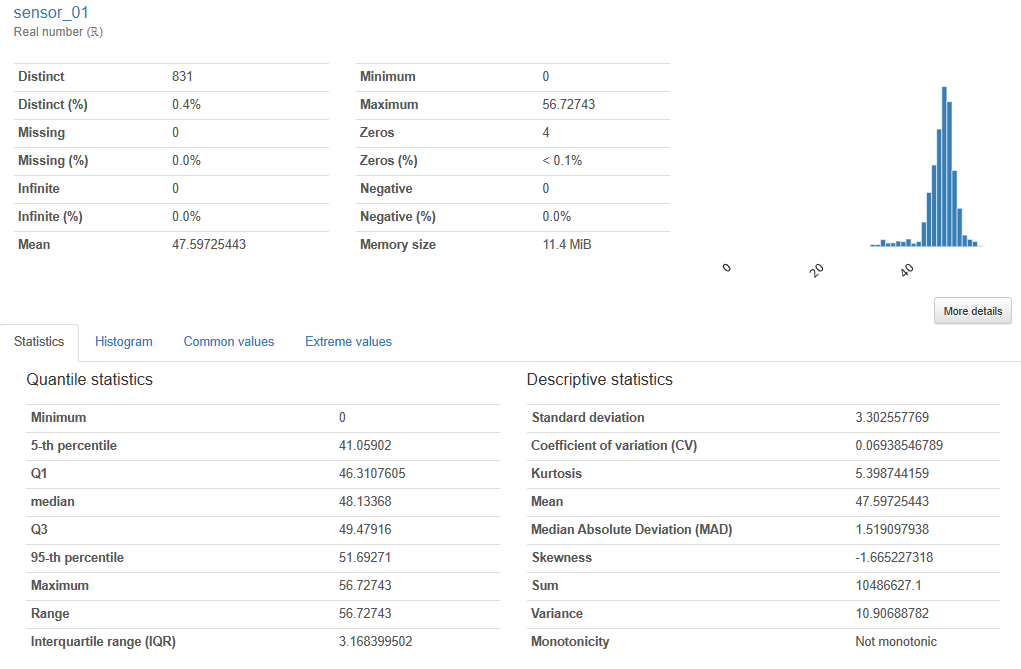


Figura Profiling com informações sobre o sensor\_01

**2.3.2 Relatório com Sweetviz**

As associações categóricas são medidas estatísticas que buscam identificar a relação entre duas variáveis categóricas. O coeficiente de incerteza, também conhecido como medida de associação de Goodman e Kruskal, é um exemplo de medida de associação categórica que varia de 0 a 1. Valores próximos de 0 indicam que não há associação entre as variáveis categóricas, enquanto valores próximos de 1 indicam forte associação.

O coeficiente de incerteza é assimétrico porque leva em consideração o grau em que uma variável categórica fornece informações sobre outra. Em outras palavras, o valor do coeficiente de incerteza pode ser diferente dependendo da ordem em que as variáveis são analisadas.

Por outro lado, as correlações numéricas buscam identificar a relação entre duas variáveis numéricas. O coeficiente de correlação de Pearson é um exemplo de medida de correlação numérica que varia de -1 a 1. Valores próximos de -1 indicam uma correlação negativa perfeita, ou seja, quando uma variável aumenta, a outra diminui. Valores próximos de 1 indicam uma correlação positiva perfeita, ou seja, quando uma variável aumenta, a outra também aumenta. Valores próximos de 0 indicam ausência de correlação.

Ao contrário do coeficiente de incerteza, a correlação de Pearson é simétrica, o que significa que o valor da correlação é o mesmo independentemente da ordem em que as variáveis são analisadas. Os quadrados são associações e os círculos são as correlações numéricas simétricas (de Pearson) de -1 a 1. A diagonal trivial é intencionalmente deixada em branco para maior clareza.

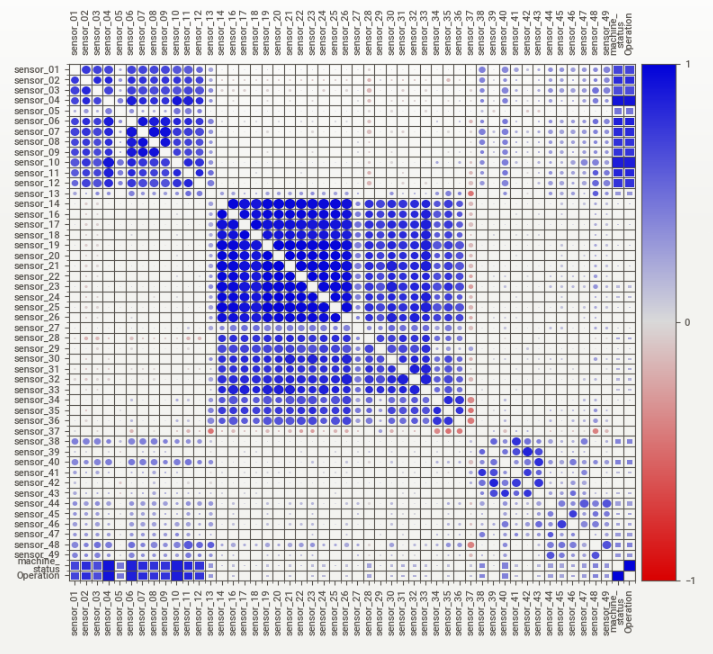


Figura Associações

**3. RESULTADOS**

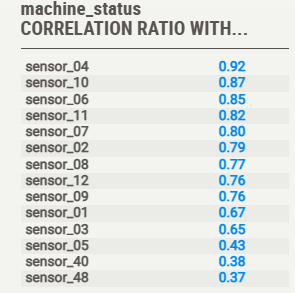
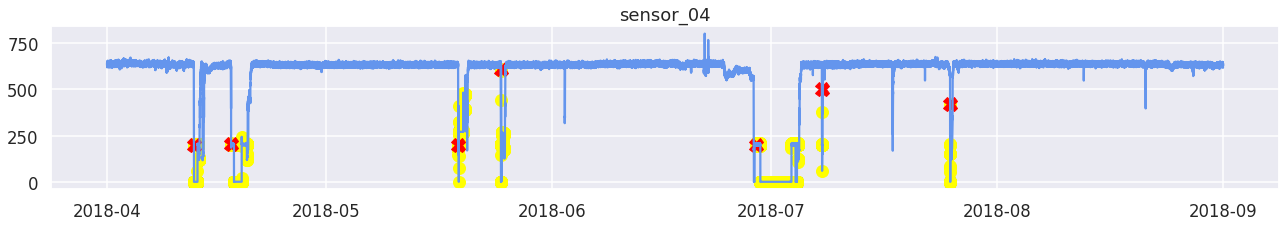


Figura Correlação entre classe machine\_status e os sensores.

A partir dessa informação sobre as correlações, uma enfase será dada nos sensores que possuirem correlação maior que 75% com a classe alvo, tendo em vista que tais sensores podem ser mais relevantes para determinar os comportamentos das classes. A seguir, uma nova sequencia temporal, na ordem do maior para o menos correlacionado, de como esses sensores se comportam nos momentos de BROKEN ( x em vermelho) e RECOVERING (pontos amarelos) 

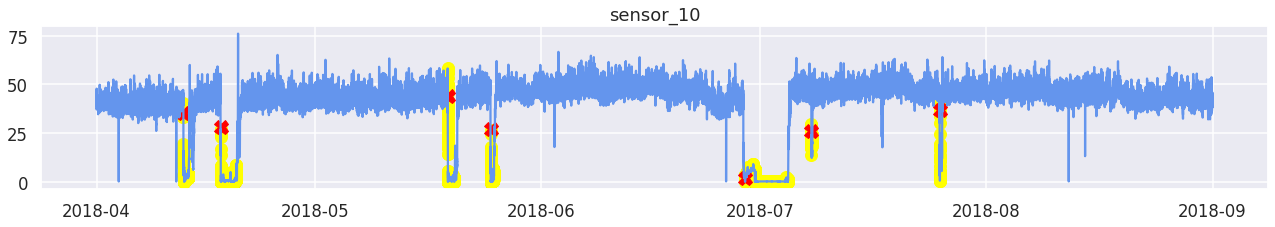
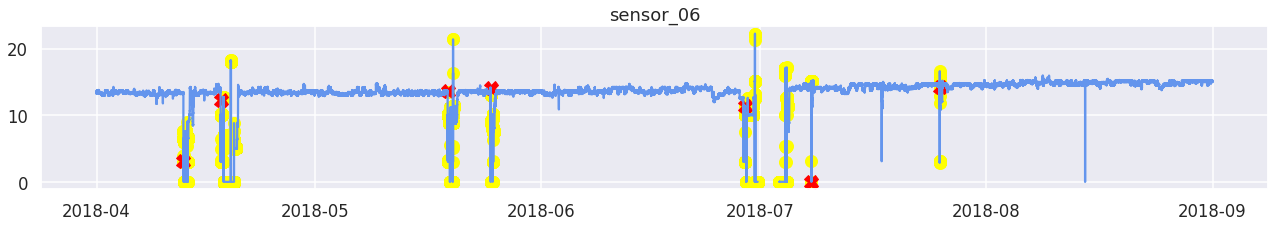
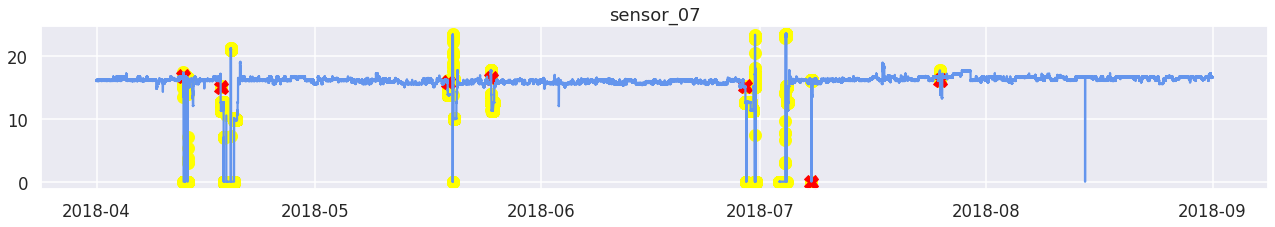
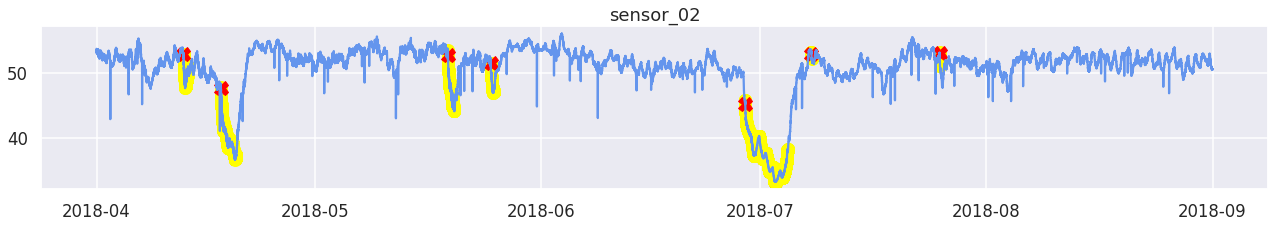
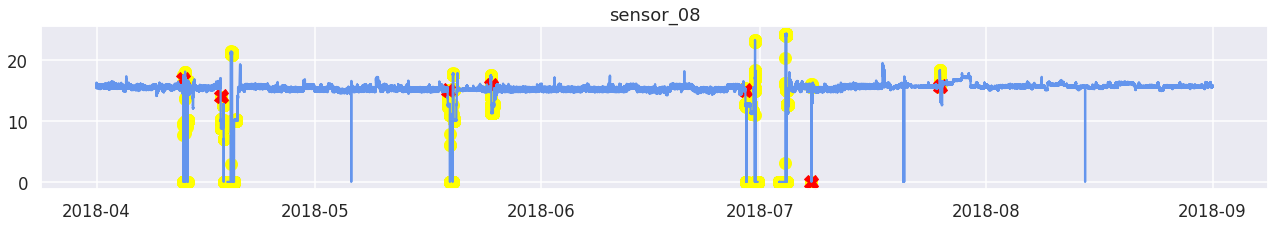
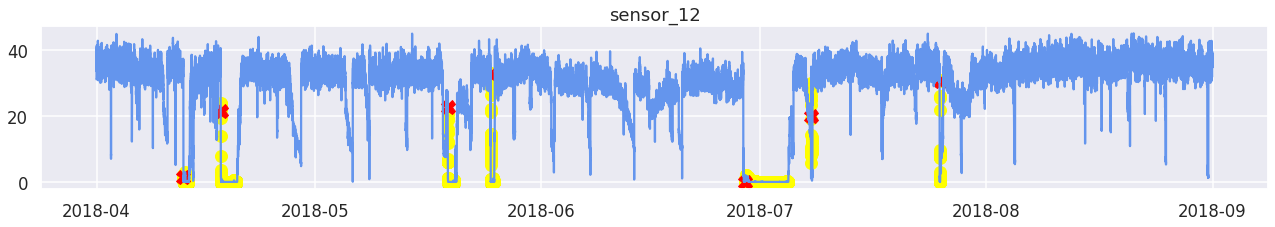
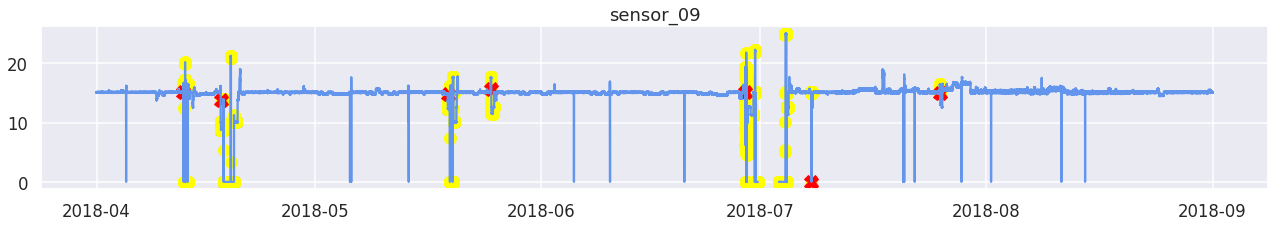
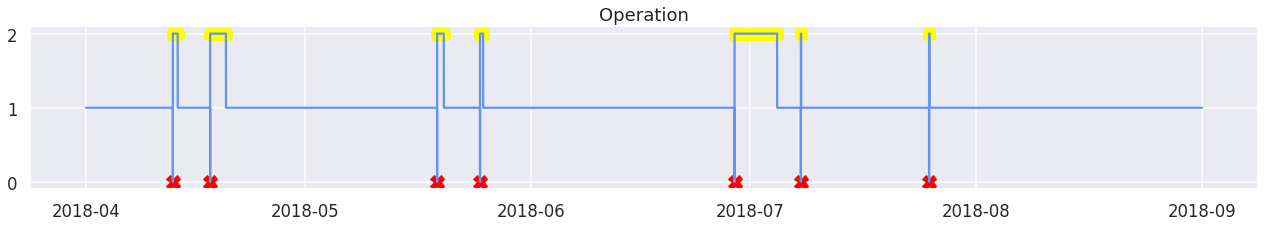
       

Figura Sensores e classe Operation exibindo momentos de BROKEN e RECOVERING

Mesmo com poucos detalhes fornecidos, pelo dono do dataset, sobre as grandezas mensuradas com base no levantamento exploratório dos dados, encontrou-se a correlação entre as classes de status da máquina e os sensores. Sendo possível assim, determinar quais sensores possuem maior relevância para determinar o estado de operação da máquina, e assim, viabilizando uma melhor abordagem para monitoramento preditivo de falhas.

Apesar de não se ter ido adiante com a criação de um modelo preditivo, é recomendado uma continuidade com features engineering, para se ter mais clareza dos padrões dos sensores e o que ocorre antecedendo a falha para assim ser possível uma medida preventiva. Será continuado esse desafio para se tentar criar uma regra de associação melhor entre os padrões dos sensores e as falhas.

**4. CONSIDERAÇÕES FINAIS**

Em resumo, a correlação encontrada entre as classes de status da máquina e os sensores representa um importante avanço na compreensão dos padrões de operação e possíveis falhas do sistema. A determinação dos sensores mais relevantes para a detecção de falhas e a continuidade com o desenvolvimento de recursos ajudarão a estabelecer um sistema de monitoramento mais eficiente e confiável. Este trabalho seguirá em busca de caminhos promissores para futuras pesquisas na área de manutenção preditiva, com potencial para melhorar a eficiência e a segurança de máquinas e equipamentos em diversos setores industriais.

**REFERÊNCIAS**

Código no Github[**https://github.com/jpscard/I2A2\_AI\_Industry\_Desafios/tree/main/Desafio%202**](https://github.com/jpscard/I2A2_AI_Industry_Desafios/tree/main/Desafio%202)

Dataset

[**https://www.kaggle.com/datasets/nphantawee/pump-sensor-data**](https://www.kaggle.com/datasets/nphantawee/pump-sensor-data)