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)

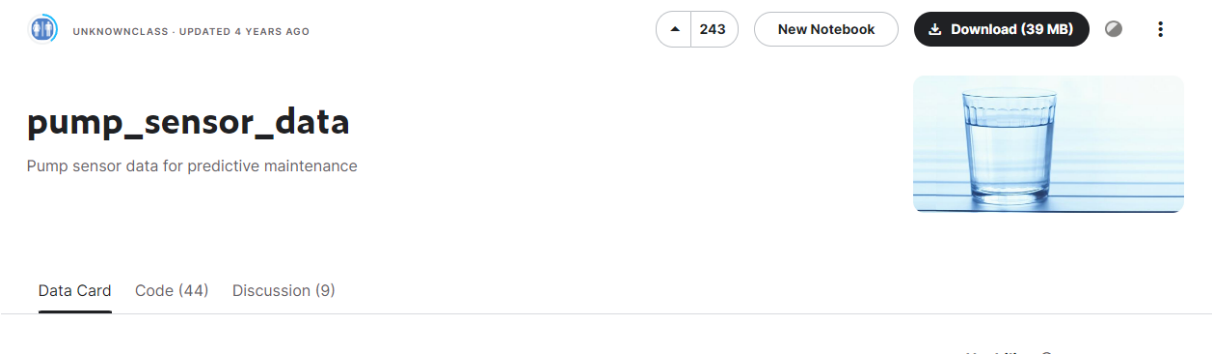


Figura 1 Imagem base de dados

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.

	Unnamed: 0	timestamp	sensor_00	sensor_01	sensor_02	sensor_03	sensor_04	sensor_05	sensor_06	sensor_07	...
0	0	2018-04-01 00:00:00	2.465394	47.09201	53.2118	46.310760	634.3750	76.45975	13.41146	16.13136	...
1	1	2018-04-01 00:01:00	2.465394	47.09201	53.2118	46.310760	634.3750	76.45975	13.41146	16.13136	...
2	2	2018-04-01 00:02:00	2.444734	47.35243	53.2118	46.397570	638.8889	73.54598	13.32465	16.03733	...
3	3	2018-04-01 00:03:00	2.460474	47.09201	53.1684	46.397568	628.1250	76.98898	13.31742	16.24711	...
4	4	2018-04-01 00:04:00	2.445718	47.13541	53.2118	46.397568	636.4583	76.58897	13.35359	16.21094	...

5 rows × 55 columns

sensor_43	sensor_44	sensor_45	sensor_46	sensor_47	sensor_48	sensor_49	sensor_50	sensor_51	machine_status
41.92708	39.641200	65.68287	50.92593	38.194440	157.9861	67.70834	243.0556	201.3889	NORMAL
41.92708	39.641200	65.68287	50.92593	38.194440	157.9861	67.70834	243.0556	201.3889	NORMAL
41.66666	39.351852	65.39352	51.21528	38.194443	155.9606	67.12963	241.3194	203.7037	NORMAL
40.88541	39.062500	64.81481	51.21528	38.194440	155.9606	66.84028	240.4514	203.1250	NORMAL
41.40625	38.773150	65.10416	51.79398	38.773150	158.2755	66.55093	242.1875	201.3889	NORMAL

Figura 2 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.

	sensor_01	sensor_02	sensor_03	sensor_04	sensor_05	sensor_06	sensor_07
count	220320.000000	220320.000000	220320.000000	220320.000000	220320.000000	220320.000000	220320.000000
mean	47.597254	50.867093	43.752337	590.664106	73.394872	13.209623	15.498316
std	3.302558	3.667314	2.418979	144.042134	17.301042	2.901954	3.152707
min	0.000000	33.159720	31.640620	2.798032	0.000000	0.014468	0.000000
25%	46.310760	50.390620	42.838539	626.620400	69.976258	13.317420	15.856480
50%	48.133680	51.649300	44.227428	632.638916	75.576430	13.628470	16.167530
75%	49.479160	52.777770	45.312500	637.615723	80.911770	14.539930	16.427950
max	56.727430	56.032990	48.220490	800.000000	99.999880	22.251160	23.596640

8 rows × 49 columns

sensor_44	sensor_45	sensor_46	sensor_47	sensor_48	sensor_49	Operation
220320.000000	220320.000000	220320.000000	220320.000000	220320.000000	220320.000000	220320.000000
42.656415	43.094291	48.017908	44.340380	150.886798	57.119821	1.065677
11.575867	12.836955	15.640575	10.441987	82.243950	19.143759	0.247846
25.752316	26.331018	26.331018	27.199070	26.331018	26.620370	0.000000
36.747684	36.747684	40.509258	39.062500	83.912030	47.743060	1.000000
40.509260	40.219910	44.849540	42.245370	138.020800	52.662040	1.000000
45.138890	44.849540	51.215280	46.585650	208.333300	60.763890	1.000000
1000.000000	320.312500	370.370400	303.530100	561.632000	464.409700	2.000000

Figura 3

2.3 Visualização

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

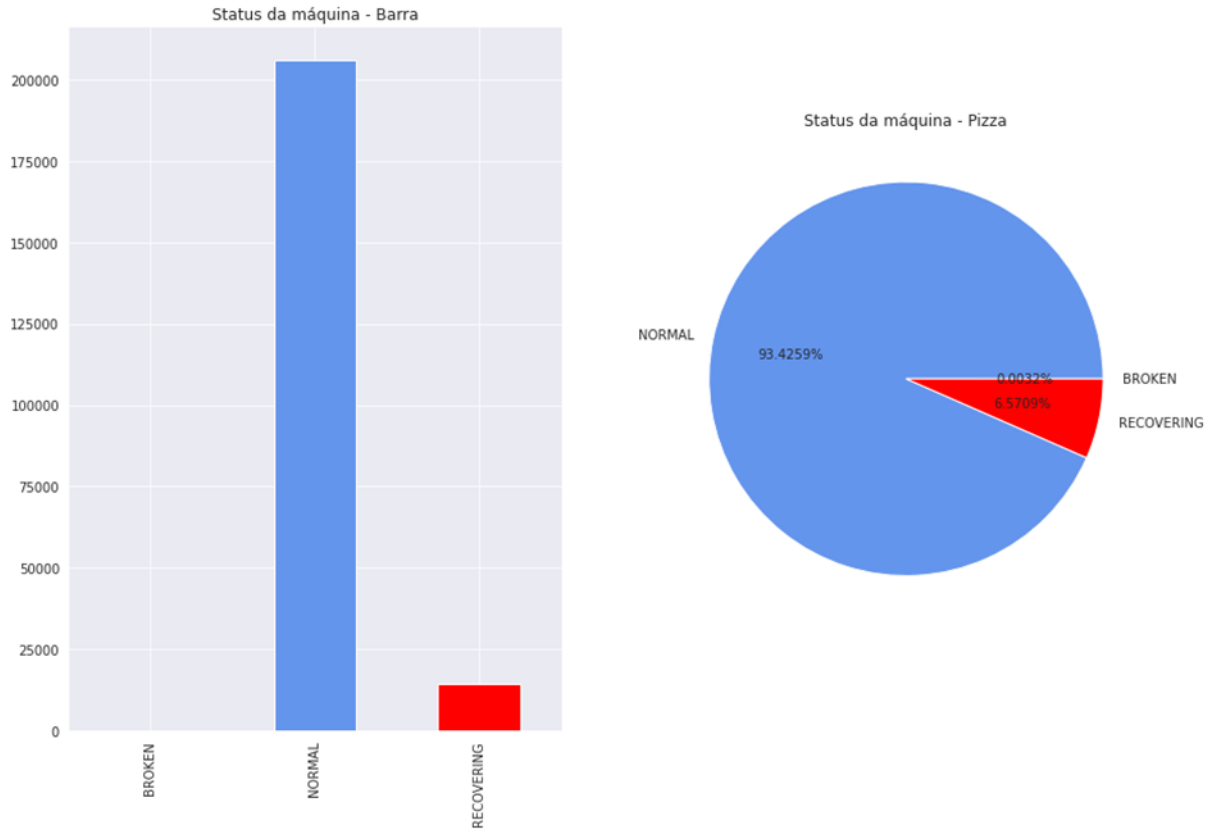
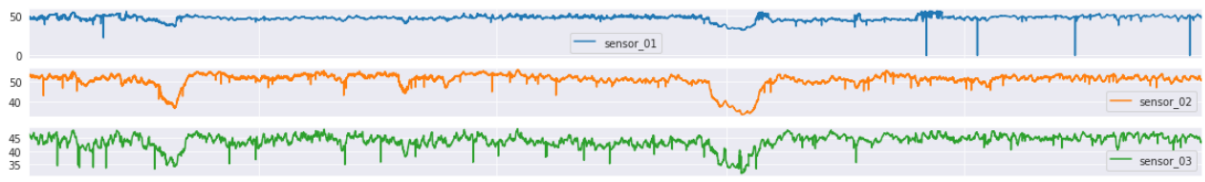


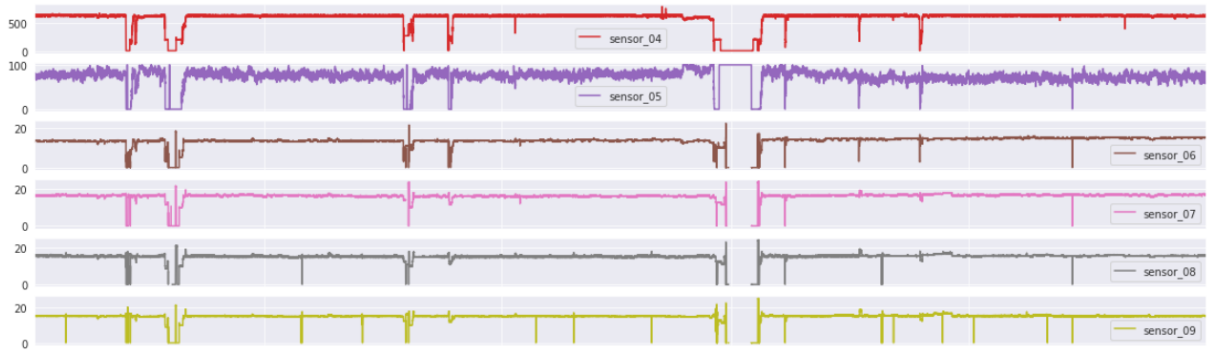
Figura 4

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.

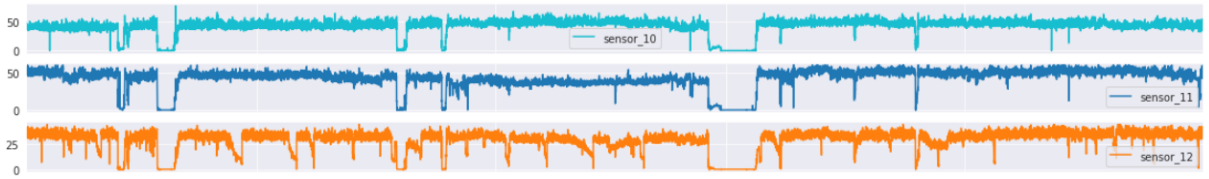
Sensores (1,2,3)



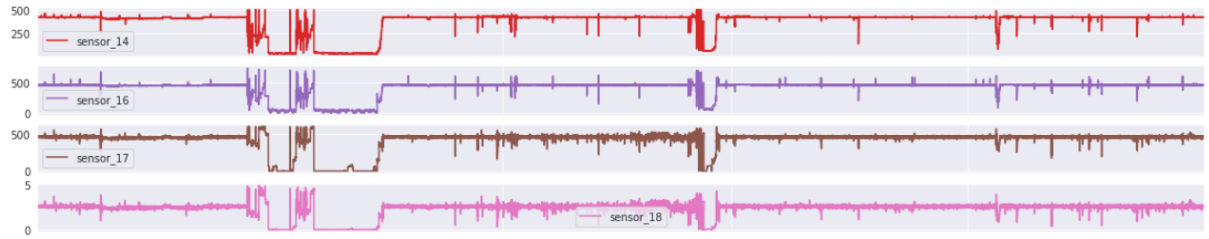
Sensores (4,5,6,7,8,9)



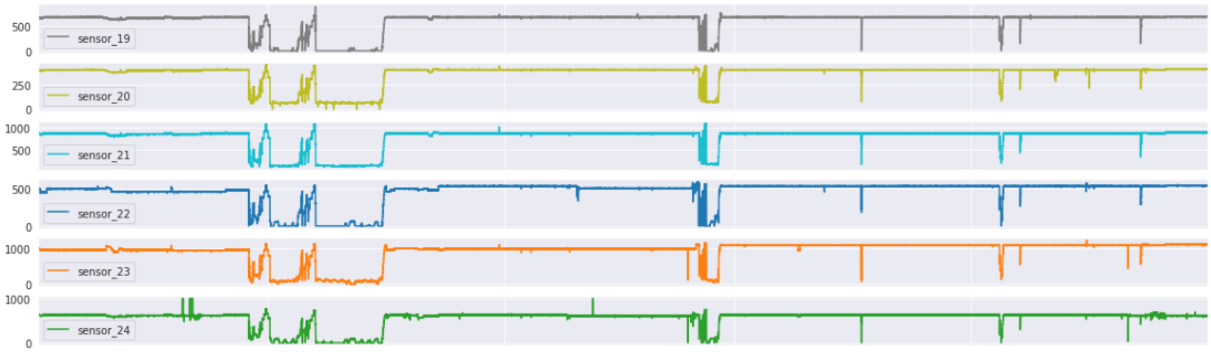
Sensores (10,11,12)



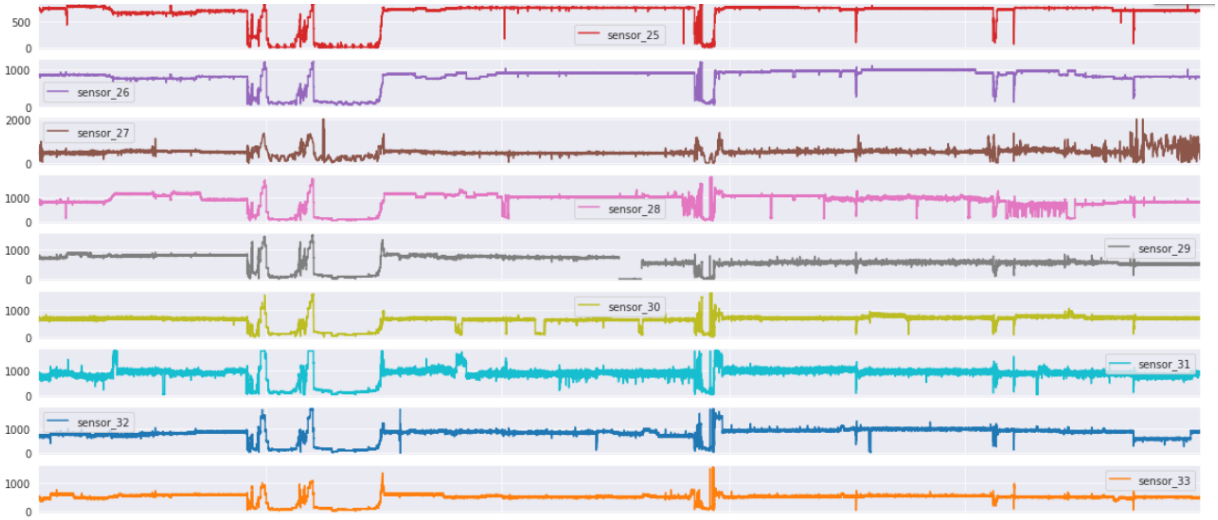
Sensores (14,16,17,18)



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

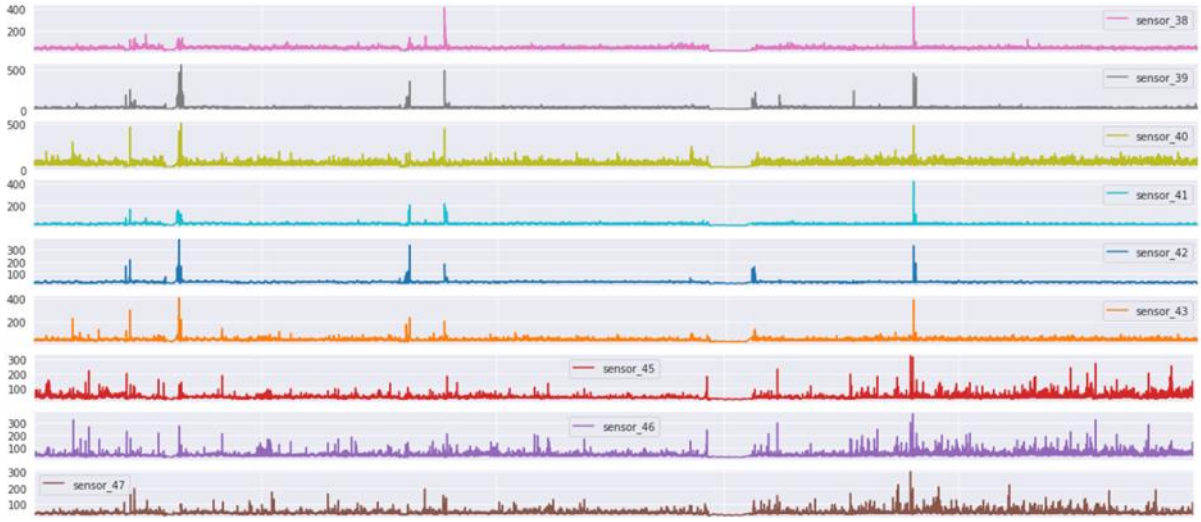


Sensores (25,26,28,29,30,31,32,33)



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.

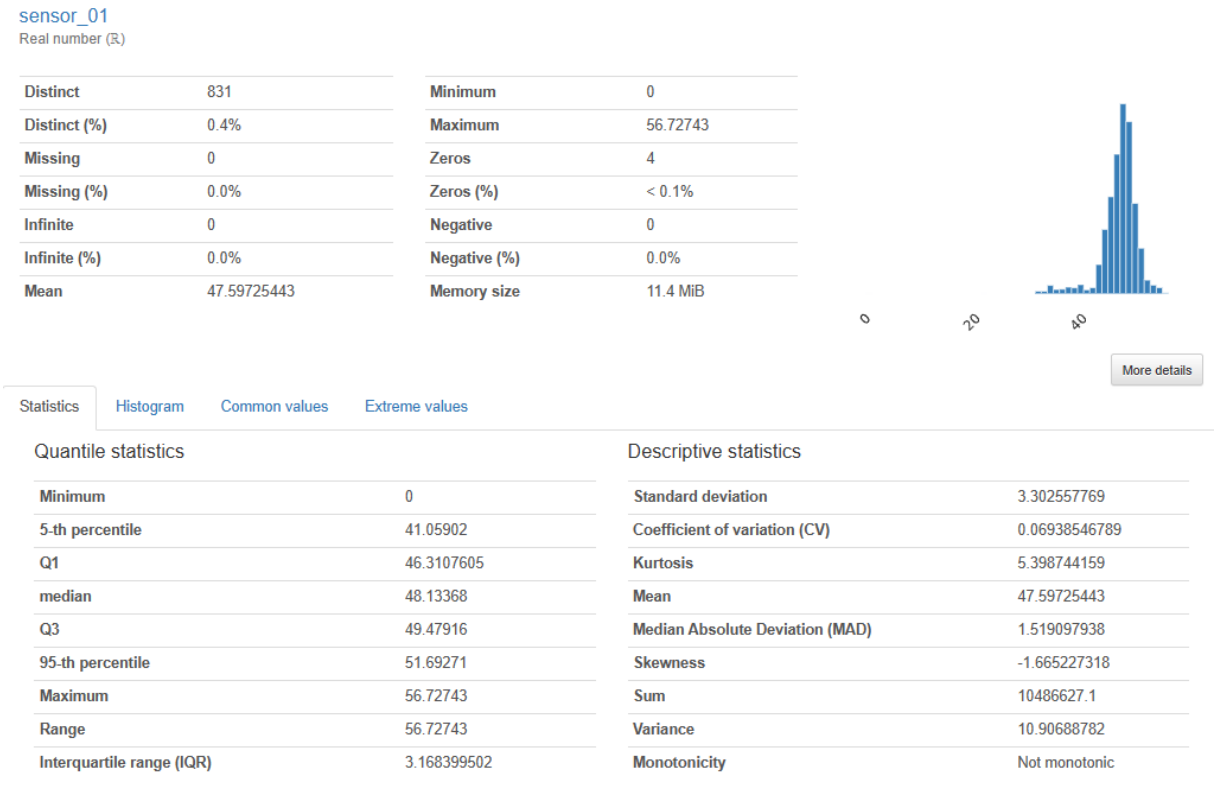


Figura 5 Profiling com informações sobre o sensor_01

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.

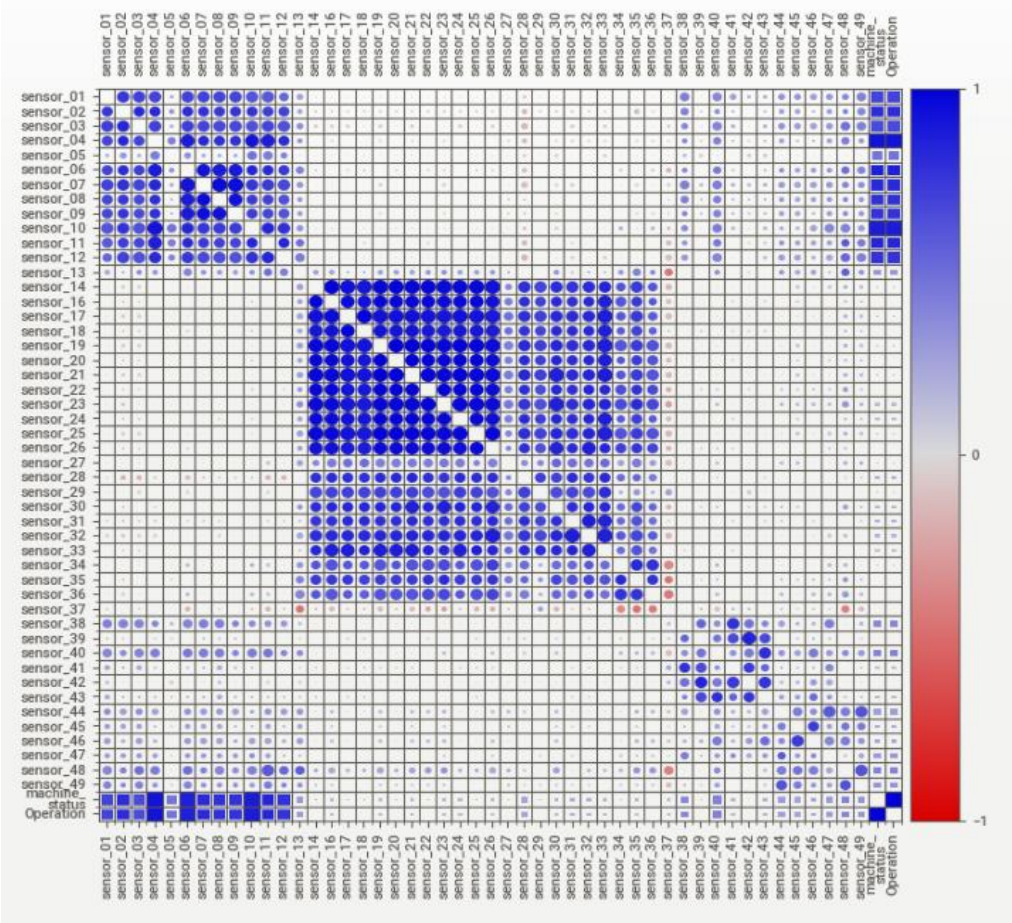


Figura 6 Associações

3. RESULTADOS

machine_status CORRELATION RATIO WITH...	
sensor_04	0.92
sensor_10	0.87
sensor_06	0.85
sensor_11	0.82
sensor_07	0.80
sensor_02	0.79
sensor_08	0.77
sensor_12	0.76
sensor_09	0.76
sensor_01	0.67
sensor_03	0.65
sensor_05	0.43
sensor_40	0.38
sensor_48	0.37

Figura 7 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)

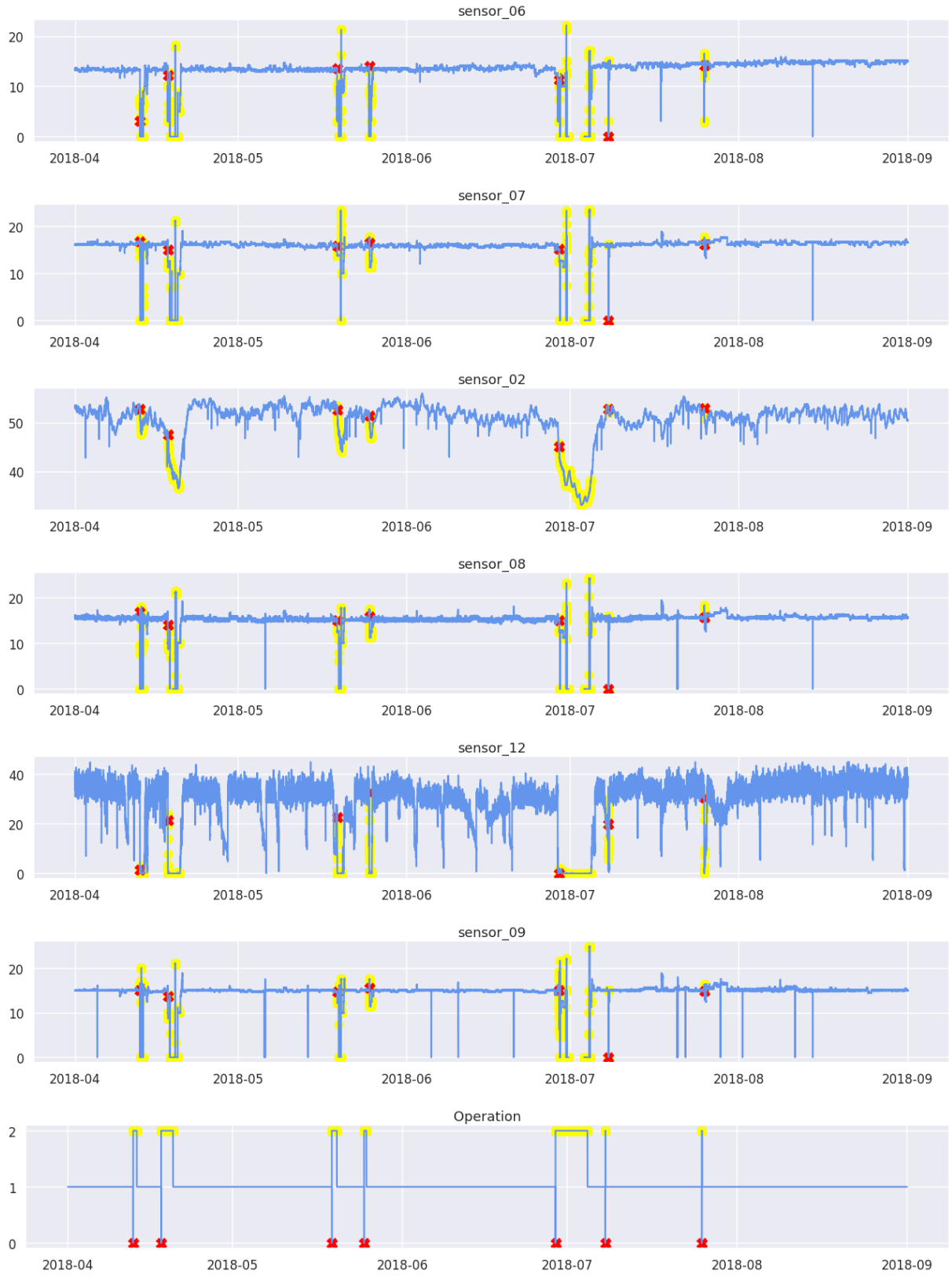


Figura 8 Sensores e classe Operation exibindo momentos de BROKEN e RECOVERING

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

Dataset

https://www.kaggle.com/datasets/nphantawee/pump-sensor-data