UNIVERSIDADE FEDERAL DA FRONTEIRA SUL

CAMPUS CHAPECÓ

CURSO DE CIÊNCIA DA COMPUTAÇÃO

Atividade 1: Análise Exploratória de Dados (EDA)

GEX1090

TÓPICOS ESPECIAIS EM COMPUTAÇÃO XL

ALESSANDRO LUIGI FERREIRA CORREA
(20230003860)

RHUAN LEHMEN DE SOUZA LEITE
(20230001285)

VICTOR NEYMAR DE CONTO

(20230004269)

CHAPECÓ, SANTA

CATARINA 2025

1 INTRODUÇÃO

A proposta da atividade era realizar uma análise dos dados fornecidos pelo professor pelo arquivo ‘df_full.csv’, neles contendo as medições dos sensores de fibra óptica em um processo de separação de óleo, água e gás no petróleo. Objetivo principal era entender os dados para próximos projetos da disciplina.

2 DESENVOLVIMENTO

Para desenvolvermos as soluções seguimos os passos: primeiramente definimos a quantidade qual seria nosso método de análise, com auxílio do da LLM Gemini AI, criamos um código em python que importava os dados do arquivo fornecido fazia os cálculos necessários para a análise dos dados.

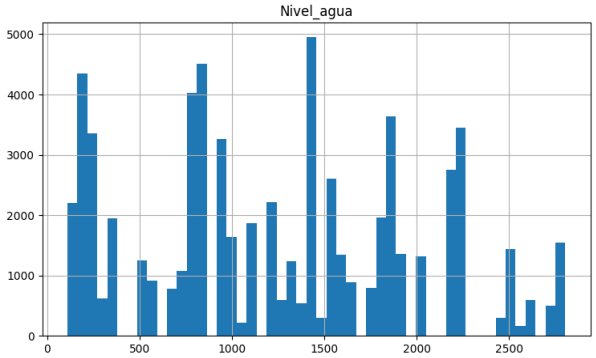
A partir da primeira análise tivemos as seguintes observações:

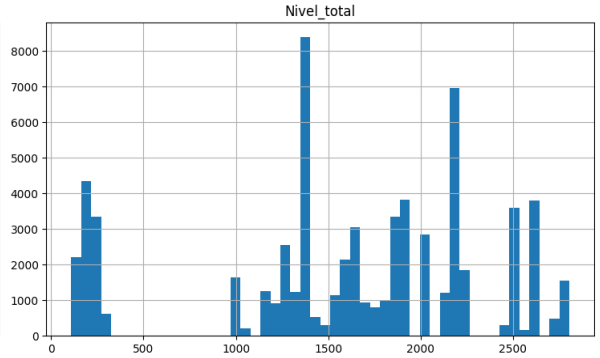
Variabilidade dos Sensores: O desvio padrão (std) da maioria dos sensores é muito baixo (ex: S1_1 com std de 0.054), indicando que suas medições são bastante estáveis e com pouca variação ao longo do tempo. A exceção notável é o sensor S4_2, que apresenta um desvio padrão de 4.58, significativamente maior que os outros. Isso sugere que S4_2 pode ser mais sensível às mudanças no processo ou talvez mais ruidoso.

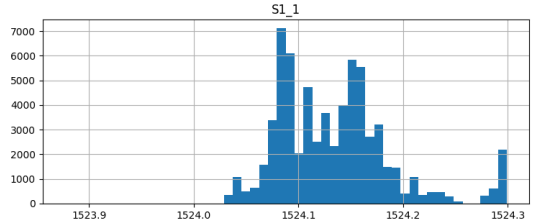
Faixa de Valores (Mín e Máx): A faixa entre os valores mínimos e máximos para a maioria dos sensores é estreita. Por exemplo, S1_1 varia apenas de 1523.87 a 1524.29. Isso reforça a ideia de baixa variabilidade.

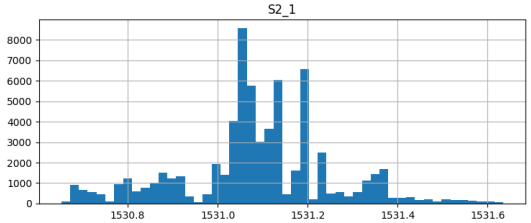
Variáveis-Alvo (Nivel Agua, Nivel Total): Ao contrário dos sensores, as variáveis-alvo Nivel Agua e Nível Total mostram uma ampla faixa de valores e um desvio padrão elevado (acima de 700). Isso é esperado, pois elas representam os níveis que variam durante o processo de medição.

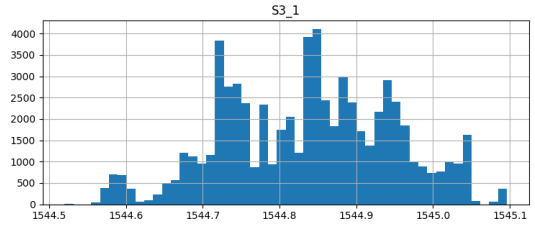
2.1 GRÁFICOS

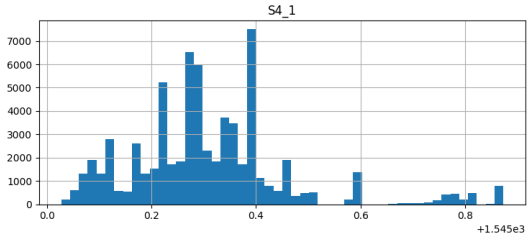

Nivel_agua

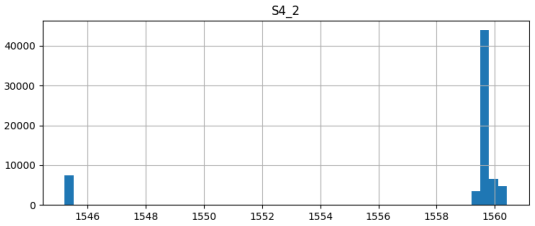

Nivel_total

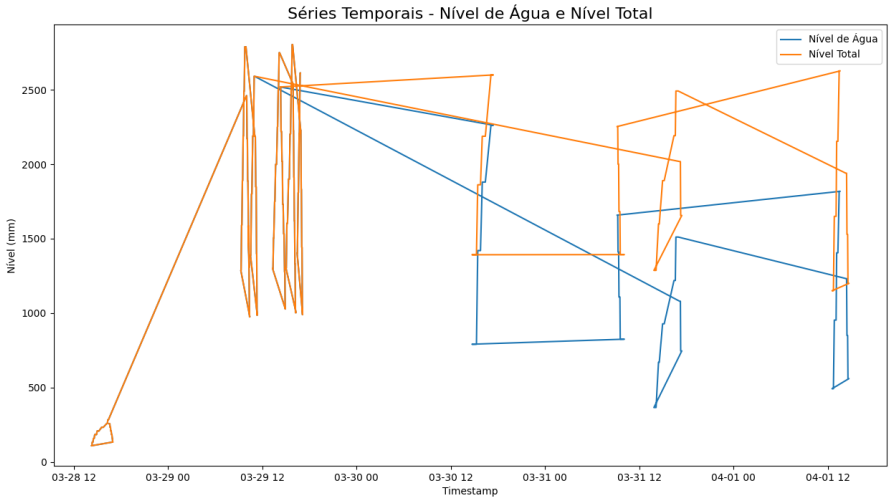

S1_1

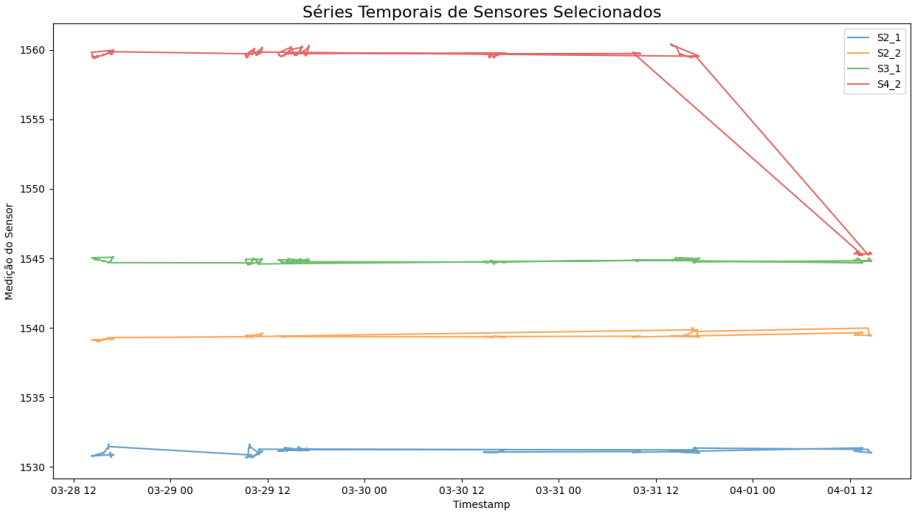

S2_1

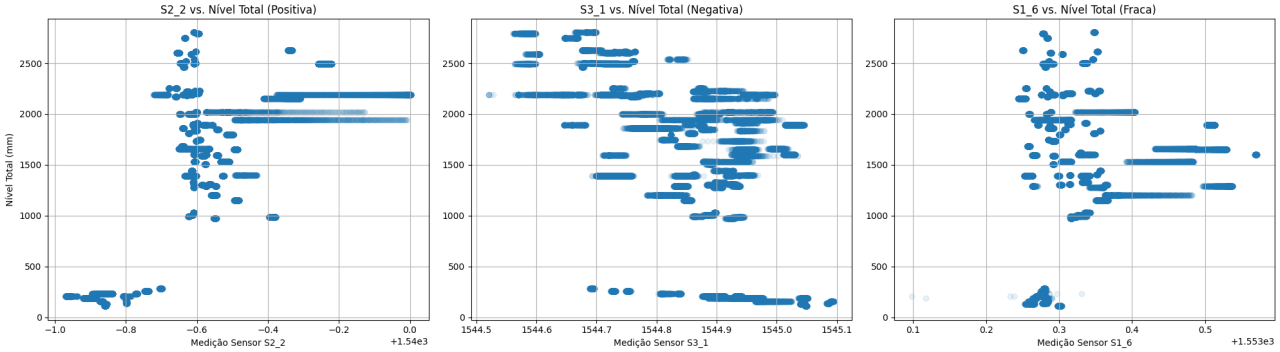

S3_1

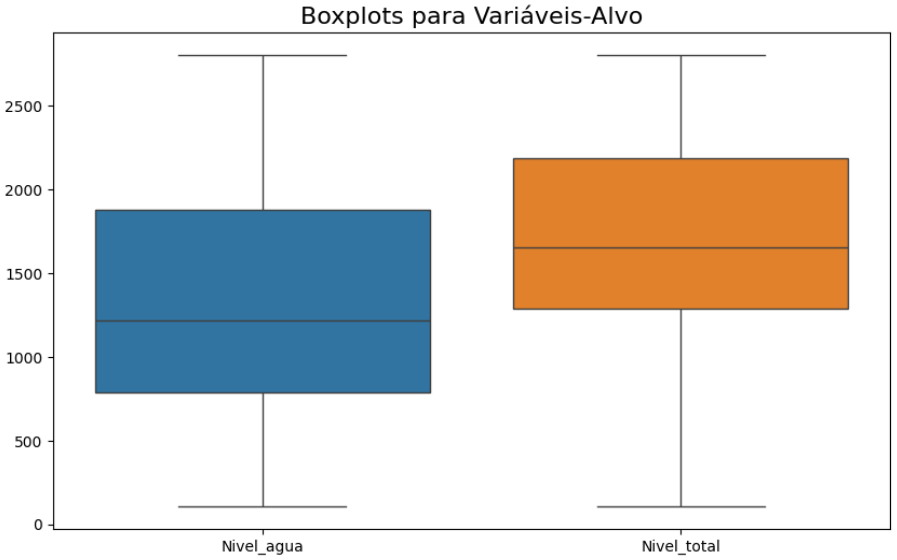

S4_1
+1.545e3

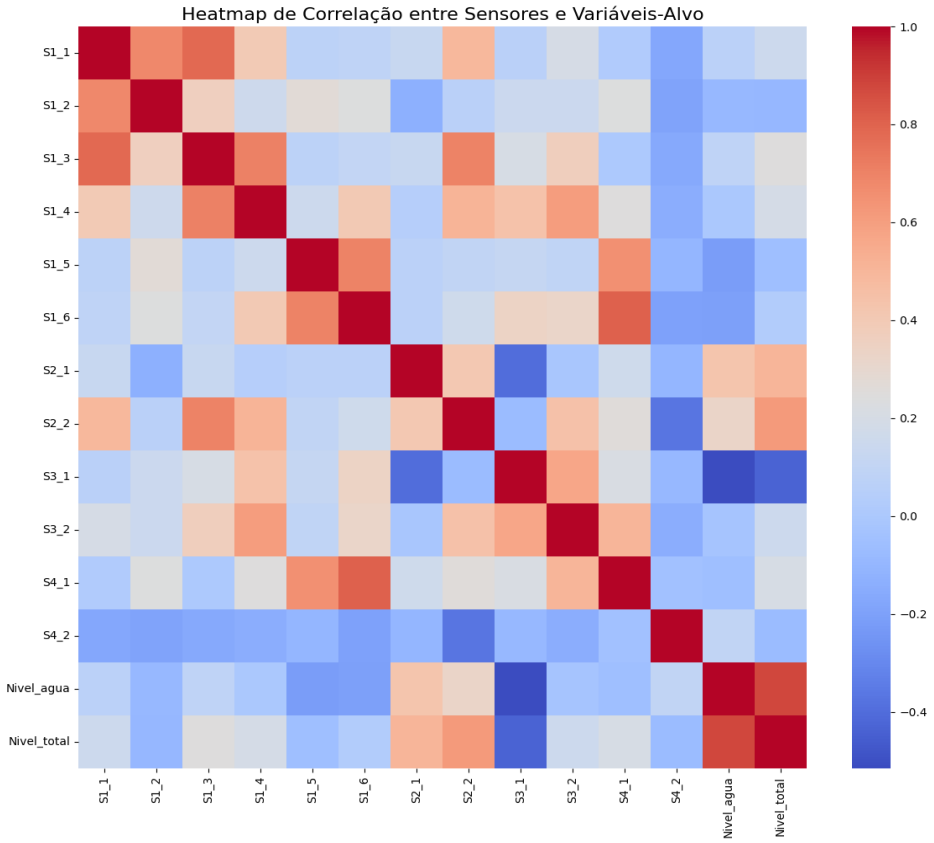

S4_2


Séries Temporais - Nível de Água e Nível Total
Nível de Água
Nível Total
Nível (mm)
Timestamp
03-28 12
03-29 00
03-29 12
03-30 00
03-30 12
03-31 00
03-31 12
04-01 00
04-01 12


Séries Temporais de Sensores Selecionados
S2_1
S2_2
S3_1
S4_2
Medição do Sensor
Timestamp
1560
1555
1550
1545
1540
1535
1530
03-28 12
03-29 00
03-29 12
03-30 00
03-30 12
03-31 00
03-31 12
04-01 00
04-01 12


S2_2 vs. Nível Total (Positiva)
S3_1 vs. Nível Total (Negativa)
S1_6 vs. Nível Total (Fraca)
Nível Total (mm)
Medição Sensor S2_2
Medição Sensor S3_1
Medição Sensor S1_6
+1.54e3
+1.553e3


Boxplots para Variáveis-Alvo
2500
2000
1500
1000
500
0
Nivel_agua
Nivel_total



Análise dos Histogramas:

- **Sensores:**

A maioria dos sensores (S1_1, S2_1, S3_1, S4_1) apresenta uma distribuição que se assemelha a uma distribuição normal (formato de sino), mas com múltiplos picos muito próximos. Isso indica que os sensores operam em torno de alguns estados ou regimes de medição muito específicos e estáveis.

O sensor S4_2 tem uma distribuição mais achatada e larga, confirmando o que vimos no resumo estatístico: ele possui maior variabilidade em suas medições em comparação com os outros.

- **Variáveis-Alvo;**

As distribuições de Nivel Agua e Nível Total não seguem um padrão claro como o normal. Elas são multimodais, ou seja, possuem vários picos de frequência.

Isso sugere que o processo opera em diferentes faixas de nível por períodos de tempo distintos. Por exemplo, pode haver um pico de ocorrências em torno de 250, outro em 1000, e assim por diante. Não há um "nível médio" que seja o mais comum; em vez disso, existem vários "níveis de operação" frequentes.

Análise dos Boxplots:

- **Sensores :**

Os boxplots dos sensores são muito "achatados", com a caixa do interquartil (IQR) sendo extremamente pequena. Isso reforça visualmente a baixíssima variabilidade dos dados da maioria dos sensores.

Quase todos os sensores apresentam uma grande quantidade de **outliers** (pontos individuais fora dos "bigodes" do gráfico). No entanto, dado que as medições dos sensores são muito consistentes, esses "outliers" podem não ser erros, mas sim pequenas flutuações ou mudanças reais no processo que se destacam devido à estabilidade geral dos dados. É importante não removê-los prematuramente, pois podem conter informações valiosas.

- **Variáveis-Alvo:**

Os boxplots para Nivel Agua e Nível Total são bem mais "altos", mostrando a grande faixa de operação desses níveis.

Não são observados outliers nesses gráficos, pois a variação dos níveis é o comportamento esperado e natural do processo.

Análise do Heatmap e da Tabela de Correlação

O heatmap nos dá uma visão geral rápida das relações entre todas as variáveis.

Alta Correlação entre Sensores do Mesmo Canal:

Observamos "quadrados" de cores quentes (vermelho) na diagonal do heatmap. Isso indica que sensores do mesmo canal (ex: S1_1 a S1_6) são **altamente correlacionados** entre si. Por exemplo, quando a medição de S1_1 sobe, a de S1_2 tende a subir também. Isso é esperado, pois eles estão medindo o mesmo processo em pontos próximos. Essa alta correlação sugere a existência de **redundância** nos dados; talvez não seja necessário usar todos os 6 sensores do canal S1 para modelar o sistema.

Correlação com as Variáveis-Alvo (Nivel_total e Nivel_agua):

A correlação entre Nivel_total e Nivel_agua é muito alta (0.877), o que é esperado, já que o nível de água é o principal componente do nível total.

Correlações Positivas Fortes:

S2_2 (0.62) e S2_1 (0.51) têm a correlação positiva mais forte com Nivel_total. Isso significa que, quando as medições desses sensores aumentam, o nível total no tanque também tende a aumentar. **Estes parecem ser os sensores mais informativos para prever o nível total.**

Correlações Negativas Fortes:

S3_1 tem uma correlação negativa considerável com ambas as variáveis (-0.44 com Nivel_total e -0.52 com Nivel_agua). Isso indica uma relação inversa: quando a medição de S3_1 aumenta, os níveis tendem a diminuir. Este também é um sensor muito importante.

Correlações Fracas:

Muitos sensores, especialmente do canal S1 (S1_5, S1_6) e o S4_2, têm correlação muito próxima de zero com Nivel_total. Isso sugere que eles têm pouca ou nenhuma relação *linear* com o nível e podem ser menos úteis para um modelo preditivo simples.

3 CONSIDERAÇÕES FINAIS

Principais Padrões Encontrados

O processo de separação trifásica opera em diferentes patamares de nível, em vez de variar suavemente, como visto nas distribuições multimodais de Nivel_total e Nivel_agua. As medições da maioria dos sensores são extremamente estáveis, mas alguns deles respondem de forma clara e linear às variações de nível. Especificamente, os sensores do canal 2 (S2_1, S2_2) possuem uma forte correlação positiva com os níveis, enquanto o sensor S3_1 apresenta uma forte correlação negativa, tornando-os os preditores mais promissores.

Problemas Identificados

O conjunto de dados é de alta qualidade, sem valores ausentes ou duplicados. O principal problema identificado é a alta redundância entre os sensores de um mesmo canal (especialmente o Canal S1), o que sugere que usar todos eles em um modelo pode ser ineficiente. Além disso, os boxplots revelaram muitos outliers nos sensores, que, devido à baixa variabilidade geral, provavelmente representam flutuações reais e informativas do processo, e não erros de medição, devendo ser tratados com cuidado.

Hipóteses para Próximos Passos

1. **Modelagem Simplificada:** É possível construir um modelo de regressão preciso para prever o Nivel_total utilizando um subconjunto muito pequeno de sensores (ex: S2_2 e S3_1), descartando os demais sem grande perda de informação.
2. **Engenharia de Atributos:** A redundância dos sensores do Canal 1 pode ser tratada criando um único atributo, como a média das medições (S1_mean), para representar o canal inteiro.

Código:

https://colab.research.google.com/drive/1pguKm796nD9kSCG5x03aTxagjBDjvsh0?usp=sharing