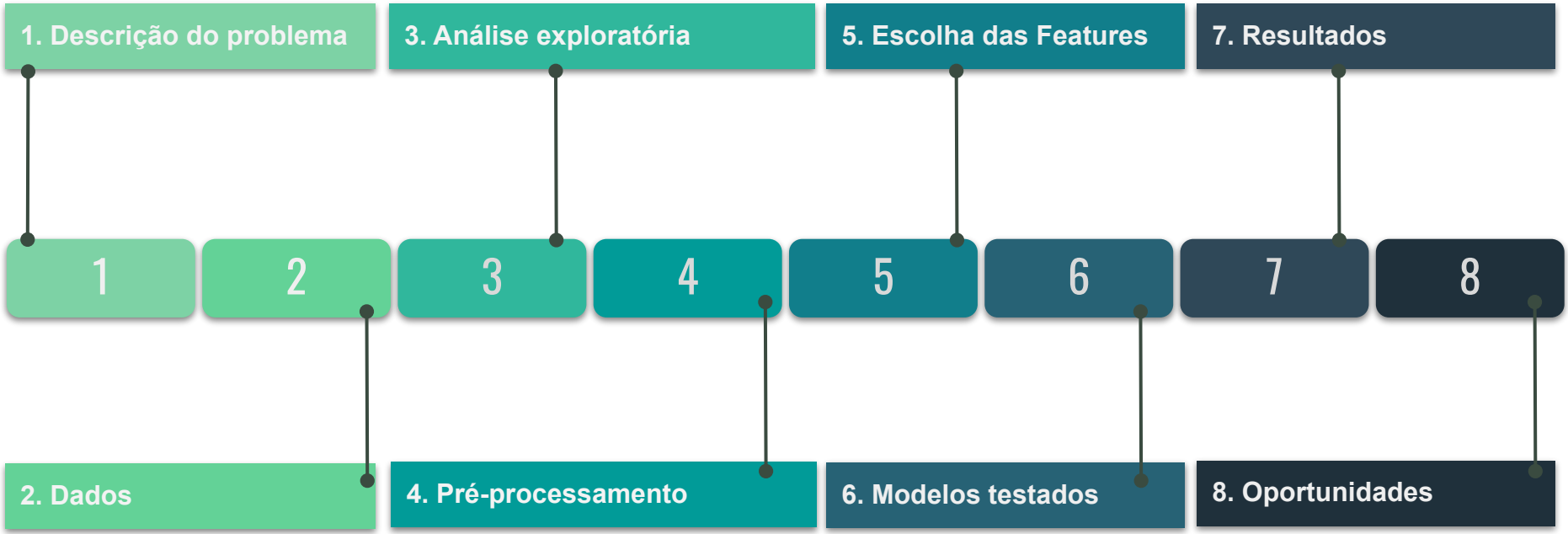
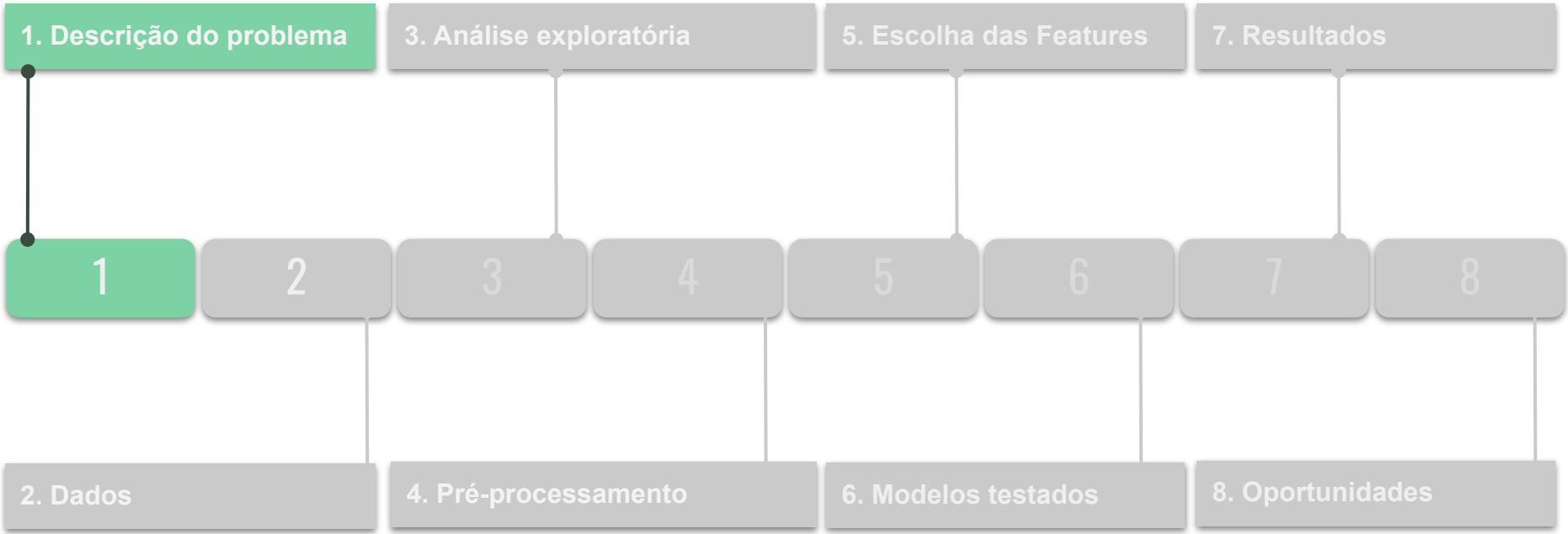
Desafio de Dados

Totmés Scheffer - Aspirante a Cientista de Dados

Roteiro (Roadmap)



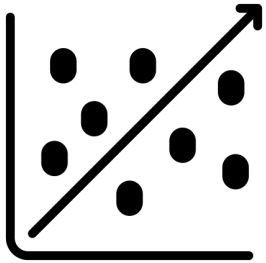
Roteiro (Roadmap)



Planta de Flotação

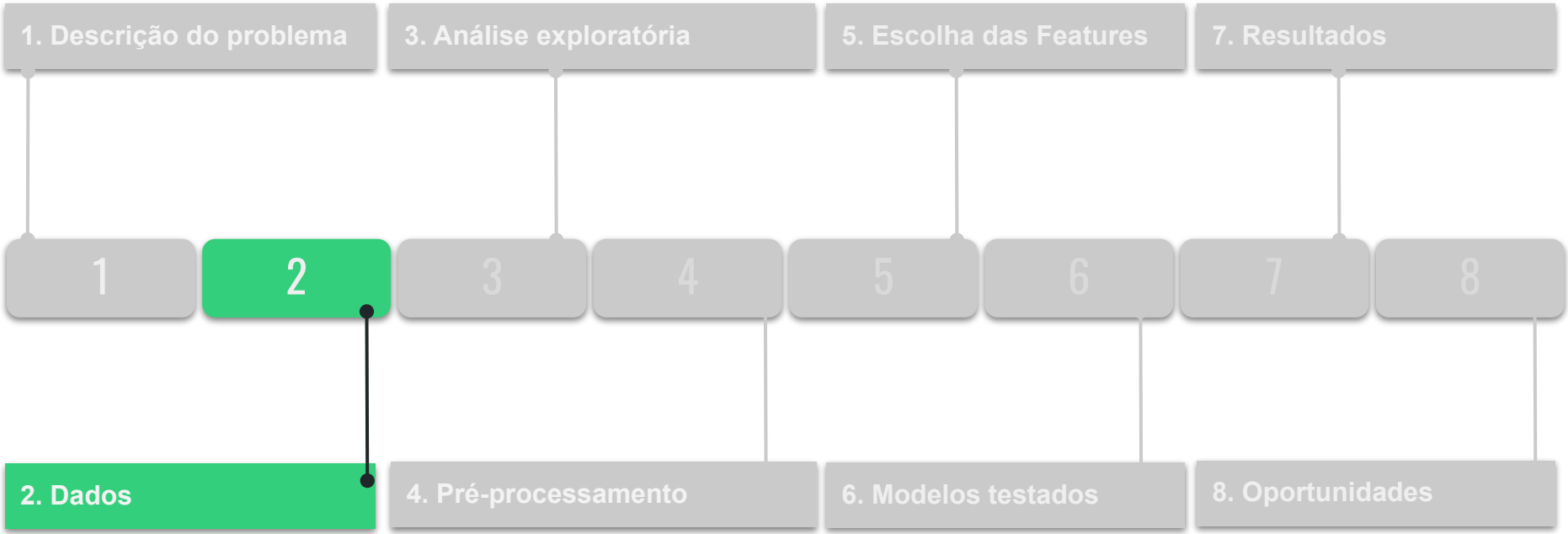
- **Motivação:** O Minério concentrado de ferro, **produto final** do processo de flotação, deve ser obtido com uma concentração de Ferro e Sílica **controlada** dentro do especificado.
- **Dificuldade:** As concentrações são medidas em **laboratório** a cada hora, porém os resultados saem com um atraso de 2 horas.

Proposta para Solução

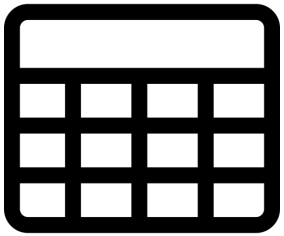


Elaborar um modelo de **Regressão** utilizando Machine Learning, com **objetivo** de estimar os valores de concentração de **Sílica** no **produto final** para antecipar flutuações do processo, buscando melhorar a eficiência de seu controle.

Roteiro (Roadmap)

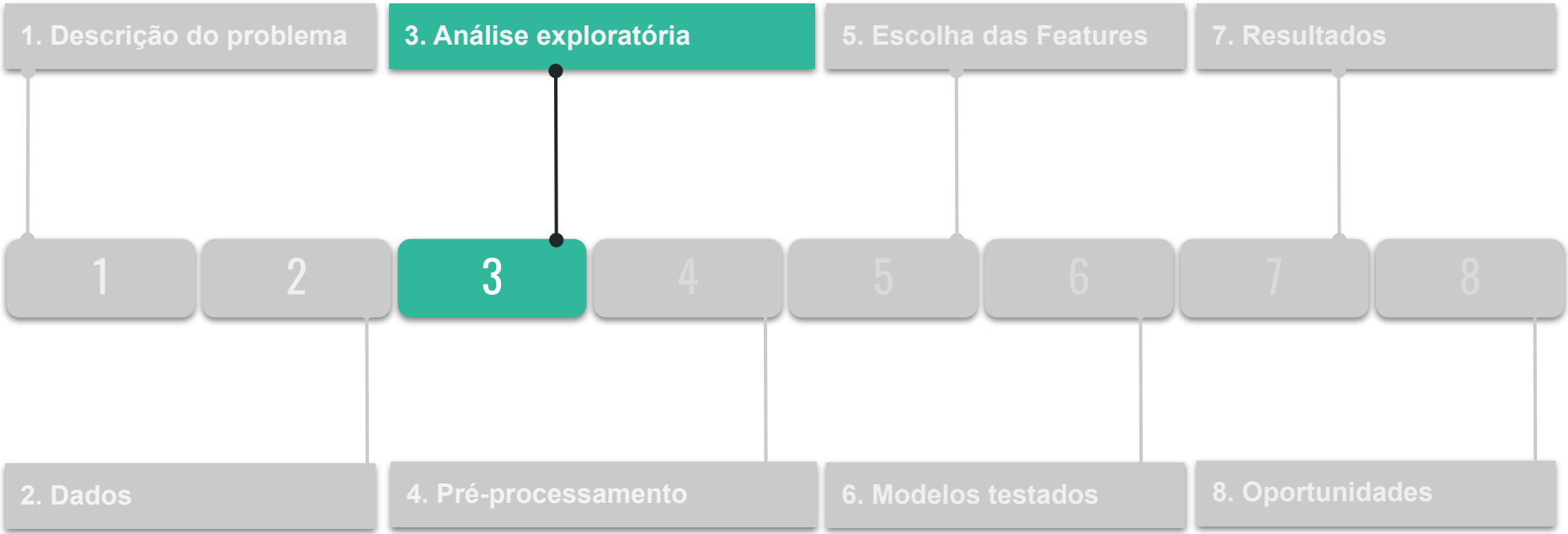


Dataset disponível no Kaggle



- Tabela em csv contendo 737.453 entradas numéricas de medições para 24 variáveis.
- Nenhum nulo.
- Uma coluna de *TimeStamp* para as medições.
- Novos dados a cada 20s.

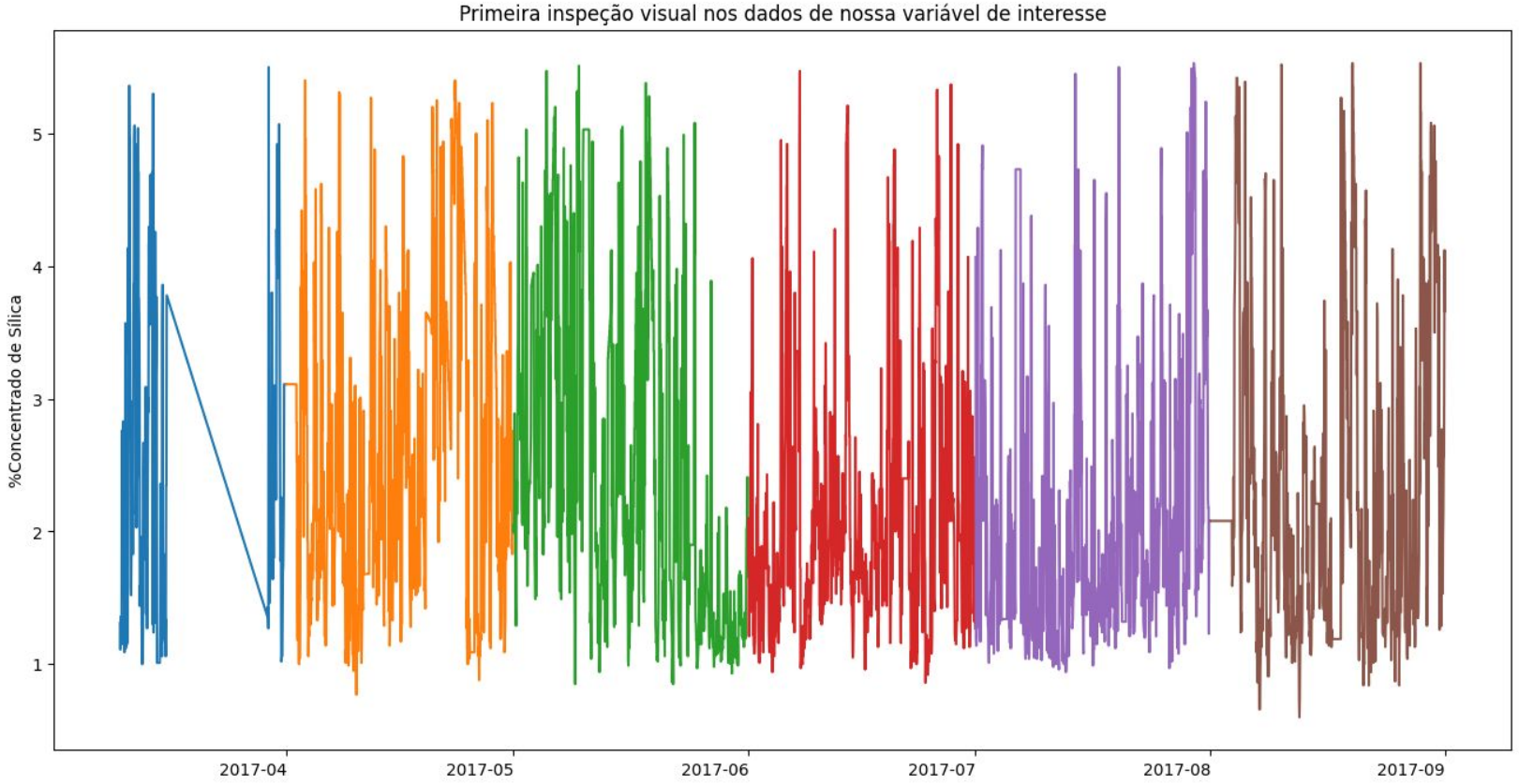
Roteiro (Roadmap)



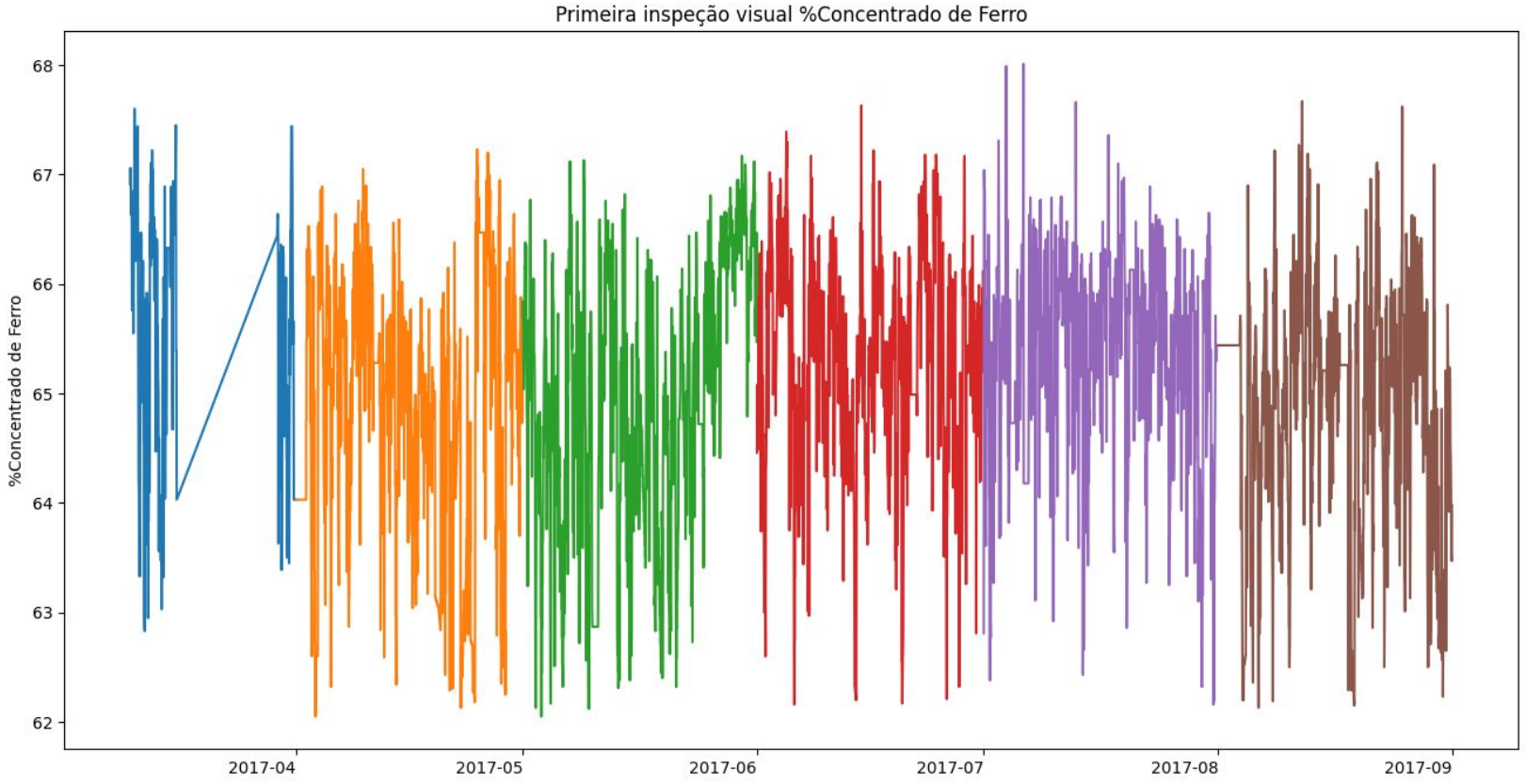
Análise exploratória

- Dados para 184 dias, entre 10/03/2017 e 09/09/2017.
- Existem lacunas de dados entre alguns períodos.
- O primeiro dia não apresentou os 4320 dados esperados.
- Novos dados a cada 20s.
- Novos valores de porcentagem de alimentação de Ferro e Sílica foram inseridos menos de 2 vezes por dia.
- Possível falha na amostragem dos dados medidos em laboratório. Foram encontradas, em média, 321 novos valores por dia para % Concentrado de Sílica, onde eram esperados 24.
- Em 14% dos dados de % Concentrado Sílica um novo valor foi inserido com atraso superior a 1 hora.

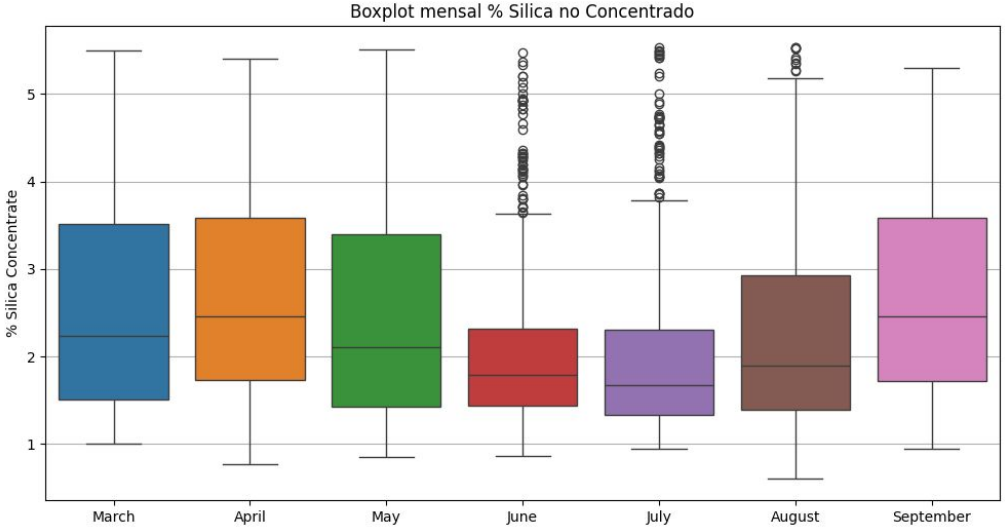
Análise exploratória

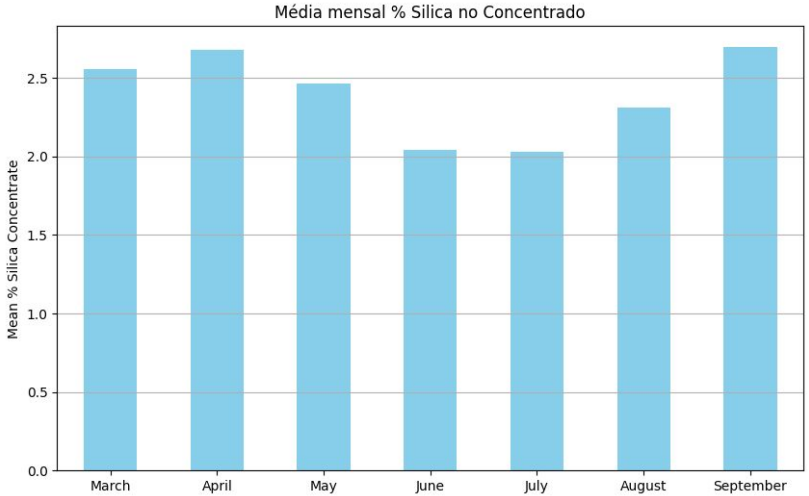


Análise exploratória

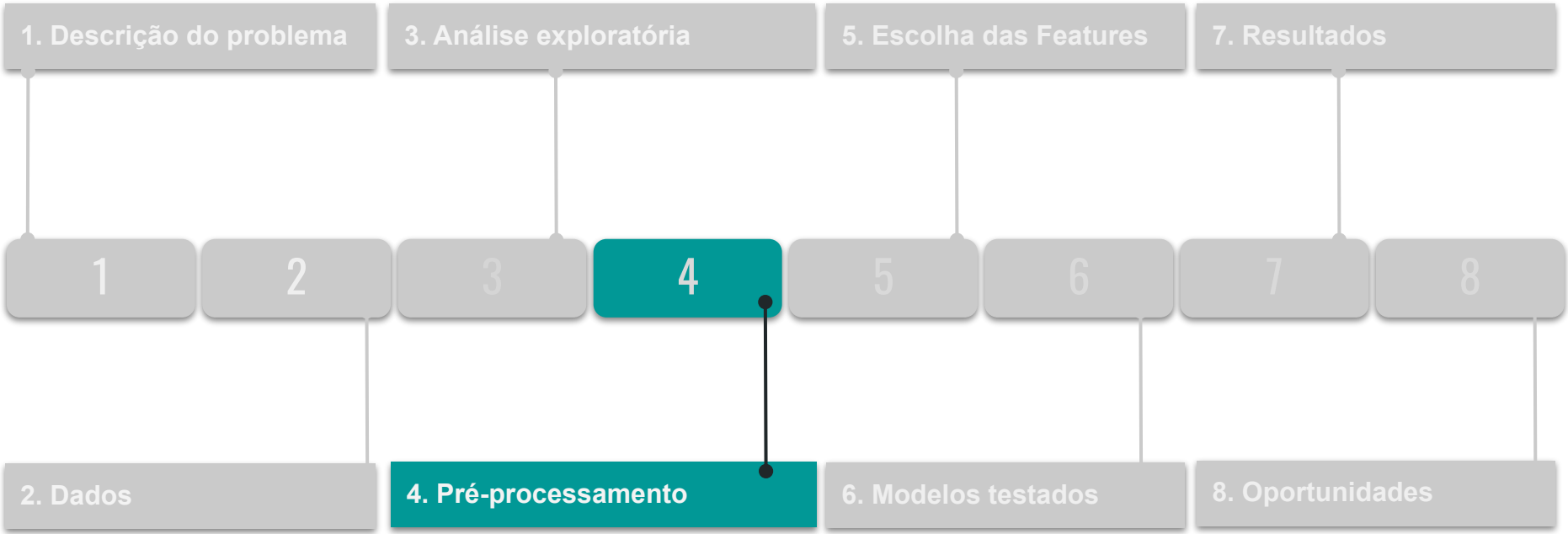


Análise exploratória



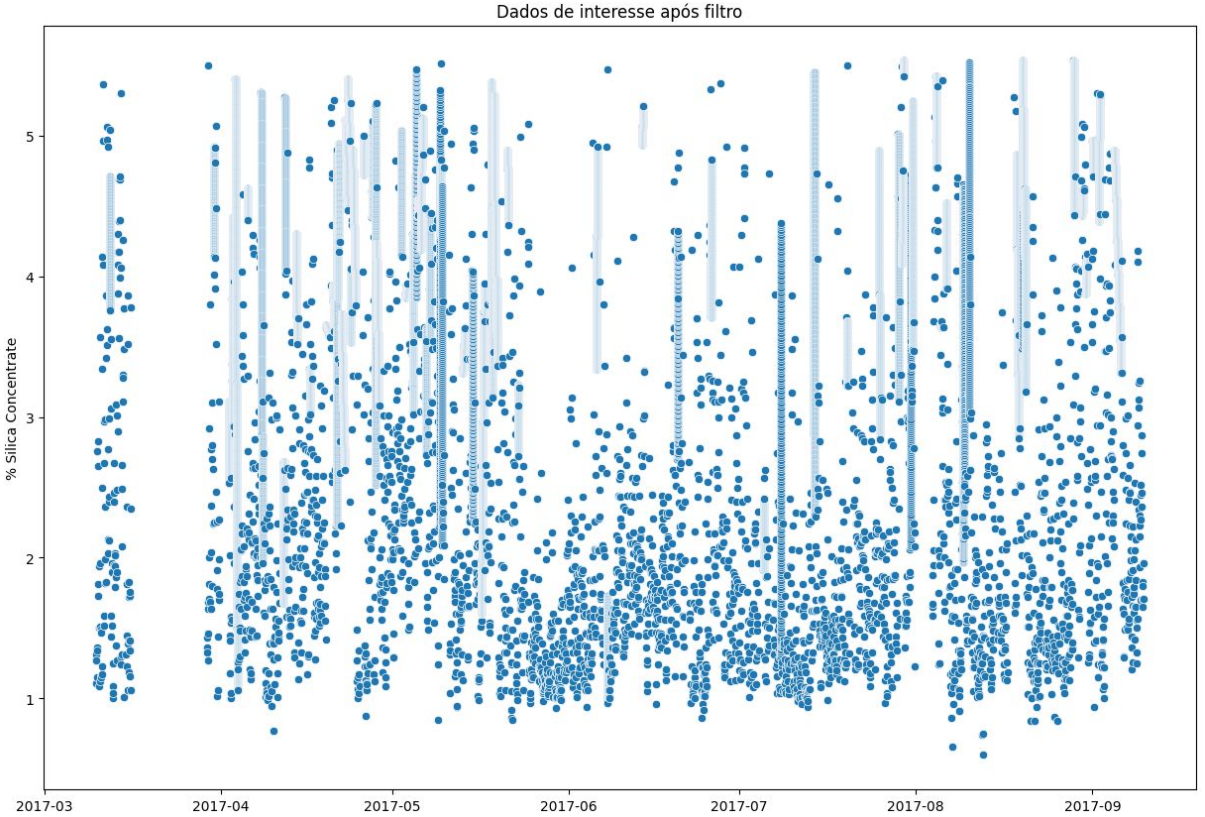


Roteiro (Roadmap)



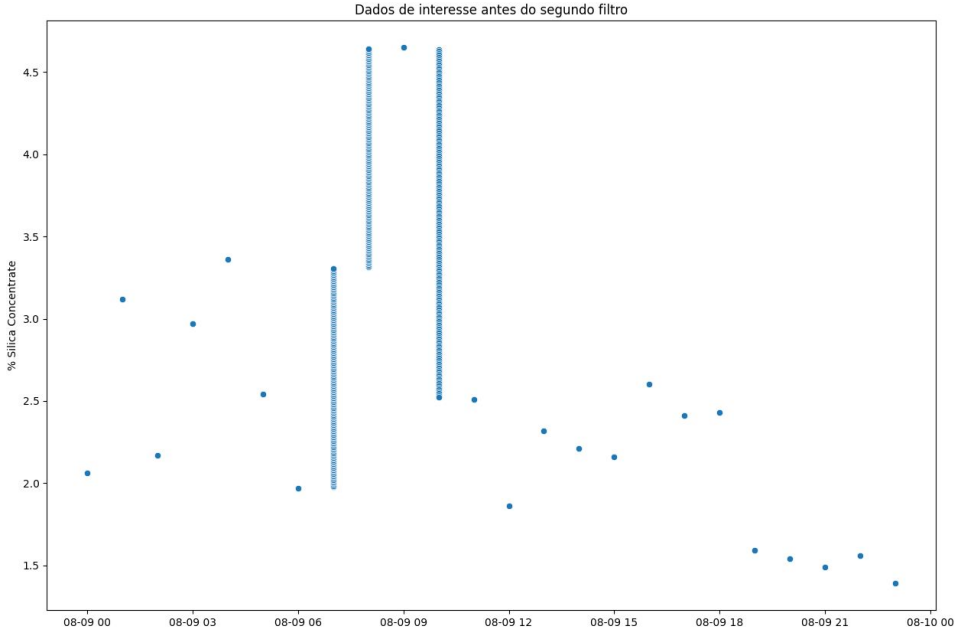
Pré-processamento

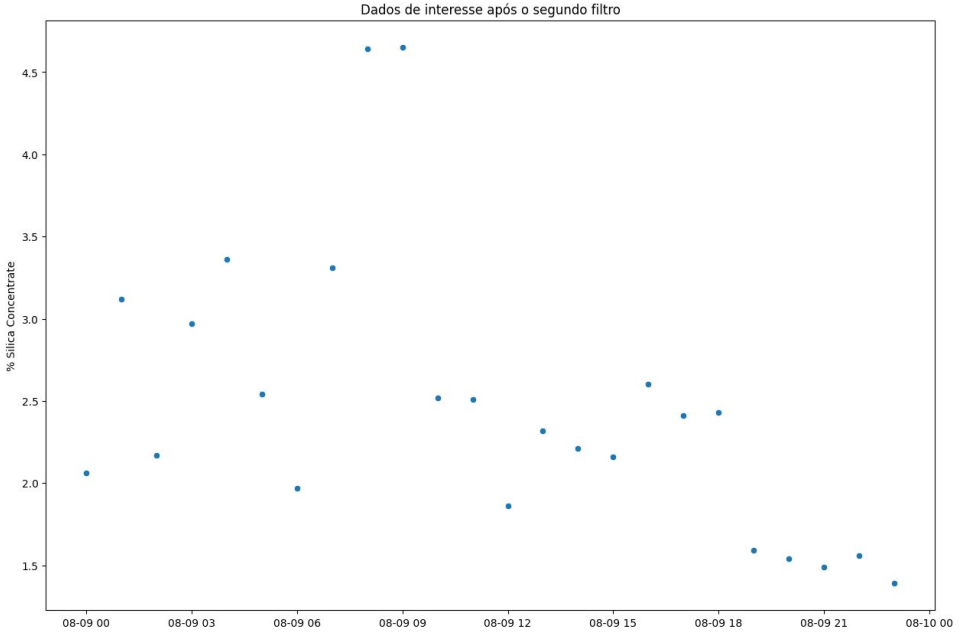
- Como a variável de interesse é a % Sílica no concentrado, apenas momentos em que o valor **é atualizado** são **mantidos**.
- Após **inspeção visual** dos pontos restantes, foi possível observar algum **desvio**, que comprova que existem valores sendo alterados dentro de uma mesma hora.
- O mesmo comportamento foi encontrado para a % Ferro no concentrado, também medido em **laboratório**.



Pré-processamento

- Os dados foram re-amostrados por hora, utilizando o último valor e removendo os valores nulos.



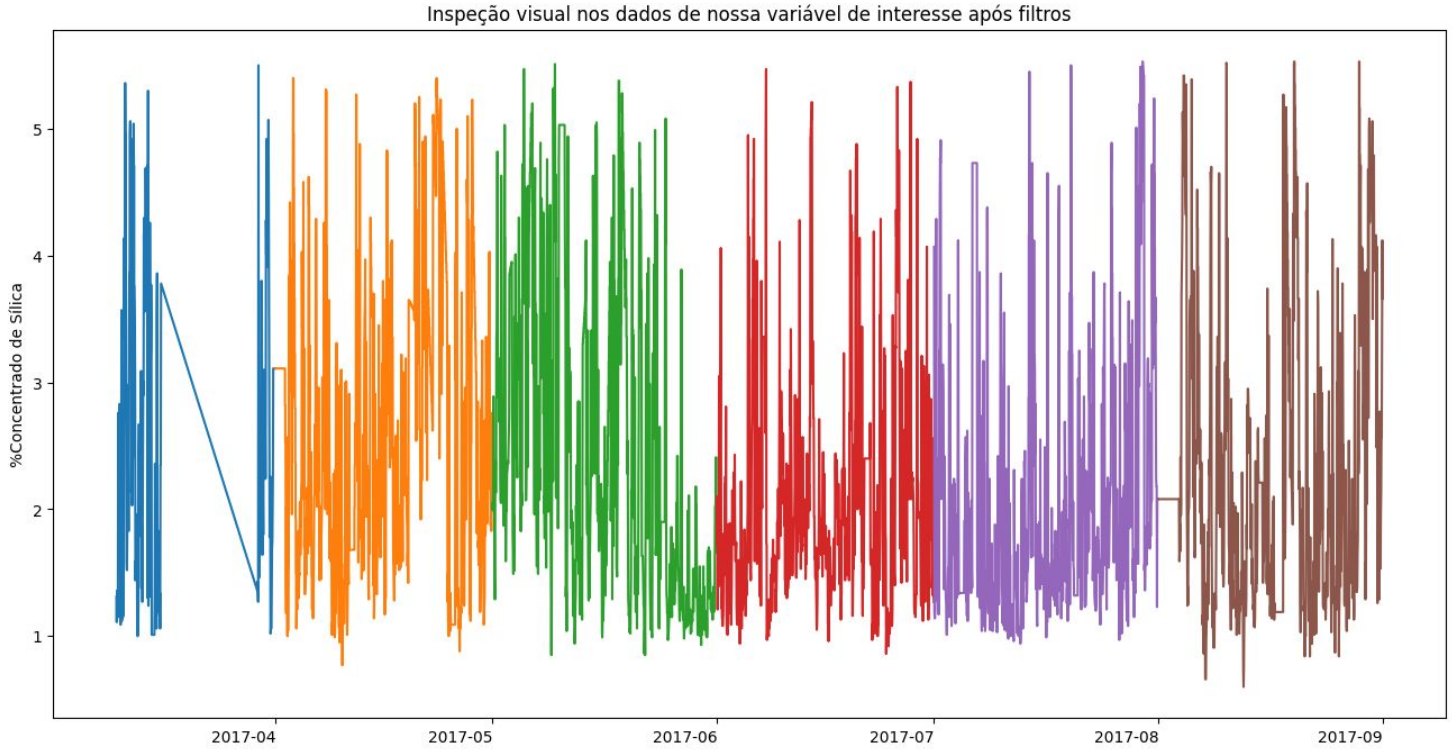


Pré-processamento

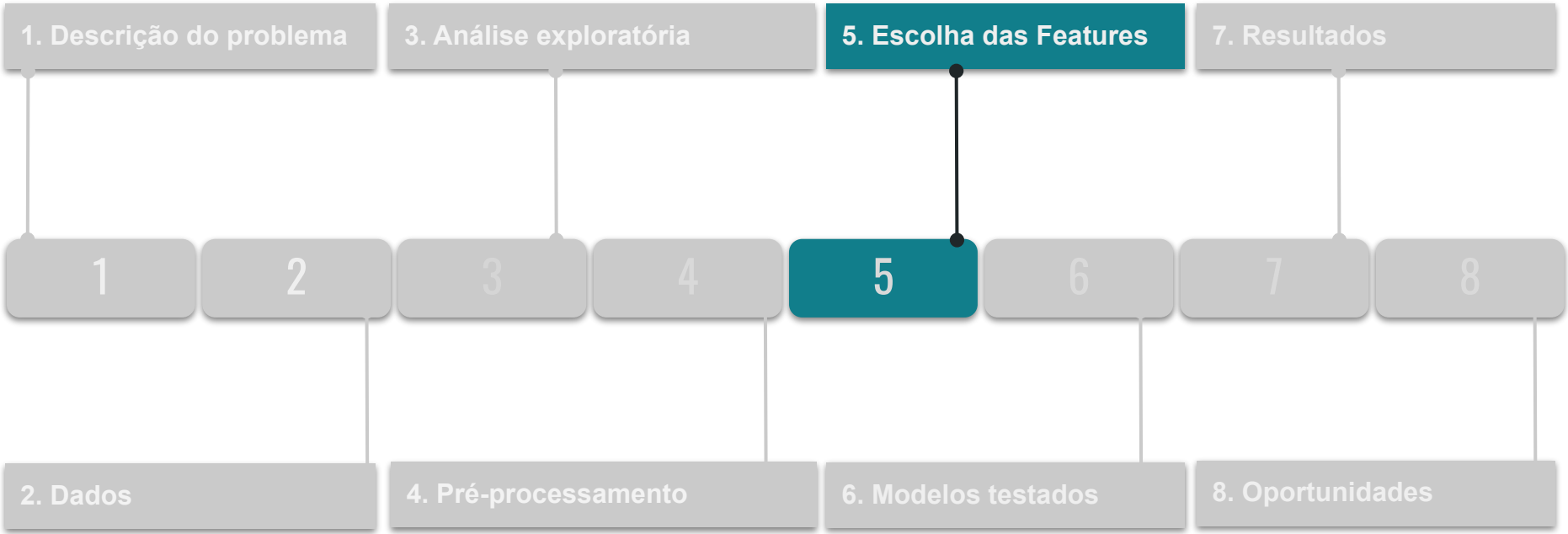
- # de dados removidos no primeiro filtro: 678312
- % de dados removidos no primeiro filtro: 92.0%
- # de dados removidos: 733802
- % de dados removidos: 99.5%
- 3651 Entradas de dados restantes
- Também foi aplicado um atraso de 2 horas nas variáveis de Laboratório.

Pré-processamento

- 3651 Entradas de dados restantes, sem perda de informação no gráfico temporal



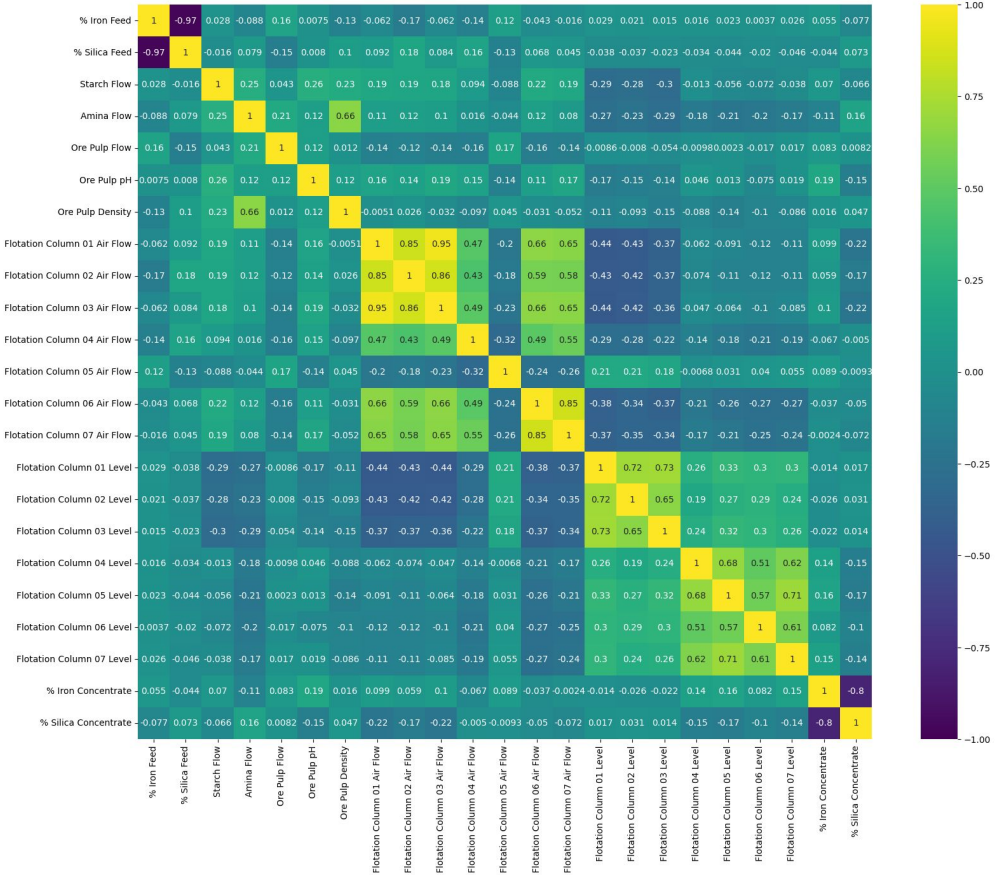
Roteiro (Roadmap)



Escolha das Features

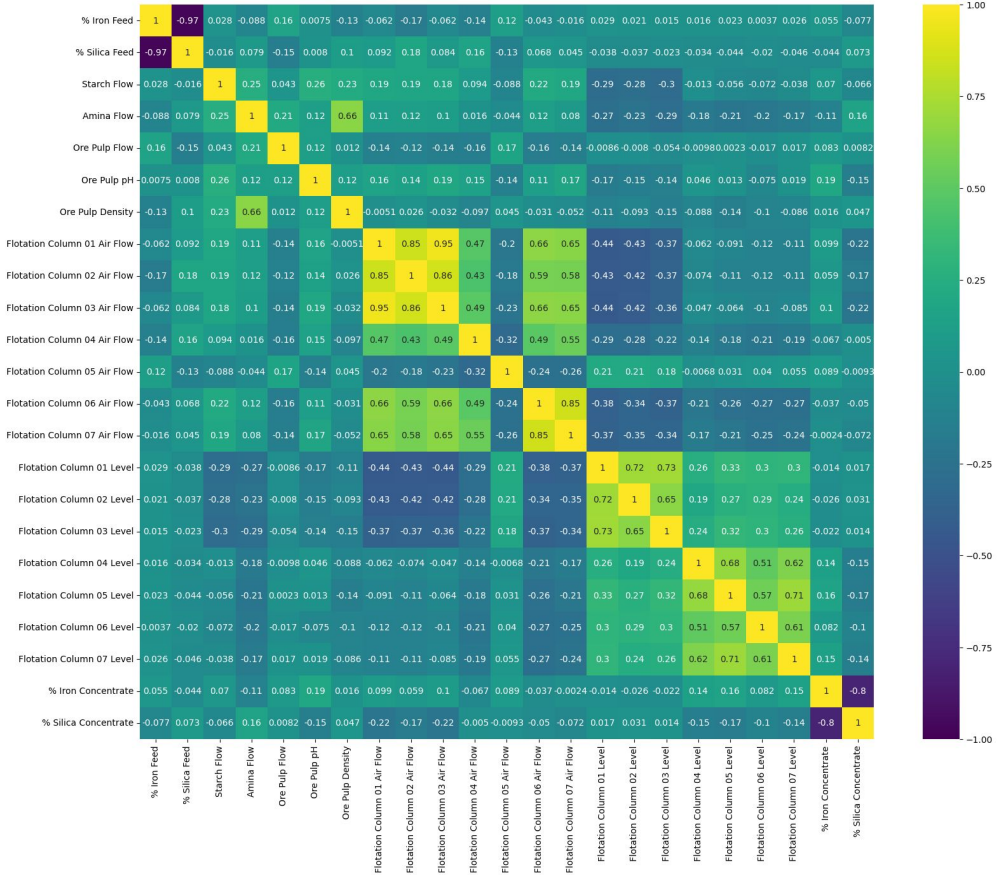
- Com base na matriz de correlação, as features de maior correlação com nossa variável de interesse foram:

- Amina Flow,
- Ore Pulp pH,
- Flotation Column 01 Air Flow,
- Flotation Column 02 Air Flow,
- Flotation Column 03 Air Flow,
- Flotation Column 04 Level,
- Flotation Column 05 Level,
- Flotation Column 06 Level,
- Flotation Column 07 Level,
- % Iron Concentrate



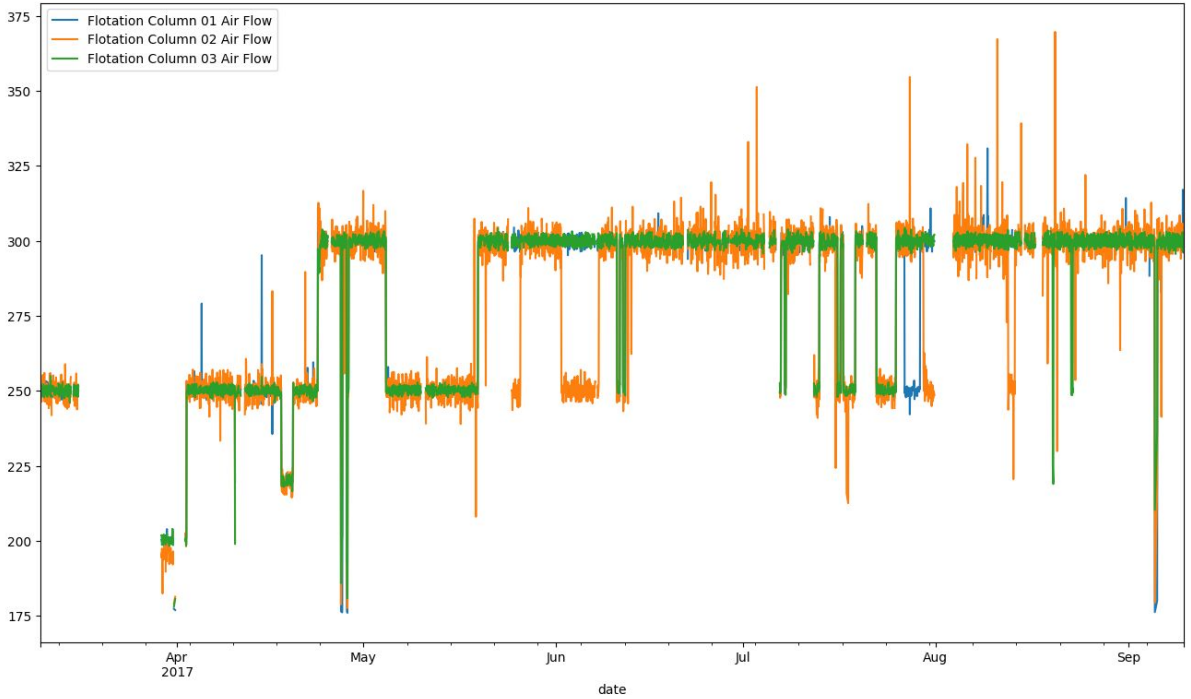
Escolha das Features

- Por conta da alta correlação entre as features escolhidas de **Air Flow** e **Level**, também foram criadas features agregadas pela **média** dos seus valores, como alternativas para o desenvolvimento de modelos com redução na dimensionalidade.



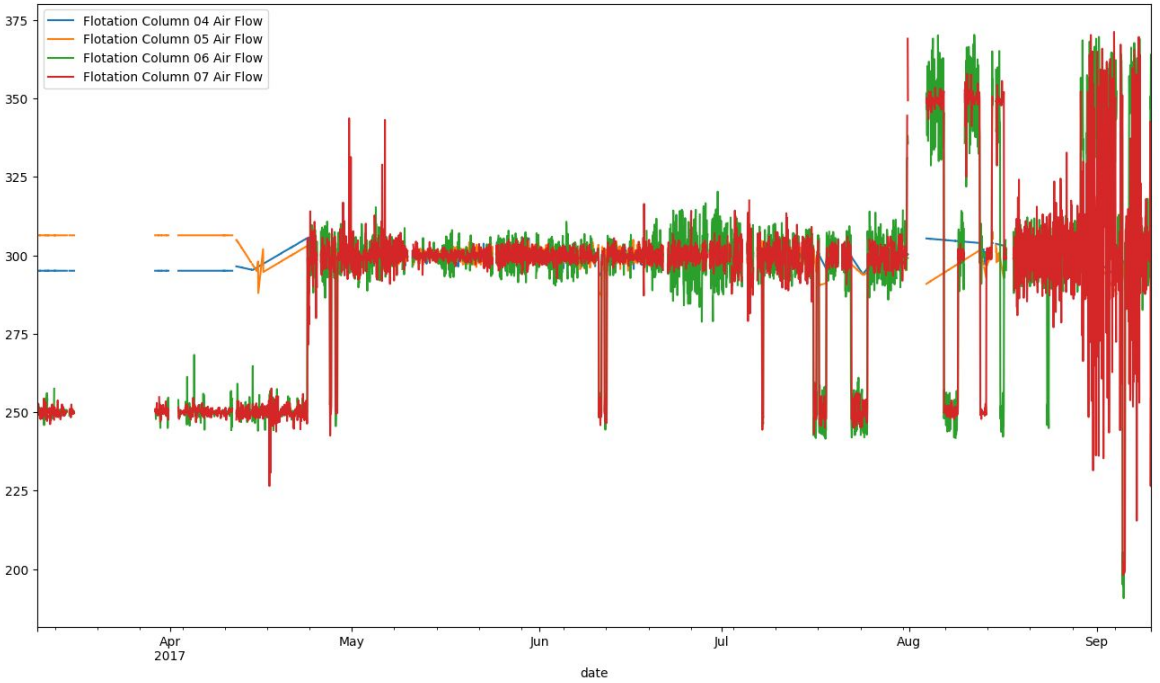
Escolha das Features

- Analisando Especificamente as variáveis de Air Flow foram observados alguns desvios.

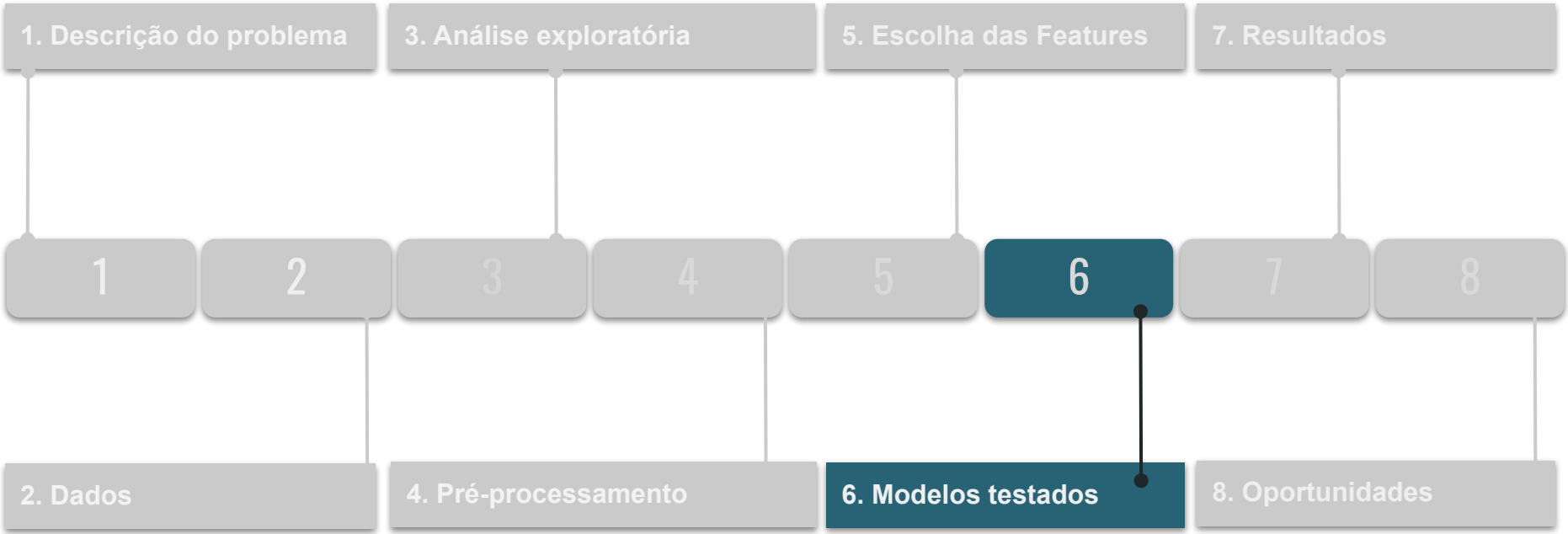


Escolha das Features

- Analisando Especificamente as variáveis de Air Flow foram observados alguns desvios.



Roteiro (Roadmap)



Modelos testados

Modelos de Regressão Simples, utilizando SciKit-Learn

- Linear
- Polinomial
- SVR Linear
- SVR rbf
- SVR rbf 2
- Random Forest
- MLP

Modelos de Regressão Complexos, utilizando TensorFlow/Keras

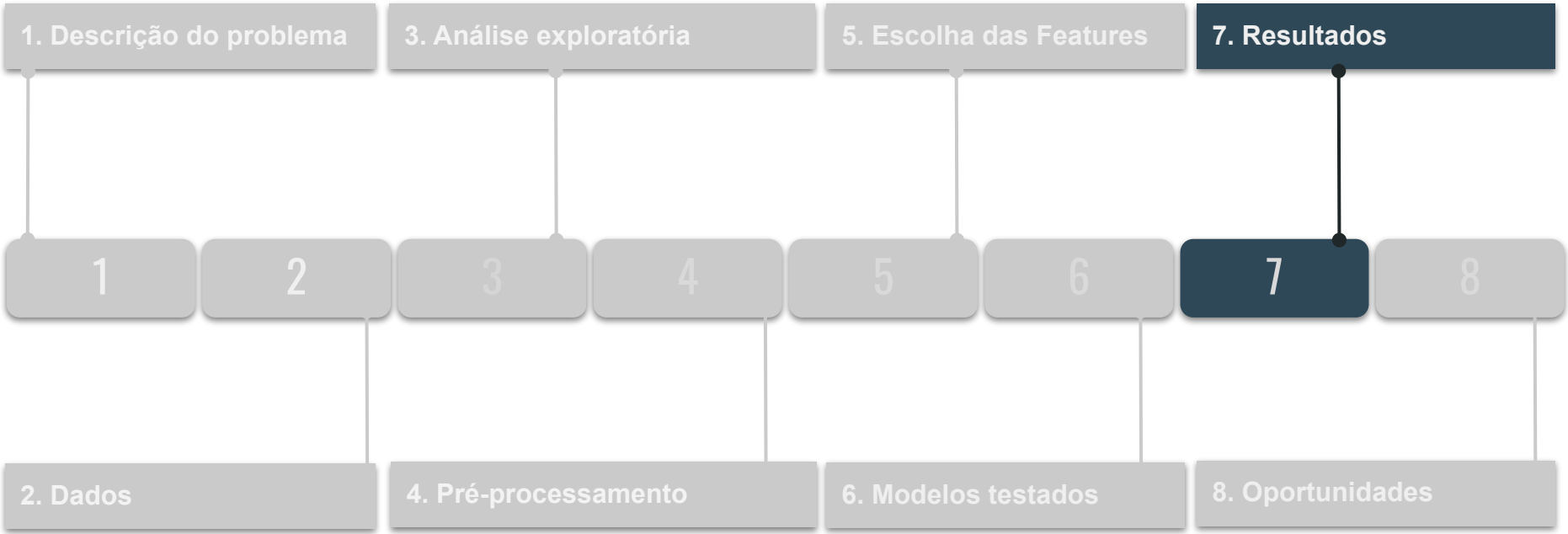
- Neural Network (NN)
- Recurrent Neural Network (RNN)
- Long short-term Memory (LSTM)

Configurações de treinamento para os modelos complexos

- Total de épocas = 1000
- Early Stop = 30 patience
- Reduce Lr on Plateau = 20 patience
- Otimizador = SGD
- Loss = Mean Squared Error(MSE)

*SVR = Support Vector Machine for Regression
MLP = Multilayer Perceptron - a simple feedforward artificial neural network

Roteiro (Roadmap)

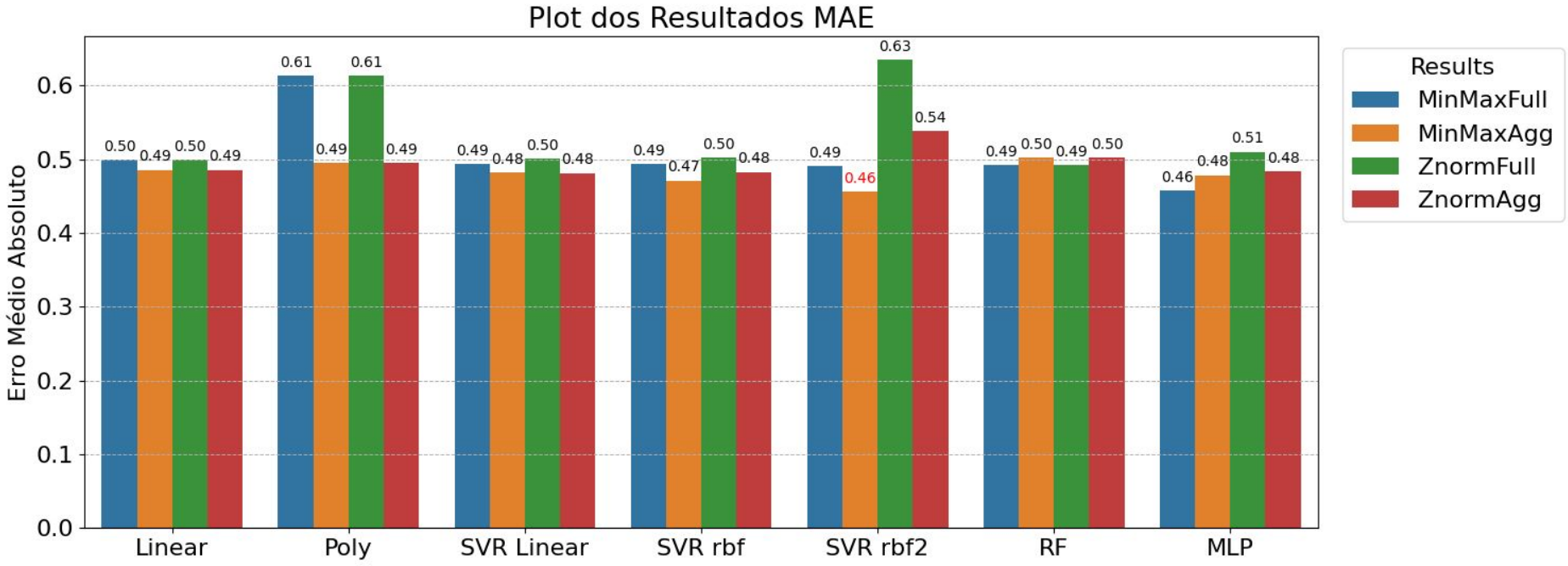


Resultados

- Foram utilizadas 4 configurações diferentes de treinamento para todos os modelos:
  - Normalização MinMax todas as features.
  - Normalização MinMax dimensionalidade de features reduzida.
  - Normalização ZNorm todas as features.
  - Normalização ZNorm dimensionalidade de features reduzida.

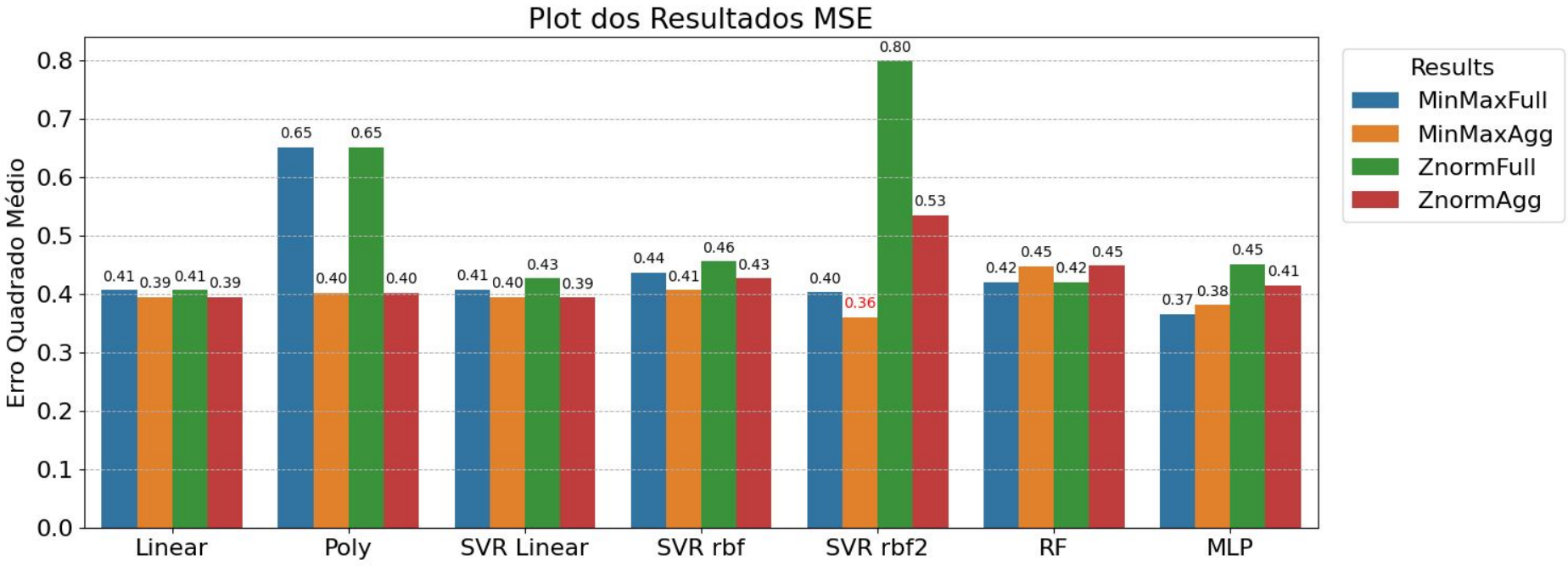
Resultados

- Modelos Simples



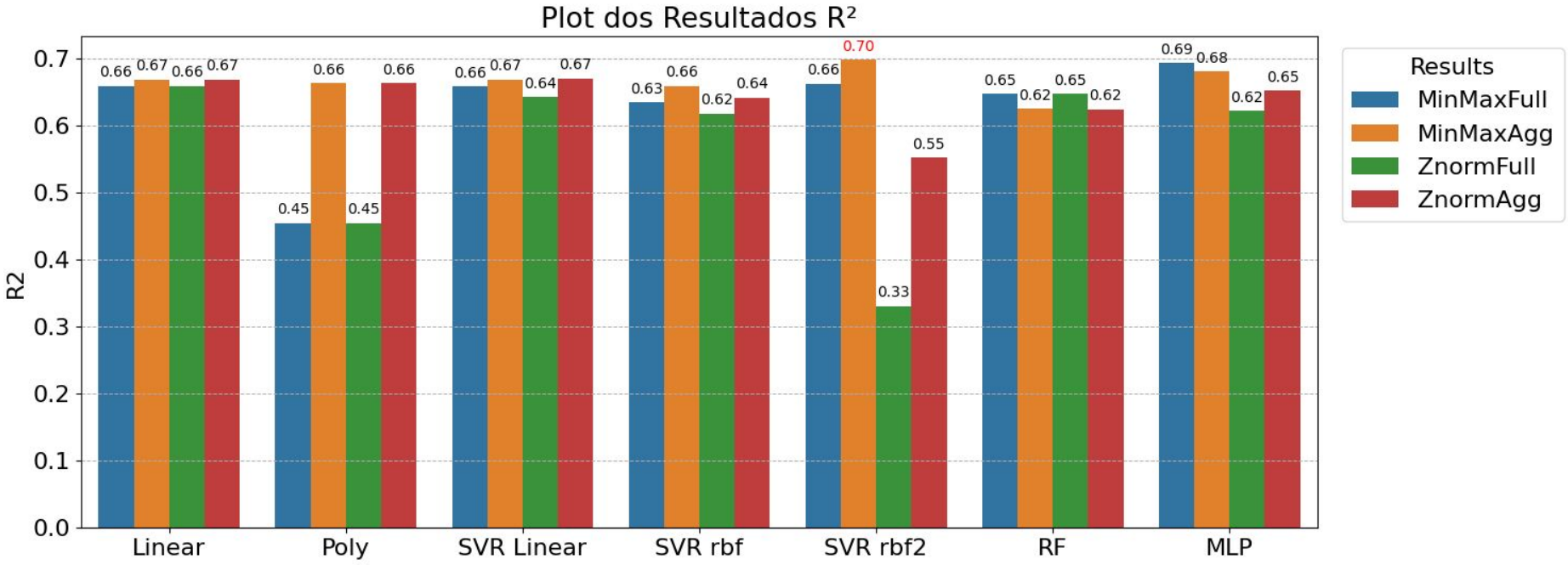
Resultados

- Modelos Simples



Resultados

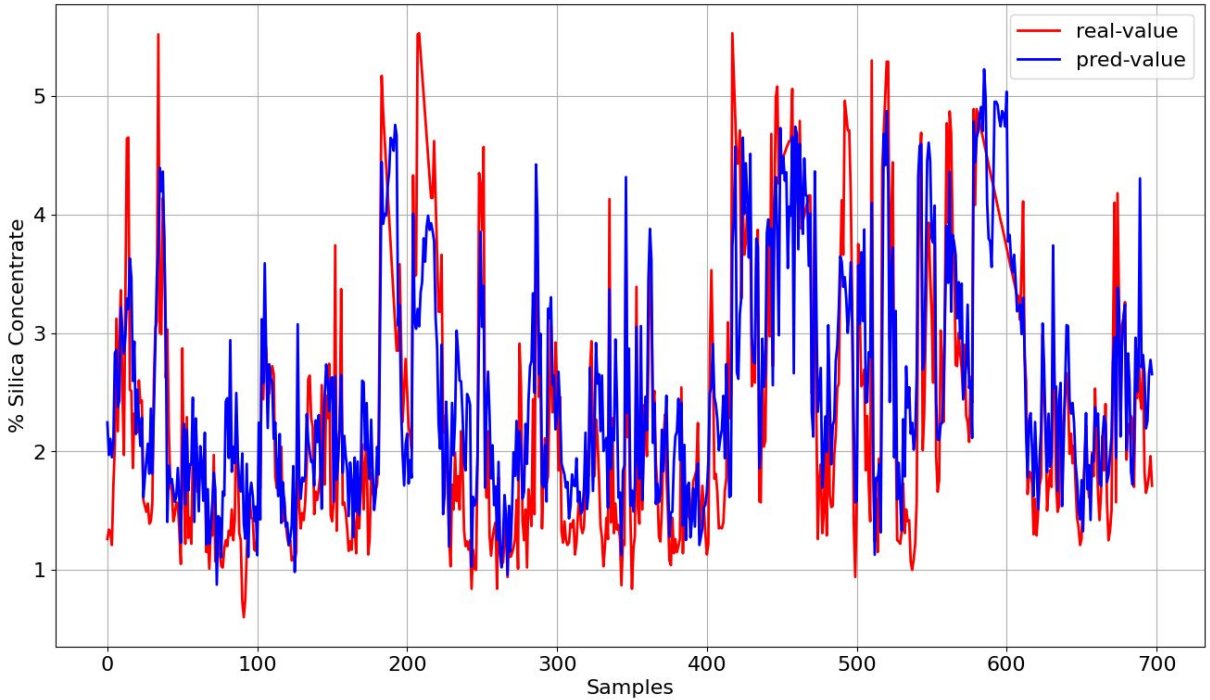
- Modelos Simples



Resultados

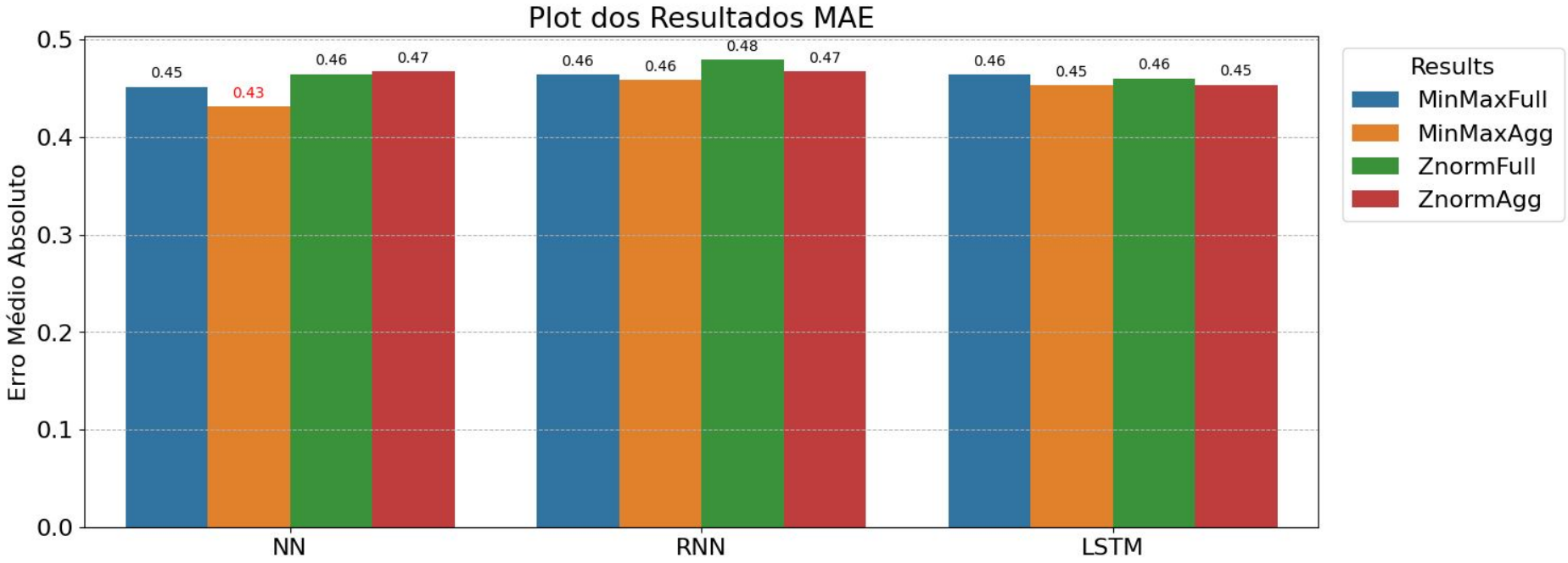
- TESTE Modelos Simples
  - Melhor modelo foi SVR rbf2, utilizando MinMaxAgg.

MAE = 0.50
MSE = 0.41
R2 = 0.70



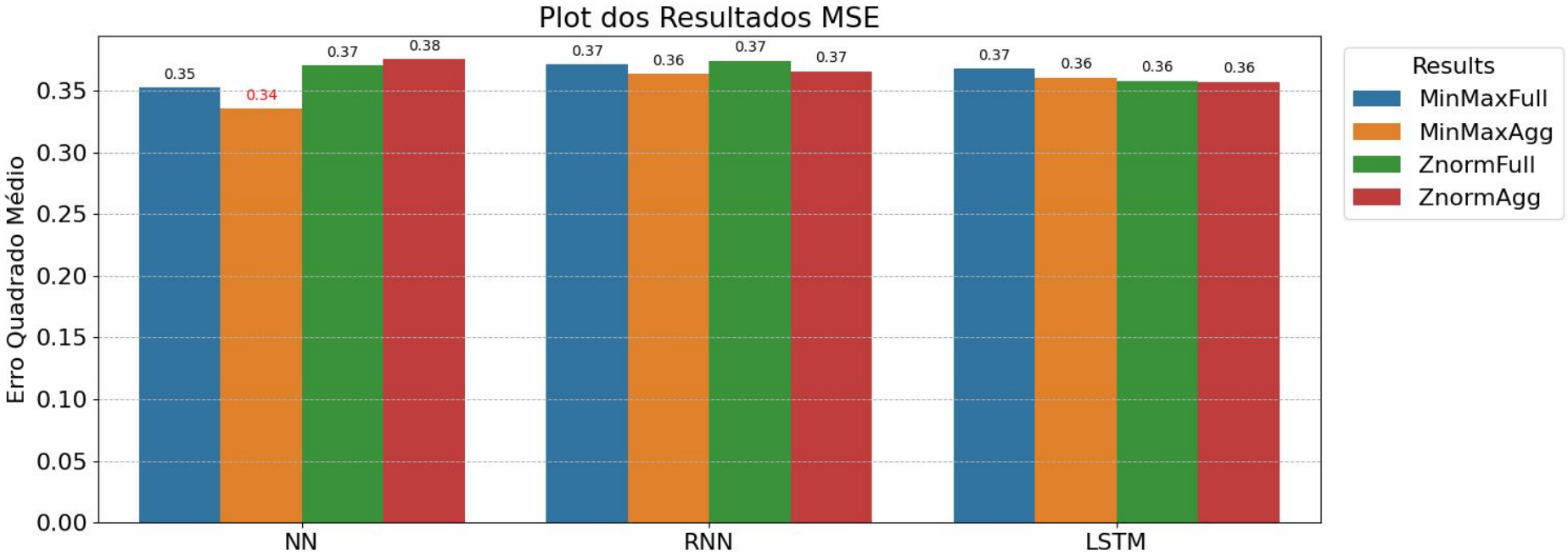
Resultados

- Modelos Complexos



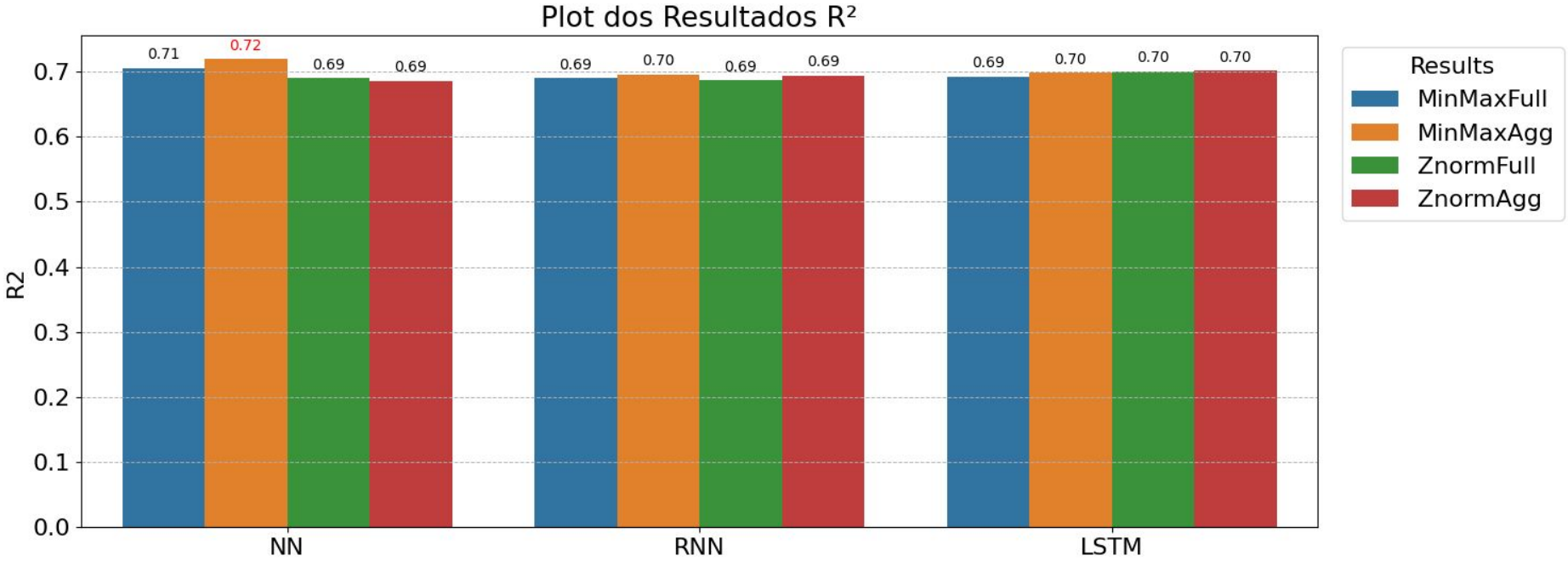
Resultados

- Modelos Complexos



Resultados

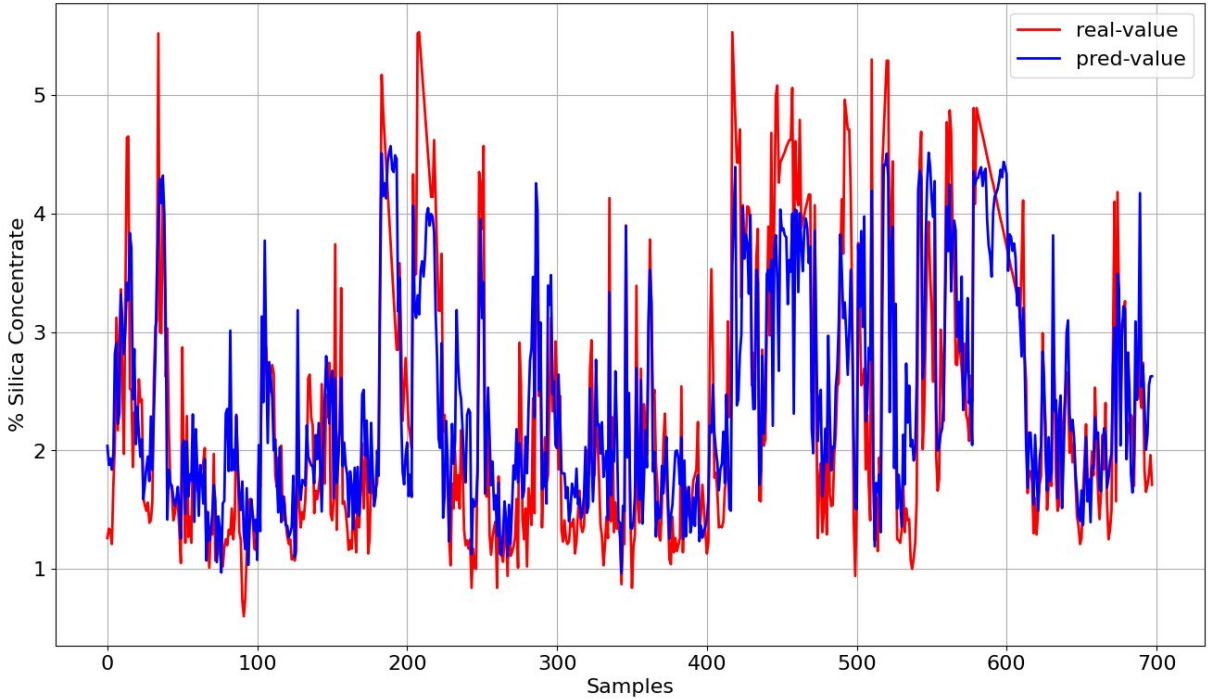
- Modelos Complexos



Resultados

- TESTE Modelos Complexos
  - Melhor modelo foi NN, utilizando MinMaxAgg.

MAE = 0.47
MSE = 0.38
R2 = 0.72



Considerações Finais

- Complexidade dos modelos melhorou o resultado.

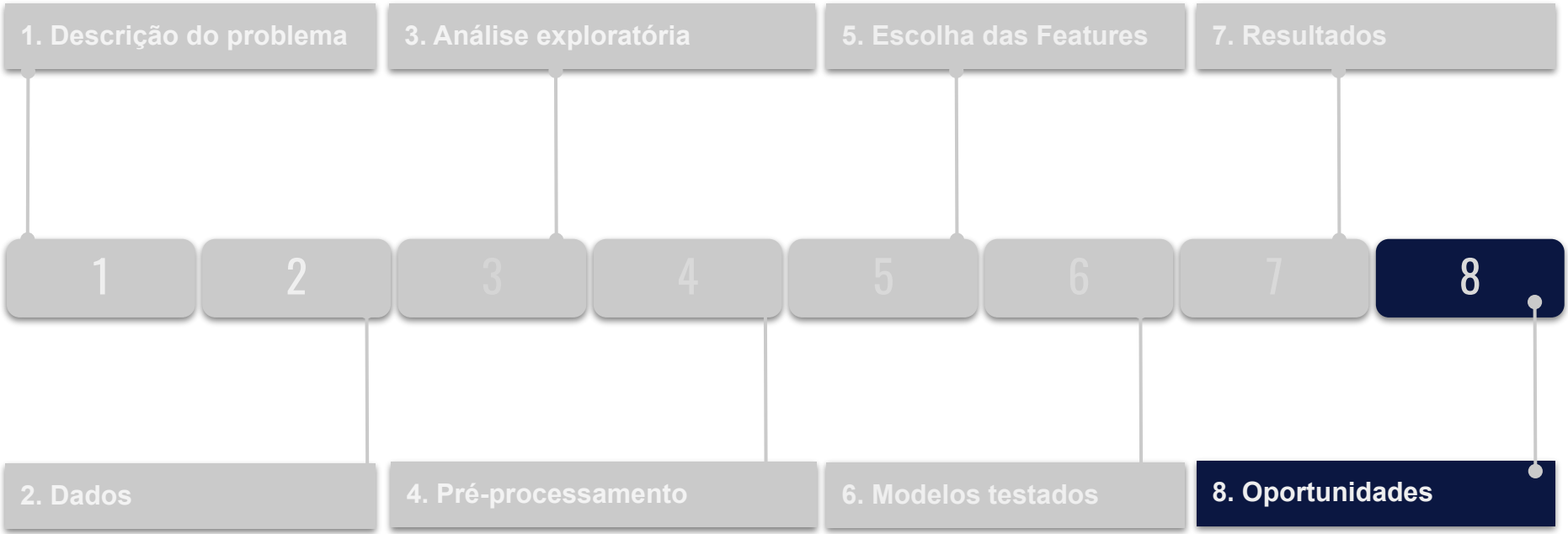
- Redução da dimensionalidade melhorou o resultado.

- Variável de laboratório como feature.

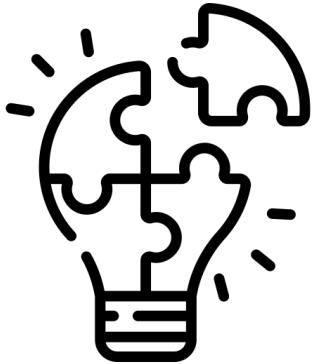
- Poucas features com boa correlação.

- **Resultado satisfatório**

Roteiro (Roadmap)



Oportunidades



- Implementar outro regressor para a % Ferro no concentrado.
- Ensemble de modelos.
- Melhorias na aquisição e quantidade das features.
- Explorar outros métodos de treinamento.
- Explorar Modelos mais complexos .

Muito Obrigado!

Totmés Scheffer