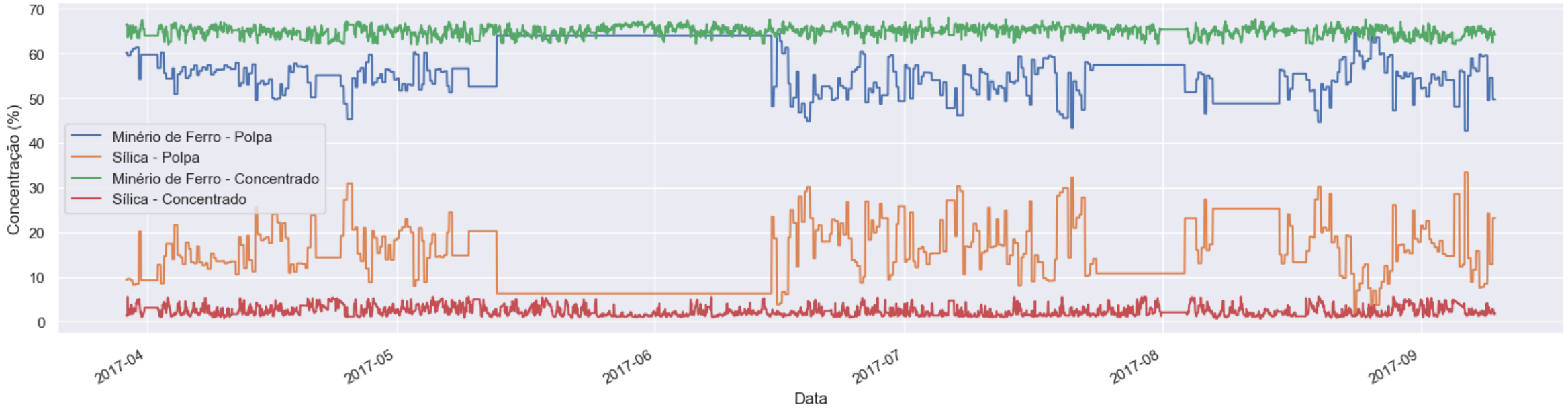
AVALIAÇÃO DE OPORTUNIDADES

Flotação

02 de junho de 2025

Processo analisado

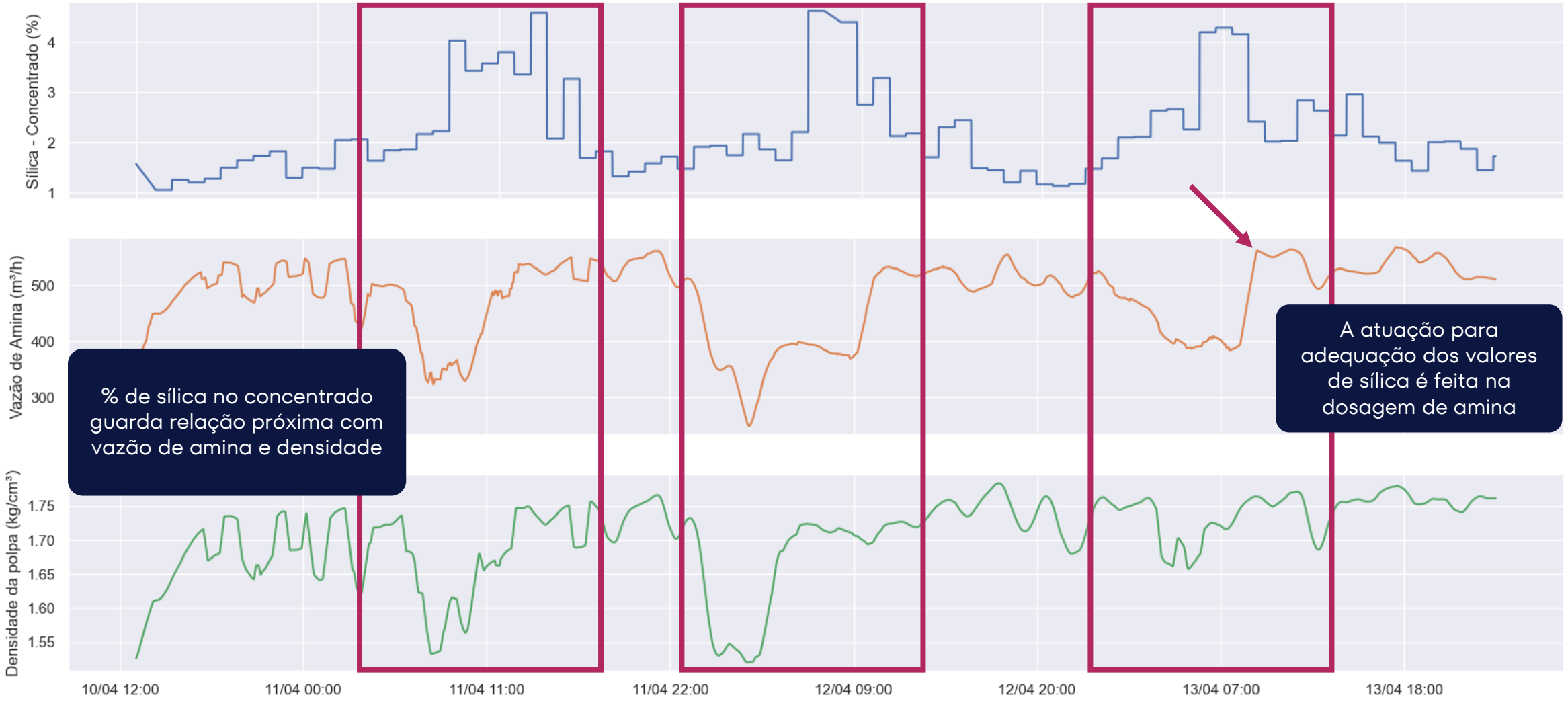
- Flotação de minério de ferro com sílica como principal mineral de ganga.
- **Objetivo:** aumentar a concentração de minério de ferro e reduzir a concentração de sílica.
  - Concentração média de entrada (polpa): 54,6% de minério de ferro e 16,5% de sílica.
  - Concentração média de saída (concentrado): 65,2% de minério de ferro e 2,2% de sílica.



Processo analisado

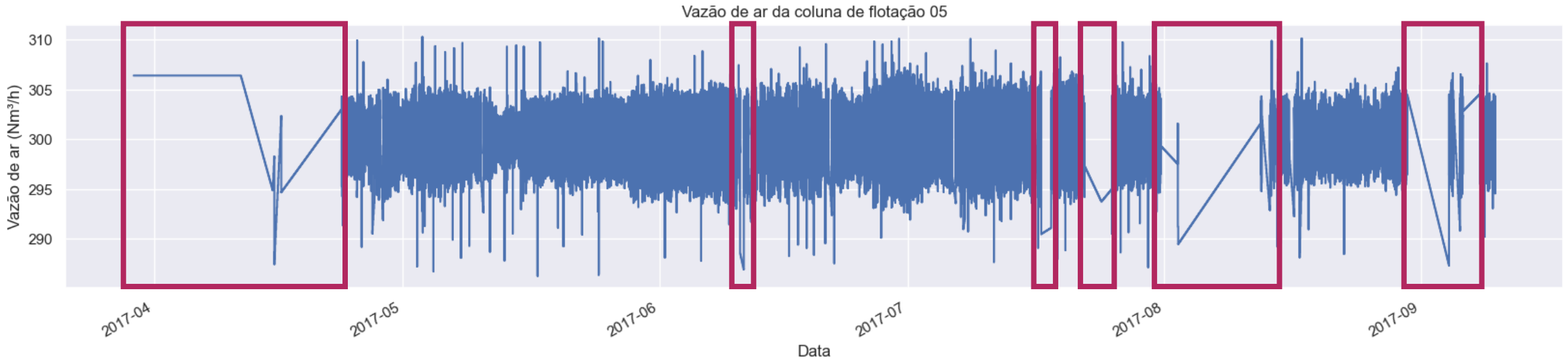
- Amina e amido são utilizados como coletor e depressor, respectivamente.
- Concentrações de minério de ferro e sílica são inversamente proporcionais e complementares (quando uma aumenta, outra abaixa).
- **Dados disponíveis:**
  - O perfil do concentrado é medido em laboratório, com amostras a cada 1h;
  - O perfil da polpa também é medido em laboratório, mas de forma mais irregular;
  - Os demais dados de processo são medidas online, amostradas a cada 20s. Estão disponíveis dados de:
    - Vazão, pH e densidade da polpa;
    - Vazão de amina;
    - Vazão de amido;
    - Nível das colunas de flotação;
    - Vazão de ar nas colunas de flotação.

Processo analisado



Avaliação e limpeza dos dados

- Dados de 10/03/2017 às 01h até 09/09/2017 às 23h → Total de 183 dias e 22 horas (4.414 horas).
  - Não há dados entre 16/03/2017 às 06h até 29/03/2017 às 12h → 19 dias e 11 horas (467 horas).
  - Como o restante do dataset está completo, considerou-se apenas os dados a partir de 29/03/2017, resultando num dataset de 164 dias e 11 horas (3.947 horas).
- Muitas das variáveis online possuem momentos em que estão travadas ou em que há interpolação diretamente nos dados, como pode ser visto na vazão de ar da coluna 05.



Avaliação e limpeza dos dados

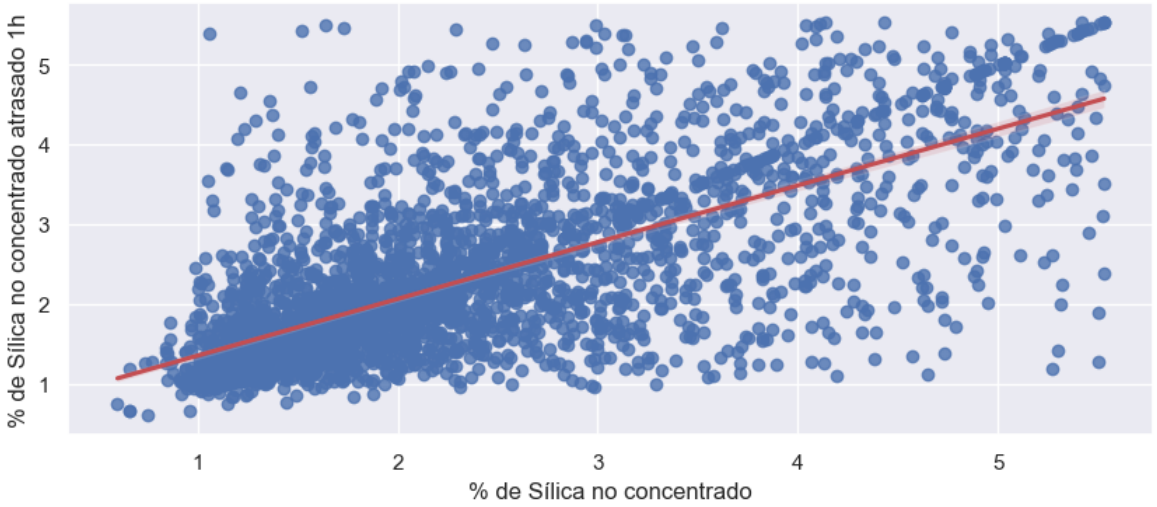
- Comportamento de interpolação também foi encontrado nas variáveis de laboratório.
- O tratamento das variáveis foi dividido entre as variáveis online e de laboratório:
  - Variáveis de laboratório:
    - Foram removidos (considerados como nulos) períodos em que as variáveis ficaram travadas por um tempo muito maior que a amostragem esperada (1-2h);
      - Pelo menos um turno completo (8h) para concentrado, dois (16h) para alimentação;
    - Também foram desconsiderados momentos em que as variáveis tinham seu valor modificado a cada ponto (20s), indicando interpolação;
    - Com essas considerações, restaram 89,9% e 87,8% dos pontos de ferro e sílica no concentrado, respectivamente, e 64,4% dos pontos do ferro e sílica na alimentação.
  - Variáveis online:
    - Foram removidos (considerados como nulos) períodos nos quais as variáveis ficavam travadas por mais de 1h, considerando a amostragem da variável de interesse;

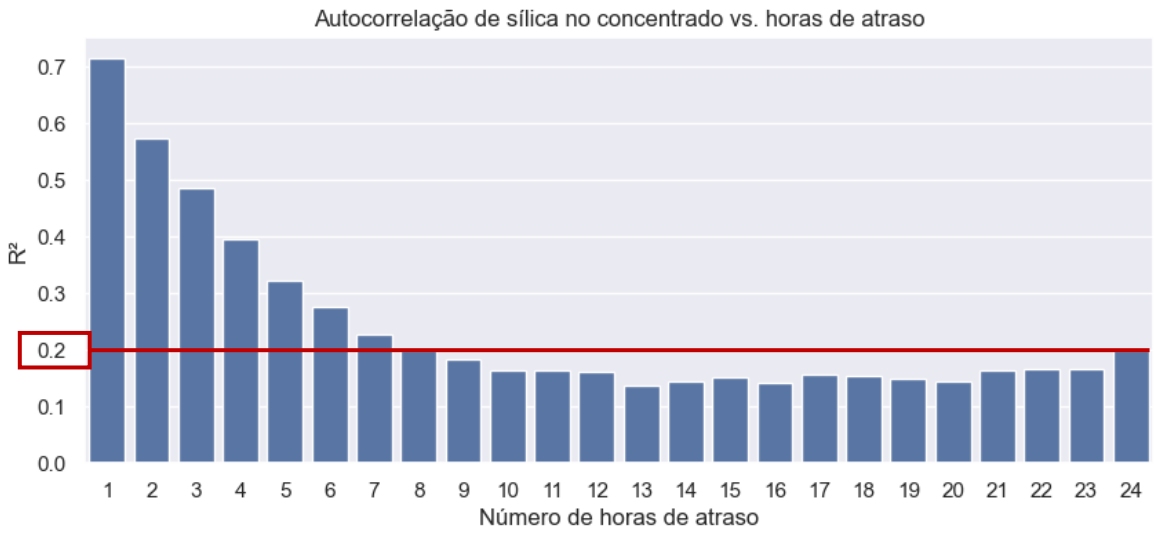
Avaliação e limpeza dos dados

    - Já em termos de interpolação, variações lineares inesperadas foram identificadas através da análise das derivadas (diferença entre os pontos), removendo períodos em que a segunda derivada era nula;
    - Na maioria das variáveis analisadas, restaram mais de 95% dos dados originais. Apenas a vazão de ar das colunas 04 e 05, que possuíam maior período travadas, ficaram com 64% dos dados originais após o filtro.
- Por fim, considerando que o % de sílica no concentrado, dado de interesse, é amostrado de maneira horária, todos os outros dados foram reamostrados utilizando as médias horárias, para permitir análises de correlação e treinamento dos modelos.
    - Pontos em que a variável de interesse havia sido filtrada foram desconsiderados, uma vez que não é possível treinar o modelo sem considerá-la;
    - Restaram, no total, 3.535 horas a serem analisadas.
- Os pontos em que as variáveis foram filtradas foram considerados como nulos (NaN), para permitir o trabalho com diferentes modelos.

Avaliação inicial de variáveis

- Após tratamento, foi realizada uma análise de correlação das variáveis presentes do dataset com a sílica do concentrado;
- Além do minério de ferro no concentrado, guardam correlações com índice acima de 0.1 a vazão de amina, níveis das colunas 5, 6 e 7, bem como fluxo de ar das colunas 1, 2, e 3, e o pH da polpa;
- Elevado índice de autocorrelação do % de sílica no concentrado, chegando a 0.77 na variável atrasada em 1h. Diversos valores de atraso mantém alta autocorrelação.
- Foram criadas variáveis atrasadas para todos os atrasos com índice maior que 0.2 (1h a 7h).





Treinamento e teste dos modelos iniciais

- Considerando que queremos prever o valor de uma variável pelo comportamento de outras, devemos utilizar algoritmos de regressão;
- Para teste, foram escolhidos 3 algoritmos, um mais simples, de regressão linear (linreg), e dois mais complexos, o Random Forests (RF) e o Histogram-based Gradient Boosting (HGB);
- Os dados horários foram separados em grupos de treinamento e teste.
  - 80% dos dados foram separados para treinamento (2828h), e o restante para teste (707h);
  - Uma vez que há relações temporais entre as variáveis, os dados foram separados mantendo sua ordem temporal, não utilizando algoritmos randomizados de separação.
- Para avaliar a performance dos modelos desenvolvidos, foram utilizados o erro quadrático médio (MSE) e o R^2.
- Todos os modelos foram testados na amostra de teste, contendo dados horários, e em seguida em toda a base dos dados brutos (a cada 20s), para avaliar a performance em ambos cenários.
- A variável de % de minério de ferro no concentrado foi desconsiderada como variável dependente, por ter comportamento proporcional e complementar à sílica.

Treinamento e teste dos modelos iniciais

- Para o teste inicial, foram criados cenários utilizando todas as variáveis presentes nos dados brutos e variando a utilização das variáveis temporais:
  - Cenário 1: apenas dados originais;
  - Cenário 2: dados originais + % de sílica no concentrado atrasado em 1h;
  - Cenário 3 : dados originais + % de sílica no concentrado atrasado de 1h até 7h (dados com índice de correlação > 0.2).
- Em todos os casos, a melhor performance foi obtida utilizando o cenário 3. Os melhores MSE e R^2 para dados de teste e dados brutos serão utilizados como baseline para aprimoramento.

Melhor Performance (MSE)

	Dados teste	Dados brutos
Cenário 1	RF (1,1549)	HGB (1,1170)
Cenário 2	HGB (0,5806)	HGB (0,5549)
Cenário 3	**linreg (0,5691)**	**HGB (0,4891)**

Melhor Performance (R^2)

	Dados teste	Dados brutos
Cenário 1	linreg (-0,2727)	HGB (0,1183)
Cenário 2	HGB (0,4858)	HGB (0,5620)
Cenário 3	**linreg (0,4960)**	**HGB (0,6140)**

Aprimoramento dos modelos

- Com o baseline em mente, e considerando a melhor performance do cenário 3, as variáveis desse cenário foram submetidas a um algoritmo de eliminação recursiva de features, que indica a melhor quantidade de features, bem como as melhores a serem utilizadas.
- Foram utilizadas 3 rodadas desse algoritmo:
    1. RFECV utilizando modelo de regressão linear (linreg): mesmo resultando objetivando otimizar tanto MSE quanto R^2 → RFECV linreg;
    2. RFECV utilizando Random Forests (RF) objetivando otimizar MSE → RFECV RF MSE;
    3. RFECV utilizando Random Forests (RF) objetivando otimizar MSE → RFECV RF R^2.
- Apesar de sua performance nos cenários iniciais, pelos resultados encontrados e otimização de tempo, o modelo HGB não foi submetido a esse algoritmo.

Aprimoramento dos modelos

- Os melhores resultados obtidos com as features selecionadas, tanto nos dados de teste quanto brutos, foram para o cenário RFECV linreg.
- Dessa forma, foi realizada uma nova rodada de eliminação recursiva de features utilizando as variáveis restantes, para os modelos de regressão linear e Random Forests.
    - Utilizando a regressão, não houve alteração nas variáveis;
    - Utilizando RF, houve a redução de uma variável.
- A seleção do modelo RF foi testada e obteve os melhores resultados nos dados brutos.

Baseline inicial

	Dados teste	Dados brutos
MSE	0,5691	0,4891
R^2	0,4960	0,6140

RFECV linreg

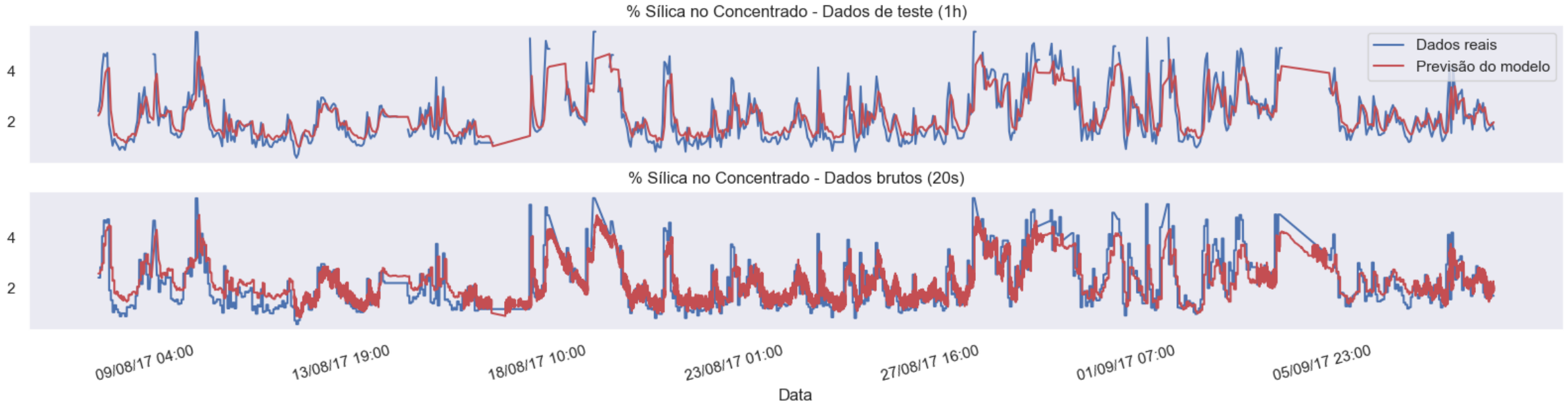
	Dados teste	Dados brutos
MSE	0,5197	0,4420
R^2	0,5398	0,6511
Modelo	linreg	RF

RFECV linreg – nova iteração

	Dados teste	Dados brutos
MSE	0,5213	0,3360
R^2	0,5384	0,7348
Modelo	linreg	RF

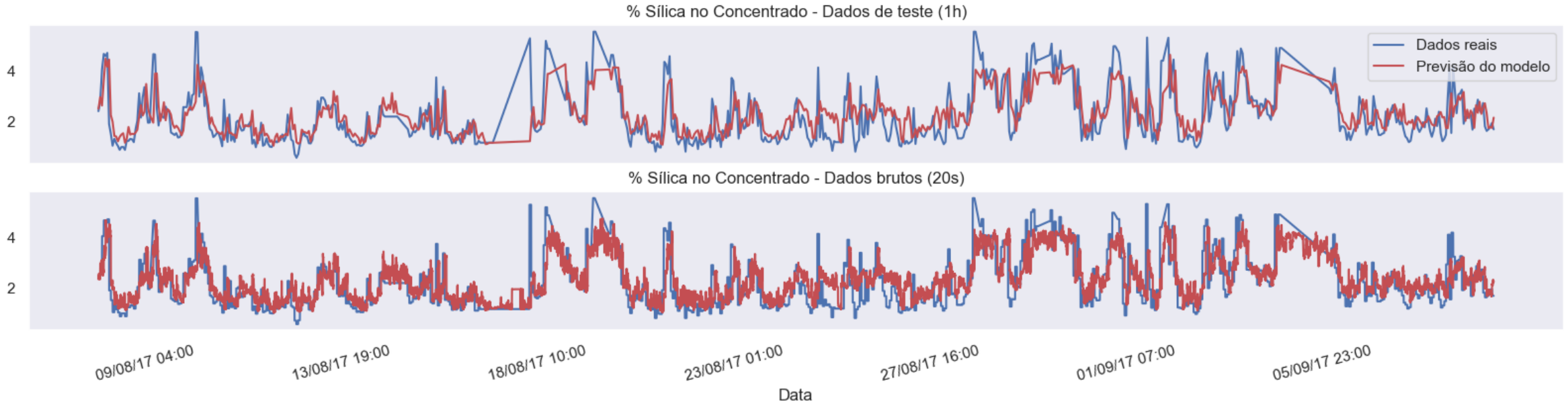
Melhor modelo – dados de teste

- A melhor performance nos dados horários de teste foi obtida pelo modelo de regressão linear, utilizando as features do cenário RFECV linreg: pH da polpa, densidade da polpa, fluxo de ar da coluna 04 e % de sílica no concentrado atrasada de 1 a 3h;
- MSE 0,5197 (22,3%) e R^2 0,5398 no teste - MSE 0,5271 (22,7%) e R^2 0,5839 nos dados brutos;
- Maior peso nas variáveis de sílica atrasada em 1h (58,4%), densidade da polpa (49,5%) e sílica atrasada em 2h (11,4%). Pesos menores que 10% nas outras variáveis.



Melhor modelo – dados brutos

- A melhor performance nos dados brutos foi obtida pelo modelo Random Forest, utilizando as features do cenário RFECV linreg – segunda iteração: pH da polpa, densidade da polpa, fluxo de ar da coluna 04 e % de sílica no concentrado atrasada de 1 a 3h;
- MSE 0,6072 (26,1%) e R^2 0,4622 no teste - MSE 0,3360 (14,4%) e R^2 0,7348 nos dados brutos;
- Maior importância para as variáveis de sílica do concentrado de 1 a 3h;
- Importância muito elevada para o valor do concentra atrasado 1h.



Conclusões e próximos passos

- Podemos notar que, mesmo com modelos simples e pouca otimização, os dados demonstram alto poder preditivo, considerando a mensuração da sílica no concentrado como objetivo;
- Ainda existem possibilidades de otimização dos modelos para redução do erro e aumento do índice de correlação:
  - Otimização de hiperparâmetros;
  - Combinação de modelos, de forma a distribuir melhor o peso/importância das variáveis.
- O modelo desenvolvido pode ser implementado online, trazendo diversos benefícios:
  - Maior assertividade no controle manual do operador, que agora consegue prever o valor atual da variável de processo;
  - Possibilidade de implementação de uma malha de controle utilizando a previsão do % de sílica no concentrado para manipular automaticamente a dosagem de amina.