



CENTRO
UNIVERSITÁRIO

Machine Learning

Data Science

Arthur Brietzke, Gabriel Miranda e Lucas Valim

Objetivo do projeto

Este projeto tem como objetivo desenvolver um pipeline completo de machine learning para prever o desempenho acadêmico de estudantes com base em variáveis comportamentais, socioeconômicas e contextuais. Ao longo do processo, foram aplicadas técnicas de pré-processamento, modelagem, tuning de hiperparâmetros e avaliação visual e estatística dos resultados.

Análise dos Dados

Etapa essencial para entender o comportamento dos dados

Modelos Treinados

Regressão Linear

Treinada com dados pré-processados.

Métricas obtidas:

- $R^2 \approx 0.746$
- $MAE \approx 0.656$
- $RMSE \approx 1.894$

Random Forest (padrão)

Métricas:

- $R^2 \approx 0.672$
- $MAE \approx 1.073$
- $RMSE \approx 2.154$

Random Forest (GridSearchCV)

Busca pelos hiperparâmetros:

- `n_estimators`
- `max_depth`
- `min_samples_split`
- `min_samples_leaf`

Resultados do teste:

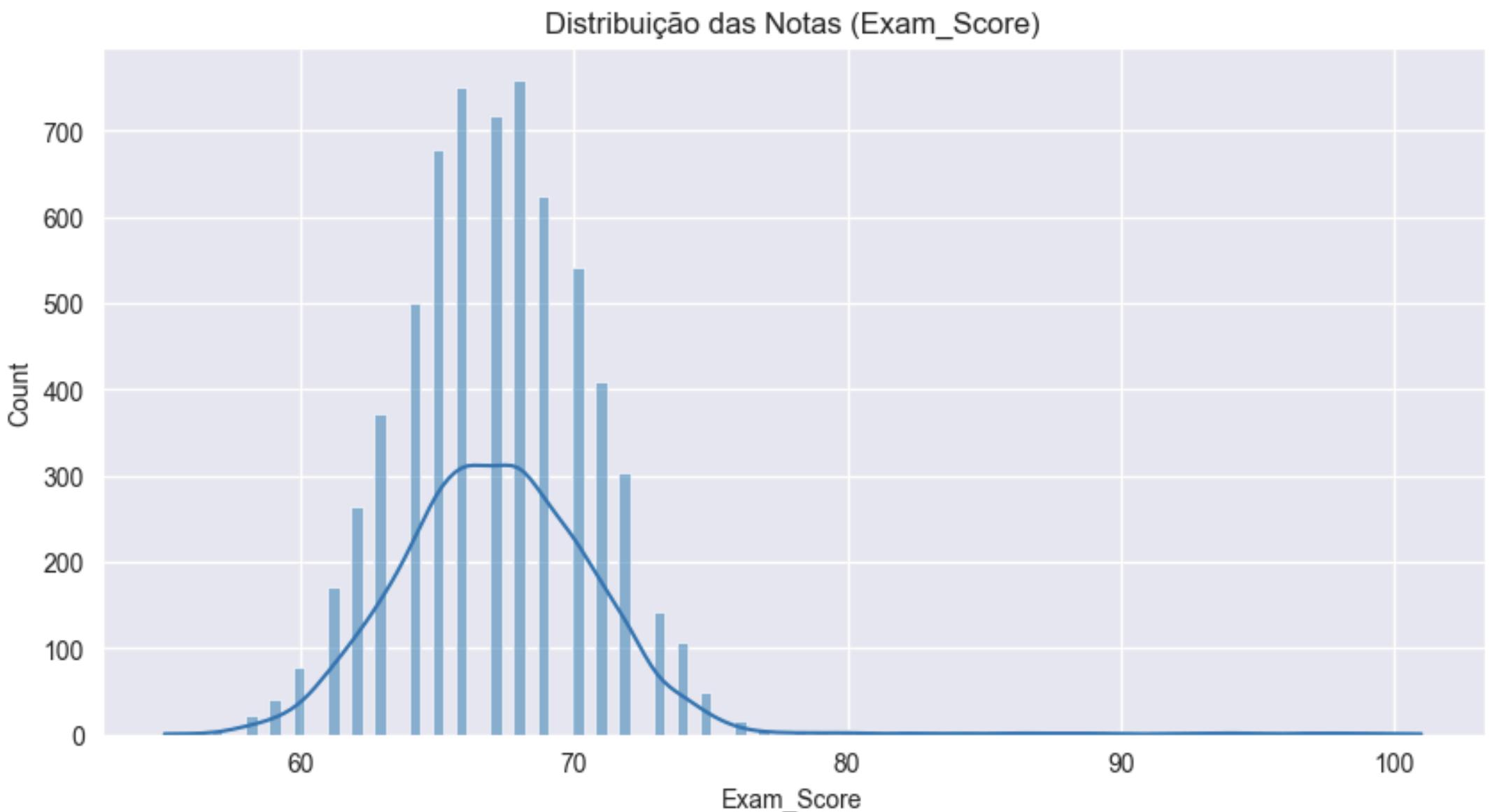
- $R^2 \approx 0.687$
- $MAE \approx 1.05$
- $RMSE \approx 2.10$

Distribuição das Notas

O gráfico ao lado mostra a distribuição das notas, permitindo avaliar se há concentração, outliers ou padrões incomuns. Isso é essencial para entender o comportamento do alvo antes de aplicar modelos de machine learning.

A distribuição do *Exam_Score* está assimétrica, inclinada para a esquerda, isso significa:

- Maioria dos estudantes tem notas mais altas
 - Os valores se concentram entre 65 e 72, indicando desempenho relativamente bom na maior parte dos casos.
- Poucos estudantes tiraram notas muito baixas
 - A cauda esquerda representa uma minoria que tirou entre 55 e 62, puxando a média para baixo.
- Indica bom desempenho geral da turma
 - Distribuições left-skew normalmente aparecem quando:
 - Existe preparação adequada;
 - Há tutorial reforçado;
 - O exame não é excessivamente difícil;
 - A turma é homogênea em desempenho;



Feature's com mais importância

Os resultados do modelo de Random Forest mostram que a presença (Attendance) e as horas de estudo (Hours_Studied) são os maiores determinantes da nota prevista.

Pontuações anteriores e sessões de tutoria aparecem em seguida com influência moderada, enquanto fatores como acesso a recursos e envolvimento dos pais contribuem menos.

O padrão geral reforça a predominância de fatores comportamentais e de desempenho passado na previsão da nota.

In [9]:

```
importances = best_rf.feature_importances_
feature_names = preprocessor.get_feature_names_out()

rf_importances = (
    pd.DataFrame({
        'feature': feature_names,
        'importance': importances
    })
    .sort_values('importance', ascending=False)
)

rf_importances.head(15)
```

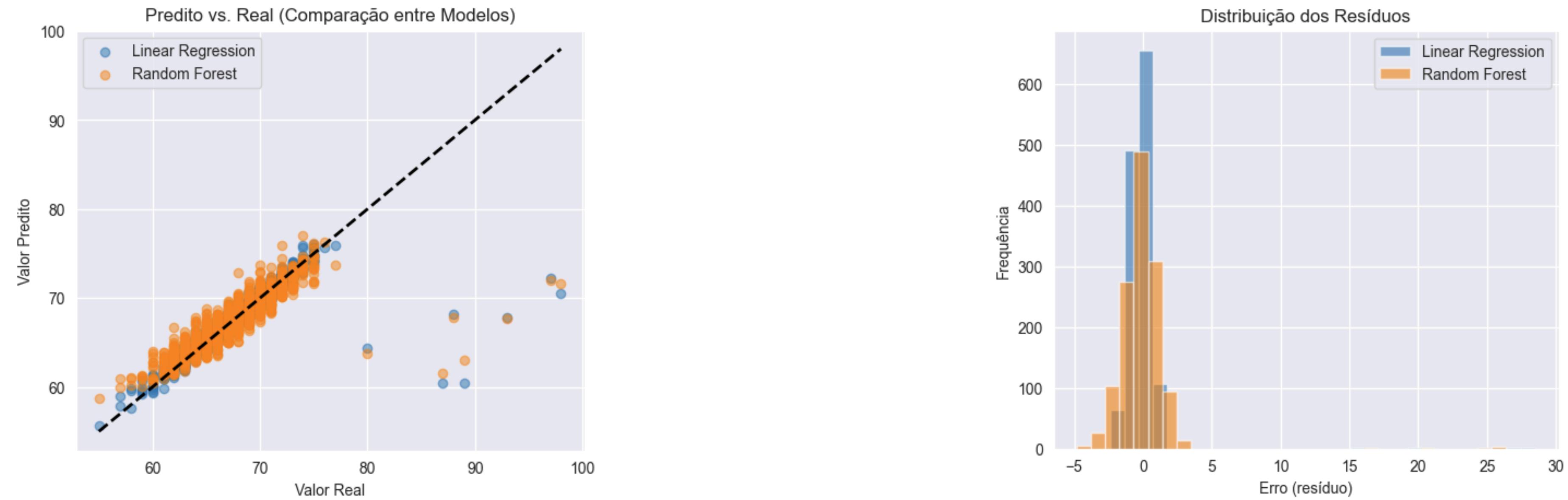
Out[9]:

	feature	importance
1	num_Attendance	0.445067
0	num_Hours_Studied	0.266552
4	num_Previous_Scores	0.074134
6	num_Tutoring_Sessions	0.031042
21	cat_Access_to_Resources_0	0.022197
24	cat_Parental_Involvement_0	0.021762
3	num_Sleep_Hours	0.014582
10	num_Physical_Activity	0.013150
9	num_Peer_Influence	0.012196
7	num_Family_Income	0.011797
25	cat_Parental_Involvement_1	0.010973
22	cat_Access_to_Resources_1	0.010387
5	num_Motivation_Level	0.009528
11	num_Distance_from_Home	0.009474
8	num_Teacher_Quality	0.008340

Aplicação do Machine Learning

Etapa onde os modelos após treinados retornam os resultados

Avaliação Visual entre Modelos de Machine Learning



Preditivo vs Real (Comparação entre modelos)

Comparando:

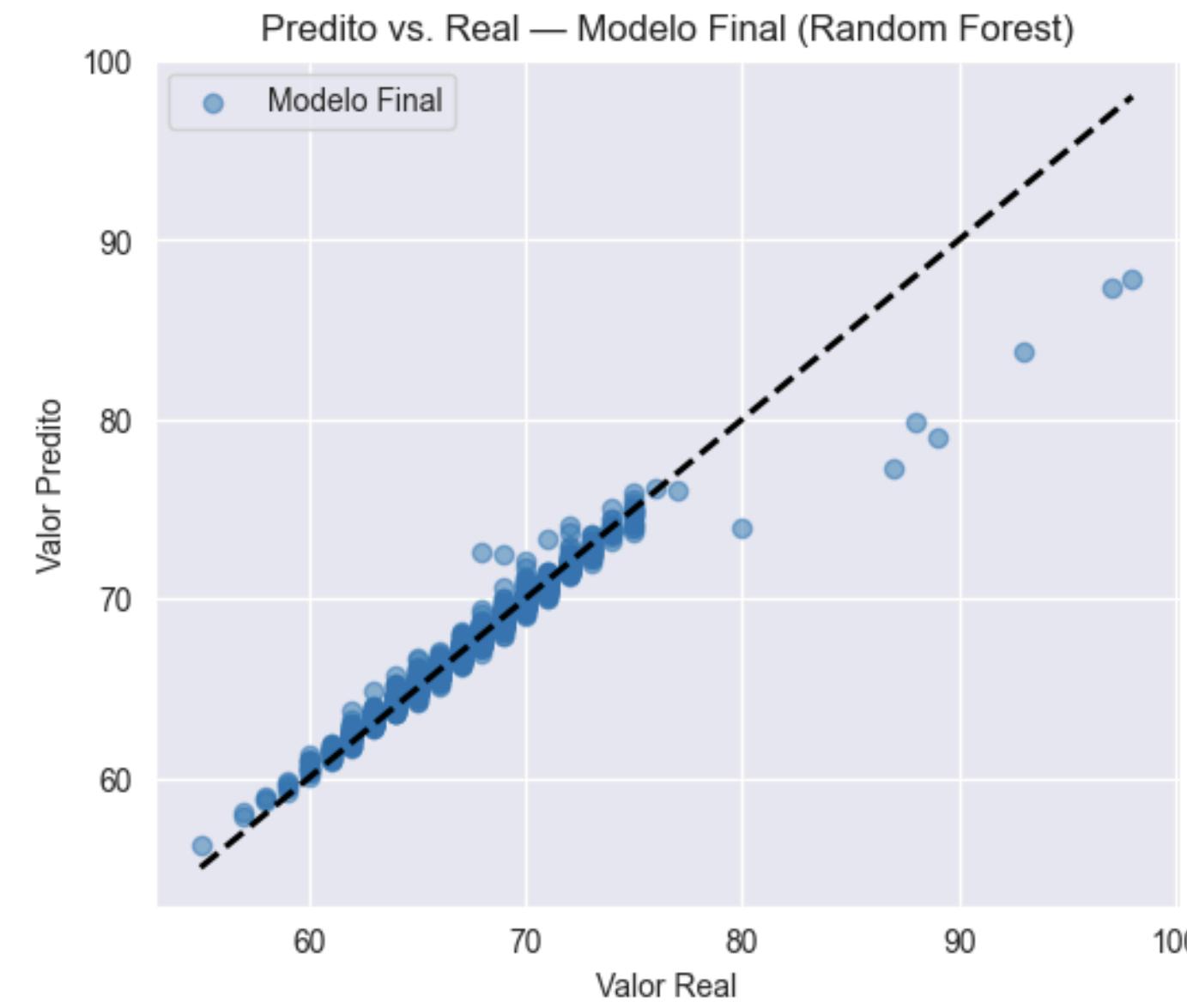
- Regressão Linear
- Random Forest

Distribuição de Resíduos

Para ambos os modelos.

As análises mostraram que a Regressão Linear apresentou resíduos mais concentrados e previsões mais próximas da linha ideal.

Avaliação Visual entre Modelos de Machine Learning



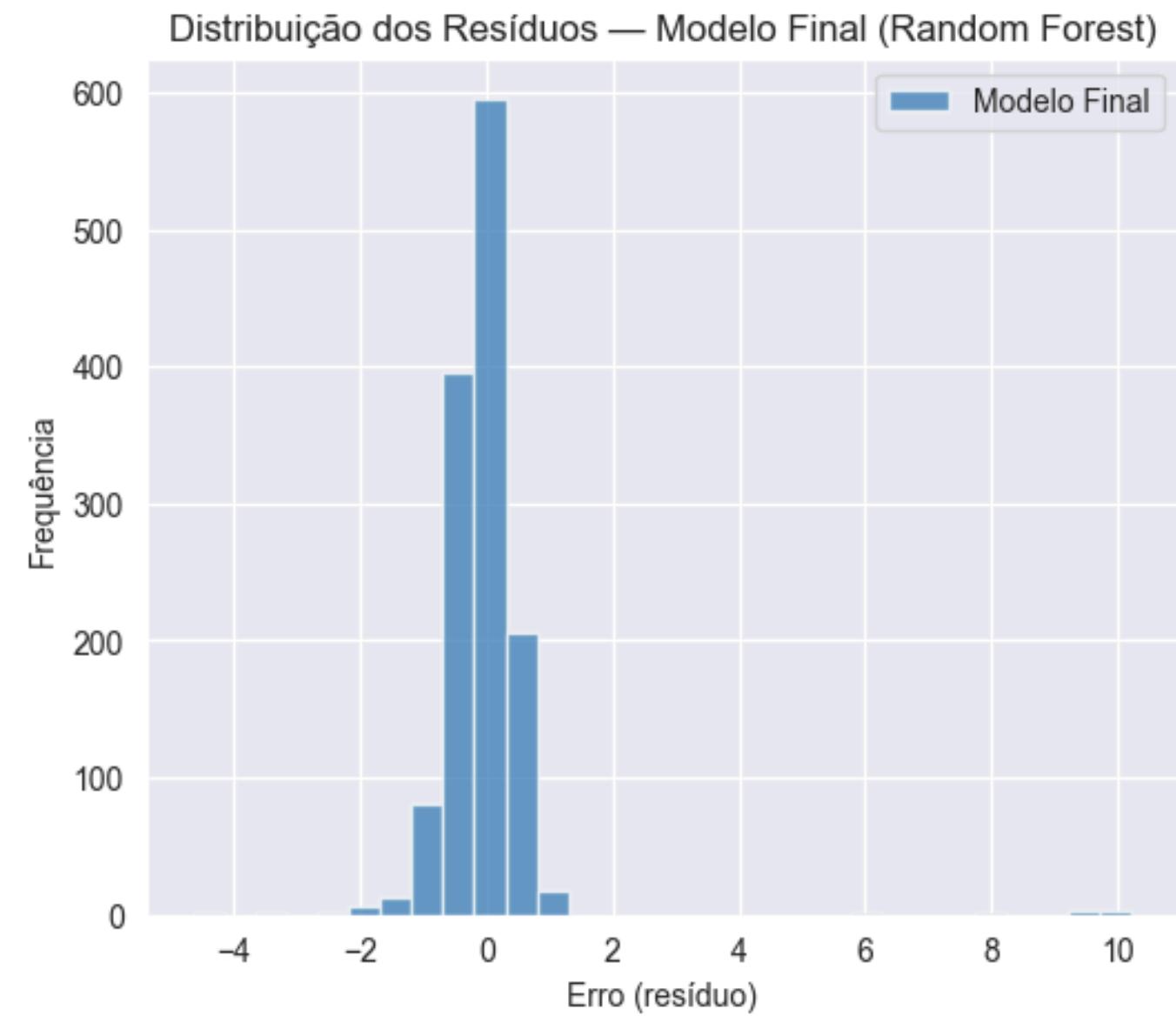
Predito vs Real - Modelo Final (Random Forest)

O modelo foi treinado com todos os dados disponíveis.

Métricas finais:

- $R^2 \approx 0.953$
- $RMSE \approx 0.816$

Observação: valores muito altos indicam possível overfitting ou vazamento de dados.



Distribuição de Resíduos - Modelo Final (Random Forest)

Dúvidas?

