Guia para Elaboração da Etapa de Modelagem com Random Forest

Nesta etapa, você deverá aplicar o algoritmo Random Forest Regressor para construir um modelo capaz de prever os níveis do separador trifásico a partir dos dados tratados e preparados na fase de pré-processamento.

O objetivo é treinar o modelo, otimizar seus hiperparâmetros, avaliar seu desempenho e interpretar os resultados obtidos.

Passo 1: Configuração Inicial do Experimento

Antes de treinar o modelo:

a) Crie um novo notebook e nomeie-o como 03-RandomForest.

b) Carregue os dados já pré-processados (divididos em treino, validação e teste, com scaling aplicado).

c) Defina as métricas que serão usadas para avaliação. Para regressão, recomenda-se:

- RMSE (Root Mean Squared Error): erro médio quadrático em unidades do nível (cm).
- MAE (Mean Absolute Error): erro médio absoluto.
- R^2 (Coeficiente de Determinação): proporção da variância explicada.

Passo 2: Treinamento do Modelo Baseline

a) Instancie um modelo de RandomForestRegressor com parâmetros padrão.

b) Treine-o apenas com os dados de treino.

c) Avalie seu desempenho em treino e validação.

d) Registre as métricas como referência inicial (baseline).

Passo 3: Otimização de Hiperparâmetros

Para melhorar o desempenho do modelo, explore diferentes configurações. Os principais hiperparâmetros são:

- n_estimators: número de árvores (100, 200, 500).
- max_depth: profundidade máxima das árvores (ex.: None, 5, 10, 20).
- min_samples_split: número mínimo de amostras para dividir um nó.
- min_samples_leaf: número mínimo de amostras em cada folha.

- max_features: número de variáveis consideradas a cada divisão (ex.: auto, sqrt, log2).

a) Use técnicas de busca sistemática, como GridSearchCV ou RandomizedSearchCV, com validação cruzada (ex.: k=5), ou ainda o Optuna.

b) Registre as combinações testadas e as métricas de validação correspondentes.

c) Escolha o melhor conjunto de hiperparâmetros.

Passo 4: Avaliação Final do Modelo

a) Treine novamente o Random Forest com os hiperparâmetros otimizados, usando treino + validação.

b) Avalie no conjunto de teste, **que deve ser usado apenas agora**.

c) Reporte os resultados finais (RMSE, MAE, R^2) e compare com o baseline.

Passo 5: Análise de Importância das Variáveis

O Random Forest permite extrair a importância relativa das features.

a) Gere o gráfico de importâncias das variáveis.

b) Interprete: quais sensores são mais relevantes para prever os níveis?

c) Discuta se o resultado faz sentido em relação ao conhecimento físico do sistema.

Passo 6: Discussão dos Resultados

No relatório, discuta:

O modelo otimizou significativamente em relação ao baseline?

Houve sinais de overfitting (treino >> validação/teste)?

As variáveis mais importantes coincidem com a intuição sobre o processo físico?

Sugestões de melhorias: aumentar a base de dados, testar outros algoritmos (ex.: Gradient Boosting, XGBoost), aplicar técnicas de regularização.

Boas Práticas

Fixe uma semente aleatória (random_state) para garantir reprodutibilidade.

Documente as configurações testadas, mesmo as que não deram bons resultados.

Evite usar o conjunto de teste antes da avaliação final.

Salve o modelo final em arquivo (modelo_rf.pkl) para uso posterior.

Boa sorte!!!