Natan Victor de Deus Silva
email: natan.silva@ctec.ufal.br

Variáveis de entrada (features):

1. **Aeration rate (Fg:L/h):** Taxa de aeração em litros por hora. Controla a quantidade de oxigênio fornecida ao sistema.
2. **Agitator RPM (RPM:RPM):** Velocidade de agitação. Influencia a mistura e homogeneidade do reator.
3. **Sugar feed rate (Fs:L/h):** Taxa de alimentação de açúcar. Fonte de carbono para o crescimento da biomassa.
4. **Acid flow rate (Fa:L/h):** Taxa de fluxo de ácido. Ajuda a controlar o pH do sistema.
5. **Base flow rate (Fb:L/h):** Taxa de fluxo de base. Também usada para controle de pH.
6. **Heating/cooling water flow rate (Fc:L/h):** Taxa de fluxo de água de aquecimento/resfriamento, para manter a temperatura ideal.
7. **Heating water flow rate (Fh:L/h):** Taxa de fluxo de água de aquecimento.
8. **Water for injection/dilution (Fw:L/h):** Taxa de fluxo de água para injeção ou diluição.
9. **Air head pressure (pressure:bar):** Pressão do ar no sistema.
10. **Dumped broth flow (Fremoved:L/h):** Taxa de remoção de caldo residual.
11. **Substrate concentration (S:g/L):** Concentração de substrato (açúcares ou nutrientes).
12. **Dissolved oxygen concentration (DO2:mg/L):** Concentração de oxigênio dissolvido, fundamental para a respiração celular.
13. **Vessel Volume (V:L):** Volume do reator.
14. **Vessel Weight (Wt:Kg):** Peso do reator.
15. **pH (pH:pH):** pH do sistema, essencial para a produção.
16. **Temperature (T:K):** Temperatura do processo em Kelvin.
17. **Generated heat (Q:kJ):** Calor gerado durante a reação.
18. **Carbon dioxide percent in off-gas (CO2outgas:%):** Percentual de CO2 no gás residual.
19. **PAA flow (Fpaa:PAA flow (L/h)):** Fluxo de PAA (ácido peracético).
20. **PAA concentration offline (PAA_offline:PAA (g L^{-1})):** Concentração de PAA medida offline.
21. **Oil flow (Foil:L/hr):** Taxa de fluxo de óleo.
22. **NH3 concentration offline (NH3_offline:NH3 (g L^{-1})):** Concentração de NH3 medida offline.
23. **Oxygen Uptake Rate (OUR:(g min^{-1})):** Taxa de consumo de oxigênio.
24. **Oxygen in percent in off-gas (O2:O2 (%)):** Porcentagem de oxigênio no gás residual.

25. **Ammonia shots (NH3_shots:kgs):** Quantidade de NH3 adicionada em "tiros".
26. **Viscosity (Viscosity_offline:centPoise):** Viscosidade do caldo, medida offline.

Variáveis de saída possíveis (targets):

1. **Penicillin concentration (P:g/L):** Concentração de penicilina no sistema, medida online.
2. **Offline Penicillin concentration (P_offline:P(g L^{-1})):** Concentração de penicilina medida offline, alvo primário de interesse.
3. **Offline Biomass concentration (X_offline:X(g L^{-1})):** Concentração de biomassa, que reflete o crescimento celular no processo.

Este projeto utiliza um conjunto de dados gerados pela simulação matemática avançada IndPenSim, que representa um sistema de fermentação de penicilina de 100.000 litros. IndPenSim é a primeira simulação a integrar um dispositivo de espectroscopia Raman realista, possibilitando o desenvolvimento e a avaliação de soluções de controle inovadoras para instalações de biotecnologia. O conjunto de dados contém informações de 100 lotes controlados por diferentes estratégias e inclui medições de processo e espectroscopia Raman, totalizando aproximadamente 2,5 GB. Este rico conjunto de dados é ideal para a aplicação de técnicas de aprendizado de máquina (ML) e inteligência artificial (IA), promovendo avanços na indústria biofarmacêutica.

Projeto de Machine Learning para Predição da Concentração de Penicilina (P: g/L)

Introdução

A produção de penicilina é um processo biotecnológico complexo que requer monitoramento constante de parâmetros para garantir eficiência e qualidade. Este projeto utiliza dados simulados do sistema IndPenSim, que simula uma planta industrial de 100.000 litros, para construir um modelo preditivo da concentração de penicilina ao longo do processo de fermentação. A meta é prever a variável-alvo **Penicillin concentration (P: g/L)**, otimizar o controle do processo e aumentar o rendimento.

Metodologia

1. Leitura e Carregamento dos Dados

Os dados foram carregados e inspecionados para garantir consistência. A base inclui múltiplas variáveis operacionais (e.g., taxa de aeração, concentração de oxigênio dissolvido) e resultados offline, totalizando cerca de 100 batches simulados.

2. Preparação dos Dados

- **Limpeza de Dados**: Tratamos valores ausentes em variáveis como "PAA_offline" e "NH3_offline".
- **Padronização**: Todas as variáveis foram normalizadas para melhorar a performance dos algoritmos de regressão.
- **Aleatoriedade:** A amostra do dataset utilizado foi aleatorizado para promover uma melhor performance dos dados treinados

3. Análise Exploratória de Dados (EDA)

- Identificamos as variáveis com maior correlação com a concentração de penicilina, incluindo "dissolved oxygen concentration" e "sugar feed rate".
- Visualizações como heatmaps e gráficos de dispersão mostraram relações não lineares entre variáveis.

4. Seleção de Features

- Utilizamos métodos baseados em correlação e árvores de decisão para reduzir o conjunto de variáveis.
- A seleção final incluiu variáveis que explicavam a maior parte da variabilidade na produção de penicilina.

5. Divisão dos Dados

Os dados foram divididos em conjuntos de treino (80%) e teste (20%), mantendo a distribuição da variável-alvo.

6. Treinamento dos Modelos de Regressão

Testamos os seguintes algoritmos:

- **Ridge Regression**: $R^2 = 0.8272$
- **Lasso Regression**: $R^2 = 0.4743$
- **ElasticNet**: $R^2 = 0.1796$
- **Gradient Boosting**: $R^2 = 0.9281$
- **Random Forest**: $R^2 = 0.9452$

7. Validação Cruzada e Ajuste Fino

- O algoritmo escolhido foi o **Gradient Boosting**, por equilibrar alta acurácia e capacidade de generalização.
- Resultados de validação cruzada: $R^2 = 0.9452$.
- Parâmetros ajustados:
    - **Learning Rate**: 0.1
    - **n_estimators**: 30
    - **Subsample**: 0.9
    - **Max Depth**: 5

8. Testes e Avaliação

- **Conjunto de Teste e treino**:
    - R^2: 0.9394 teste
    - R^2 Ajustado: treino
- **Conjunto de dados completo**:
    - R^2: 0.9781
    - MAE: 1.1043
    - MSE: 2.3091

Conclusão

O modelo baseado em Gradient Boosting apresentou alta capacidade preditiva para a concentração de penicilina, com R^2 próximo de 0.978. Os resultados mostram que o modelo pode ser usado para monitorar e otimizar a produção em tempo real, reduzindo desperdícios e aumentando o rendimento. Essa abordagem pode ser ampliada para prever outros parâmetros críticos na planta industrial, garantindo controle eficiente e alta produtividade.