MODELAGEM - DBO/DQO DE UMA LAGOA AERADA EM UMA FÁBRICA DE CELULOSE E PAPEL

Renê Gibran e Silva Nery[1]

Thiago Franklin da Mata Souza Silva[2]

Resumo

Uma fábrica de papel e celulose precisa controlar a Demanda Biológica de Oxigênio (DBO) de uma lagoa aerada, entretanto, a resposta da análise laboratorial sobre as amostras coletadas nesse efluente são extremamente lentas, cerca de 5 dias, e impossibilita ações corretivas de controle eficazes. A alternativa é utilizar métodos de Machine Learning e formular modelos estatísticos para uma predição precisa da DBO. Assim, é possível utilizar outros parâmetros para a obtenção do resultado desejado. Os modelos escolhidos foram de regressão linear e a regressão lasso. A linguagem de programação Python foi a ferramenta utilizada para a implementação dos processos.

Palavras-Chave: Regressão Linear, Lasso, Demanda Biológica de Oxigênio, Python, Machine Learning.

Introdução

Em resposta ao aumento desenfreado da temperatura global, alto índice de poluição, queimadas e etc, as lideranças mundiais estabeleceram a formulação de leis e restrições ambientais como absolutamente necessárias para a qualidade de vida das futuras gerações. O que forçará as indústrias e empresas contemporâneas a produzir de maneira sustentável. Dessa forma, ao buscar a rentabilidade, a redução de custos só é possível com o avanço científico e tecnológico.

A forma mais efetiva de aferir os níveis de poluição de um efluente é através da Demanda Bioquímica de Oxigênio (DBO), que representa a quantidade de oxigênio consumida por bactérias (aeróbias). Essas, realizam a degradação da matéria orgânica no ambiente (rios, esgotos e etc) através de reações oxidativas, como a respiração. Desse modo, quanto maior a DBO, maior a quantidade de microrganismos degradando a matéria orgânica despejada (poluição), o que pode causar um desequilíbrio nos níveis de oxigênio e causar a morte da vida aquática. Um baixo nível de DBO indica um efluente limpo com baixo índice de poluição.

Existem diversos fatores que podem influenciar a DBO, entre eles: demanda química de oxigênio (DQO), pH, sólidos em suspensão, nitrogênio nitrato, nitrogênio amoniacal,

[1]Universidade Federal da Bahia (UFBA), renenery@ufba.br

[2]Universidade Federal da Bahia (UFBA), thiago.franklin.s.silva@gmail.com

fósforo, cor, temperatura, condutividade, vazão do efluente do sistema de lagoas aeradas. Entretanto, o tempo entre coleta de dados no efluente e análise laboratorial dos resultados é lenta, podendo levar de três a cinco dias para a conclusão.

Um modo mais eficiente de conseguir a DBO é obtido empiricamente, com a coleta de dados suficiente, é possível uma estimação através de modelos estatísticos lineares ou não-lineares, que melhor se ajustam ao problema para prever os valores desejados. Com a predição é possível ter informações suficientes para saber quando um determinado parâmetro irá afetar a DBO, permitindo uma ação corretiva. A maior eficiência ocorre quando é possível definir os parâmetros mais importantes, economizando na coleta de dados que pouco interferem e priorizando os mais relevantes.

Objetivos

Com o objetivo de tratar melhor os efluentes, a indústria começou a observar alguns parâmetros diariamente, no entanto, por conta da demora na análise da demanda biológica de oxigênio, decisões para melhor controle desse afluente são dificultadas. É necessário propor um mecanismo de previsão da demanda biológica de oxigênio (DBO) a partir de parâmetros relacionados, para um efluente.

Materiais e Métodos

Para o desenvolvimento do projeto, foi utilizado o ambiente virtual Google Colaboratory junto com a linguagem de programação Python, as bibliotecas Pandas, Matplotlib.pyplot, Seaborn, Numpy, sklearn.linear_model, sklearn.model_selection, sklearn.metrics e a base de dados com os parâmetros do sistema de lagoas aeradas fornecida pela professora Karla Patricia Oliveira Esquerre.

A indústria de papel e celulose estava com dificuldade em tomar medidas de controle da qualidade do efluente tratado, visto que o tempo da análise da demanda biológica de oxigênio (DBO) é de cinco dias. Para tentar solucionar esse problema, um algoritmo de machine learning foi implementado.

Os dados foram importados da base de dados do github utilizando o Python no ambiente virtual Google Colaboratory. Foi necessário um tratamento dos dados para que fosse possível utilizar o dataframe. As datas, por exemplo, estavam como objeto e não no formato adequado. De maneira semelhante, alguns parâmetros numéricos também foram importados como objeto. Após análise, foi constatado que os números foram escritos com vírgula e não estavam sendo considerados como números por isso. Para resolver o problema, todos os parâmetros tiveram vírgula substituída por ponto.

Para a filtragem dos dados, foram feitos gráficos de dispersão e boxplots com o intuito de observar o comportamento dos parâmetros e a presença de possíveis outliers. A partir da observação e da literatura recomendada (OLIVEIRA-ESQUERRE, 2004) somente um valor de outlier foi removido do DataFrame (pHin = 0,85). Ele foi removido por ser um valor muito

discrepante dos outros e por ter sido o único com mudanças drásticas naquele dia, indicando erro de medida ou de digitação. Apesar da existência de outros outliers, esses não foram removidos, já que podem auxiliar na construção de um modelo menos “engessado”, ou seja, com um viés um pouco mais baixo.

Gráfico 1: BoxPlot(pHin) com outlier

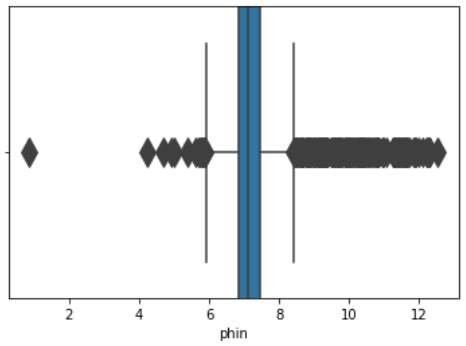
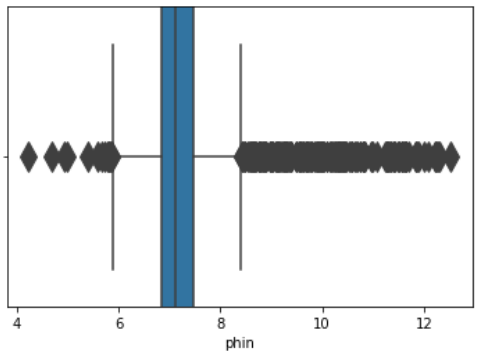
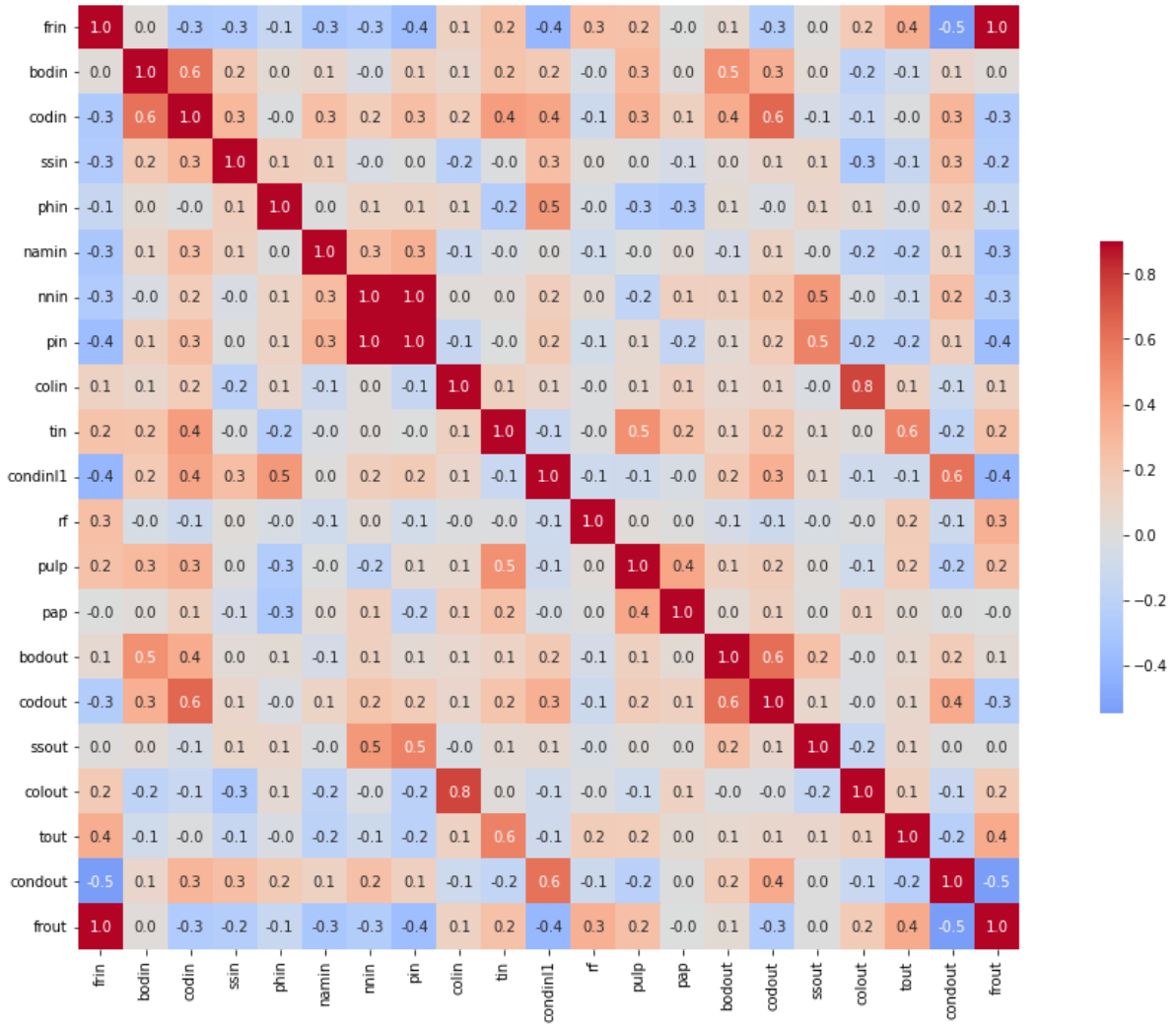


Gráfico 2: BoxPlot(pHin) sem outlier



Após o tratamento inicial, foi feita uma análise de correlação entre as variáveis com o objetivo de selecionar os melhores parâmetros para o modelo. Além disso, foi observada a quantidade de dados faltantes, já que não poderiam ser utilizados para a construção do programa.

Gráfico 3: Mapa de calor correlações das variáveis



Os parâmetros relevantes para o modelo aplicado foram aqueles que apresentaram uma correlação acima de 0.1 e menos de 7% de dados faltantes.

Tabela 1: Variáveis independentes Multi Regressão Linear

Variáveis	codout	bodin	codin	condinl1	ssout
Correlação	0.6	0.5	0.4	0.2	0.2
Dados faltantes	5%	6%	6%	3%	86%
Relevante	Sim	Sim	Sim	Sim	Não

Apesar da quantidade de dados faltantes dos parâmetros escolhidos ser baixa quando comparada à quantidade em "SSout", ainda não é possível implementar um modelo. Como a porcentagem de dados faltantes nessas variáveis é menor (mas não insignificante), optou-se por utilizar a interpolação para preencher os valores vazios.

Por fim, após o tratamento dos dados e da seleção das variáveis, foi implementado o modelo de multi regressão linear utilizando as bibliotecas sklearn.linear_model, sklearn.model_selection e sklearn.metrics. Também foi feito um modelo utilizando a regressão Lasso (com um parâmetro a mais que a multi regressão linear) e as mesmas bibliotecas.

Tabela 2: Variáveis independentes Regressão Lasso

Variáveis	codout	bodin	codin	condinl1	pulp
Correlação	0.6	0.5	0.4	0.2	0.1
Dados faltantes	5%	6%	6%	3%	7%
Relevante	Sim	Sim	Sim	Sim	Sim

Resultados e Discussão

A partir da implementação dos modelos de regressão linear múltipla e regressão Lasso, foi possível comparar os resultados obtidos com ambos os modelos. Para os dois casos, o programa foi rodado cem vezes com o propósito de chegar a um valor de R^2 mais próximo da realidade, já que os treinos e testes foram feitos com diferentes grupos de dados.

Tabela 3: Coeficientes da Multi Regressão Linear

Variáveis	codout	bodin	codin	condinl1	INTERCEPT
Coeficiente	0.25916671	0.25923306	-0.09951894	0.00303246	-9.4141521

Tabela 4: Valores dos erros da Multi Regressão Linear

R^2 único	R^2 médio (100x)	MAE	MSE
0.518443393887664	0.524329213423037	13.53966002402742	323.8849685421776

Tabela 5: Coeficientes da Regressão Lasso

Variáveis	codout	bodin	codin	condinl1	pulp	INTERCEPT
Coeficiente	0.2459417	0.2602485	-0.097077	0.0018881	0.0060926	-10.1333788

Tabela 6: Valores dos erros para Regressão Lasso

R^2 único	R^2 médio (100x)	MAE	MSE

0.515233482309972	0.535036602846642	26.07349210918531	1118.909719538615

O MAE é a distância entre o valor médio das predições e o valor real, enquanto o MSE avalia a proximidade da linha de regressão do conjunto de dados. O valor de R^2 significa a proporção de variância na variável dependente explicada pela variável independente. O ajuste nos modelos foi de aproximadamente 52%, ou seja, é possível perceber que o modelo não se ajustou tão bem quanto esperado. Isso se deu, pela baixa correlação entre as variáveis utilizadas para predição e a BOD, pelo tamanho do banco de dados, pela quantidade de dados faltantes, modelagem insuficiente e pela incerteza com relação às medidas feitas. A partir da análise das tabelas de erros dos modelos aplicados, utilizando praticamente as mesmas variáveis, é possível perceber que regressão lasso possui um melhor desempenho no R^2 médio de quase 1%.

O programa foi construído considerando que as medidas foram feitas de maneira semelhante durante todo o período do banco de dados. Medidas feitas de maneira muito diferente ou até errada podem acarretar na variação da capacidade de ajuste do modelo.É possível perceber que o valor de R^2 nos dois modelos foi próximo. Os gráficos comparativos dos valores reais(azul) com os valores preditos(vermelho) estão dispostos abaixo.

Gráfico 4: Multi Regressão Linear - Predito x Real

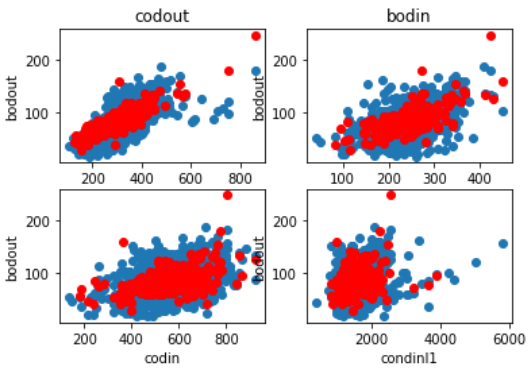
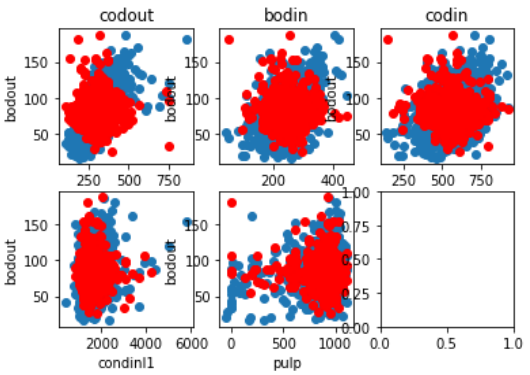


Gráfico 5: Regressão Lasso - Predito x Real



Conclusão

A partir do banco de dados disponível, foi possível criar um modelo de predição da DBO com base em parâmetros correlacionados a ela. A análise exploratória, de correlações, distribuições e verificações de outliers trouxeram informações relevantes para o problema, já que permitiram tratar os dados de maneira mais eficiente, além de permitir uma melhor e mais adequada escolha dos parâmetros a serem utilizados. A informação temporal foi útil para observar o comportamento das variáveis com o passar do tempo. Foi possível

perceber variação da vazão do efluente com o passar do tempo, por exemplo. Em outras situações, seria bastante útil para observar variações sazonais. Sem essas avaliações, a implementação do modelo se tornaria muito mais difícil e propensa a muito mais erros.

Seria possível criar outros códigos para prever outros parâmetros. Nesse caso, os parâmetros selecionados para serem preditos seriam os que têm análises mais demoradas, mais caras ou até mesmo as que requerem grande quantidade de material ou de pessoas para serem feitas.

É possível também propor diversos modelos para agrupamento de variáveis. Os modelos podem variar em precisão, complexidade e aplicabilidade, dependendo do que o formulador do modelo achar mais efetivo para a predição da variável estudada. Os dados faltantes foram interpolados de forma simples para gerar uma estimativa do valor real, mas existem vários tipos de interpolações que poderiam ser mais efetivas, como a spline e a cúbica. Entretanto, a melhor solução seria uma real amostragem sem dados faltantes, o que permitiria a predição do modelo sobre dados reais sobre o problema, diminuindo a quantidade de resíduos.

É importante salientar que talvez seja possível propor outro modelo, e que dados de qualidade são fundamentais para o sucesso do programa. A utilização de métodos de Machine Learning é uma ótima solução para a predição da DBO. O aperfeiçoamento dos modelos com técnicas mais complexas de predição ainda é possível. Assim, uma melhoria significativa do desempenho é possível e deve ser explorada futuramente.

Referências

Karla Patricia Oliveira-Esquerre , Dale E. Seborg, Roy E. Bruns, Milton Mori. Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Part I. Linear approaches. Chemical Engineering Journal 104 (2004) 73–81.

Karla Patricia Oliveira-Esquerre, Dale E. Seborg, Milton Mori, Roy Edward Bruns. Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill Pat II. Nonlinear approaches. Chemical Engineering Journal 105 (2004) 61–69.

O Que é Demanda Bioquímica de Oxigênio (DBO)?. Acesso em 15 de novembro de 2022. Link: https://www.fusati.com.br/o-que-e-demanda-bioquimica-de-oxigenio-dbo/

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013

PYTHON SOFTWARE FOUNDATION. Python Language Site. Disponível em: <https://www.python.org/doc/>. Acesso em: 09 de nov. de 2022.

Bibliotecas Python utilizadas:

- pandas
- matplotlib.pyplot
- seaborn
- numpy
- sklearn.linear_model
- sklearn.model_selection
- sklearn.metrics

Anexo

O código está disponível no seguinte link: Projeto DBO - Renê e Thiago - Colaboratory (google.com)