Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Fernanda Cardoso Farias

Avaliação de modelos de aprendizagem de máquinas para a otimização de produção de enzimas

Curitiba
2019

Fernanda Cardoso Farias

Avaliação de modelos de aprendizagem de máquinas para a otimização de produção de enzimas

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Daniel Weingaertner

Curitiba
2019

Avaliação de modelos de aprendizagem de máquinas para a otimização de produção de enzimas

Fernanda Cardoso Farias[1]

[1]Aluna do programa de Especialização em Data Science Big Data

Resumo

A produção de enzimas em larga escala é uma atividade de grande importância comercial e para os objetivos de sustentatibilidade de diversos ramos da indústria. Entretanto, a otimização dos processos envolvidos na fermentação para produção destas enzimas é até hoje uma atividade extretamente empírica, baseada na experiência das pessoas que trabalham neste tipo de processo industrial. Por esta razão, o presente trabalho utilizou a abordagem de aprendizado de máquinas para a identificação dos principais parâmetros que influenciam a produção de um determinado tipo de enzimas em escala industrial. Com a avaliação estatística por Pearson e otimização de um modelo de Random forest, foram identificados dois parâmetros como essenciais para obtenção de uma boa produtividade: o tempo de fermentação, e a dosagem do químico "B". Assim como dois parâmetros que prejudicam a produção desta enzima: O peso total da batelada de produção e a potência de agitação.

Palavras-chave: aprendizado de máquinas, seleção de parâmetros, otimização de processos, produção de enzimas, Random Forest

Abstract

Large-scale enzyme production is an activity of great commercial importance and for the purposes of sustainability of various branches of industry. However, the optimization of the processes involved in Fermentation for the production of these enzymes is still an extremely empirical activity, based on the experience of people working in this type of industrial process. For this reason, the present work used the machine learning approach to identifying the main parameters that influence the production of a particular type of enzyme on an industrial scale. With statistical evaluation by Pearson and optimization of a Random forest model, two parameters were identified as essential for obtaining good productivity: the fermentation time, and the dosage of the chemical "B". Just like two Parameters that impair the production of this enzyme: The total weight of the production batch and the potency of agitation.

Keywords: machine learning, parameter selection, process optimization, production of Enzymes, Random Forest

1. Introdução

Enzimas são, em sua maior parte, proteínas com a função de catalisar biologicamente reações químicas. Essas proteínas são produzidas por todos os tipos de seres vivos, dos mais simples como as bactérias do domínio arquea, até os animais vertebrados [1].

Por conta da sua capacidade em catalisar reações de todos os tipos, de modo seletivo e em condições brandas de processo, a utilização de enzimas em processos para as mais diversas indústrias é grande e tende a crescer ainda mais nos próximos anos [2]. Além disso, o uso de enzimas traz ainda a vantagem de processos mais sustentaveis para diversas indústrias, por substituir muitas vezes processos químicos clássicos que são altamente poluentes.

A produção em larga escala de enzimas, obedecendo às exigências de controle de qualidade e segurança microbiológica que as indústrias necessitam, somente foi possível com o advento da microbiologia, engenharia de proteínas, transgenia e conhecimento sobre fermentação microbiológica em escala industrial. Este tipo de produto, assim como qualquer outro tipo de fermentação, depende de basicamente três fatores: a)

o microrganismo produtor da enzima em questão, b) o meio de cultivo e c) os parâmetros de fermentação [3].

Os microrganismos mais utilizados são bactérias ou fungos. Eles podem produzir de forma nativa a enzima desejada, ou ainda ser um organismo geneticamente modificado (OGM) para ser capaz de produzir a enzima desejada. E as exigências destes microrganismos são responsáveis por determinar os outros fatores necessários para a produção de enzimas. O segundo fator determinante, o meio de cultivo, é o alimento para o crescimento do microrganismo e produção das enzimas, mas também pode servir como indutor para a produção de uma quantidade maior dos produtos de interesse da fermentação.

Por sua vez, os parâmetros de fermentação são um conjunto de características que ajudam a dar as condições necessárias para o crescimento dos microrganismos e produção das enzimas, como o pH do meio, temperatura, pressão, quantidade de ar injetado, agitação, químicos adicionados e outras condições exigidas pelo microrganismo. A otimização deste conjunto de parâmetros consiste no maior desafio industrial para o escalonamento da produção de enzimas.

Com isso, no contexto de indústrias de biotecnologia produtoras de enzimas, boa parte da otimização da produção e dos parâmetros de fermentação são basicamente da ordem de controle de produção, onde se tem o objetivo de maximizar certos alvos, como a produtividade, e minimizar outros, como por exemplo o uso de água ou energia. No controle diário dos parâmetros envolvidos é possível encontrar empiricamente a melhor combinação de todas as variáveis envolvidas, e geralmente é desta forma que os parâmetros são otimizados, com base na experiência e conhecimento das pessoas responsáveis que trabalham há anos com o mesmo processo.

Essa abordagem lembra a ideia de aprendizado de máquinas supervisionado e esta semelhança é um forte indicativo de que essas técnicas podem ser empregadas nesse tipo de indústria para otimização dos processos. Portanto, ao analisar uma quantidade considerável de dados históricos é possível que os algoritmos identifiquem relações complexas entre os diferentes parâmetros e identificar pontos ótimos de operação [6, 7].

Modelos de otimização de processos baseados em aprendizado de máquina podem fornecer um cenário da produtividade, com seus picos e vales representando alta e baixa produção. Os algoritmos de otimização multidimensional (contando com os diversos parâmetros que influenciam o processo) se movimentam nesse contexto procurando o pico mais alto que represente a maior produtividade possível [11]. Desta forma, os algoritmos são capazes de fornecer recomendações sobre como atingir melhor esse pico, ou seja, quais variáveis de controle ajustar e para que intervalos de controle as ajustar [8, 9, 10].

Idealmente a otimização de processos por meio de aprendizado de máquina passa por três etapas: a) identificação de um algoritmo que seja capaz de predizer bem seu processo, seja a produtividade, o consumo de insumos ou a qualidade do produto final; b) a otimização dos diferentes parâmetros do processo pode ser feita com o o algoritmo identificado; c) as recomendações recebidas pelo algoritmo escolhido podem redefinir os parâmetros de processo para efetiva otimização [12].

O presente trabalho se propõe a identificar o melhor algoritmo para a classificação de bateladas possivelmente boas ou não na produção de um tipo de produto contendo enzimas produzido por uma indústria de biotecnologia instalada em Araucária PR.

2. Materiais e Métodos

2.1. Programas e pacotes utilizados

A linguagem de programação Python foi utilizada para manipular os dados e implementar os modelos estudados. Dentro desta linguagem os pacotes de manipulação de dados, cálculos, vizualizacao de dados e aprendizado de maquinas scikit learn, seaborn, pandas, matplotlib, pandas e numpy foram utilizados [13, 14, 15, 16].

2.2. Fonte dos dados

Os dados obtidos são fundamentalmente compostos pelos parâmetros que são monitorados por sensores que atuam em tanques de fermentação, e também por dados inseridos no sistema SAP de monitoramento de controle de qualidade da empresa em questão. Todos os dados foram gerados durante diferentes lotes de produção de produtos enzimáticos proveniente de fermentação microbiana.

Esses dados são contínuos ao longo da fermentação, e relativos ao seu lote de produção e ao tempo de fermentação. Além destes dados contínuos ao longo da fermentação, também foram coletados os dados relativos a cada lote para (a) rendimento e (b) produtividade. Esses atributos são a resposta aos atributos

do processo e podem ser classificados diretamente em classes de qualidade do produto gerado.

Os dados correspondentes ao período de janeiro de 2016 a março de 2019, foram resgatados do banco de dados da empresa produtora de enzimas. O software que, registra, gerencia, faz a comunicação com a interface do SAP utilizado pelo laboratório de controle de qualidade e extrai os dados diretamente para um arquivo de “xlms” ou “cvs” é o PiSystem da Osysoft [17].

2.3. Seleção e limpeza dos dados

Para fins de simplificação foram escolhidos somente os lotes de fabricação de somente um tipo de produto. A razão desta escolha é que os parâmetros podem ter influências diferentes para cada produto, assim como sua faixa de trabalho vai ser diferenciada. Desta forma o modelo poderia ficar prejudicado pelas diferentes soluções possíveis. Com isso, o universo de amostras para este único tipo de produto foi reduzido a 225 lotes.

Como as variaveis resposta para cada lote são pontuais e não contínuas como os parâmetros monitorados, um resumo para cada lote foi feito dos parâmetros. Neste resumo, os valores dos parâmetros com vazão de entrada de componentes e ar inseridos no tanque de fermentação foram integrados para cada lote, assim como foi feita uma média dos parâmetros controlados, como temperatura, pH e potência de agitação. Com isso restaram 24 parâmetros de controle e apontamento para cada lote. O que significa que o modelo teria que lidar com um problema de alta dimensionalidade e com poucas amostras coletadas.

Para iniciar a análise exploratória destes dados, foi assegurado que nenhuma das características tivesse uma porcentagem muito alta de dados faltantes, assim como que todos os dados que devam ser tratados como números do tipo “float” estivessem neste formato. Após isso, as informações que fossem de classe, como turno ou tanque utilizado foram transformadas em características de números inteiros para melhor manipulação dos dados.

2.4. Análise exploratória dos dados

Para conhecer melhor os dados obtidos sobre os 225 lotes de fermentação para a produção de enzimas, algumas análises iniciais foram empregadas. Inicialmente, a porcentagem de dados faltantes foi calculada para todos parâmetros para a produção de enzimas e somente um apresentou 23,6

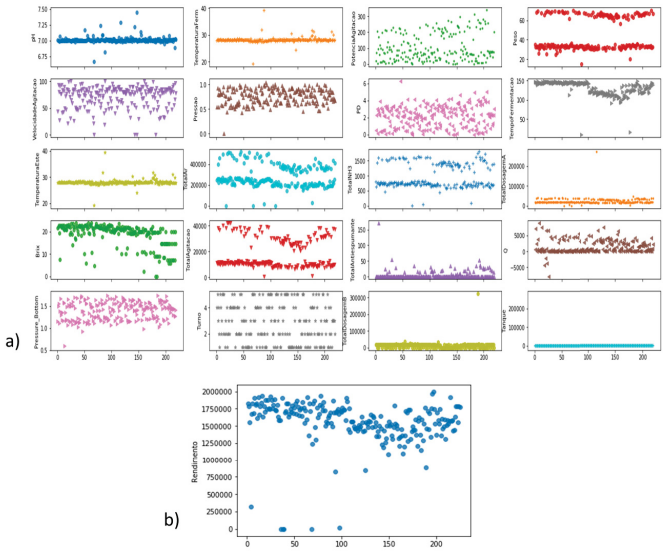

Figura 1: a) Gráficos de dispersão para todas as variáveis envolvidas no processo de fabricação de enzimas e b) para a variável resposta ’Rendimento’.

Observando a distribuição dos dados, alguns pontos se destacaram por estarem muito distantes do resto da distribuição (Figura1a e 1b). A anormalidade destes dados muitas vezes não condiz também com uma possibilidade real dentro de uma fermentação, como por exemplo lotes em que o rendimento final de enzima era igual a zero, ou a dosagem de algum químico superior em muitas ordens de grandeza em um lote em relação a todos os outros.

2.5. Remoção de outliers

Para a remoção destes pontos anormais seguirem um padrão e não serem arbitrários, foram removidas as linhas contendo somente os dados que estivessem muito abaixo ou muito acima da distribuição de cada variável. Para isso definimos que seriam removidos os dados que, em uma distribuição dos dados, estivessem três vezes a distância interquartil abaixo do primeiro quartil, ou três vezes a distância interquartil acima do terceiro quartil.

Removendo os pontos anormais para a variável resposta (Rendimento), e outras três variáveis (‘Dosagem total de reagente A’, ‘Dosagem total de reagente B’ e ‘Tempo de fermentação’), os dados ficaram distribuídos como demonstrado na Figura 2.

2.6. Seleção de variáveis

Com a base de dados limitada, e uma dimensão grande de variáveis (24 variáveis ao todo), foi necessária reali-

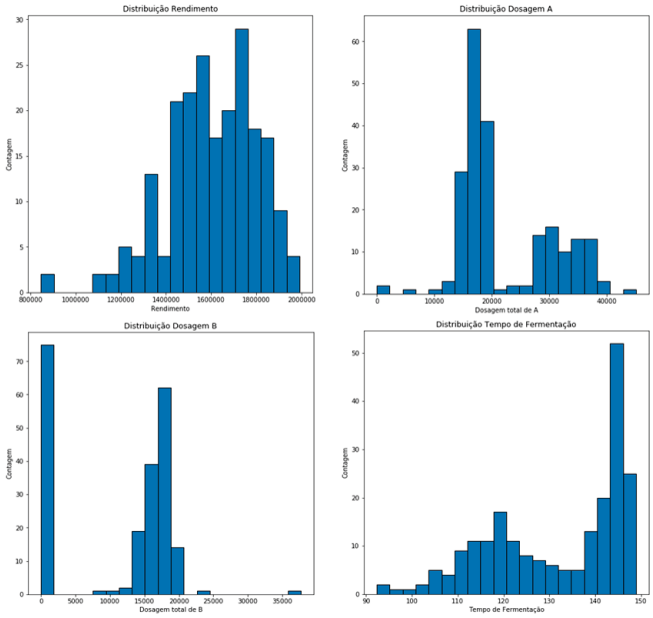

Figura 2: Gráficos de distribuição de dados para as variáveis que tiveram pontos anormais removidos.

zar uma pré-seleção das variáveis. Foram removidas as variáveis que tivessem uma grande colinearidade com outras variáveis que não a resposta. Removendo variáveis colineares estamos removendo a redundância que pode ser acrescentada ao modelo e melhorando a capacidade de generalização do modelo.

Uma métrica simples foi utilizada para remover as variáveis com coeficiente de correlação superior à 0,6. Com essa estratégia restaram 22 variáveis para descrição do modelo.

Como a ideia central desta análise é relacionar os parâmetros de fermentação com a qualidade da produção por meio do rendimento obtido, neste ponto a variável "Qualidade" foi criada, onde foi atribuído o valor de "0" para todos os lotes que obtiveram um rendimento inferior ao primeiro quartil de distribuição da variável "Rendimento", e para todos os outros lotes foi atribuído o valor de "1".

2.7. Normalização dos dados

A normalização é necessária pois cada parâmetro determinante para o modelo é medido em unicades e escalas de grandeza diferentes. Com essa técnica é possível deixar todas as características de um modelo em uma mesma escala numérica. Embora métodos como Regressão Linear e Random Forest não necessitem de uma normalização, como desejamos uma comparação entre diferentes modelos, a normalização foi necessária.

Neste estudo o MinMaxScaler do pacote Scikit-Learn foi utilizado para deixar todos os parâmetros em um intervalo entre 0 e 1, certificando-se de que os dados sejam divididos entre treinamento e teste antes da normalização.

2.8. Modelos empregados

Foram avaliados seis modelos usualmente aplicados em aprendizado de máquinas e disponibilizados na biblioteca do Scikit-Learn: 'método linear clássico (LDA)', 'support vector machine', 'indução de árvore de decisão classificatória', 'método de k-vizinhos', 'random forest' e 'gradient boosting'.

3. Resultados

3.1. Correlação entre variáveis e alvo

Para identificar a correlação entre as variáveis de processo selecionadas e o rendimento alvo foi utilizado o coeficiente de correlação de Pearson, que mede a força de correlação linear de cada variável com o alvo resposta. Este coeficiente também evidencia o sentido desta correlação, se é positivo ou negativo. Lembrando que essa correlação medida é linear, então as variáveis podem ainda ter uma correlação não-linear com o rendimento alvo.

Pela correlação de Pearson, o "Tempo de fermentação" e a "Dosagem total do químico B" foram apontados como os fatores que mais influenciam positivamente o rendimento (Figura3). Enquanto o "Peso do fermentador" e a "Potência de agitação" são os fatores que influenciam mais negativamente o rendimento.

3.2. Avaliação dos modelos e otimização

Em uma primeira etapa, a performance de base dos modelos foi avaliada sem realizar um tunning dos parâmetros de cada modelo. A métrica escolhida para avaliar os modelos foi o erro quadrático médio.

O modelo com melhor desempenho para a classificação dos lotes com boa ou má qualidade de rendimento de produção de enzimas, foi o 'Random Forest', com erro quadrático médio de 0,2037.

Considerando que o modelo de "Random forest" teve o melhor desempenho na avaliação inicial, seus parâmetros de trabalho foram otimizados, com uma busca randômica pelos melhores hiperparâmetros para este problema de classificação. Com a otimização dos parâmetros utilizados o modelo alcançou a acurácia de 85,19

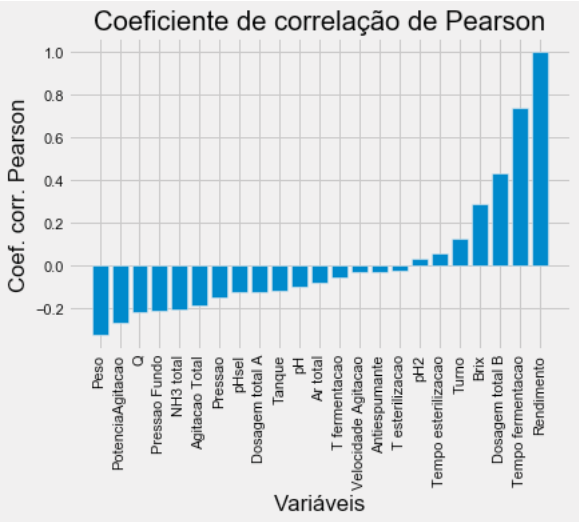

Figura 3: Coeficiente de correlação linear de Pearson para as variáveis a serem empregadas no modelo

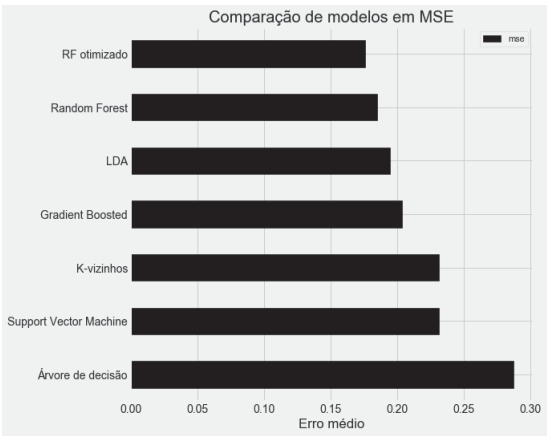

Figura 4: Comparação de modelos de aprendizado de máquina com base no erro quadrático médio (MSE).

O modelo melhor modelo de Random Forest encontrado pelo grid aleatório utiliza de "booststrap"para gerar as árvores, com no máximo 1400 árvores, cada árvore com profundidade de no máximo 110, número mínimo de amostras para criar um nó de 10 e com o critério de gini para criação de nós em cada árvore (Figura4).

A importância das variáveis empregadas no modelo de Random forest foi calculada e está apresentada na Figura 5. Essas importâncias reforçam mais uma vez a análise pelo coeficiente de Pearson, tendo o “Tempo de fermentação”, assim como a “Dosagem total do químico B” como variáveis com maior influência na qualidade de produção deste tipo de enzima.

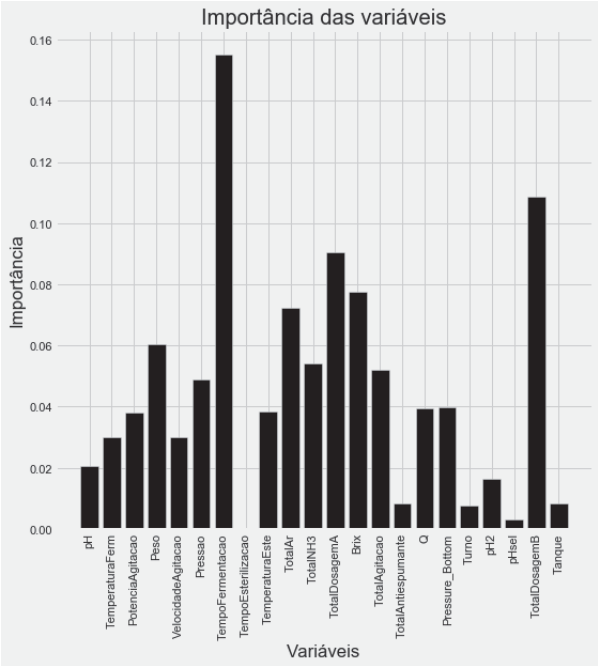

Figura 5: Índice da importância das variáveis empregadas no modelo de ’Random Forest’ para classificação de lotes de produção de enzima.

4. Conclusões

A avaliação dos dados de fermentação para produção de enzimas evidenciou que os fatores mais importantes para a obtenção de uma produção boa são a maximização do tempo de fermentação, e a maior dosagem de químico B. O primeiro fator pode indicar que a fermentação em alguns casos é interrompida antes de alcançar o pico de produtividade. Além disso, bateladas menores (com menor peso final do fermentador) e com menor agitação também contribuem para melhor produção da enzima de interesse. Esta indicação de menor agitação pode ser resultado de sensibilidade do microrganismo em questão à taxas altas de cisalhamento.

Uma das árvores extraídas do modelo de Random Forest otimizado, na Figura 6, exemplifica essas influências dos principais parâmetros identificados para uma boa produção de enzimas.

Agradecimentos

Gostaria de agradecer aos professores do programa de Especialização em Data Science e Big Data, assim como à Universidade Federal do Paraná pela oportunidade em participar deste programa de pós-graduação.

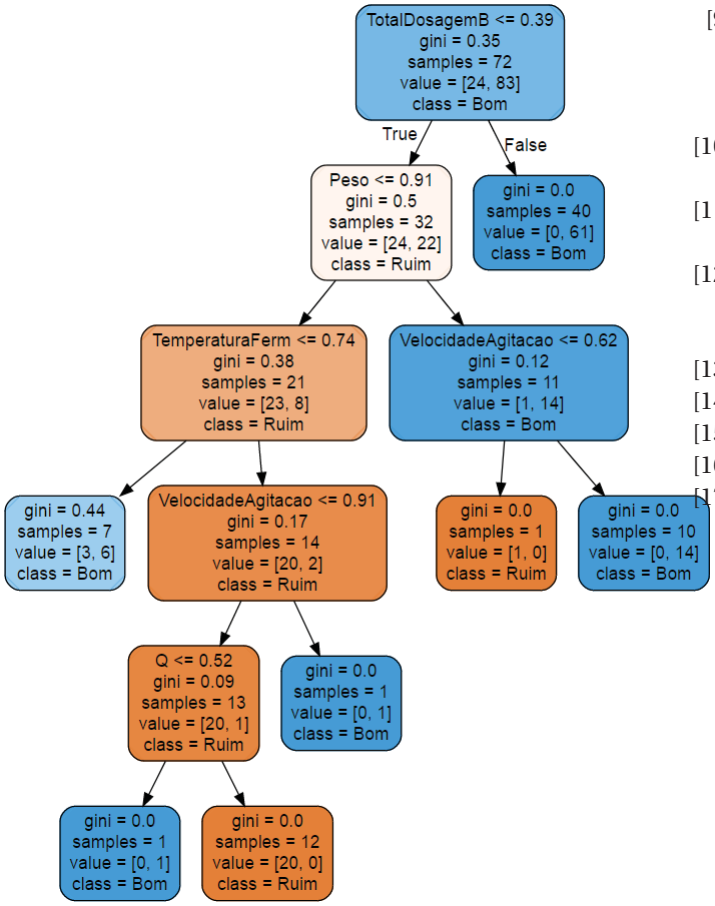

Figura 6: Árvore extraída do modelo de Random Forest otimizado para a seleção de parâmetros de fermentação e produção de determinada enzima.

Referências

[1] A. Lehninger, D. L. Nelson, M. M. COX, *Princípios de bioquímica*, (Artmed, Porto Alegre, 2011), 6. ed.

[2] M. A. Z. Coelho e A. Medeiros, *Tecnologia Enzimática*, (EPUB, Petrópolis, 2008).

[3] W. Borzani, W. Schmidell, U. A. Lima, E. Aquarone, *Biotecnologia Industrial - Vol. 3*, (Blucher, 2001).

[4] W. Borzani, W. Schmidell, U. A. Lima, E. Aquarone, *Biotecnologia Industrial - Vol. 2*, (Blucher, 2001).

[5] P.A. Ferreira, P. Vargas, M. P. Costa, I. L. Machado, F. C. Costa, J. C. S. Borges, *Revista Processos Químicos, v.3, n.5, ano 3*, (SENAI, Goias, 2009)

[6] J. Wang, Y. Ma, L. Zhang, R. X. Gao, *Deep Learning for Smart Manufacturing: Methods and Applications*

[7] J. Dobsa, B. Sobotoa-Soloman, V. Mihokovic, *Modelling food industry process by decision tree*, (Conference Paper, September 2007)

[8] K. Siddharth, A. Pathak, A. K. Pani, *Real-time quality monitoring in debutanizer column with regression tree and ANFIS*, (Journal of Industrial Engineering International (2019) 15:41–51)

[9] E. Ikonomovska, *Algorithms for Learning Regression Trees and Ensembles on Evolving Data Streams*, (Doctoral Dissertation, Jozef Stefan International Postgraduate School Ljubljana, Slovenia, October 2012)

[10] W. Koehrsen, *Hyperparameter Tuning the Random Forest in Python*, (Towards Data Science, Jan, 2018)

[11] V. Flovik, *How to use machine learning for production optimization*, (Towards Data Science, Aug, 2017)

[12] Y. Zhou, *Data driven process monitoring based on neural networks and classification trees*, (Doctoral Dissertation, Texas AM University, Aug, 2004)

[13] *https://www.python.org/*

[14] *https://matplotlib.org/*

[15] *https://numpy.org/*

[16] *https://scikit-learn.org/stable/index.html*

[17] *https://www.osisoft.com/pi-system/*