Monitoramento de uma Estação de Tratamento de Efluentes Industriais utilizando Machine Learning

Octavio Bomfim Santiago

Projeto Final em Engenharia Química

Orientador:

Prof. Bruno Didier Olivier Capron, D.Sc.

Agosto de 2019

Monitoramento de uma Estação de Tratamento de Efluentes Industriais utilizando Machine Learning

Octavio Bomfim Santiago

Projeto Final em Engenharia Química submetida ao Corpo Docente da Escola de Química, como parte dos requisitos necessários à obtenção do grau de Engenheiro Químico.

Aprovador por:

_____

Verônica Maria de A. Calado, D.Sc.

_____

Marcellus Guedes Fernandes de Moraes, M.Sc.

_____

Yuri Pereira Ribeiro Silveira, Eng.

Orientado por:

_____

Bruno Didier Olivier Capron, D.Sc.

Rio de Janeiro, RJ - Brasil

Agosto de 2019

Santiago, Octavio Bomfim.

Monitoramento de uma Estação de Tratamento de Efluentes Industriais utilizando Machine Learning / Octávio Bomfim Santiago. Rio de Janeiro: UFRJ/EQ, 2019.

Projeto Final – Universidade Federal do Rio de Janeiro, Escola de Química, 2019.
Orientador: Bruno Didier Olivier Capron, D.Sc..
1. Machine Learning. 2. Ciência de Dados. 3. Tratamento de Efluentes Industriais. 4. Monografia. (Graduação – UFRJ/EQ).

5. Bruno Didier Olivier Capron. I. Monitoramento de uma Estação de Tratamento de Efluentes Industriais utilizando Machine Learning

Resumo do Projeto Final apresentada à Escola de Química como parte dos requisitos o necessários para obtenção do grau de Engenharia Química.

Monitoramento de uma Estação de Tratamento de Efluentes Industriais utilizando Machine Learning

Octavio Bomfim Santiago

Agosto, 2019

Orientador: Bruno Didier Olivier Capron, D.Sc.

O processo de tratamento de efluentes industriais por via biológica é bastante complexo e de difícil controle e monitoramento por meios convencionais, por possuir muitas variáveis e dinâmicas diferentes. O presente trabalho, busca destacar a implementação de uma tecnologia emergente que pode ser aplicada em diversos setores industriais e com grande potencial de crescimento, o *machine learning*, e ressaltar a oportunidade de investimento em sua implementação na área de tratamento de efluentes industriais. Ao longo do trabalho, analisou-se o potencial de aplicação de um algoritmo de *machine learning* supervisionado para a predição da qualidade do efluente final de um processo de tratamento biológico por lodo ativado junto com uma solução de monitoramento dos principais indicadores de eficiência da estação de tratamento. Este trabalho foi desenvolvido em parceria com uma empresa da área petroquímica a fim de demonstrar para a comunidade científica a aplicabilidade da solução proposta e sua viabilidade técnica e econômica. O resultado encontrado mostra que a previsibilidade da qualidade do efluente de forma antecipada seria totalmente viável e sua implementação poderia gerar uma operação rentável com grandes ganhos de economia em análises laboratoriais e qualidade do efluente, além de uma tomada de decisão mais rápida e precisa do que por meios convencionais.

Palavras-chave: machine learning, tratamento de efluentes industriais, controle e monitoramento, ciência de dados

ABSTRACT

The biological wastewater treatment process is quite complex and hard to control and monitor by conventional means because it has many different variables and dynamics. This article aims at emphasizing the implementation of an emerging technology, machine learning, that can be applied to various industrial sectors with high growth potential and highlight the investment opportunity in its implementation in the area of industrial wastewater treatment. Throughout the work, it was analyzed the potential of applying a supervised machine learning algorithm to predict the effluent quality of an activated sludge biological treatment process along with a monitoring solution of the key performance indicators of the wastewater treatment. This work was developed in partnership with a petrochemical company in order to demonstrate to the scientific community the applicability of the proposed solution and its technical and economic viability. The result shows that the predictability of effluent quality would be totally feasible and its implementation could generate a profitable operation with great savings in laboratory analysis, as well as faster decision making than by conventional means.

Keyword: machine learning, wastewater treatment, process control and monitoring, data science

Índice

Índice de Figuras

Índice de Tabelas

I. Introdução

Desde o desenvolvimento da primeira ferramenta na história de humanidade, os seres humanos utilizam de artifícios externos para aumentar a eficiência de suas ações e dar poder a suas ideias e inovações. O uso em larga escala da tecnologia de maneira descontrolada causa efeitos adversos ao meio ambiente. O desenvolvimento da ciência e tecnologia da inovação é primordial para manter o controle dos processos e o equilíbrio com o meio ambiente.

Uma das tecnologias mais utilizadas para a diminuição do impacto poluidor industrial no meio ambiente é o tratamento biológico de efluentes líquidos que utiliza de uma população bacteriana e processos físico-químicos para degradar a matéria orgânica e componentes poluidores do efluente industrial. O processo de tratamento de efluentes industriais por via biológica é bastante complexo e de difícil controle e monitoramento por meios convencionais por possuir muitas variáveis e dinâmicas diferentes. (VON SPERLING, 1997)

Com o desenvolvimento da tecnologia computacional e estatística surge a área de Ciência de Dados que utiliza do aprendizado de máquina, do inglês *machine learning*[1], para elaborar modelos matemáticos com capacidade preditiva que permite uma tomada de ação adiantada para o controle da estação de tratamento, visto que o processo completo com um ciclo dura 24 horas e a análise de concentração poluidora no efluente pode durar até 5 dias.

[1] Machine Learning – Aprendizado de Máquina: um método de análise de dados que automatiza a construção de modelos analíticos. É um ramo da inteligência artificial baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana. (SAS, 2019)

II. Objetivo

O objetivo do trabalho é obter um modelo de *machine learning* adequado ao processo que consiga prever os casos de Demanda Química de Oxigênio (DQO) fora dos limites legais com uma fidelidade satisfatória para ser implementado como uma ferramenta de processo que junto com um dashboard de monitoramento dos indicadores auxiliará na tomada de decisão dos funcionários da fábrica.

No geral os objetivos do projeto foram:

- Melhorar a eficiência operacional
- Redução de custos por análises laboratoriais
- Predição com a mais alta acurácia da qualidade do efluente final no menor tempo possível
- Relatório de conhecimento do processo
- Sistema de monitoramento operacional e gerencial da estação e tratamento

II.1. Viabilidade econômica do projeto

O objetivo deste estudo é alcançar a viabilidade econômica para colocá-lo em operação na fábrica e futuramente em outras unidades da mesma companhia. Partindo deste princípio, foi realizado um estudo básico de mercado com os valores necessários para a realização dos testes laboratoriais "in loco" e a economia que teria em adotar um controle automatizado.

De acordo com uma pesquisa de mercado, mais especificamente no site Forlab Express, um kit de testes de DQO tem um valor médio de R$ 400,00 para 20 testes porém, cada kit cobre uma faixa específica de DQO e para cobrir o espaço com 3 faixas diferentes, tem-se o valor de R$ 60,00 por teste.

Fazendo um cálculo básico para uma estação de tratamento com 5 pontos de captação de amostras, tem-se um gasto de R$ 300,00 por dia com reagentes e R$ 109.500,00 ao ano, sem contar os custos com a manutenção do equipamento, salários dos analistas do laboratório, custos operacionais do laboratório e estes custos podem ser maiores se os exames forem feitos em um laboratório terceirizado.

Reduzindo os pontos de captação para 1 ponto que seria o do afluente principal (tanque de equalização) e eventualmente o efluente final utilizando o modelo de *machine learning* estabelecido, que de acordo com uma análise estatística dos dados se mostra 10% de casos em não conformidade no ano, temos cerca de 50 amostras anuais com risco de não conformidade. Isso se traduz em R$ 21.900,00 ao ano com uma amostra e R$ 3.000,00 com os testes no efluente final e mais R$ 3.000,00 para a extensão no efluente da SAO. No total o gasto fica em R$ 27.900,00 ao ano, o que reduz em 74,5% os custos com análises laboratoriais de efluentes, R$ 81.600,00 em valor ao ano.

III. Estação de tratamento de Efluentes (ETE)

Este capítulo objetivou, com base em revisão bibliográfica, apresentar uma visão sobre a estação de tratamento de efluentes industriais utilizada para o estudo de modo que facilite a compreensão global do contexto de sua tecnologia e processos.

Está descrito na Resolução CONAMA 357, de 17 de Março de 2005, a todo efluente gerado, seja líquido ou sólido, deve ser submetido a um processo de tratamento antes da sua disposição final no meio ambiente. Para o caso dos efluentes líquidos a disposição é feita em uma estação de tratamento de efluentes (ETE) própria ou destinada para uma empresa terceira que fará o tratamento em uma ETE externa. O efluente então é tratado de acordo com suas características e lançado em um corpo hídrico de acordo com os limites impostos pelo órgão ambiental regional e nacional.

A empresa citada neste trabalho possui uma estação de tratamento de efluentes própria e devidamente projetada para o tratamento dos efluentes gerados pela fábrica.

III.1. Sistema de tratamento de Efluentes – Lodo Ativado

A estação de tratamento de fluentes presente na fábrica é composta por um sistema de lodo ativado com um reator biológico de operação intermitente (batelada) de aeração prolongada. A classificação do processo de tratamento ne estação de efluentes segue alguns critérios como (VON SPERLING, 1997):

- Divisão quanto a idade do lodo:
    - Lodos ativados convencional
    - Aeração prolongada
- Divisão quanto ao fluxo
    - Fluxo contínuo
    - Fluxo intermitente (batelada)
- Divisão quanto ao afluente à etapa biológica do sistema de lodos ativados
    - Efluente industrial/esgoto bruto
    - Efluente de decantador primário
    - Efluente de reator anaeróbio
    - Efluente de outro processo de tratamento

E quanto a idade do lodo temos a segunda classificação do sistema (VON SPERLING, 1997):

***Tabela 1**- Classificação do sistema por Idade do lodo*

Idade do lodo	Carga de DBO aplicada por unidade de volume	Faixa de idade do lodo	Denominação
Muito reduzida	Muito alta	Inferior a 3 dias	Aeração modificada
Reduzida	Alta	4 a 10 dias	Lodos ativados convencional
Intermediária	Intermediária	11 a 17 dias	-
Elevada	Baixa	18 a 30 dias	Aeração prolongada

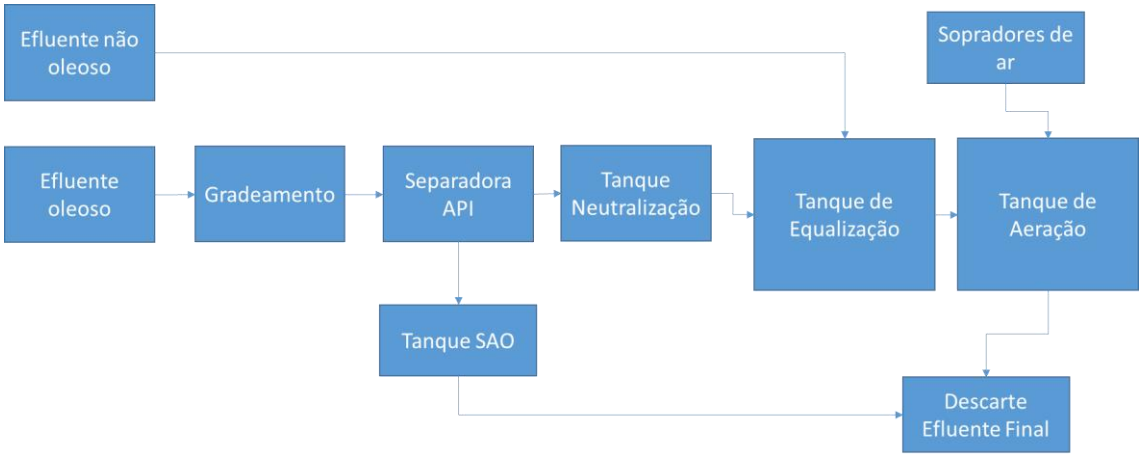
Fonte: Von Sperling, 1997.

Esta classificação se aplica tanto para fluxo contínuo quanto para fluxo intermitente (batelada).

III.1.1. Fluxograma e Equipamentos

Por um acompanhamento do processo e do estudo das conexões no sistema de tratamento de efluentes industriais da empresa, foi desenvolvido um fluxograma básico do processo de tratamento na fábrica que terá seus processos detalhados neste seção.

Figura 1 *- Fluxograma de Processos da ETE*



Fonte: O autor

III.1.1.1. Gradeamento

O sistema de gradeamento é composto de grades de ferro igualmente espaçadas e em uma inclinação específica, geralmente de 45 graus, que vedam todo o caminho do efluente a fim de separar os componentes sólidos maiores contidos no efluente como folhas, garrafas, tampas entre outros.

Figura 2*- Sistema de gradeamento de ETE*

Fonte: Habitissimo

O seu maior atrativo para o mercado consiste nas duas últimas características expostas, alta resistência mecânica e leveza, quando comparada com possíveis

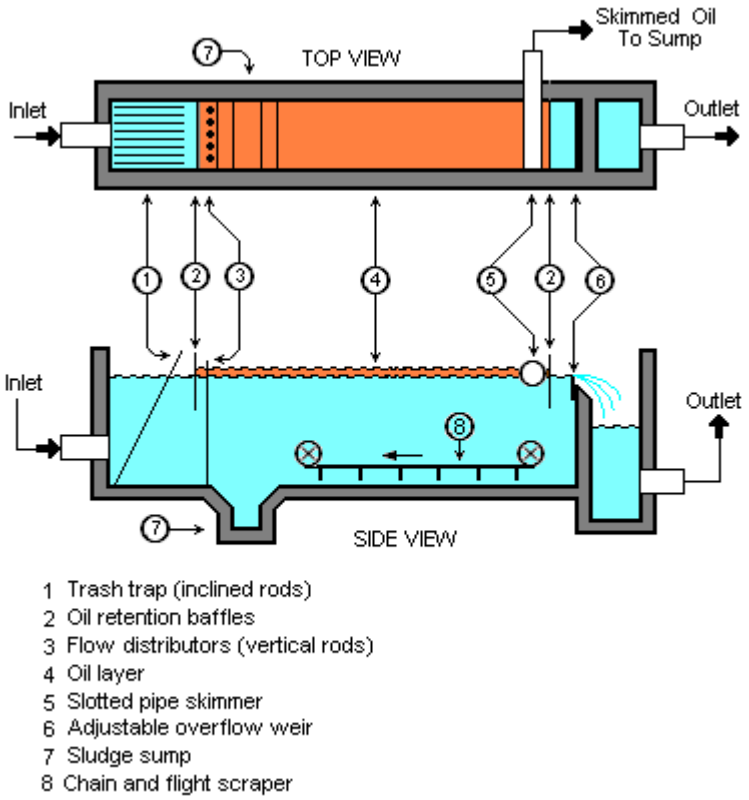
concorrentes. Logo, ela se torna uma opção extremamente interessante para indústrias que visam diminuir o peso de seus produtos e de máquinas sem perder a resistência necessária, como é o caso da indústria aeroespacial e automobilística. Essa última busca reduzir o peso dos carros para diminuir o consumo de combustíveis e a emissão de poluentes.

III.1.1.2. Separador API

O nome separadora API vem do fato que esta tecnologia é desenhada de acordo com as normas do American Petroleum Institute (API) e é muito utilizado em indústria química e petroquímicas para a separação e captação do óleo de um efluente oleoso.

A separação se dá por gravidade utilizando o princípio da lei de Stokes e separando os efluentes por sua diferença de gravidade específica. Um separador API necessita de um fluxo laminar e de baixa velocidade de escoamento para uma boa separação e para isso ele é projetado segundo a norma com uma razão mínima de 5:1 de comprimento para largura e 0,3: 0,5 de profundidade para largura.

Figura 3 *- Separador API*

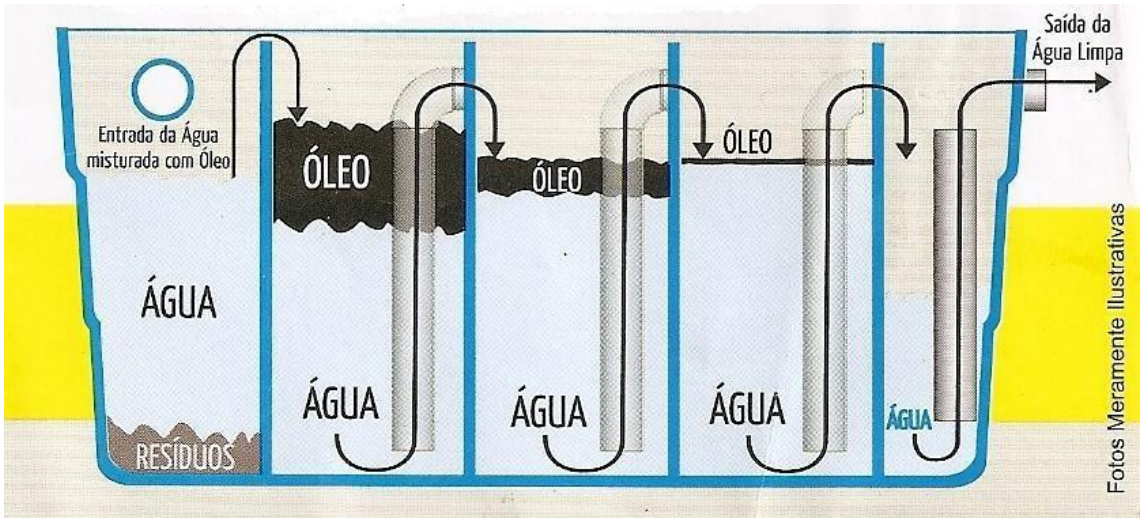


Fonte: Wikipedia

III.1.1.3. Tanque SAO

Um tanque SAO, separadora de água e óleo, utiliza também da diferença entre a gravidade específica do óleo com a água mas de uma maneira diferente do separador API. O tanque SAO é divido em partes para tornar o fluxo de entrada laminar e reduzir a velocidade a fim de melhorar a separação, com isso a água é retirada por baixo enquanto o óleo sobrenadante é retirado por cima até no fim da caixa restar somente água. O tanque opera em um fluxo contínuo de baixa vazão.

***Figura 4**- Tanque SAO*

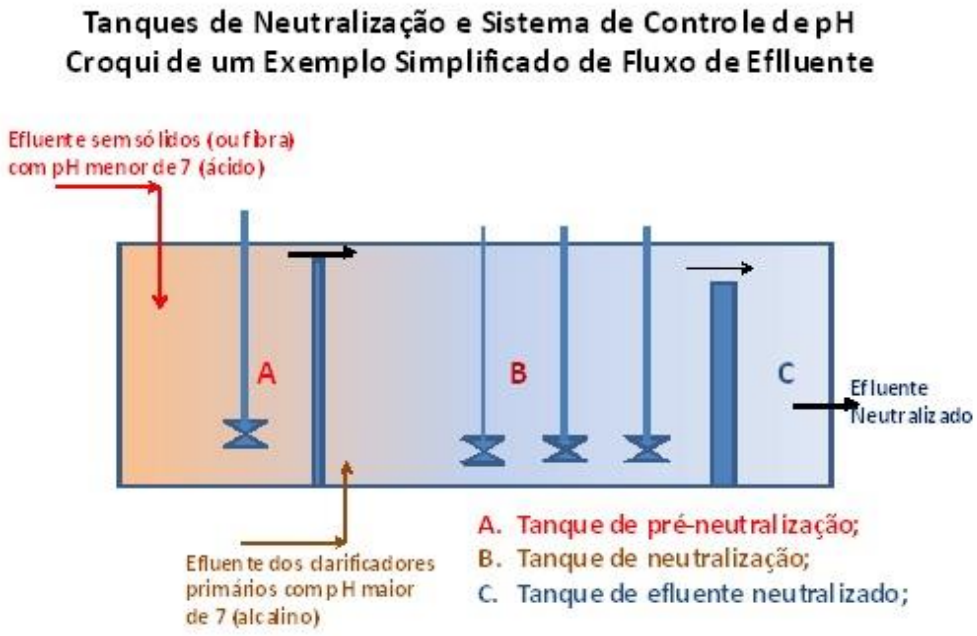


Fonte: Dedambiental

III.1.1.4. Tanque de Neutralização

O Tanque de neutralização apresenta separações entre a entrada, tratamento e saída, onde a parte da entrada apresenta o efluente com o pH mais irregular proveniente do afluente original (A na Figura 5), a parte do tratamento apresenta um regime transiente de pH e é onde o pH normalmente é regulado pela adição de ácido ou base e apresenta misturadores para homogeneizar a mistura (B na Figura 5), e na parte de saída tem-se o efluente homogeneizado com o pH uniforme e tratado (C na Figura 5).

***Figura 5**- Tanque de Neutralização*



Fonte: Celuloseonline

III.1.1.5. Tanque de Equalização

O tanque de equalização é muito utilizado quando o afluente possui uma variabilidade alta nas suas características, mudando muito o DQO/DBO[2] de entrada e para uniformizar a vazão de afluente, também variável, antes da entrada no reator biológico. O tanque de equalização então tem o intuito de amortecer a variação de vazão e carga para o tratamento biológico posterior. Estes tanques normalmente são bem largos e profundos e possuem agitadores mecânicos ou aeradores para homogeneizar a mistura.

[2] DQO e DBO – Demanda química de oxigênio e Demanda bioquímica de oxigênio

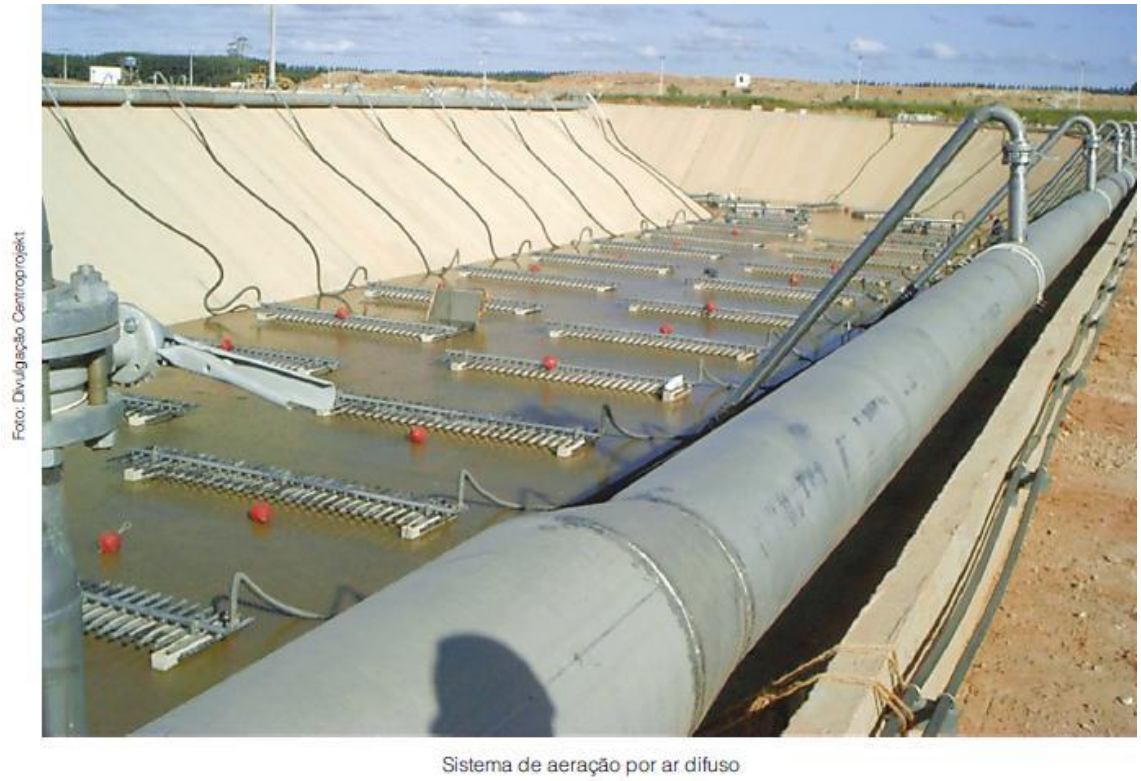
***Figura 6**- Tanque de Equalização*

Fonte: Meiofiltrante

III.1.1.6. Sistema de aeração

Os sopradores de ar são usados para injetar oxigênio no reator biológico e controlar o oxigênio dissolvido no sistema. Além disso, os sopradores também cumprem a função de homogeneizar a mistura no tanque de aeração. Existem diversos métodos de aerar e homogeneizar o sistema, porém o uso de aeradores através de ar comprimido é o método mais utilizado por ser menos destrutivo para a população microbiológica no reator.

***Figura 7**- Sistema de aeração por ar difuso*



Fonte: Revista Tae

III.1.1.7. Tanque de aeração

O tanque de aeração é o reator biológico onde fica o lodo ativado. É normalmente o processo mais importante e crítico da estação, pois costuma ser a última barreira antes do lançamento do efluente final no corpo hídrico. O intuito do tratamento bioquímico do efluente através de uma população microbiológica é majoritariamente a redução do DBO presente no afluente, mas também apresenta bons resultados na remoção de DQO, nutrientes, coloração entre outros.

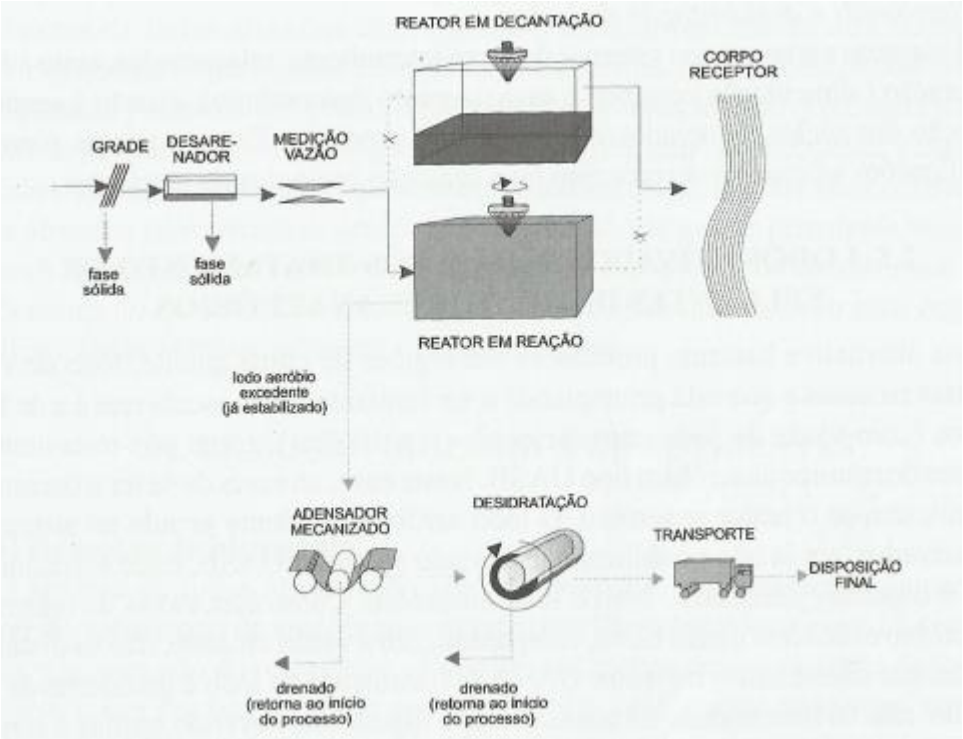
***Figura 8**- Tanque de aeração*

Fonte: Ymgerman

III.1.2. Operação Intermitente – Batelada

O processo de tratamento por lodo ativado em operação intermitente consiste na unificação de todas as unidades, processos e operações presentes no tratamento tradicional em um único reator, unificando assim a oxidação biológica com a decantação primária e decantação secundária no mesmo tanque. Utilizando as operações em um único tanque os processos passam a ser distintos através do tempo e não mais de sistemas físicos, como ocorre no fluxo contínuo (VON SPERLING, 1997).

***Figura 9**- Fluxograma de operação Intermitente - Batelada*



Fonte: Von Sperling, 1997

O processo de tratamento de efluentes industriais com lodo ativado por operação intermitente (batelada) tornou-se popular a partir da década de 1980 e vem sido utilizado amplamente no Brasil por se aplicar a um tratamento de efluentes mais diversos e ter um melhor conhecimento do sistema para o desenvolvimento da instrumentação de processo. Um fator de vantagem importante quanto ao sistema em batelada é o maior controle do sistema e um sistema mais simples para lançamento de efluentes, podendo interromper facilmente o lançamento se necessário e controlar melhor a vazão de lançamento.

Os reatores biológicos de operação intermitente (batelada) possuem uma operação cíclica onde cada ciclo segue as etapas de enchimento, reação, sedimentação, esvaziamento e repouso. Dependendo da carga afluente na estação o processo pode apresentar um ou mais ciclos durante o dia.

Caso o enchimento ocorra com os aeradores desligados pode haver uma possibilidade de remoção de nitratos remanescentes. Após a etapa da reação aeróbia, durante a sedimentação e descarte tem-se uma etapa anóxica, ocorrendo novamente a desnitrificação porém em condições endógenas.

As características do processo de lodo ativado em batelada consiste nas seguintes etapas (VON SPERLING, 1997):

1. Enchimento – entrada do efluente bruto no reator
2. Reação – aeração/mistura da massa líquida contida no reator
3. Sedimentação – sedimentação e separação dos sólidos em suspensão do esgoto tratado
4. Esvaziamento – retirada do efluente tratado do reator
5. Repouso – ajuste de ciclos de remoção de lodo excedente

Figura 10 *- Ciclo de operação típico em Batelada*

Fonte: Von Sperling, 1997

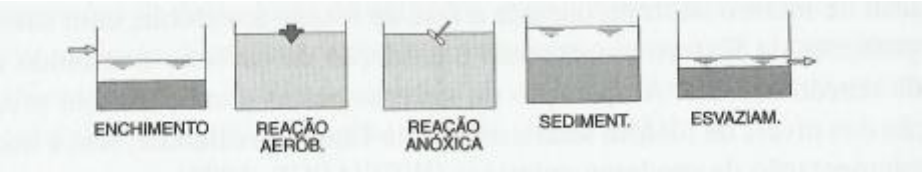
No processo fabril analisado neste estudo, temos os seguintes tempos para cada processo e seus respectivos horários. O processo tem como padrão o ciclo de 1 dia, e algumas etapas ocorrem em paralelo, como o enchimento e a reação aeróbica, pois o enchimento é feito com os aeradores ligados para melhorar o processo de homogeneização.

***Tabela 2** - Ciclo de operação em batelada*

Processo	Duração	Horário do dia
Enchimento	2 horas	14:30 as 16:30
Reação Aeróbica	17 horas	14:30 as 7:30
Sedimentação	2 horas	7:30 as 9:30
Esvaziamento	5 horas	9:30 as 14:30

Fonte: O autor

***Figura 11**- Ordem de etapas no ciclo de operação em Batelada*



Fonte: Von Sperling, 1997

III.1.3. Características dos Efluentes

A indústria em questão possui dois tipos de efluente, um efluente majoritariamente oleoso e com um DQO que varia bastante entre 50 e 3000 mas ficando na maioria do tempo entre 100 e 1000 além de um pH majoritariamente ácido, e outro efluente não oleoso que tem o DQO entre 100 e 8000 e pH do neutro para o básico.

O efluente oleoso passa por tratamento de separação e remoção do óleos e graxas e correção de pH no tanque de neutralização, após isto ele é destinado para o tanque de equalização. O efluente não oleoso é destinado direto para o tanque de equalização.

III.1.4. Controle operacional

Os objetivos principais para controle operacional do processo de tratamento de efluentes podem ser (VON SPERLING, 1997):

- Produzir um efluente final com uma qualidade que satisfaça os padrões de lançamento
- Reduzir a variabilidade da qualidade do efluente
- Evitar grandes falhas do processo
- Reduzir os custos de operação
- Aumentar a capacidade de tratamento sem a expansão física do sistema
- Implementar uma operação com eficiência varável, de forma a acomodar variações sazonais

Uma estação de tratamento de efluentes apresenta uma variabilidade muito grande na carga dos afluentes e fica maior quando a fábrica possui produtos com características diferentes. Além disso, o processo bioquímico é muito complexo para modelar e controlar em comparação com os processos químicos convencionais. A dinâmica dos sistemas de tratamento biológico contém (a) não linearidades, (b) faixas bem amplas de constantes no tempo, (c) uma cultura heterogênea de microrganismos metabolizando um substrato heterogêneo, (d) imprecisão e (e) estabilidade interrompida por falhas abruptas (VON SPERLING, 1997).

Levando em consideração todas as complexidades de modelagem da dinâmica do sistema, é fácil de compreender a dificuldade para se elaborar um controle operacional automatizado em uma estação de tratamento de efluentes. Além do mais outras dificuldades adicionais reduzem sua aplicação de maneira mais ampla, tais como (Lumbers, 1982; Markantonatos, 1988, von Sperling, 1997):

- As características do afluente são de natureza dinâmica e estocástica, com distúrbios desconhecidos e ruídos de medição superpostos a variações de processo
- O processo apresenta efeitos de tempo morto e magnitude de resposta bem diferentes dentre as variáveis de processo controladas
- Há carência de sensores confiáveis e de tempo real para algumas variáveis do processo

- Não é possível medir todas as variáveis de processo diretamente
- Em grande parte das estações de tratamento a possibilidade de controle é limitada à um projeto pouco flexível e antigo
- Há dificuldades em se incorporar modelos de processos complexos nos algoritmos de controle convencionais e limitações nas estratégias de modelos simples

Devido à tamanha complexidade, propõem-se estratégias centradas na área de Ciência de Dados que fazem uso da Inteligência Artificial com técnicas de *Machine Learning*. Esta técnica possui algoritmos capazes de modelar dinamicamente os dados de forma precisa e com a capacidade de utilizar somente as variáveis mais correlacionadas ao sistema, reduzindo a necessidade de se aplicar uma instrumentação em todo o processo.

O presente trabalho visa trazer um estudo real de caso em uma indústria que possui uma estação de tratamento de efluentes industriais. As variáveis mais importantes para controle do sistema, monitoramento da estação e a modelagem do algoritmo de *Machine Learning* serão explicitadas a seguir com o embasamento teórico de diversas fontes presentes na revisão bibliográfica do mesmo.

III.1.5. Variáveis de Processo

Para o entendimento dos processos a seguir, alguns conceitos básicos de engenharia de controle serão descritos.

De maneira geral, o controle operacional em uma estação de tratamento de efluentes pode ser classificado de acordo com o seu grau de automação nas seguintes categorias (ANDREWS, 1972):

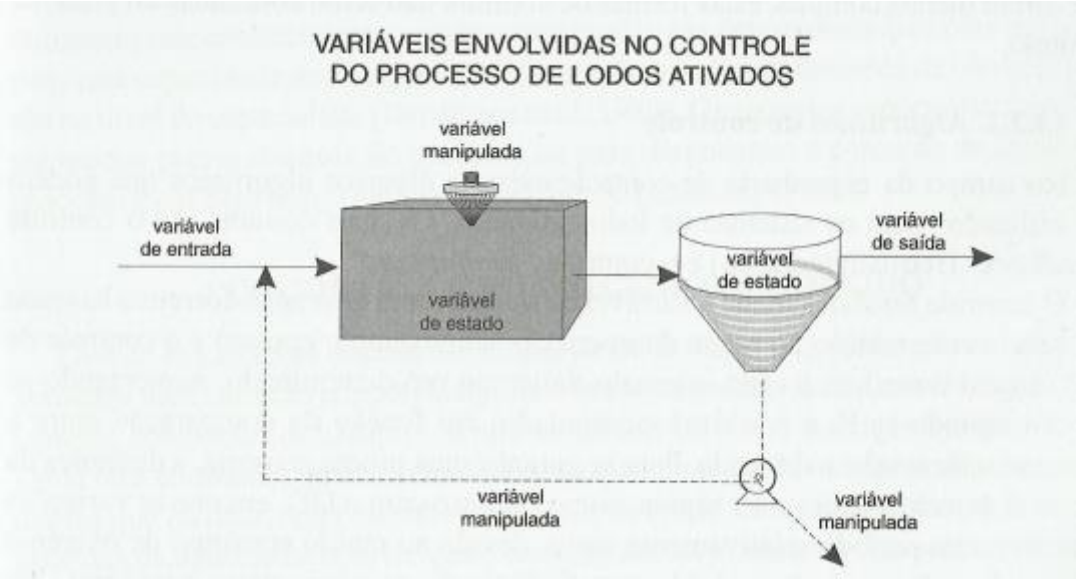
1. Operação manual, com (a) avaliação do desempenho através de sentidos humanos e (b) controle manual do processo
2. Operação manual, com (a) avaliação do desempenho por análises ou instrumentos indicadores registradores e (b) controle manual do processo
3. Controle automático, com (a) avaliação do desempenho por sensores automatizados e (b) controle automatizado do processo

A operação que mais se adequa ao processo da indústria citada no presente trabalho é a operação de classificação número 2, e todos os detalhes a seguir serão referentes a essa classificação.

Como primeira parte do controle de um sistema, devem-se identificar as variáveis envolvidas no processo que se pode distinguir dentre quatro tipos (VON SPERLING, 1997):

1. *Variáveis de entrada*
2. *Variáveis de controle (variáveis de estado e/ou variáveis de saída)*
3. *Variáveis medidas (variáveis de entrada e/ou variáveis de controle)*
4. *Variáveis manipuladas*

***Figura 12**- Variáveis de processo em Operação Batelada na ETE*



Fonte: Von Sperling, 1997

Variáveis de entrada são as que vêm diretamente do processo anterior e são externas ao processo avaliado. Essas não podem ser diretamente controladas na maior parte das estações de tratamento de efluentes. Tem-se como exemplo de variáveis de entrada as características do afluente, como DQO, DBO, pH, vazão e Sólidos Suspensos (SS).

Variáveis de controle são as que necessitam ser controladas para manipular o sistema. Dentre essas podem-se citar as *variáveis de estado* como SST[3], OD[4] e volume de lodo. Um caso particular dentre as variáveis de estado são as *variáveis de saída* que definem a qualidade do efluente final como DQO, DBO, SS, pH e nutrientes.

Variáveis medidas são as variáveis de entrada, saída, controle ou outras que possuem informações importantes para o controle do sistema e que podem apresentar um sistema de medição. A escolha de quais variáveis devem ser medidas depende da viabilidade da sua medição, do algoritmo de controle e dos resultados das análises operacionais.

Variáveis manipuladas são as que podem ser alteradas para manter as variáveis de controle no nível desejado e alcançar os resultados pretendidos nas variáveis de saída. Um sistema de tratamento de efluentes possui poucas variáveis manipuladas comparado aos demais processos na indústria. Algumas variáveis manipuladas são vazão de ar dos sopradores, tempo de aeração, pH, vazão de afluente (havendo um tanque de equalização) e volume no reator.

Variáveis disponíveis na estação de tratamento de efluentes da indústria selecionada e suas características:

***Figura 13**- Variáveis disponíveis para modelagem*

Variável	Descrição	Tipo de variável	Tipo de Medição	Frequência	Unidade
Equalização	DQO do tanque de equalização	Entrada	Instrumento -Manual	Diária	mg/L
SAO	DQO do tanque SAO	Saída	Instrumento -Manual	Diária	mg/L
Afluente	DQO no afluente industrial oleoso	Entrada	Instrumento - Manual	Esporádica	mg/L
LZAM	DQO no afluente industrial não oleoso	Entrada	Instrumento -Manual	Esporádica	mg/L

[3] Sólidos Suspensos totais

[4] Oxigênio dissolvido

Aeração	DQO no tanque de aeração	Estado	Instrumento -Manual	Diária	mg/L
Efluente Final	DQO no efluente final de descarte	Saída	Instrumento -Manual	Diária	mg/L
Vol Eq	Volume no tanque de Equalização	Entrada	Manual	Diária	m^3
Vol Aera	Volume no tanque de aeração	Controle	Manual	Diária	m^3
phEntrada	pH do Afluente industrial oleoso	Entrada	Instrumento - Auto	Horária	-
phNeutralização	ph do tanque de Neutralização	Controle	Instrumento - Manual	Horária	-
phAeração	Ph no tanque de aeração	Controle	Instrumento - Manual	Horária	-
OD	Oxigênio dissolvido no tanque de aeração	Controle	Instrumento – Manual	Horária	mg/L
VLodoSD30	Volume de Lodo sedimentado em 30 minutos	Estado	Instrumento -Manual	Diária	mL/L
DiasDescarte	Dias desde o último descarte do lodo	Controle	Manual	Diária	dias

Fonte: O autor

Todas as variáveis que apresentam (D-1) representam o valor da variável no dia anterior.

III.1.5.1. Característica das variáveis

É de extrema importância ter o conhecimento teórico e a representação físico química das variáveis e como eles interagem com o sistema como um todo. Nesta seção, será desenvolvido um pouco da característica de cada variável junto com algumas

recomendações e observações da literatura e da experiência de operação na estação de tratamento de efluentes industriais.

III.1.5.1.2. Oxigênio Dissolvido (OD)

A concentração de oxigênio dissolvido (OD) no reator biológico é de extrema importância no controle do sistema de lodos ativados. Em excesso, o oxigênio pode causar a morte de parte de população e em falta pode resultar em um fator limitante no crescimento dos microrganismos.

A faixa de concentração ideal de oxigênio dissolvido é de 1,0 a 2,0 mg/L (CLAAS, 2007). A NBR 12209 (1992), recomenda a concentração de 1,5 mg/L quando a idade do lodo for igual ou superior a 18 dias e 2,0 mg/L quando a idade do lodo for menor que 18 dias.

Class (2007) cita que teores baixos de OD podem causar perdas de massa biológica causando odores desagradáveis e variação negativa na eficiência, e valores acima de 4 mg/L prejudicam o desenvolvimento da população bacteriana e também a sedimentação do lodo.

Valores recomendados: Entre 1,5 e 3,0 mg/L.

III.1.5.1.3. Idade do lodo (Dias de descarte)

Idade do lodo é o tempo de permanência da biomassa no sistema, onde na operação em batelada pode ser considerado como o tempo decorrente entre o descarte do lodo, visto que o descarte visa retirar em sua maior parte os microrganismos com uma idade mais elevada através da sedimentabilidade. No sistema atual tem-se a variável de Dias de descarte representando a idade do lodo.

Como citado na Tabela 1, no sistema de aeração prolongada a biomassa permanece mais tempo no sistema e os tanques de aeração são maiores. Com isto, há menos DBO/DQO disponível para as bactérias, o que faz com que elas se utilizem da matéria orgânica do próprio material celular para a sua manutenção. Em decorrência, o lodo excedente retirado já sai estabilizado. No entanto, o sistema processa uma carga menor de DQO/DBO afluente pelo mesmo período de tempo no ciclo do que o sistema

convencional de lodo ativado deixando o sistema mais exposto a altos valores de DQO/DBO de entrada e podendo apresentar valores altos no efluente final, principalmente se o descarte do lodo for severo e a relação Alimento/Microrganismo for reduzida.

Valores recomendados: entre 10 e 30 dias (VON SPERLING, 1997)

III.1.5.1.5. DQO

A Demanda Química de Oxigênio, ou DQO, é um indicador que mede a matéria orgânica suscetível a oxidação química que é presente em um sistema através do oxigênio dissolvido no mesmo. Esta medição é realizada através da oxidação química da matéria orgânica utilizando o dicromato de potássio (K2Cr2O7) e quantificando a concentração reagida. O processo de análise completo leva certa de 2 a 3 horas.

A quantificação de DQO em mg/L é muito utilizada para determinar o grau de poluição de um efluente visto que ela reflete a quantidade total de componentes quimicamente oxidáveis, seja carbono, hidrocarbonetos, nitrogênio, enxofre, detergente ou fósforo. O DQO também é bastante utilizado pois o processo de análise é muito mais rápido que o DBO e seu valor é sempre superior ao DBO por oxidar também a matéria não biodegradável.

Segundo a diretriz DZ-205.R-5 1991 do estado do Rio de Janeiro, indústrias Químicas e Petroquímicas devem apresentar um DQO menor que 250 mg/L no efluente a ser lançado no corpo hídrico.

III.1.5.1.6. DBO

A Demanda Bioquímica de Oxigênio, ou DBO, é o indicador que mede a quantidade de matéria orgânica que pode ser decomposta por processos bioquímicos através do oxigênio dissolvido no efluente. A DBO é o parâmetro mais utilizado para a quantificação de poluição e no Brasil utiliza-se o indicador DBO 5,20 que representa o oxigênio consumido na degradação da matéria orgânica durante 5 dias a 20ºC, medido em mg/L.

O indicador DBO é medido através de uma amostra do efluente por inoculação em um ambiente saturado de oxigênio e frasco âmbar com uma quantidade fixa de microrganismos e nas mesmas condições (20ºC) por 5 dias, no caso do DBO 5,20. Por fim, calcula-se a diferença entre o OD inicial e o OD final da reação. Como a medição de DBO dura 5 dias o mais indicado para fim de controle e monitoramento é a DQO, que representa indiretamente a DBO e pode ser medida em apenas 2 horas.

Segundo a diretriz DZ-205.R-5 1991 do estado do Rio de Janeiro no item 5.1, indústrias com um processo biológico convencional incluindo o sistema de lodos ativados em aeração prolongada devem apresentar uma redução de DBO mínima de 90% entre o afluente e o efluente a ser lançado no corpo hídrico.

III.1.5.1.7. pH

Claas (2007) descreve que o pH representa um papel muito importante no sistema de lodos ativados e, por isso, deve ser constantemente monitorado. A faixa de pH considerada boa é de 6,0 a 8,0 com valores ótimos entre 7 e 7,5 próximo do pH neutro, valores abaixo ou acima destes causam efeitos prejudiciais ao sistema. Variações bruscas de pH podem causar efeito tóxico para os microrganismos, além de afetar as reações enzimáticas e diminuir a velocidade das reações existentes no sistema.

Como o tanque de aeração tende a apresentar naturalmente um pH levemente ácido, o pH no tanque de Equalização deve ser levemente básico para controlar bem o pH no reator biológico a cada enchimento, e recomenda-se um valor próximo de 8. Valores menores podem aumentar a acidez do tanque de aeração e valores maiores levam a basicidade o que pode causar a morte de microrganismos e prejudicar a reação.

Valores recomendados: pH ótimo é em torno de 8,0 na Equalização e entre 7 e 7,5 na Aeração.

IV. Metodologia

A ciência de dados é uma área recém desenvolvida que visa analisar os dados estruturados ou não, de maneira multidisciplinar para gerar informação de maneira a auxiliar uma tomada de decisão. O desenvolvimento de um projeto na área de ciência de dados envolve o conhecimento de diversas áreas como estatística, programação, álgebra e entendimento do negócio (HAYAHI, 1998).

Atualmente, existem diversas metodologias para se trabalhar com dados e neste projeto foi utilizada a metodologia TDSP[5] da Microsoft. Neste capítulo, serão apresentados todos os fundamentos por trás da metodologia, como ela foi aplicada no projeto e as análises estatísticas preliminares para o treinamento do modelo.

IV.1 Metodologia TDSP

A metodologia utilizada no desenvolvimento do modelo de Machine Learning segue o padrão TDSP (Processo de Ciência de Dados de Equipe) da Microsoft. O método TDSP é focado em um padrão de metodologia ágil e iterativa para ciência de dados que melhora a colaboração e aprendizagem e facilita a implementação efetiva de soluções de Machine Learning.

O TDSP faz uso de um ciclo de vida para estruturar os projetos de ciência de dados. O ciclo de vida apresenta várias etapas que atuam em conjunto de modo iterativo visando uma solução ótima para o problema. Essas etapas são:

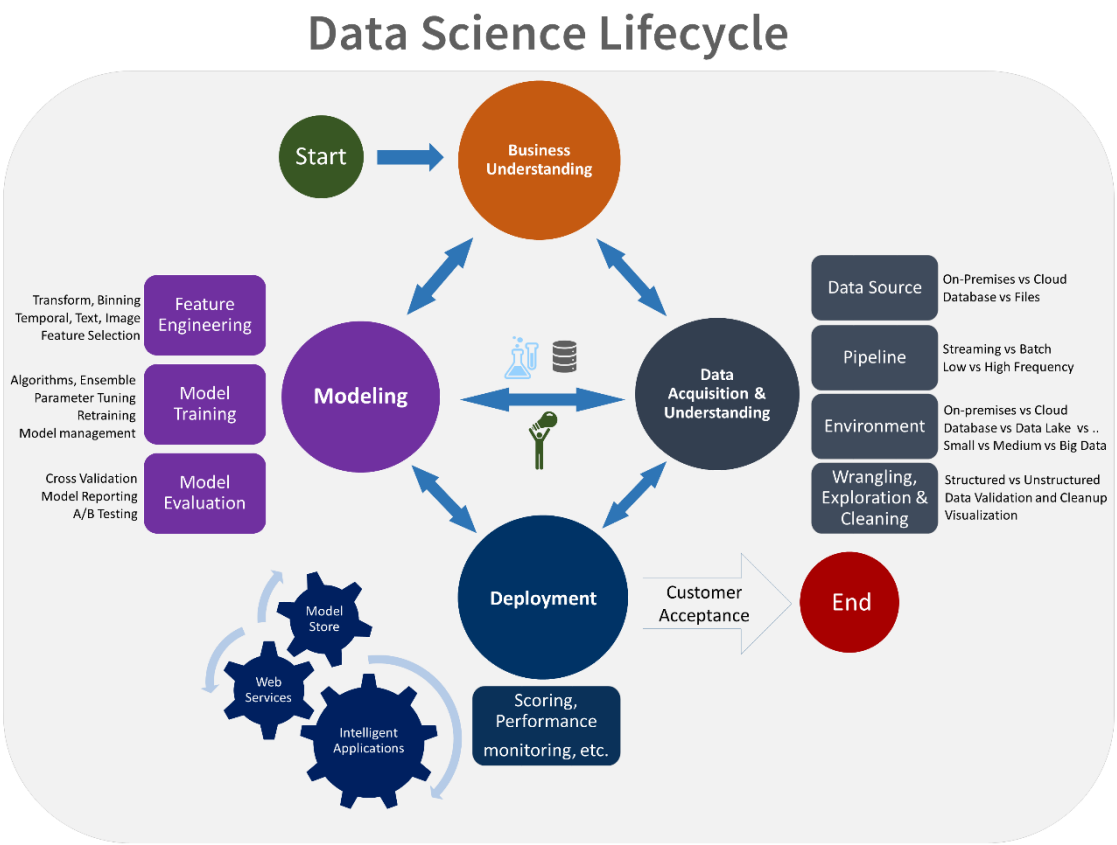
- Business Understanding - Conhecimento do negócio
- Data Acquisition and Understanding - Aquisição de dados e preparação
- Modeling - Modelagem
- Deployment - Implantação

Essa metodologia da Microsoft é focada para qualquer tipo de projeto na área de Ciência de Dados e apresenta conceitos similares a outras metodologias fundamentadas

[5] TDSP - Team Data Science Process Documentation (Processo de documentação para um tine de Ciência de Dados).

na área, como a CRISP-DM[6]. Por ter um conceito geral, a metodologia pode ser usada para trabalhos acadêmicos também, mesmo tendo sido desenvolvida com o foco em projetos comerciais.

***Figura 14**- Ciclo de Ciência de Dados - TDSP*



Fonte: Microsoft

É importante citar que o estudo foi feito para um ambiente de *Machine Learning* supervisionado, em que se possuem os dados de treino e teste previamente categorizados. O processo iterativo do ciclo seguiu do conhecimento do negócio até a modelagem, e a implantação será realizada na fábrica em outro momento.

[6] CRISP-DM é a abreviação de Cross Industry Standard Process for Data Mining, que pode ser traduzido como Processo Padrão Inter-Indústrias para Mineração de Dados. É um modelo de processo de mineração de dados que descreve abordagens comumente usadas por especialistas em mineração de dados para atacar problemas. (SHEARER, 2000)

Nos tópicos seguintes, será introduzida uma visão mais detalhada das etapas presentes na metodologia.

IV.1.1 Conhecimento do Negócio

Como primeira etapa do processo TDSP, o Business Understanding, ou Entendimento do Negócio, é fundamental para conhecer o processo e analisar todas as etapas de modelagem e predição do modelo, com a finalidade de evitar modelos enviesados, resultados que não têm sentido físico/químico ou modelos que possuam uma aplicabilidade inviável.

O processo de entendimento do negócio foi feito durante 7 meses e consistiu no embasamento teórico da operação da estação de tratamento de efluentes por meio de experiências, conhecimento compartilhado, consulta na literatura e em artigos. O processo contou também com a operação "in loco" de maneira prática com acompanhamento diário.

Todo o estudo foi realizado com o foco na viabilidade da solução. Muitos conceitos e insights gerados pelas análises descritas neste trabalho já estão sendo utilizadas pela empresa e apresentaram excelentes resultados.

IV.1.2 Aquisição e Entendimento dos Dados

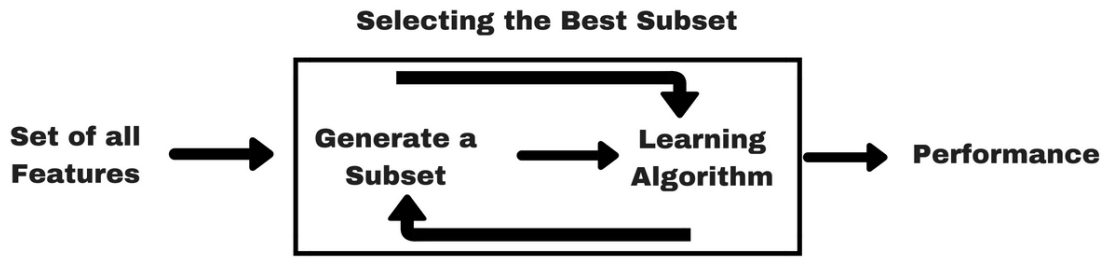
A segunda etapa de Data Acquisition and Understanding, aquisição e entendimento dos dados, consiste em diversos processos e costumam levar mais tempo porém, eles são fundamentais para garantir a qualidade de solução. Um modelo de *machine learning* é tão bom quanto os dados que são usados para alimentá-lo.

IV.1.2.1 Feature Selection

O processo de Feature Selection, ou seleção de recursos, é uma etapa que visa selecionar um subconjunto principal de recursos de dados originais para tentar reduzir a dimensionalidade do problema de treinamento. O tamanho da dimensionalidade é um

parâmetro muito importante para adequar a dimensionalidade do seu problema com a quantidade de dados de treino disponível para o modelo. A seleção de recursos pode ser feita usando a experiência de processo da etapa anterior, além da estatística com a covariância entre os dados e os dados da literatura entre outros métodos. O processo de feature selection é iterativo em que se seleciona um conjunto de variáveis, treina-se e se valida o modelo, e dependendo do resultado, retorna-se à etapa de seleção novamente. Em muitos casos tem-se que a inserção ou remoção de uma variável melhora ou piora o desempenho do modelo e é sempre válida uma conferência de resultados do modelo após uma iteração de seleção de variáveis.

Figura 15 *- Processo de Feature Selection*



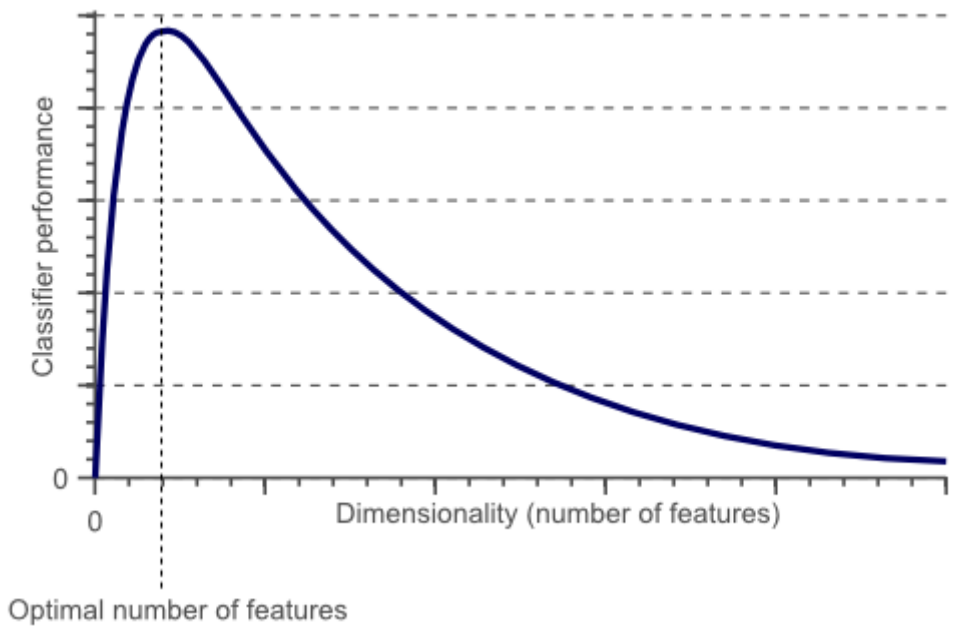
Fonte: Analytics Vidhya

Quando seu conjunto de dados apresenta uma dimensionalidade muito elevada, normalmente acima de 100 variáveis, ou com poucos dados, pode ocorrer o fenômeno conhecido como a maldição da dimensionalidade.

IV.1.2.1.1 Maldição da dimensionalidade

A maldição da dimensionalidade refere-se a vários fenômenos que surgem ao analisar e organizar dados em espaços de alta dimensão que não ocorrem em ambientes de baixa dimensão. A expressão foi cunhada por Richard E. Bellman ao considerar problemas na otimização dinâmica. Em outras palavras, a maldição da dimensionalidade cita que a quantidade de dados necessária para uma boa generalização aumenta exponencialmente com o número de variáveis no modelo. Pode-se ver abaixo o decaimento da performance de um modelo de classificação com o número de variáveis no modelo e um valor constante de dados.

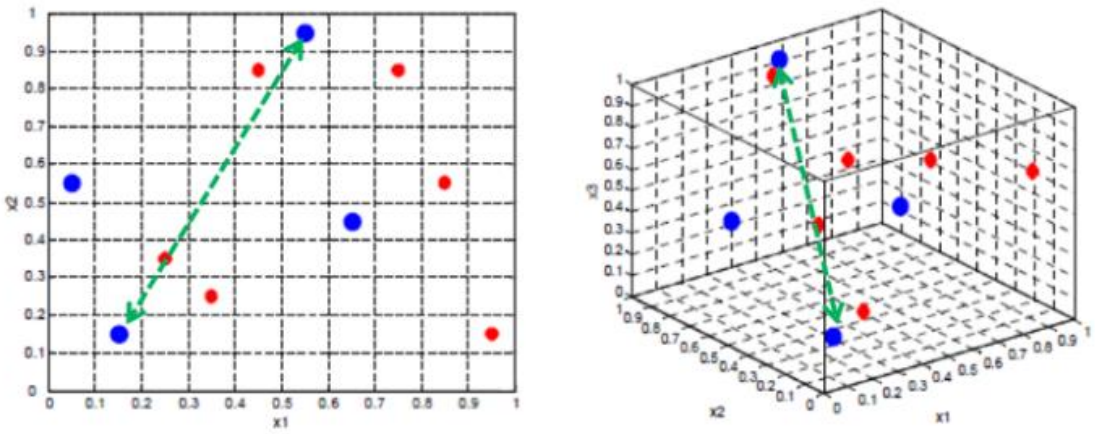
***Figura 16**- Maldição da dimensionalidade*



Fonte: Facom UFU

O fato em comum desse problema é que, quando a dimensionalidade aumenta, o volume do espaço aumenta tão rapidamente que os dados disponíveis se tornam esparsos. Essa dispersão é problemática para qualquer método que exija significância estatística. Além disso, modelar dados geralmente depende da detecção de áreas onde os objetos formam grupos com propriedades semelhantes, como no k-NN ou um kMeans. Em dados dimensionais elevados, todos os objetos parecem ser esparsos e desiguais, o que inviabiliza essas estratégias comuns de organização de dados. Na figura a seguir tem-se uma abstração do aumento de um espaço de dimensão 2D para 3D mantendo a mesma quantidade de dados e a modelagem visual da distância e distribuição de espaço entre os dados.

Figura 17 *- Aumento da dimensionalidade - 2D para 3D*

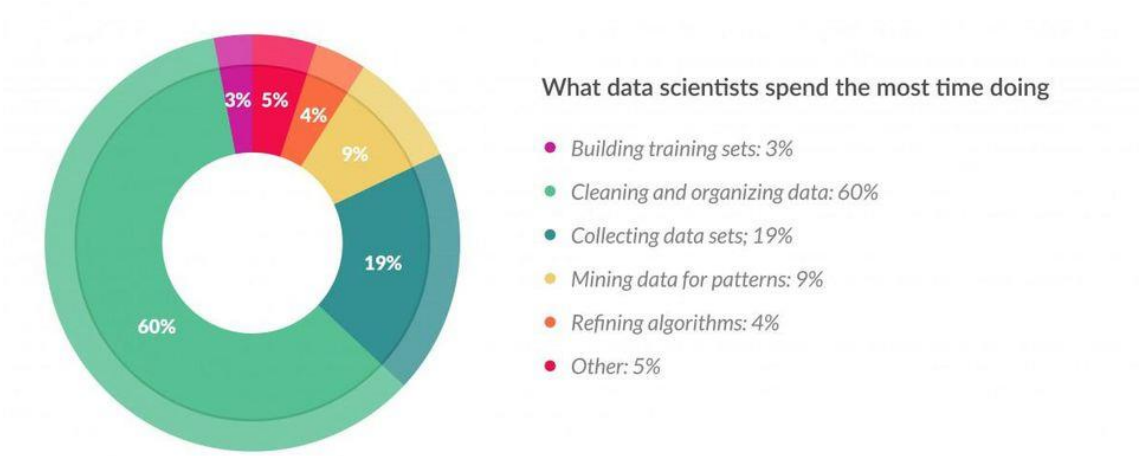


Fonte: Facom UFU

IV.1.2.2 Limpeza dos dados

O artigo da Forbes de 2016 cita que 80% de todo o trabalho de um cientista de dados é gasto na preparação dos dados e dentro disto 60% somente com limpeza dos dados. Quando se inicia um trabalho de ciência de dados, é comum se deparar com um dataset cheio de problemas operacionais, seja na entrada deste dados no sistema com um erro humano ou em algum erro do próprio sistema. Dentre os problemas mais recorrentes nos dados, tem-se dados com valores nulos, dados com o tipo errado, dados inseridos sem nenhuma precisão ou conferência do usuário, dados com valores longe do esperado etc. Sendo assim, a etapa de limpeza dos dados se inicia logo após o entendimento do negócio e se estende até a modelagem e suas interações.

***Figura 18**- Contabilização de esforços em projetos de Ciência de Dados*



Fonte: Forbes (2016)

V. Model Training

Nesta seção será abordado o conceito de machine learning e sua aplicabilidade na área de controle e monitoramento de processos além da explicação da metodologia utilizada e dos tipos de algoritimos, modos de treino e tipos de avaliação.

A parte de treino do modelo é uma das etapas mais importantes do processo iterativo, e uma boa escolha dos modelos iniciais de treino deve ser feita para ganhar eficiência e direcionar o projeto para o caminho mais assertivo e fiel ao processo real. Para isto, a experiência, bom senso e apoio da literatura são fundamentais.

O objetivo do treino é obter um modelo de machine learning adequado ao processo que consiga prever os casos de DQO fora dos limites legais com uma fidelidade satisfatória para ser implementado como uma ferramenta de processo que auxiliará a tomada de decisão dos funcionários da fábrica.

V.1 Conceito de Machine Learning

O Machine Learning, ou aprendizado de máquina, é o estudo de algoritmos e modelos estatísticos que os sistemas computacionais utilizam para melhorar

progressivamente seu desempenho em uma tarefa previamente especificada. Algoritmos de machine learning constroem um modelo matemático a partir de dados de amostra, conhecido como "dados de treino", para fazer previsões ou decisões sem ser explicitamente programado por um desenvolvedor.

Em outras palavras, os algoritmos de machine learning usam da matemática, estatística e computação para modelar os dados fornecidos como entrada do algoritmo e generalizá-los com o menor erro possível com o objetivo de construir modelos preditivos, classificatórios ou agregativos. No caso de modelagem de processos industriais, os algoritmos são comumente usados para modelar o processo e fornecer variáveis de saída condizentes com o processo real dado um conjunto de dados de entrada e um perfil de treino. Em casos mais específicos como os de saída de processo, os algoritmos de machine learning podem ser usados para detectar anomalias no funcionamento do processo, no caso do tratamento dos efluentes em ordem de predizer um valor futuro desenquadrado antes do lançamento e tomar a decisão de realizar ou não a análise laboratorial mais detalhada.

V.2 Modelos de Machine Learning

Tem-se hoje uma gama enorme de algoritmos disponíveis para serem usados em modelos de predição com diferentes tipos de abordagens e resultados. Desenvolver uma aplicação de machine learning em um conjunto de dados se trata de modelar um processo com combinações de equações matemáticas e estatísticas a fim de obter um erro aceitável entre a modelagem e a realidade, para no fim este modelo se traduzir em uma capacidade preditiva aceitável de um futuro que não é totalmente incerto, onde as variáveis de processo não se diferem tanto do passado. Em outras palavras, os modelos de machine learning bem treinados têm plena capacidade de prever o futuro desde que este futuro não esteja muito diferente do passado, caso as variáveis de processo apresentem uma grande mudança o modelo deve ser treinado novamente para continuar apresentando uma validade aceitável.

V.2.1 Logistic Regression

A regressão logística, ou logistic regression, é um algoritmo estatístico que visa predizer valores categóricos normalmente binários a partir de um conjunto de observação. Devido à natureza do modelo tem-se uma saída de probabilidade entre 0 e 1 e que de acordo com um corte previamente definido, normalmente de 0,5, um dos valores categóricos binários é selecionado na predição.

O modelo de regressão logística é amplamente utilizado em modelos de Machine Learning para predição de variáveis categóricas e possui uma boa capacidade de interpretação por apresentar um valor de probabilidade associado as categorias da predição. Este modelo normalmente é bem útil para modelar a probabilidade de um evento ocorrer por dependência de outros fatores, assim como o modelo Bayesiano e faz parte do grupo de Modelos Lineares Generalizados (em inglês GLM) com uma função de ligação logit.

Devido ao número pequeno de casos não conformes na amostra de dados e a não dependência de uma alta precisão na predição na medição de DQO de saída do efluente da estação de tratamento pode-se usar um modelo de predição categórica para a predição. Neste modelo, verificado se o DQO de saída está desenquadrado (>= 250 mg/L) ou enquadrado (< 250mg/L). Muitos casos como este são binários (1 ou 0) e tem uma distribuição de Y especificada por probabilidades $P(Y=1) = \pi$ de sucesso e $P(Y=0) = (1-\pi)$ de falha onde Y é a categoria do DQO de saída, se Y=1 o DQO está desenquadrado e Y=0 o DQO está enquadrado na legislação e x é o vetor que representa as variáveis de entrada do modelo. Na maioria dos casos a relação entre P(Y) e x é não linear e uma mudança em x pode ter menos impacto quando P(Y) está próximo de 0 ou 1 do que quando P(Y) está próximo a média. (Fonte: livro *An Introduction to Categorical Data Analysis*)

Como se trata de um modelo de Machine Learning Supervisionado, deve-se rotular todos os dados de treino e categorizá-los entre sucesso e falha. Após categorizados estes dados são dispostos na distribuição de Bernoulli que é a distribuição discreta das variáveis em um espaço amostral {0,1} a fim de se obter a probabilidade de cada valor da distribuição. Estas probabilidades são inicialmente desconhecidas e deve-se usar um método para obtê-las de acordo com as suas devidas categorias. (Fonte: livro *"Binomial distribution", Encyclopedia of Mathematics*)

Sendo assim, o método de Regressão Logística utiliza-se da função logit, onde π é a probabilidade, x_i são as variáveis do modelo e β_i são os parâmetros de ajuste da função, que são habitualmente estimados através do método da máxima verossimilhança.

***Equação 1** – Logit (Logistic Regression)*

$$\operatorname{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

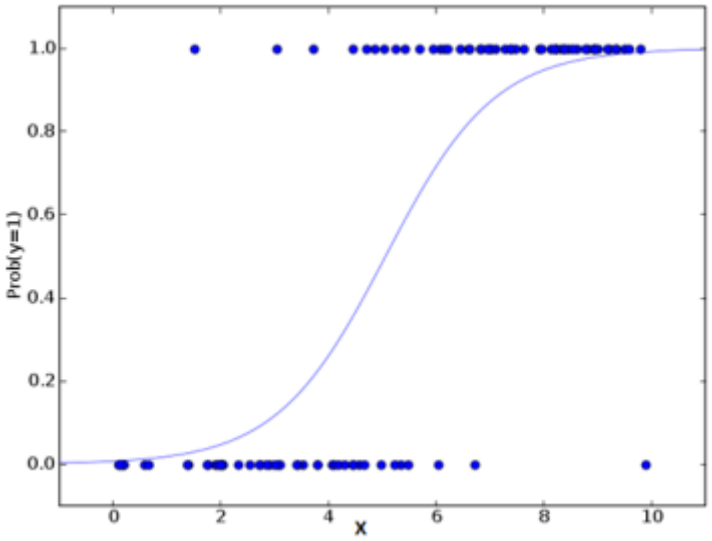
Assim, tem-se a probabilidade através da dedução abaixo.

***Equação 2** - Probabilidade Logistic Regression*

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}.$$

Vale ressaltar que esta função de probabilidade é normalmente identificada como um perceptron[7] de camada simples, ou seja, um modelo de rede neural de uma só camada. E por fim, tem-se a representação gráfica de uma modelagem por logit abaixo:

***Figura 19**- Distribuição da função probabilidade - Logistic Regression*



Fonte: Analyticsvidhya

[7] O perceptron é o tipo mais simples de rede neural artificial feedfoward inventada em 1957 por Frank Rosenblatt no Cornell Aeronautical Laboratory. (Fonte: Wikipedia)

V.2.2 Random Forest

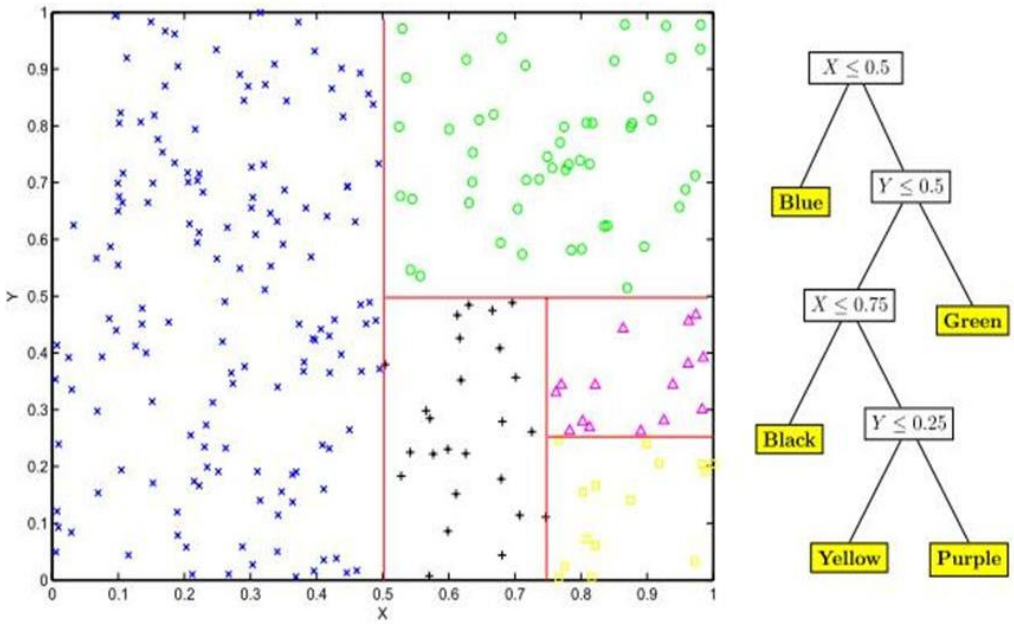
O modelo de Random Forest ou Random Decision Forest, em português Floresta de Árvores de decisão, é um modelo da classe dos Ensemble Learning Models usado para classificação ou regressão. A classe dos Ensemble Learning engloba modelos individualmente treinados (como árvores de decisão ou redes neurais) que são combinados para classificar da melhor maneira novas instâncias. Os modelos combinados (Ensemble) costumam ser mais acurados do que seus modelos individuais (OPITZ,1999).

Árvores de Decisão

Os algoritmos baseados em árvores de decisão, ou Decision Tree, são modelos já estabelecidos na área de Machine Learning. O modelo de árvore de decisão é bastante utilizado por ser invariante sob escalonamento e transformação de variáveis, robusto sob inclusão de variáveis irrelevantes e produz modelos facilmente inspecionáveis, no entanto, estes modelos são raramente precisos (HASTIE, 2008).

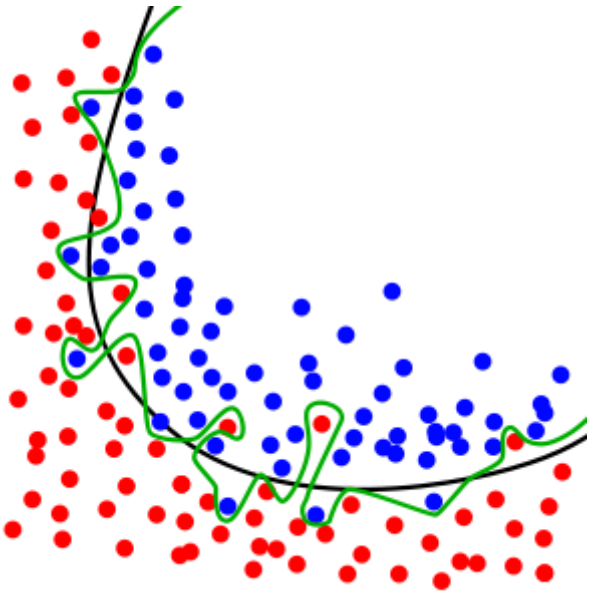
As árvores de decisão são modelos que visam traçar limites no espaço amostral partindo de regras que melhor separam e categorizam os dados por ordem decrescente de relevância do topo para o fundo da árvore. Este comportamento pode ser melhor visto na imagem descrita a seguir.

Figura 20 *- Indução de árvores de decisão*



Modelos de árvore de decisão, particularmente as árvores que são muito longas, tendem a aprender muitos padrões irregulares e causam um overfitting (sobre ajuste) nos dados de treino. Este overfitting reduz o viés mas apresentam uma alta variância que reduz a precisão na previsão de dados novos. Para isto, o modelo combinado de Random Forest visa obter uma média entre várias árvores longas que reduzem a variância, aumentam um pouco o viés mas resultam em uma performance final melhor do modelo.

***Figura 21**- Conceito visual de overfitting*



Fonte: Wikipedia

V.3 Treinamento e Validação

O treinamento supervisionado de um modelo de Machine Learning visa generalizar da melhor forma o modelo baseando-se nos dados previamente categorizados. Para ter uma generalização em um ambiente similar ao ambiente real de predição, deve-se separar o conjunto de dados entre conjunto de treino e conjunto de teste, a fim de medir a eficiência da predição e ajustar os parâmetros do modelo se necessário, conhecido como método de validação cruzada.

V.3.1 K-fold

Um dos métodos mais utilizados é o k-fold que divide o conjunto total em k subconjuntos de mesmo tamanho, onde um subconjunto é utilizado para testes e os k-1 conjuntos restantes para treino. Esta abordagem é realizada de forma iterativa por k vezes abrangendo todos os subconjuntos e gerando métricas como acurácia e precisão (KOHAVI, 1995).

V.3.2 Leave one out

O modelo leave-one-out é uma variação mais robusta do método k-fold onde k é igual a todos os dados N da amostra. Neste método são realizados N=k analises iterativas onde um simples valor é deixado para testes e N-1 são usados para treino. O método de leave-one-out tem a vantagem de percorrer todo o conjunto de dados e investigar toda a variação da amostra porém o método apresenta um elevado custo computacional e apresenta maiores ganhos para conjuntos com poucos dados (KOHAVI, 1995).

Todos os treinos foram feitos no regime Leave one out que favorece dados desbalanceados e com datasets pequenos, característica do dataset utilizado nesse trabalho.

V.4 Métricas de Avaliação

Algumas métricas são usadas para avaliar os modelos além da acurácia, como a especificidade e a sensitividade, o F1 score, a precisão e a curva ROC[8] ou AUC[9], que serão melhor explicados nesta seção. A definição dos parâmetros que serão usados na

[8]ROC - Receiver Operating Characteristic Curve (Curva característica de operação do receptor)

[9] AUC – Area under the curve – Área representada pela curva ROC

avaliação do modelo com suas justificativas são extremamente importantes para a implementação do modelo em casos reais.

V.4.1 Sensitividade, Especificidade e Precisão

Sensitividade se traduz em geral na capacidade de predizer um resultado positivo quando este resultado realmente é positivo, um verdadeiro positivo. Como um algoritmo que sempre prevê positivo acerta todos os casos de positividade e, sendo assim possui sensitividade perfeita 1, não se deve avaliar esta métrica isoladamente e avaliamos ela em conjunto com a especificidade. Especificidade é a habilidade de um modelo não prever positivo quando o resultado é negativo. Em resumo, alta sensibilidade é a previsão de verdadeiros positivos e alta especificidade é a previsão de verdadeiros negativos. Logo a matriz de confusão pode ser interpretada da seguinte maneira.

***Tabela 3** - Matriz de Confusão - Sensitividade e Especificidade*

	Resultado Positivo	Resultado Negativo
Previsão Positiva	Verdadeiro Positivo (VP)	Falso positivo (FP)
Previsão Negativa	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Com isto pode-se unificar os dois conceitos e se chega em:

***Equação 3** – Cálculo da Sensitividade*

$$Sensitividade = \frac{VP}{(VP + FN)} = Recall$$

Esta medida é também chamada de Taxa de verdadeiro positivo, ou Recall que consiste na primeira coluna da matriz. Ao mesmo tempo se tem a outra medida abaixo:

Equação 4- *Cálculo da Especificidade*

$$Especificidade = \frac{VN}{(VN + FP)}$$

E esta medida também é conhecida como Taxa de verdadeiro negativo, que demonstra a segunda coluna da matriz.

Outra maneira de determinarmos a Especificidade é conhecida como o cálculo de Precisão, que traduz a primeira linha da matriz:

Equação 5 *- Cálculo de Precisão*

$$Precisão = \frac{VP}{(VP + FP)}$$

Neste caso, ao contrário das medidas de taxas de verdadeiro positivo e verdadeiro negativo, a precisão depende da prevalência dos dados, logo alta prevalência gera uma alta precisão mesmo quando está adivinhando os resultados.

Ao analisar o caso em questão, tem-se como parâmetro mais importante a especificidade do modelo e tem-se como prioridade sua maximização com a redução dos casos de falso positivo, visto que predizer um falso negativo leva a considerar uma amostra fora dos limites da legislação como uma amostra dentro dos limites, o que acaba sendo infinitamente mais prejudicial do que prever uma amostra fora dos limites quando esta está dentro do padrão, logo devemos visar a especificidade e a precisão no modelo.

V.4.2 F1 Score

O recomendado é que sempre seja estudado especificidade e sensitividade juntos, porém em casos de otimização e avaliação usa-se em larga escala medidas únicas como a média entre as duas, chamado de Acurácia Balanceada ou Balanced Accuracy ou uma outra medida mais popular chamada de F1 Score, que consiste na média harmônica da especificidade e sensitividade.

Equação 6 *- Cálculo do F1 Score*

$$F1\ Score = \frac{1}{\frac{1}{2}(\frac{1}{recall} + \frac{1}{precisão})}$$

A escolha entre a maximização de uma das medidas depende muito do caso que é avaliado. Por exemplo, em um caso de segurança na aviação é muito mais importante maximizar a sensitividade pois falhar na previsão de um mal funcionamento antes do avião quebrar é muito mais custoso do que forçar um pouso em perfeitas condições. Porém, em um caso de julgamento criminal de assassinato o inverso é preferível, maximizar a especificidade é melhor pois um falso positivo pode levar a morte de um inocente. Para a otimização do F1 Score utiliza-se a letra grega beta que representa a importância da sensitividade comparada com a especificidade neste fórmula:

Equação 7*- Cálculo do F1 Score 2*

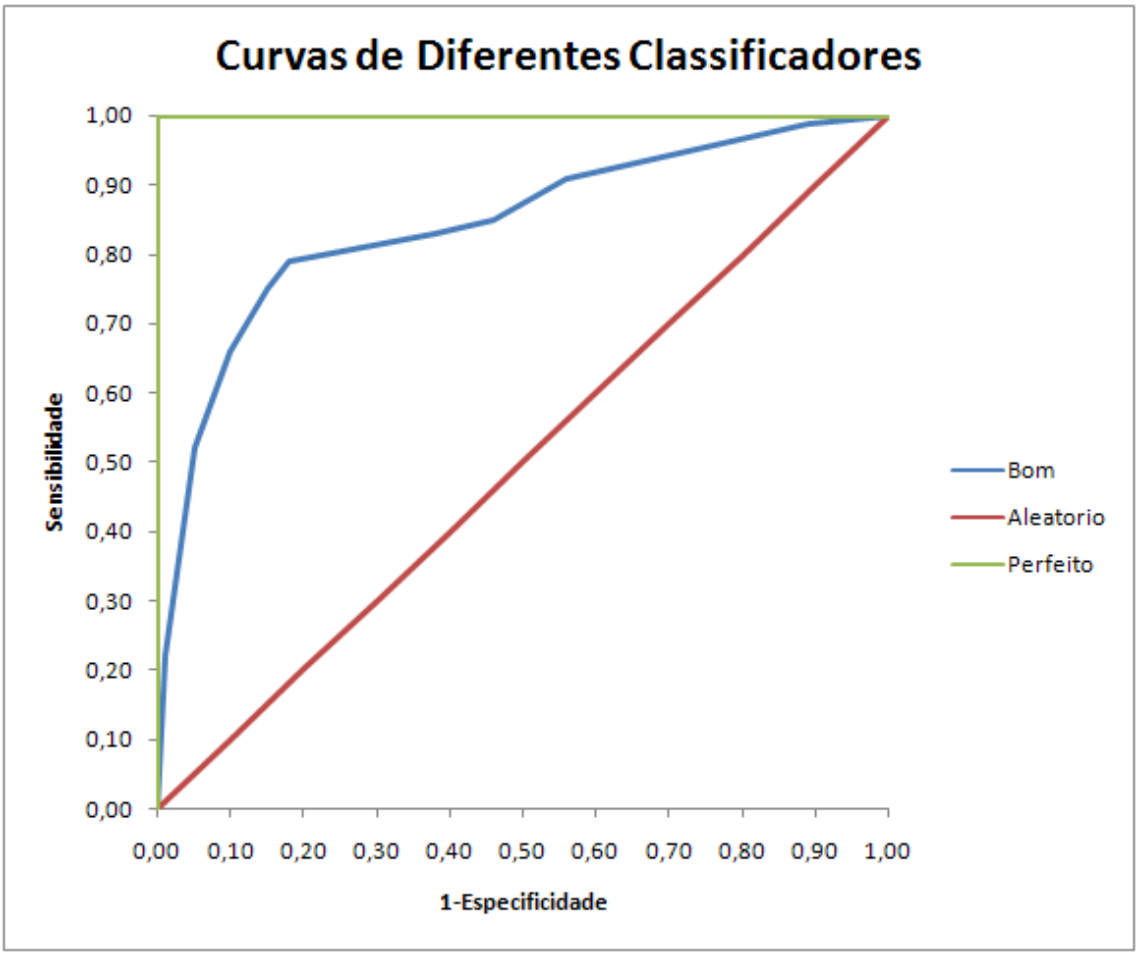
$$F1\ Score = \frac{1}{\frac{\beta^2}{1+\ \beta^2} * \frac{1}{recall} + \frac{1}{1+\ \beta^2} * \frac{1}{precisão}}$$

V.4.3 Curva ROC e AUC

Outro parâmetro bastante usado compara os diferentes casos de especificidade e sensitividade graficamente é conhecido como Receiver Operating Characteristic, ou curva ROC. Esta curva nos mostra o quão bem o modelo pode distinguir casos de classificação binária, como positivo e negativo e o AUC (area under the ROC curve) é uma forma de quantificar o gráfico calculando a área sob a curva ROC. Este valor vai de 0 a 1 onde 1 é o caso ideal onde o modelo possui previsões 100% corretas e 0.5 se traduz como o modelo resultando similar a uma adivinhação, visto que 50% das precisões são corretas. A curva ROC plota a sensitividade sobre 1 – especificidade, ou a taxa de verdadeiro positivo sobre a taxa de falso positivo.

Após a introdução dos parâmetros de avaliação de um modelo de machine learning com o uso de um modelo simples pode-se fazer uso de modelos mais complexos a fim de desenvolver um modelo com alta acurácia e alta especificidade.

***Figura 22**- Curva ROC (AUC)*



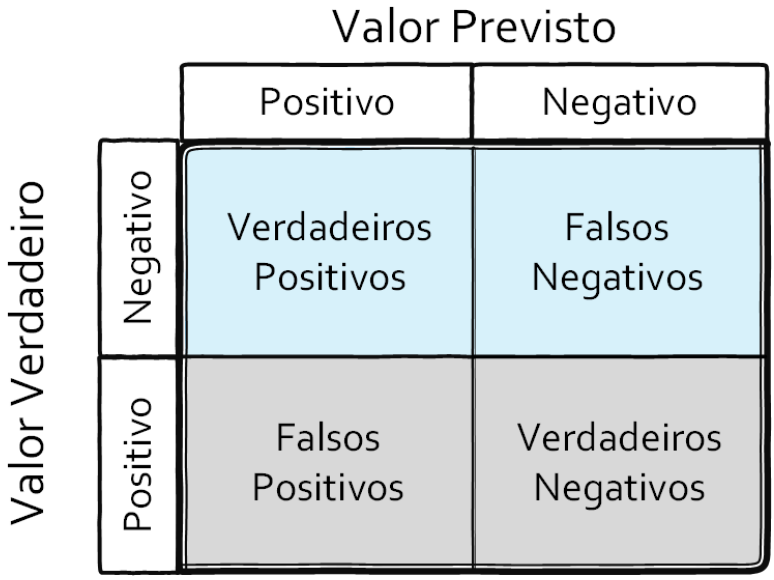
Fonte: Cezar Souza

V.4.4 Matriz de Confusão

Um dos indicadores mais utilizados em problemas de classificação com algoritmos de Machine Learning é a Matriz de Confusão. A matriz de confusão é apresentada de forma tabular e tem normalmente como colunas o valor previsto pelo algoritmo e como linhas o valor verdadeiro do dataset, com isso tem-se:

- Verdadeiros Positivos que mostra o quanto de classe Positivo o algoritmo acertou.
- Verdadeiros Negativos mostra o quanto de classe Negativo que o algoritmo acertou.
- Falsos Positivos e Falsos Negativos mostram o quanto que o algoritmo errou.

***Figura 23**- Matriz de Confusão*



Fonte: Medium

V.4.5 Information Gain – Fisher Information

O conceito de ganho de informação, ou Information Gain, elaborado por Fisher combina valores extremos de probabilidade de cada teste de hipótese (p-value) para cada variável em um teste estatístico (X^2) (FISHER, 1955).

***Equação 8** - Information Gain*

$$X^2 = -2 \sum_{i=1}^{k} \ln(pi)$$

Onde pi = p-value para cada ith teste de hipótese.

- Quando p-value é pequeno = X^2 é grande.
- Dependência por testes estatísticos.
- Quando X^2 é muito pequeno não há dependência significativa entre as variáveis.

Teste de hipóteses, teste estatístico ou teste de significância é um procedimento estatístico que permite tomar uma decisão (aceitar ou rejeitar a hipótese nula H_0) entre duas ou mais hipóteses (hipótese nula H_0 ou hipótese alternativa H_1), utilizando os dados observados de um determinado experimento. Há diversos métodos para realizar o teste de

hipóteses, dos quais se destacam o método de Fisher (teste de significância), o método de Neyman–Pearson e o método de Bayes.

O conceito de Fisher Information é às vezes denominado Information Gain, porque representa a quantidade de informação que uma variável fornece sobre algum parâmetro desconhecido do qual ela depende.

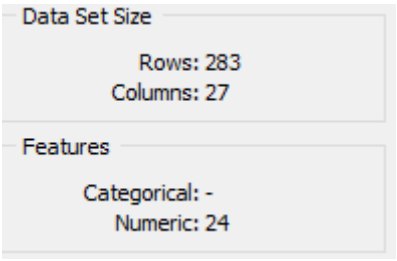
O valor numérico é calculado medindo a variância entre o valor esperado da informação e o valor observado. Quando a variação é minimizada, a informação é maximizada. Como a expectativa do score é zero, a informação de Fisher é também a variância do score.

VI. Resultados

Após todo o processo de tratamento e limpeza dos dados ser realizado o data set passou pelo processo de modelagem e avaliação dos modelos de machine learning, onde os resultados estão presentes neste capítulo.

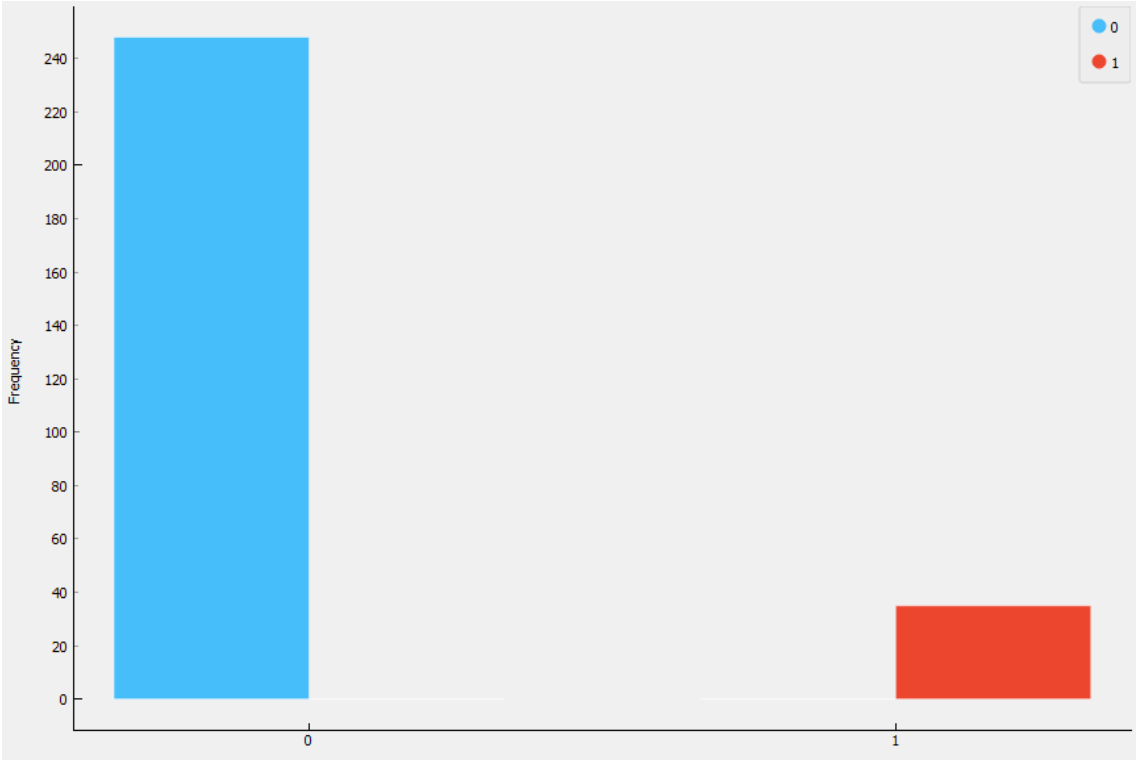
O tamanho inicial do data set limpo é de 283 registros, o que representa quase um ano de dados e 27 variáveis.

Figura 24 *– Características do Data set*



Ao conferir a frequência da variável resposta nota-se um grande desbalanceamento entre as os valores de não conformidade (1) e conformidade (0) de DQO. Esta não conformidade é natural e benéfico para o processo, visto que o objetivo da empresa é atingir zero casos de não conformidade. No entanto, este desbalanceamento nos dados com 241/31 é ruim para a modelagem pois tende a deixar o algoritmo de *machine learning* enviesado e sem saber definir bem a categoria com menos amostras.

Figura 25 *- Distribuição da frequência na resposta categórica*



A escolha dos modelos para treinamento foram modelos mais simples e robustos para ter uma boa rastreabilidade e aplicabilidade no processo, além da pouca quantidade de dados disponível e o data set muito desbalanceado. Os modelos escolhidos para as análises foram: Logistic Regression e Random Forest.

O modelo Random Forest foi usado com hiperparâmetros de 10 árvores com 5 folhas cada uma delas e o Logistic Regression foi usado com a regularização de Ridge (L2) com o parâmetro C=1.

O objetivo é selecionar modelos com as melhores métricas de avaliação ao mesmo tempo em que preveem com mais precisão a variável de não conformidade (1) e erram menos casos de conformidade, pois um Falso negativo é muito mais prejudicial do que um Falso Positivo neste modelo.

Após o feature selection também foram selecionadas as variáveis que irão fazer parte do modelo final. Como o intuito do modelo é realizar uma previsão do valor de DQO do efluente final no dia do lançamento a maioria das variáveis vem no regime D-1 pois a transferência é realizada no dia anterior. Essas variáveis são:

- Eq D-1

- Efluente D-1
- phNeutralização D-1
- pHEqualização D-1
- pHAeração D-1
- OD D-1
- pHEqualização
- pHAeração
- OD
- VLodoSD30
- VAeração
- VEqualização
- DiasDescarte
- pHEntrada D-1

Por fim, o treinamento e avaliação foi feito em código Python utilizando algumas bibliotecas como Scikit-learn, Orange, Matplotlib, Pandas, Numpy entre outras.

VI.1. Análise Estatística

A abordagem ideal é selecionar as variáveis de processo que tenham sentido físico e químico no processo de tratamento de efluentes e ao mesmo tempo estejam disponíveis no dia da tomada de decisão. Para isso foi realizada uma análise estatística a fim de entender a distribuição das variáveis e sua correlação com os dados de saída da estação, no caso o DQO do efluente final.

O processo em batelada começa no dia anterior (D-1) ao dia do descarte e finaliza no dia/horário do descarte. Sendo assim, a transferência é feita no dia anterior com as variáveis de processo em (D-1).

As variáveis de processo foram submetidas a uma análise estatística por meio de distribuições de histograma, box plot e correlações de Pearson agrupadas pela variável categórica de não conformidade indicadas em vermelho e em azul os casos de conformidade.

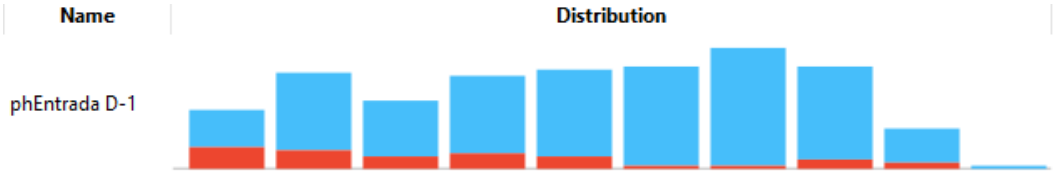
VI.1.1. pH de entrada

Como explicitado anteriormente, o pH de entrada representa o pH no afluente oleoso que deságua continuamente no tanque API.

Com a análise dos dados disponíveis nas Figuras 19 e 20, observa-se que o pH está centralizado em 6,62, o que é aceitável como média, perto de 7,0. O valor mínimo de pH foi de 0,4 o que é muito abaixo do ideal e o máximo de 12,7 muito mais alto, mostrando uma variabilidade grande para o pH de entrada do afluente.

Pode-se notar que a distribuição do pH nos casos de não conformidade é muito maior para os pHs baixos e nos casos de conformidade, esta distribuição é equilibrada. Conclui-se que o pH baixo no afluente está diretamente ligado à não conformidade do DQO no efluente final.

***Figura 26**- Histograma do pH de entrada*



Conformidade e Não Conformidade

Fonte: O autor

***Figura 27**- Estatísticas do pH de entrada*

Center	Dispersion	Min.	Max.
6.62	0.48	0.40	12.70

Fonte: O autor

Pela correlação de Pearson, observa-se uma correlação do inversamente proporcional de -0,127 do pH de entrada com o DQO do efluente final, visto que um menor pH de entrada resulta em um Efluente Final com um maior DQO.

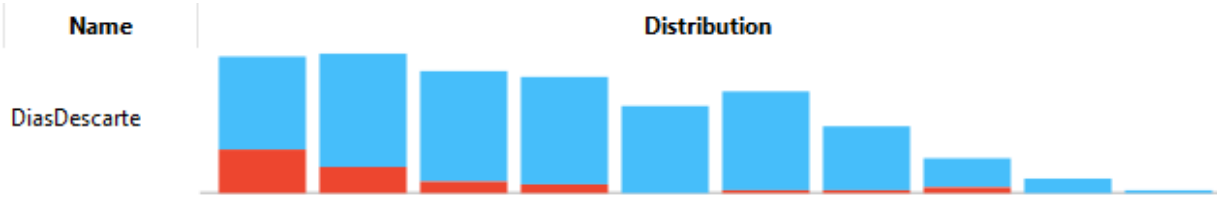
Observa-se também que em 10% dos casos não se tem o valor de pH do dia anterior. Em muitos casos isso se dá ao fato de domingo não haver operação na ETE e assim, segunda não apresenta valor.

VI.1.1.2 Dias de descarte

Idealmente, os dias de descarte de lodo devem ir até 40 dias no máximo e o indicador mostra quantos dias se passaram desde o último descarte. A distribuição em ambientes de não conformidade é levemente similar em ambientes de conformidade. Porém em ambientes de não conformidade temos uma alta predominância quando descarte do efluente é feito no dia ou a poucos dias do descarte do lodo. Uma das causas deste fato pode indicar um descarte de lodo em um volume maior que o ideal, reduzindo bastante a população bacteriana e deixando exposta a uma alta carga de DQO de entrada. A variação de população pode ser verificada em testes laboratoriais de SSV[10].

O máximo está em 34 dias dentro dos 40 dias ideais, o que mostra que o lodo tem sido descartado dentro dos limites propostos.

***Figura 28**- Distribuição dos Dias de descarte*



Conformidade e Não Conformidade

[10] SSV – Sólidos em Suspensão Voláteis – Um dos indicadores da população bacteriana presente no tanque biológico

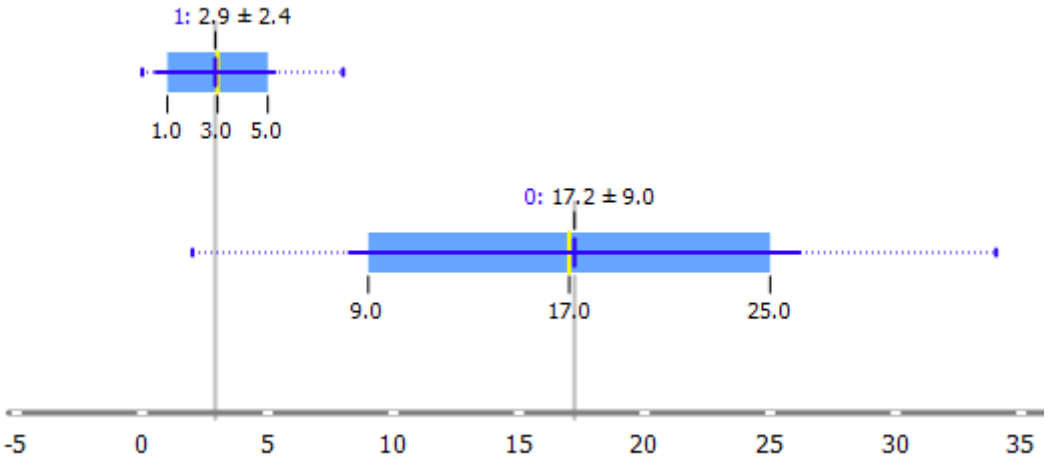
***Figura 29**- Estatísticas dos Dias de descarte*

Center	Dispersion	Min.	Max.
12.67	0.66	0.00	34.00

Pela correlação de Pearson, observa-se uma certa correlação inversamente proporcional de -0,303, visto que menos dias desde o descarte resulta em um Efluente Final com um maior DQO. Esta correlação também é mais que o dobro do pH de entrada, mostrando inicialmente ter um maior impacto no tratamento.

A avaliação de box plot mostra claramente uma concentração dos valores com não conformidade de DQO (1) entre 1 a 5 dias do descarte de lodo e valores conformes (0) entre 9 a 25 dias.

***Figura 30**- Box plot dos Dias de descarte*



O eixo x representa o número de dias, os valores 1 e 0 mostra a separação entre os dois box plot dos dados categorizados conformes (0) e não conformes (1) e tem-se os valores de média em amarelo +- 2 desvio padrão.

Fonte : O autor

VI.1.1.3 Oxigênio Dissolvido (OD)

O oxigênio dissolvido está igualmente distribuído para os dois casos tendo diferença somente em casos de OD muito baixo, o que causa mortalidade dos microrganismos e leva ao aumento no DQO do efluente final levando à não conformidade. No geral, os valores estão centrados em 2,8 o que é levemente acima do recomendado de 2,5 mg/L. O valor máximo ficou 3,42 um pouco acima e o mínimo com 1,48 um pouco abaixo do recomendado. Pela análise dos valores, observa-se que a operação pode ser ajustada para valores um pouco acima do recomendado pois tende a ser benéfico para o tratamento.

***Figura 31**- Distribuição do Oxigênio Dissolvido*

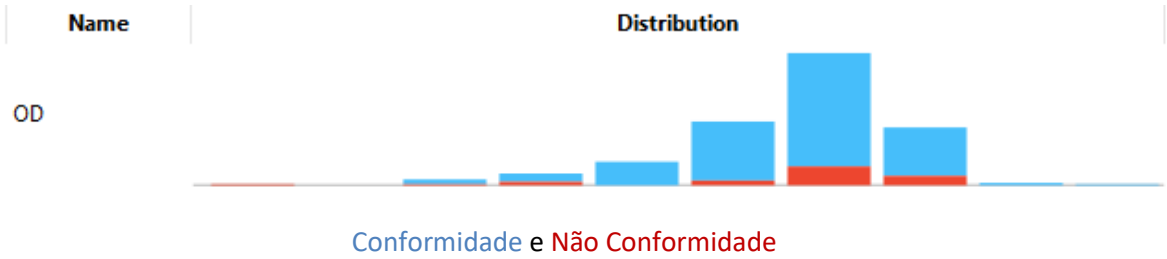


***Figura 32**- Estatísticas do Oxigênio Dissolvido*

Center	Dispersion	Min.	Max.
2.80	0.09	1.48	3.42

Por fim, o coeficiente de correlação de Pearson apresentou valores muito baixos de correlação, +0,06 o que mostra uma pouca influência com o Efluente final para a modelagem de acordo com os dados disponíveis.

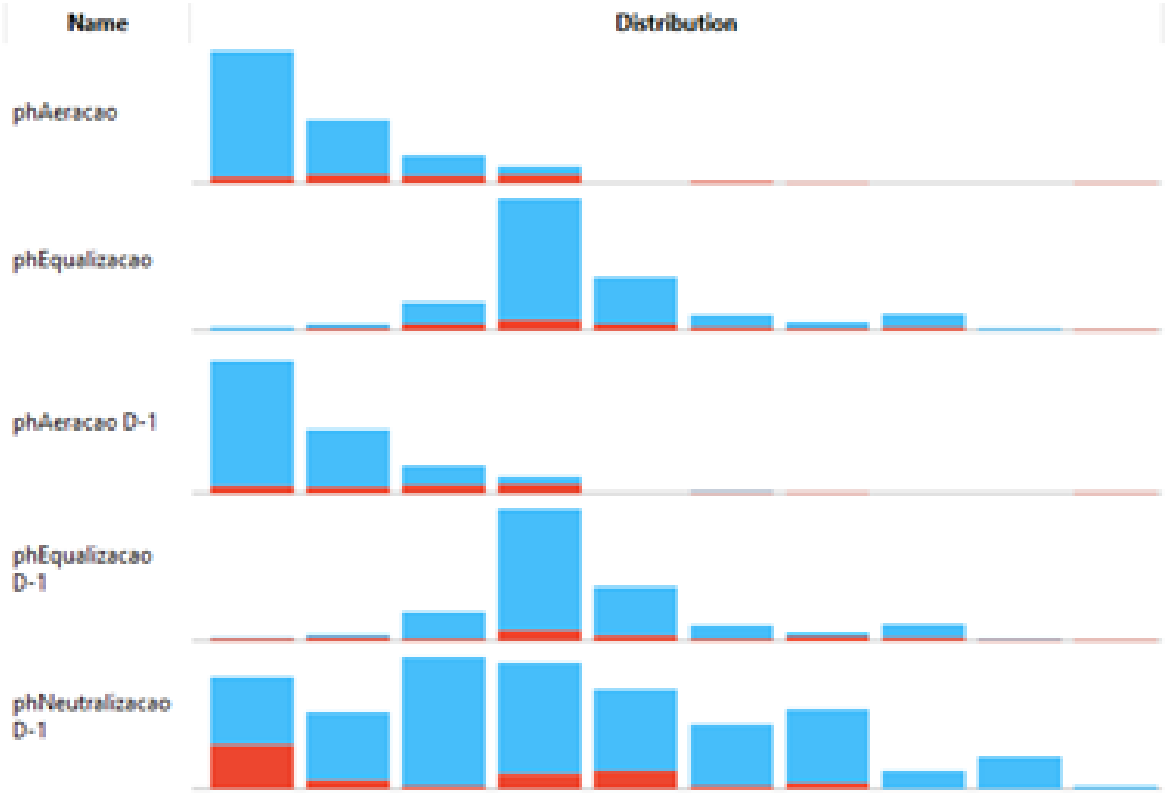
VI.1.1.4 Demais pHs

O pH de aeração tanto no dia quanto em D-1 indica um perfil maior de não conformidade com pH mais alto, acima de 7 que é o ideal. O pH mínimo é 6.8 e máximo de 8, sendo assim o pH de aeração é mais prejudicial quando está acima de 7 para 8, que encontram-se predominantemente casos de não conformidade de pH.

O pH da Equalização também encontra-se centrado perto 7 e também demonstra ser mais prejudicial em valores muito baixos ou muito altos. No caso tem-se o mínimo de 3,5 e máximo de 11,5 onde os casos de não conformidade acontecem em sua maioria. O tanque de equalização é a última fronteira antes do reator biológico e a transferência é realizada através dele, sendo assim é de profunda importância a manutenção de um pH com valores adequados no tanque.

Idealmente, o pH do tanque de Neutralização deve estar levemente básico, por volta de 8, no entanto ele está centrado em 5,8 e com valores muito concentrados para baixo com o mínimo em 1,0. Sugere-se um maior controle do pH no tanque de Neutralização pois a maior quantidade de dados com não conformidade se concentra em um pH baixo no tanque de neutralização.

***Figura 33**- Distribuição dos pHs*



Conformidade e Não Conformidade

***Figura 34**- Estatísticas dos pHs*

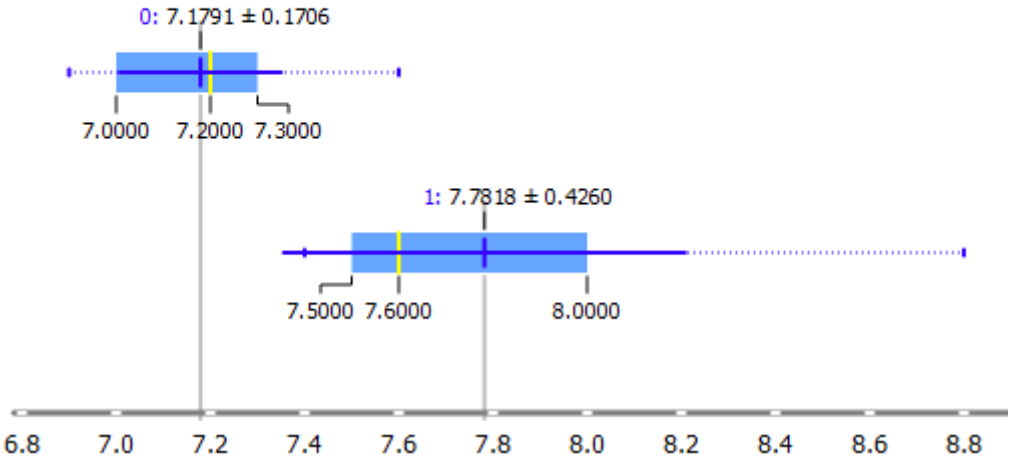
Center	Dispersion	Min.	Max.
7.12	0.03	6.80	8.80
7.16	0.17	3.50	11.50
7.12	0.03	6.80	8.80
7.16	0.17	3.50	11.50
5.89	0.51	1.00	13.10

O coeficiente de correlação de Pearson mostrou que a variável de pH da Aeração tem uma maior correlação comparada com as demais, +0,476 mostrando sua influência no DQO do Efluente final, o que segue o embasamento teórico e mostra que um controle rigoroso do pH no tanque de aeração é fundamental, os demais coeficientes estão entre 0,10 e 0,2.

Com relação aos outros valores de pH observa-se o pH de Entrada no efluente oleoso e o pH da Neutralização como os mais relacionados com o DQO de saída.

Uma análise de box plot elucida que o pH do tanque de aeração acima de 7,5 é prejudicial e aumenta bastante a chance de não conformidade de DQO no efluente de saída.

***Figura 35**- Box plot do pH no Tanque de Aeração*

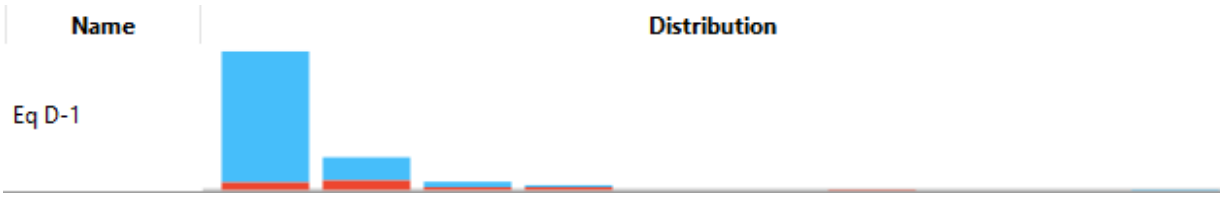


O eixo x representa o pH, os valores 1 e 0 mostra a separação entre os dois box plot dos dados categorizados conformes (0) e não conformes (1) e tem-se os valores de média em amarelo +- 2 desvio padrão.

VI.1.1.5 Equalização em D-1

A variável de Equalização em D-1 é uma das mais importantes no modelo pois ela representa o valor de DQO que foi transferido para a reação que gera o Efluente final do dia em análise. Assim, Eq D-1 nos mostra um valor da carga que foi transferida no enchimento para o tratamento do efluente.

***Figura 36**- Distribuição do DQO na Equalização em D-1*



Conformidade e Não Conformidade

Figura 37 *- Estatísticas do DQO na Equalização em D-1*

Center	Dispersion	Min.	Max.
758.99	1.13	25.00	8240.00

VI.1.1.6 Efluente em D-1

O valor de DQO no efluente D-1 mostra a influência na qualidade do efluente final entre os dias, como o volume descartado não é total pode haver uma carga remanescente de DQO no tanque biológico e a distribuição mostra este fato, visto que os valores de não conformidade se concentram em Efluentes D-1 com DQO mais alto. Também em casos de recém descarte, a população de bactérias é menor no tanque biológico e assim leva-se mais tempo para processar toda a carga de DQO remanescente no tanque biológico e no tanque de equalização.

Figura 38*- Distribuição do DQO no Efluente Final em D-1*

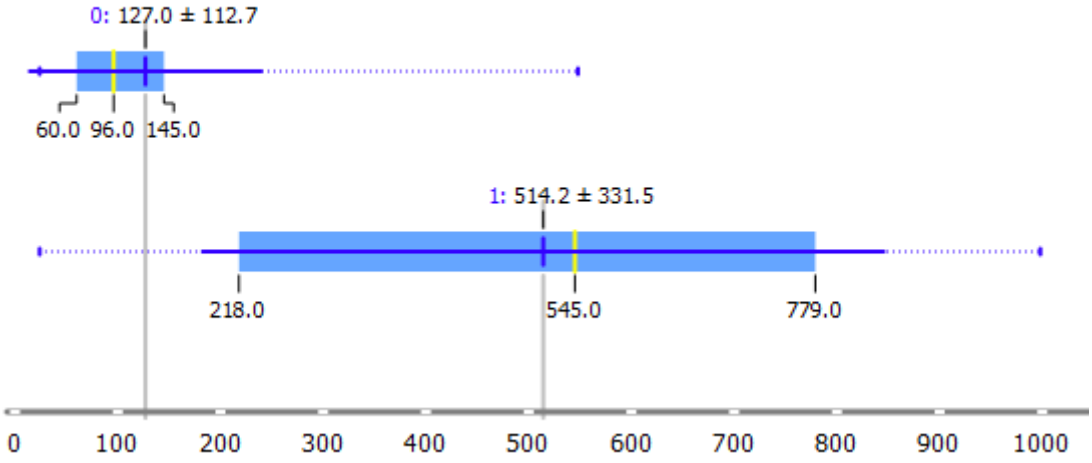
Figura 39*- Estatísticas do DQO no Efluente Final em D-1*

Center	Dispersion	Min.	Max.
165.92	1.04	0.00	1500.00

A correlação de Pearson mostra uma correlação positiva com o DQO do Efluente Final de +0,409, o que mostra que casos de DQO alto no dia anterior no DQO do Efluente Final pode refletir em um DQO maior também no dia da análise.

Nota-se no box plot que a concentração maior de dados conformes se encontra com DQO do Efluente Final D-1 entre 60 e 145 com média em 127. Quando o valor do Efluente no dia anterior chega próximo ou passa do limite de 250 as ocorrências de não conformidade aumentam com valores entre 218 e 779, e média em 514.2.

***Figura 40**- Box plot do DQO no Efluente Final em D-1*



O eixo x representa o DQO, os valores 1 e 0 mostra a separação entre os dois box plot dos dados categorizados conformes (0) e não conformes (1) e tem-se os valores de média em amarelo +- 2 desvio padrão.

VI.2. Treino sem tratamento

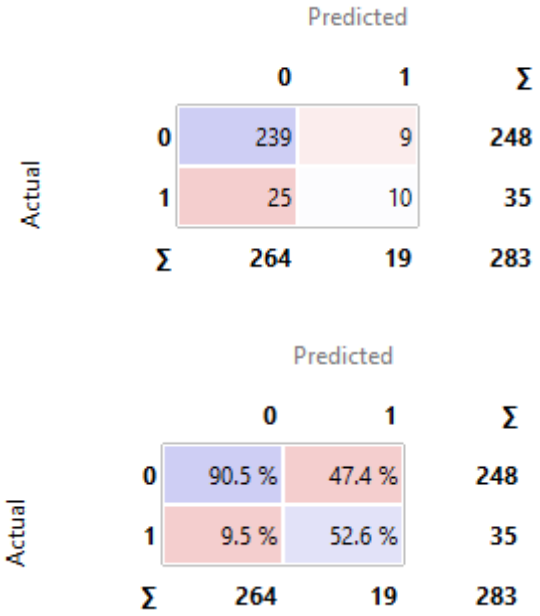
O primeiro caso de estudo foi feito utilizando o data set puro, sem nenhum tratamento adicional além do treinamento usando o método Leave one out. Os resultados estão descritos na Tabela 4:

***Tabela 4** - Resultados de Treino 1*

Modelo	AUC	Acurácia	F1	Precisão	Recall
Logistic Regression	85,6%	88%	86,4%	85,8%	88%
Random Forest	83,3%	88%	86,4%	85,8%	88%

Já de início notou-se um bom resultado, onde a curva ROC (AUC) e todas as outras métricas ficaram com valores acima de 80% pois como explicitado anteriormente, um valor acima de 50% indica previsibilidade maior do que a aleatoriedade e o indicador tem valor máximo de 100%, resultado que indica um caminho de uso correto das variáveis para predição.

***Figura 41** -Matriz de Confusão do Treino 1 – Logistic Regression*



Ao avaliar a Matriz de Confusão do melhor modelo, Logistic Regression, nota-se que o modelo errou cerca de 9,5% dos valores de conformidade apresentando 25 casos de Falso Negativo de 264 no total, o que é um bom resultado inicial. No entanto, o modelo está amplamente enviesado para a variável com maior amostragem. No caso de não conformidade o modelo previu somente 52,6%, o que representa um pouco melhor do que um chute.

VI.3 – Treino com tratamento

Na tentativa de melhorar a previsibilidade do modelo alguns tratamentos foram feitos antes da modelagem. Os resultados estão descritos nas seções seguintes.

VI.3.1 – Remoção de valores nulos

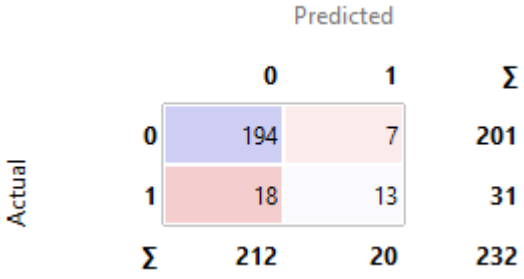
Como citado anteriormente, um bom conjunto de variáveis apresenta valor vazio por conta de dias sem operação, cerca de 10% dos valores. No entanto, tem-se duas variáveis com 74% de valores vazios e estas foram deixadas de lado no momento para não reduzirmos o data set a somente 30% dos dados originais. O primeiro tratamento foi a remoção destas colunas vazias e o data set ficou com 232 linhas na amostra. Os resultados foram elucidados na Tabela 5:

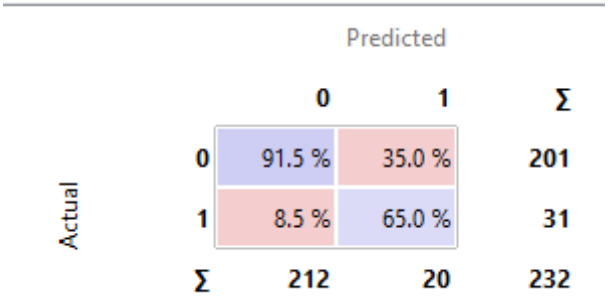
Tabela 5 *- Resultados do Treino 2*

Modelo	AUC	Acurácia	F1	Precisão	Recall
Logistic Regression	80,6%	89,2%	88,2%	88%	89,9%
Random Forest	86,4%	87,5%	86,3%	85,8%	87,5%

O modelo Logistic Regression apresentou uma visível melhora nos parâmetros de Acurácia, Precisão, F1 e Recall indicando a melhor previsibilidade da variável de não conformidade. No entanto, o modelo Random Forest permaneceu praticamente inalterado.

Figura 42*- Matriz de Confusão do Treino 2 - Logistic Regression*

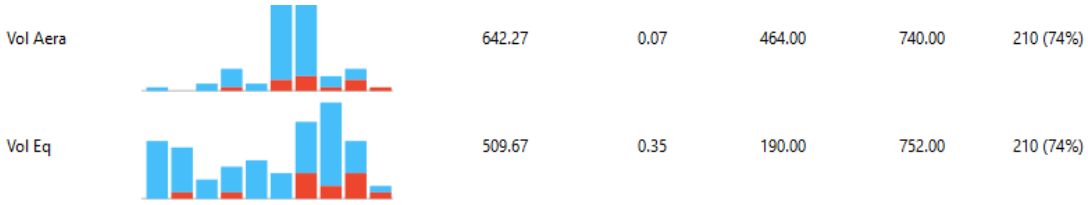




Ao avaliar a Matriz de Confusão do modelo Logistic Regression nota-se a redução dos erros entre a não conformidade para 8,5% e a melhora significativa na previsibilidade das amostras desenquadradas de DQO com 65% dos dados previstos.

Em uma segunda etapa foi retirado todos os registros onde não haviam valores para as variáveis Vol Aera e Vol Eq, que representavam os volumes nos tanques de Aeração e Equalização respectivamente, estes tinham 74% dos dados nulos.

Figura 43 *- Distribuição de variáveis com muitos dados nulos*



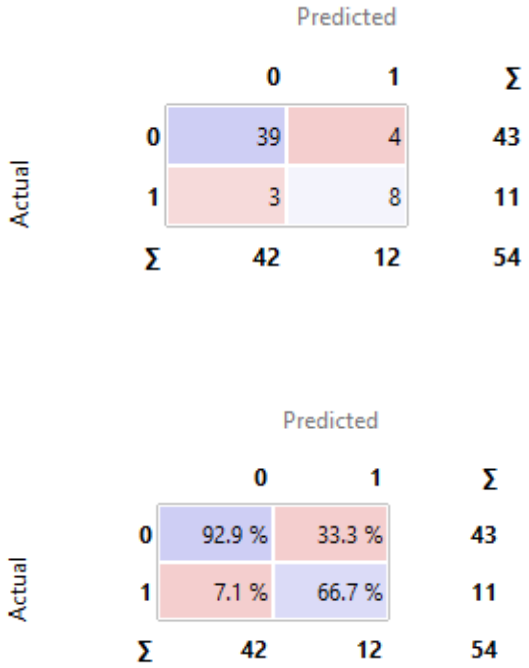
Com esta abordagem o data set foi reduzido para apenas 54 amostras o que é um número muito pequeno para treino, porém foram realizados testes a fim de avaliar os resultados e indicar a previsibilidade e uma melhoria do modelo em casos futuros de maior amostragem.

Tabela 6 *- Resultados do Treino 3*

Modelo	AUC	Acurácia	F1	Precisão	Recall
Logistic Regression	85,8%	87%	87,2%	87,5%	87%
Random Forest	92,4%	85,2%	84,6%	84,4%	85,2%

Nota-se uma melhoria na previsibilidade das amostras desenquadradas pois o data set ficou mais balanceado visto que a maior parte dos valores vazios eram de dados conformes.

***Figura 44** - Matriz de confusão do Treino 3 - Logistic Regression*



No fim do treino, pela matriz de confusão nota-se que o modelo reduziu o erro para 7,1% nos dados conformes e aumentou a previsibilidade para 66,7% dos casos de não conformidade.

VI.3.2 – Normalização

Um segundo tratamento foi feito para melhorar os resultados. As variáveis apresentam uma diferença considerável na dimensão de seus valores visto que possuem unidades diferentes. Tem-se, por exemplo, o pH com valores de 0 a 14 e DQO com valores indo até a faixa dos 10.000, assim para reduzir esta variabilidade e melhorar o treino todas as variáveis foram normalizadas pela média e escaladas pelo desvio padrão para valores entre 0 a 1. Os resultados foram expostos na Tabela 7:

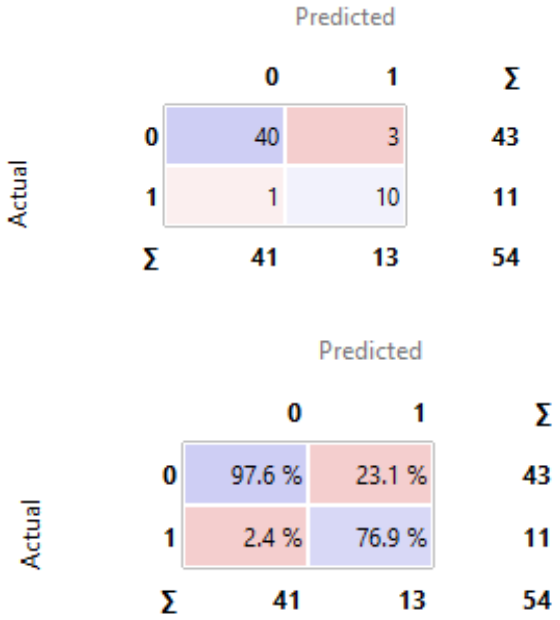
***Tabela 7**- Resultados do Treino 4*

Modelo	AUC	Acurácia	F1	Precisão	Recall
Logistic Regression	94,5%	92,6%	92,8%	93,4%	92,6%

Random Forest	81,4%	81,5%	80,8%	80,3%	81,5%

O modelo de Logistic Regression apresentou uma melhora bem significativa com a normalização das variáveis e mostrou os melhores resultados até então. Já o modelo Random Forest não apresentou uma diferença significativa devido ao baixo número de casos na amostra.

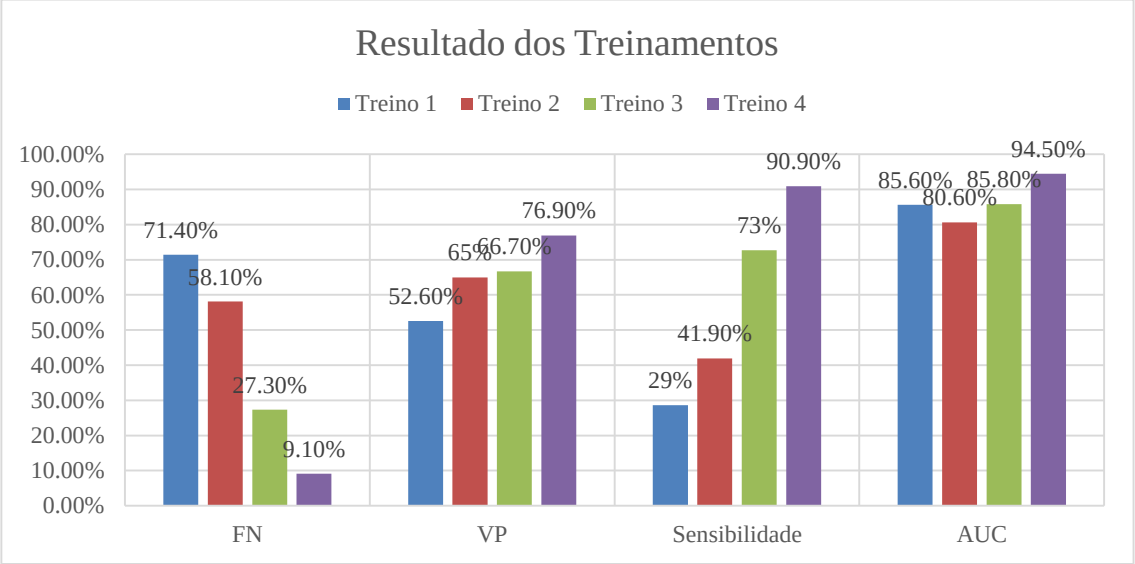
Figura 45 *- Matriz de Confusão do Treino 4 - Logistic Regression*



Através da análise das matrizes de confusão nota-se que o erro em variáveis enquadradas reduziu para o mínimo de 2,4% entre o Logistic Regression e a previsão correta dos valores não conformes chegou ao máximo de 76,9%.

VI.3.3 – Comparação entre os resultados

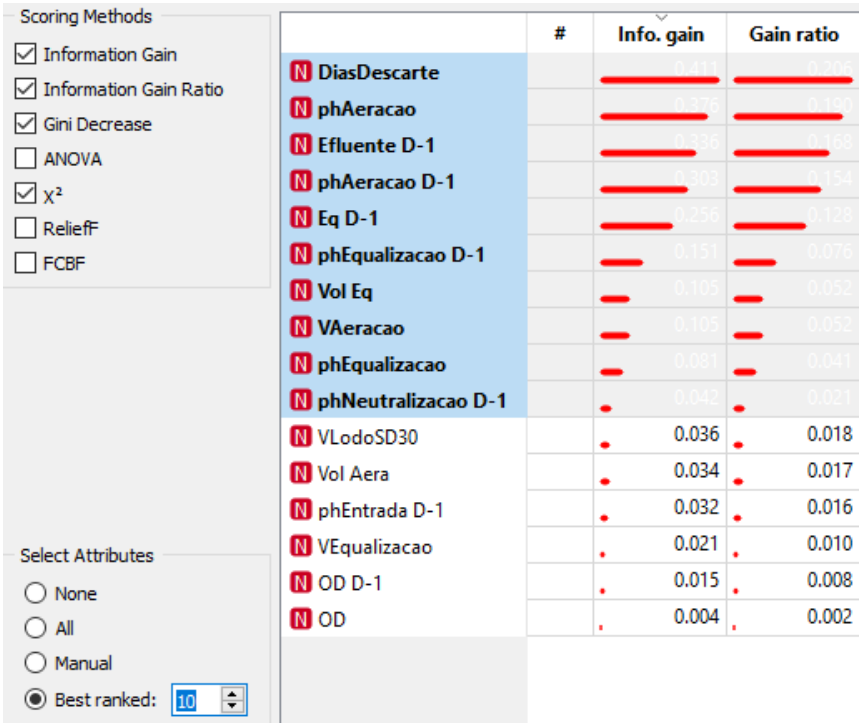
Nota-se uma visível melhora entre o treinamento do algoritmo sem nenhum tratamento com o treinamento utilizando tratamentos de refino dos dados e de modelagem do algoritmo como a retirada de dados nulos e a normalização, sendo a etapa da normalização a que apresentou melhores resultados no treinamento em comparação com as anteriores.



VI.3.3 – Importância das variáveis

Dentre as variáveis usadas para a modelagem, foi usado o conceito de Information Gain, elucidado na seção **V.4.5**, para ranquear as variáveis com maior correlação com os valores de DQO no efluente final para priorizar um estudo aprofundado e planos de ação a fim de melhorar o controle da estação de tratamento de efluentes na indústria.

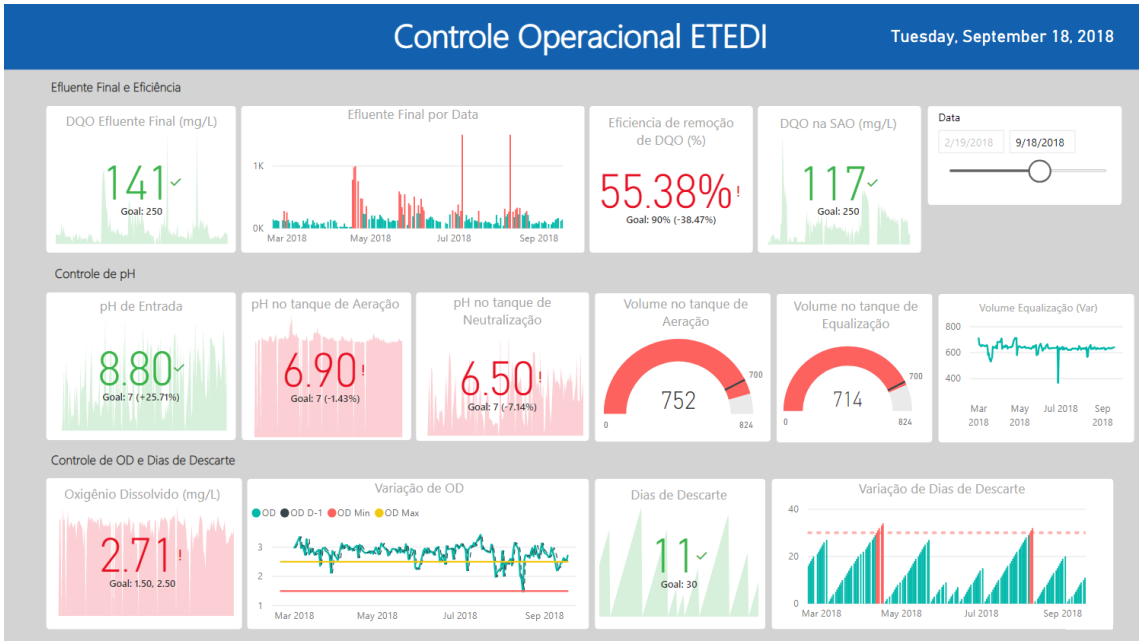
***Figura 46**- Ranqueamento das variáveis por Information Gain*



VI.4. Dashboard de operação - Power BI

Para a um controle mais avançado da estação de tratamento e uma melhor tomada de decisão e um ambiente mais qualificado e focado em análises, foi elaborado um dashboard de acompanhamento da Estação de tratamento na ferramenta de Microsoft chamada Power BI. O dashboard mostra as variáveis mais importantes para o controle e monitoramento da ETE em forma de KPI junto com suas variações com o tempo.

***Figura 47**- Dashboard de Operação da ETE em Power BI*



Fonte: O autor

O dashboard mostra os principais indicadores da estação de maneira visual, com um indicativo em cores que indica em verde as variáveis que estão dentro dos limites recomendados e em vermelho quando estão fora dos limites, além de mostra o percentual de distância do limite especificado, o que traz uma quantificação do distanciamento do padrão e patamar da variável medida.

VII. Conclusão

Este trabalho trata de um tema cuja relevância promete crescer ao longo dos próximos anos, junto com a aplicação de soluções de machine learning devido ao avanço do movimento da indústria 4.0. De acordo com os resultados demonstrados, a aplicação da solução é viável e promissora por apresentar uma boa previsibilidade e economia de processo. O modelo de machine learning desenvolvido atinge 92,6% de acurácia e 93,4% de precisão, além de prever de forma assertiva 76,9% das amostras não conformes, mesmo com a base de dados desbalanceada. O dashboard de monitoramento da estação de tratamento mostra com clareza o estado atual da estação e seus principais indicadores, sendo de grande importância no suporte para uma tomada de decisão precisa.

Como melhoria do trabalho e suas aplicações, destaca-se o uso de valores médios do dia para as variáveis medidas, a maior automatização dos sensores para aumentar a amostragem de variáveis como pH, OD e volume, a geração de dados de vazão de afluente e efluente, que são parâmetros importantes para calcular a carga, e se possível uma infraestrutura para o acompanhamento dos dados de monitoramento em tempo real no dashboard.

Por fim, destaca-se que a aplicação das análises estatísticas na indústria parceira gerou excelentes resultados que podem ser potencializados com a aplicação futura das soluções de machine learning e monitoramento. As soluções de machine learning e ciência de dados se mostram bastante eficientes e promissoras na Engenharia Química e devem ser incentivadas para o desenvolvimento contínuo da área.

VIII. Bibliografia

SAWYER, Clair N.; MCCARTY, Perry L.; PARKIN, Gene F. **Chemistry for Environmental Engineering and Science.** 5. ed. New York: McGraw-Hill, 2003.

CLAAS, Isabel Cristina. **Lodos ativados: Princípios teóricos fundamentais, operação e controle**. Porto Alegre: Evangraf, 2007.

AGRESTI, Alan. **Categorical Data Analysis**. S.l.: New York: Wiley-Interscience, 2002.

OPITZ, D.; MACLIN, R. **"Popular ensemble methods: An empirical study"**. Journal of Artificial Intelligence Research. 11: 169–198. 1999.

FISHER, R. **Statistical Methods and Scientific Induction. Journal of the Royal Statistical Society**, Series B, 1955.

HAZEWINKEL, Michiel, ed., **"Binomial distribution"**, Encyclopedia of Mathematics, Springer Science+Business Media B.V. / Kluwer Academic Publishers, 2001.

KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection.** In: International joint Conference on artificial intelligence. [S.l.: s.n.], 1995. v. 14, p. 1137–1145.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning** (2nd ed.). Springer, 2008.

FORLABEXPRESS. **Valores de testes de DQO**. Disponível em: < https://www.forlabexpress.com.br/>, Acessado em 20 jul. 2019.

VON SPERLING, Marcus. **Lodos ativados.** Belo Horizonte: Departamento de Engenharia Sanitária e Ambiental da Universidade Federal de Minas Gerais, 1997.

JORDÃO, E.P., et al. **Controle microbiológico na operação de um sistema de lodos ativados – estudo em escala piloto.** In 19° Congresso Brasileiro de Engenharia Sanitária e Ambiental, 1997.

FORBES. **Limpeza de Dados**. Disponível em: <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#3e312e026f63 >, Acessado em 15 jul. 2019.

AMORIN. L. **Análise de eficiência do sistema de lodo ativado no tratamento de efluentes de um curtume na cidade de Uberlândia MG.** Disponível em: <https://www.ibeas.org.br/congresso/Trabalhos2014/III-078.pdf>, Acessado em 20 jul. 2019.

MICROSOFT AZURE. **Ciclo TDSP.** Disponível em: <https://docs.microsoft.com/pt-br/azure/machine-learning/team-data-science-process/overview> Acessado em: 01 mai. 2019.

MICROSOFT. **Power BI.** Disponível em: <https://powerbi.microsoft.com/pt-br/>, Acessado em: 20 jul. 2019.

DZ205-RJ. **Diretriz sobre lançamento de efluentes.** Disponível em: <http://www.tesalab.com.br/site/downloads/INEA_dz205.pdf> Acessado em: 17 jul. 2019.

BRASIL. CONAMA. Resolução nº 357, de 17 de Março de 2005. **Classificação dos corpos de água e diretrizes ambientais para o seu enquadramento, bem como condições e padrões de lançamento de efluentes, e outras providências.** Diário Oficial da União, Brasília, 18 de março de 2005.

GOLDMAN, Charles R.; HORNE, Alexander J. **Limnology**. McGraw-Hill: 1983.

NT-202.T-10. **Critérios e padrões para lançamento de efluentes líquidos**. Disponível em: < http://www.inea.rj.gov.br/cs/groups/public/documents/document/bmvh/mdey/~edisp/inea012974.pdf> Acessado em 21 jul. 2019.

SHEARER, C. **The CRISP-DM model: the new blueprint for data mining.** J Data Warehousing, 2000.

HAYASHI, Chikio. **"What is Data Science? Fundamental Concepts and a Heuristic Example"**. Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan, 2008.

Anexo I

CD (“compact disk”) – Projeto em PDF e Resumo