UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE ENERGIAS ALTERNATIVAS E RENOVÁVEIS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

**Filipe Bulhões Froes**

# Comparação de modelos de desenvolvimento de sensores virtuais baseados em inteligência artificial

João Pessoa - PB
2022

Filipe Bulhões Froes

# Comparação de modelos de desenvolvimento de sensores virtuais baseados em inteligência artificial

Trabalho de Conclusão de Curso apresentado à Coordenação de Engenharia Elétrica do Centro de Energias Alternativas e Renováveis da Universidade Federal da Paraíba como parte dos requisitos necessários para a obtenção do título de Engenheiro Eletricista.

Universidade Federal da Paraíba

Centro de Energias Alternativas e Renováveis

Curso de Graduação em Engenharia Elétrica

Orientador: Prof. Dr. Juan Moises Mauricio Villanueva

João Pessoa - PB

2022

Filipe Bulhões Froes
Comparação de modelos de desenvolvimento de sensores virtuais baseados em inteligência artificial/ Filipe Bulhões Froes. – João Pessoa - PB, 2022-
43 p. : il. (algumas color.).

Orientador: Prof. Dr. Juan Moises Mauricio Villanueva

Trabalho de Conclusão de Curso – Universidade Federal da Paraíba
Centro de Energias Alternativas e Renováveis
Curso de Graduação em Engenharia Elétrica, 2022.

1. PC1. 2. PC2. 3. PC3. 4. PC4

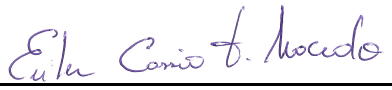
Filipe Bulhões Froes

# Comparação de modelos de desenvolvimento de sensores virtuais baseados em inteligência artificial

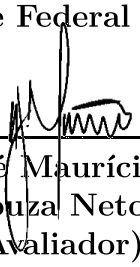
Trabalho de Conclusão de Curso apresentado à Coordenação de Engenharia Elétrica do Centro de Energias Alternativas e Renováveis da Universidade Federal da Paraíba como parte dos requisitos necessários para a obtenção do título de Engenheiro Eletricista.

**Data de Aprovação:** 01 / 12 / 2022

**Prof. Dr. Juan Moises Mauricio Villanueva**
**(Orientador)**
**Universidade Federal da Paraíba**



**Prof. Dr. Euler Cássio Tavares de Macedo**
**(Avaliador)**
**Universidade Federal da Paraíba**



**Prof. Dr. José Maurício Ramos de Souza Neto**
**(Avaliador)**
**Universidade Federal da Paraíba**

João Pessoa - PB
2022

Dedico este trabalho à minha família e amigos.

# Agradecimentos

Agradeço a Deus por proporcionar as condições para realizar este trabalho. Desde o tempo para que fosse feito até a saúde da minha família, todas as condições que garantiram a finalização deste trabalho foram graças a ti.

Agradeço também à minha família e amigos pelo apoio. Em especial à minha mãe, por ser a minha base forte e inabalável, aos meus tios, tias, primos e avós, pelo apoio incondicional, ao meu pai, pela inspiração e à minha madrinha, minha segunda mãe.

Agradeço a todos os meus amigos, minha segunda família, pela motivação e pela paciência. Àqueles que estão longe há tanto tempo, mas que sempre se fizeram presentes, àqueles que conheci durante a graduação e de quem me orgulho em dizer que não são apenas colegas de profissão, mas também valiosos amigos.

Agradeço à instituição de ensino, à qual tenho o orgulho de pertencer e a todos que fazem parte dela. A todos os professores que tive o prazer de conhecer ao longo do curso. Em especial àqueles que me orientaram em projetos, pesquisas e no trabalho de conclusão de curso. Por todas as lições, principalmente lições de vida, lições sobre respeito, lições sobre humildade e sobre amor à profissão.

Finalmente, agradeço a todos aqueles que contribuíram direta ou indiretamente para a minha formação. Professores do ensino fundamental, professores do ensino médio, profissionais de serviços gerais, que sempre me cumprimentaram com um sorriso no rosto, e a todos os outros que me ajudaram nessa caminhada.

A todos, muito obrigado!

*Faça ou não faça. Tentativa não há.*
(Mestre Yoda)

# Resumo

O rápido avanço dos sistemas de informação impacta todos os setores da economia. No setor produtivo, sua importância é tão grande que deu orígem à quarta revolução industrial. Com isso, sistemas de controle e sistemas de aquisição de dados tornam-se cada vez mais importantes na indústria e, com a digitalização dos sistemas, surge a necessidade de analisar e manipular uma quantidade crescente de informação. Uma forma de melhorar o sensoriamento de maneira ordenada e otimizada é utilizando instrumentos de medição virtuais. Os chamados sensores virtuais ou *soft sensors* são aplicações que podem substituir sensores físicos e sua utilização pode não apenas reduzir a quantidade de sensores ao longo da planta industrial, mas também tem o potencial de reduzir drasticamente os custos de uma indústria. O presente trabalho avalia a implementação de um sensor virtual por meio de dois modelos matemáticos desenvolvidos utilizando inteligência artificial. Os modelos foram comparados com base em métricas de avaliação amplamente utilizadas na indústria e, por fim, é feita a análise de incertezas do modelo de melhor desempenho.

**Palavras-chave**:Sensores virtuais. Inteligência artificial. Modelos matemáticos. Análise de incertezas.

# Abstract

The quick development of the information systems has a deep impact in all sectors of the economy. In the industry, it became so important that started the fourth industrial revolution. Therefore, the control and data acquisition systems have become progressively important in the industry and, with the digitalization of the systems, comes the urge to analyze and transform an increasing amount of information. One way to improve sensoring in an organized and optimized manner is using soft sensors. The called soft sensors are softwares that are used as sensors and that can replace physical sensors. Their use can reduce the quantity of physical sensors throughout the assembly line and has the potential to drastically decrease the costs related to manufacturing. The present study evaluates two mathematical approaches, based on artificial intelligence, in the development of a soft sensor. The models were compared using well known evaluation metrics. Finally, is made the analysis of the uncertainties of the best model.

**Keywords**: Soft sensors. Artificial intelligence. Mathematical models. Uncertainty analysis.

# Lista de ilustrações

# Lista de tabelas

# Lista de abreviaturas e siglas

| | |
|---|---|
| FCD | Diferença entre a estatística F e a correlação |
| FCQ | Quociente entre a estatística F e a correlação |
| GMM | Modelo de misturas de Gaussianas |
| IA | Inteligência artificial |
| IoT | Internet das coisas |
| MAE | Média do erro absoluto |
| MAPE | Média do erro percentual absoluto |
| MEP | Máximo erro percentual |
| MRMR | Máxima relevância e mínima redundância |
| MSE | Média do erro ao quadrado |
| PCR | Regressão de componentes principais |
| PLSR | Regressão por mínimos quadrados parciais |
| RNA | Redes neurais artificiais |
| SV | Sensores virtuais |
| SVM | *Support vector machine* |
| USICODA | Usina Central Olho D'água |

# Sumário

# 1 Introdução

Com o advento dos computadores, a indústria vem passando por um processo de digitalização, culminando no mais recente avanço no setor produtivo, a indústria 4.0. Esse termo que se refere à nova revolução industrial descreve uma série de inovações que combinam tecnologia da informação e automação industrial com o objetivo de aumentar a eficiência da produção (VAIDYA; AMBAD; BHOSLE, 2018). Dessa forma, a medição e aquisição de dados tem se tornado cada vez mais importantes no controle de processos. Assim, os sensores têm um papel fundamental nesses sistemas, possibilitando o monitoramento dos fenômenos físicos que ocorrem durante o processo de produção.

Contudo, sensores físicos podem ter um elevado custo de aquisição e manutenção. Além de necessitarem de processamento e amostragem de sinais, ainda têm limitações quanto a sua utilização em comparação aos sensores virtuais. Sensores virtuais (SV) ou *soft sensors* são algoritmos ou aplicações utilizadas como instrumentos de medição. De forma geral, são programas de computador que contêm modelos matemáticos que replicam o comportamento de sensores físicos, podendo substituí-los ou serem utilizados em conjunto com os mesmos.

Eles são empregados principalmente na medição de variáveis convencionalmente difíceis de medir por meio de inferência dada a relação entre a variável de interesse e variáveis secundárias mais fáceis de medir. Dentre algumas vantagens dos sensores virtuais, podemos destacar seu baixo custo, e a sua possibilidade de utilização na medição de variáveis com longo tempo de atraso, o que torna sensores virtuais uma solução atrativa e promissora para determinadas aplicações (LIU et al., 2018).

Dessa forma, é possível utilizar sensores virtuais como alarmes, para identificar falhas no sensor principal e até mesmo atuar provisoriamente substituindo-o durante uma manutenção. Podem ser utilizados também substituindo sensores físicos permanentemente, principalmente em situações em que o sensor físico tem um alto custo. Normalmente, esses sensores calculam o valor de uma variável, a qual se deseja medir, por meio de valores obtidos por sensores reais que monitoram variáveis secundárias. Sua aplicação torna-se imprescindível quando não é possível medir diretamente uma variável, principalmente se ela for subjetiva, como no caso do trabalho de (YIN; ZHU; KARIMI, 2013), em que os autores desenvolvem um sensor virtual para prever a qualidade de vinhos baseado em características diversas por meio de métodos estatísticos.

Os modelos matemáticos dos *soft sensors* podem ser simples, sendo apenas uma relação matemática determinística entre as variáveis, por exemplo a relação entre força e pressão em uma superfície. Podem ser também sofisticados, sendo necessário em alguns

casos a utilização de ferramentas de inteligência artificial (IA) para sua implementação.

No trabalho de (LIU et al., 2018), os autores citam várias formas de se implementar modelos matemáticos aplicados a sensores virtuais. Métodos lineares, como por exemplo Regressão de componentes principais (*Principal component regression* - PCR) e regressão por mínimos quadrados parciais (*Partial least squares regression* - PLSR) são simples e amplamente utilizados, podendo até ser adaptados para análises probabilísticas. Além de métodos não lineares, como *Support vector machines* (SVM), modelos de misturas de gaussianas (*Gaussian mixture model* - GMM) e ferramentas de inteligência artificial e aprendizado de máquina, como redes neurais artificiais - RNA e árvores de decisão, por exemplo.

No setor sucroenergético, há uma crescente demanda por sensores virtuais à medida em que o setor se moderniza. De acordo com (LIMA, 2022a), já é possível observar plantas com centros de controle operacional que utilizam redes industriais. Há também a perspectiva de implementação de conceitos como internet das coisas (IoT) e *Big data*, o que gera a necessidade de medições cada vez mais numerosas e precisas ao longo de toda a planta. Nesse sentido, devido ao seu baixo custo, sensores virtuais tornam-se bastante atrativos para o setor.

O objetivo geral deste trabalho é desenvolver um sensor virtual de vazão por meio de regressão, utilizando técnicas de inteligência artificial. No desenvolvimento do instrumento, serão utilizadas duas abordagens, aplicando árvores de decisão e redes neurais artificiais. Esses dois modelos serão comparados quanto ao seu desempenho e será calculada a incerteza associada ao modelo com melhores estatísticas. O modelo desenvolvido foi baseado no estudo de caso de uma planta industrial do setor sucroenergético, exposto no trabalho de (LIMA, 2022a).

## 1.1 Organização do trabalho

Neste trabalho, além do presente Capítulo de introdução, há também mais quatro capítulos, os quais serão apresentados a seguir.

No Capítulo dois, será apresentada a fundamentação teórica a respeito do tema. Nele, são aprofundados os conceitos a respeito dos SV e é feita uma revisão a respeito de RNA, árvores de decisão e incertezas de medição.

No Capítulo três, será apresentado o estudo de caso, com as informações sobre a planta industrial e seus respectivos processos. É também apresentado o banco de dados e o sistema utilizado para coleta de dados. Ainda no Capítulo três, é feita uma introdução a métodos de seleção de variáveis para treinamento dos modelos preditivos. E, por fim, são apresentados os algoritmos de estimação de vazão e de estimação de incertezas e as

métricas de avaliação dos modelos.

No Capítulo quatro, são apresentados a implementação e os resultados obtidos. Nele, são dispostos os modelos de estimação de vazão desenvolvidos, seus desempenhos baseados nas métricas estabelecidas, a comparações entre eles e a estimação da incerteza do melhor modelo.

Finalmente, no Capítulo cinco, são feitas as considerações finais e apresentadas as conclusões e trabalhos futuros.

# 2 Fundamentação Teórica

## 2.1 Conceitos básicos sobre sensores virtuais

Sensores virtuais são uma forma eficiente de redução de custos e de aumento na eficiência na indústria. Levando em consideração que para cada sensor físico é necessário um plano de manutenção, calibração e ajuste, substituir um sensor físico por um virtual pode reduzir drasticamente os custos de operação de uma planta. Uma das aplicações dos SV é evitar que sejam adquiridos e instalados sensores físicos além do necessário (JIANG et al., 2021).

À medida em que a quantidade de sensores em uma planta aumentam, junto com eles cresce a quantidade de informação adquirida. Com isso, surge o desafio de gerenciar esses dados, a fim de evitar redundâncias desnecessárias de informação. Os SV são uma saída viável quando uma variável de interesse pode ser obtida por meio de informações já disponíveis de outros sensores (SUN; GE, 2021). Sendo possível medir uma variável de interesse por meio de relações matemáticas entre ela e variáveis secundárias.

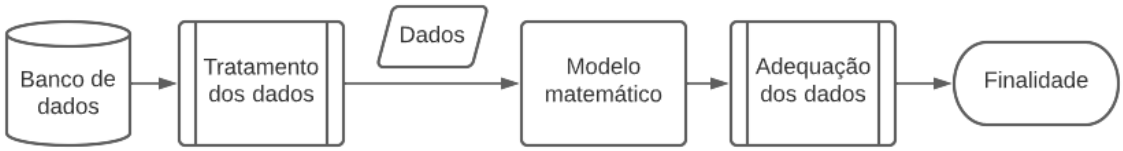
Para cada sensor físico que deixa de ser utilizado, seus custos de aquisição e manutenção também deixam de ser empregados. Isso resulta em menos paradas na linha para manutenção e ajuste, redução do estoque de peças de reposição e redes de comunicação industrial mais simples (JIANG et al., 2021). Como um exemplo de SV, podemos citar o algoritmo desenvolvido no trabalho de (LIMA, 2022b). No estudo em questão o autor desenvolve um instrumento virtual para estimação de vazão em sistemas de abastecimento de água.

### 2.1.1 - Topologia

A parte principal de um SV é o modelo matemático que irá gerar os valores do instrumento de medição. Contudo, além dele são necessárias ferramentas auxiliares para que a aplicação seja completa. A topologia geral de um SV pode ser dividida em três partes principais, sendo elas as partes de coleta e tratamento de dados, modelo matemático e adequação dos dados de saída (Figura 1). Note que a coleta dos dados se refere à obtenção dos dados diretamente de um banco de dados, rede de comunicação industrial e similares. A aquisição dos dados dos sensores desde o sensor físico até o momento em que esses dados ficam disponíveis na rede não faz parte do SV e sim do sistema de aquisição de dados da planta (JIANG et al., 2021).

A coleta e tratamento de dados tem a função de obter os dados brutos da rede ou de um banco de dados e condicioná-los ao formato compatível com o modelo matemático. É nessa etapa que serão feitas a filtragem de ruídos, remoção de *outliers*, preenchimento

Figura 1 – Topologia geral de um SV.

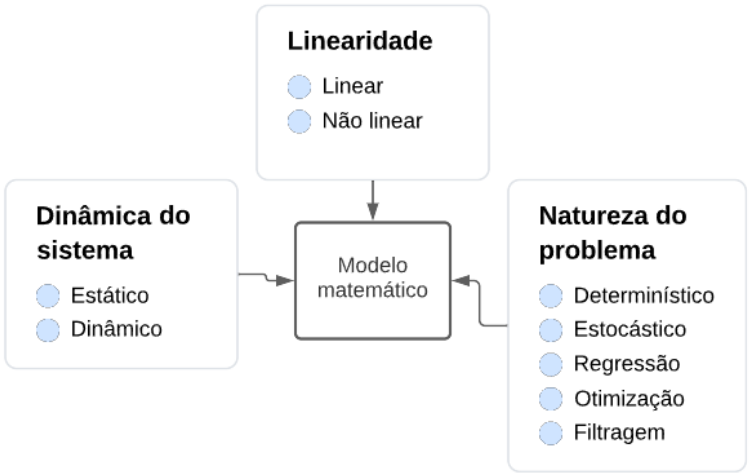



Fonte: Autoria própria.

de dados faltantes, normalização dos dados, de acordo com a necessidade.

O modelo matemático é o que permitirá relacionar os dados de entrada com a variável que se deseja medir. Esse modelo deverá ser tão complexo quanto necessário, de acordo com a natureza do problema. Há três parâmetros principais que irão determinar qual a complexidade do modelo matemático: a dinâmica do sistema, a linearidade e a natureza do problema (Figura 2) (JIANG et al., 2021).

Figura 2 – Parâmetros de modelagem.




Fonte: Autoria própria.

Quanto à dinâmica do sistema, ele pode ir de estático a dinâmico, de acordo com a variação do seu comportamento ao longo do tempo. Isso significa que, para sistemas estáticos, um mesmo conjunto de dados de entrada levariam a uma mesma saída, independentemente do tempo que levasse entre as medições, para o caso estático. Para o dinâmico, os mesmos dados de entrada em tempos diferentes levariam a resultados diferentes. Dessa forma, sistemas dinâmicos necessitam de modelos que se adaptam ao longo do tempo, como por exemplo redes neurais adaptativas.

Com relação à linearidade, sistemas lineares necessitam de modelos matemáticos mais simples que modelos não lineares. E, de forma semelhante, a natureza do problema, que pode ir de determinístico a estocástico, pode ser um problema de otimização, filtragem, regressão, classificação, entre outros.

Por fim, a adequação dos dados de saída tem o objetivo de entregar os dados ao destino de forma compatível com ele, prontos para a utilização. É nessa etapa que será feita a desnormalização e a codificação, de acordo com a necessidade, levando em consideração a sua finalidade. Assim, os valores de saída do instrumento devem se adequar a sua finalidade, seja ela informar um valor na tela para um operador ou servir de entrada para um atuador, por meio de uma comunicação de rede industrial.

### 2.1.1 Abordagens e metodologias no desenvolvimento de SV

Há duas abordagens principais para se modelar um SV: por meio de modelos determinísticos e por meio de modelos orientados por dados. Os modelos determinísticos são baseados em relações bem conhecidas entre as variáveis. Esse tipo de SV requer conhecimento aprofundado do modelo físico ou estatístico do sistema envolvido. Dessa forma, recomenda-se aplicar esse tipo de metodologia apenas a problemas simples e, de preferência, estáticos, que não variam ao longo do tempo. Para problemas complexos e dinâmicos, os modelos orientados por dados, como os de aprendizado de máquina, são os mais recomendados (SUN; GE, 2021).

Dessa forma, é possível definir um procedimento para desenvolver o SV, que pode ser descrito nos seguintes passos:

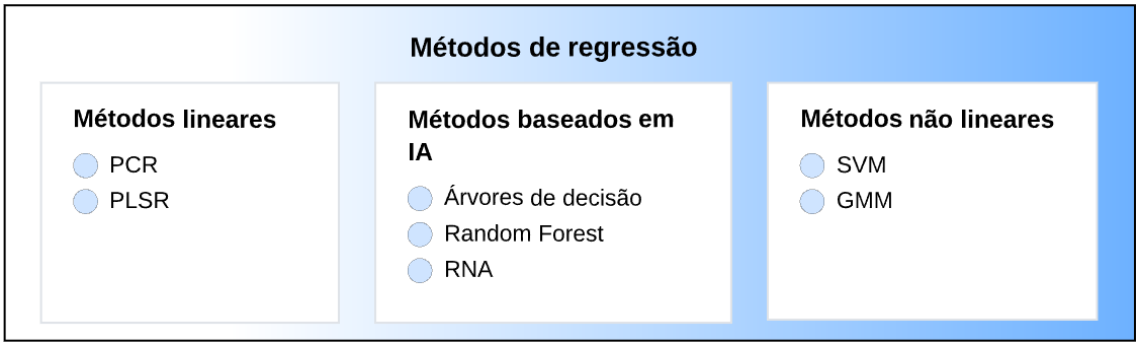
- Estudo do problema;
- Definição e desenvolvimento do modelo matemático;
- Implementação da aplicação.

Na fase de estudo do problema, deve-se estudar a variável a ser medida, quais variáveis secundárias serão utilizadas e qual a relação entre as variáveis de entrada e a de interesse. Também é necessário fazer o levantamento do formato dos dados a serem utilizados e o formato em que os dados de saída devem ser disponibilizados pela aplicação.

Em seguida, deve-se escolher o modelo matemático mais adequado ao problema com base nos três parâmetros principais mencionados na seção anterior. Ilustra-se na Figura 3, alguns métodos aplicados à regressão de acordo com a linearidade do sistema. Uma vez escolhido o modelo, este deve ser avaliado e otimizado para que se alcance um resultado satisfatório. Para isso, são aplicadas métricas de avaliação ao modelo, sendo necessário que este atinja um desempenho mínimo estipulado pelo desenvolvedor. O presente trabalho compara a aplicação de RNA e árvores de decisão no desenvolvimento de um SV.

Por fim, devem ser feitos os algoritmos relativos à coleta e tratamento dos dados de entrada e adaptação dos dados de saída, caso já não tenha sido feito durante a etapa

Figura 3 – Exemplos de métodos aplicados à regressão.




Fonte: Autoria própria.

anterior. Por fim, deve-se instalar a aplicação no computador ou sistema em que será utilizado, deixando-a pronta para uso.

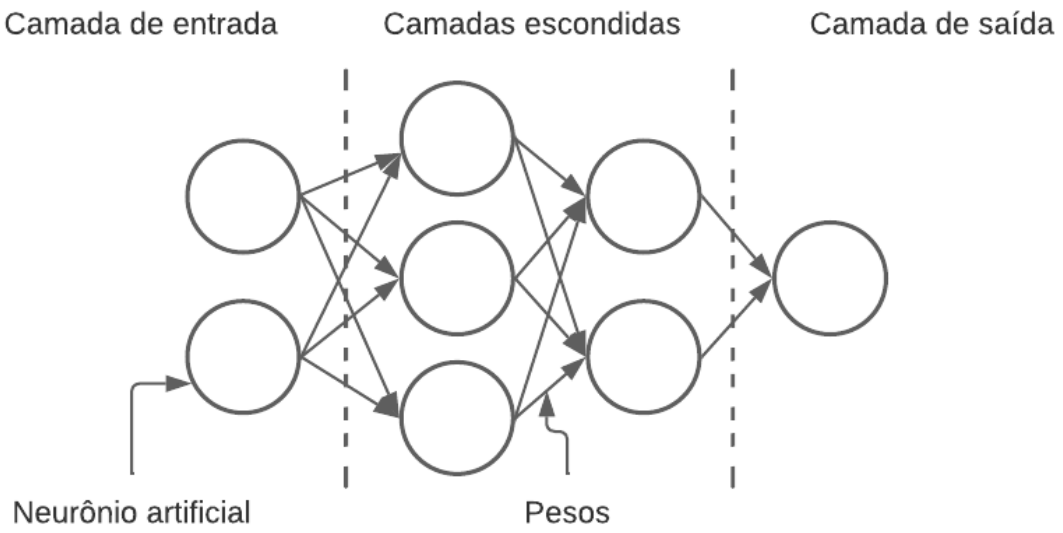
#### 2.1.1.1 RNA: Conceitos básicos

Redes neurais artificiais fazem parte do universo dos algoritmos de aprendizado de máquina. É uma técnica que visa reproduzir a capacidade de aprendizado do ser humano, de forma artificial. Por meio de um conjunto de dados, é possível treinar uma RNA, de forma que ela se ajuste a esses dados, tomando-os como estímulos e calculando pesos para as conexões entre cada neurônio artificial (RODRIGUES, 2022). Assim, ela aprende a reagir a estímulos, dados de entrada, de acordo com uma saída controlada, a qual foi exposta durante o treinamento. Dessa forma, pode substituir modelos matemáticos complexos, sem a necessidade de se compreender as relações entre as variáveis de entrada e saída, uma vez que o modelo se adapta aos dados. Note que o modelo pode até mesmo estabelecer relações para conjuntos de dados aos quais nunca foi exposto, desde que as relações entre as variáveis de entrada e saída (sistema), sejam as mesmas que as do conjunto de treinamento, sendo essa uma de suas principais vantagens.

De modo geral, uma RNA é composta por neurônios artificiais ligados a outros de camadas adjacentes. Cada ligação possui um peso, que serve para propagar a ativação de uma camada para a camada seguinte, até que ela chegue à saída, na última camada. Dessa forma, o valor de entrada de cada neurônio será calculado pela soma ponderada das saídas nos neurônios da camada anterior. Os neurônios artificiais são modelados pela equação 2.1. Onde $in_j$ é o valor de entrada do neurônio $j$, $w_{i,j}$ é o peso associado à ligação entre o neurônio $i$, da camada anterior, e o neurônio $j$ e $a_i$ é o valor de saída do neurônio $i$ (RUSSELL; NORVIG; MACEDO, 2013).

$$in_j = \sum_{i=0}^{N} w_{i,j} a_i \tag{2.1}$$

Figura 4 – Esquema de uma RNA.




Fonte: Autoria Própria.

O valor de saída do neurônio é calculado pelo valor de entrada $in_j$ aplicado à sua função de ativação 2.2. Há algumas funções que são comumente utilizadas como função de ativação ($g(in_j)$). Entre elas, há algumas que são mais indicadas para classificação, como por exemplo o seno hiperbólico, binária e sigmóide. E há também algumas das mais utilizadas para modelos de regressão, como a relu e a linear.

$$a_j = g(in_j) \tag{2.2}$$

Normalmente, a primeira camada, a camada de entrada, não possui função de ativação e, em alguns casos, a última camada também não possuirá. A RNA aprende por meio do ajuste dos pesos durante o treinamento, reduzindo gradualmente o valor do erro entre os valores de saída da RNA e os valores de saída disponibilizados para teste, por meio de uma métrica de avaliação: a função de perda. Essa métrica é escolhida pelo desenvolvedor e varia de problema a problema. Normalmente é avaliado caso a caso qual tem o melhor desempenho. Essas métricas também podem ser utilizadas pós treinamento para a avaliação de desemenho do modelo com novos dados.

Apesar da grande vantagem das RNA de se adaptarem ao conjunto de dados, sua performance nunca será perfeita, ou seja, sempre haverá um erro associado às suas previsões, seja o modelo aplicado à regressão, classificação ou reconhecimento de padrões. Por isso, algumas práticas devem ser verificadas para que o desempenho da RNA atinja parâmetros de avaliação mínimos desejáveis, mantendo ao máximo sua característica generalista. Dentre essas práticas, podemos citar a seleção de variáveis ou atributos de entrada (*feature selection*), regularização e modelagem da rede.

#### 2.1.1.2 Árvores de decisão

Árvores de decisão ou *Decision trees* são ferramentas de inteligência artificial comumente aplicadas a classificação, mas que também podem ser utilizadas para regressão com variáveis numéricas contínuas. Assim como as RNA, são um método orientado por dados, apesar de modelos mais simples poderem ser montados manualmente por um especialista.

Diferentemente das RNA, esses modelos não utilizam cálculos baseados em pesos e sim operações booleanas que classificam o valor da variável de saída de acordo com os valores das variáveis de entrada.

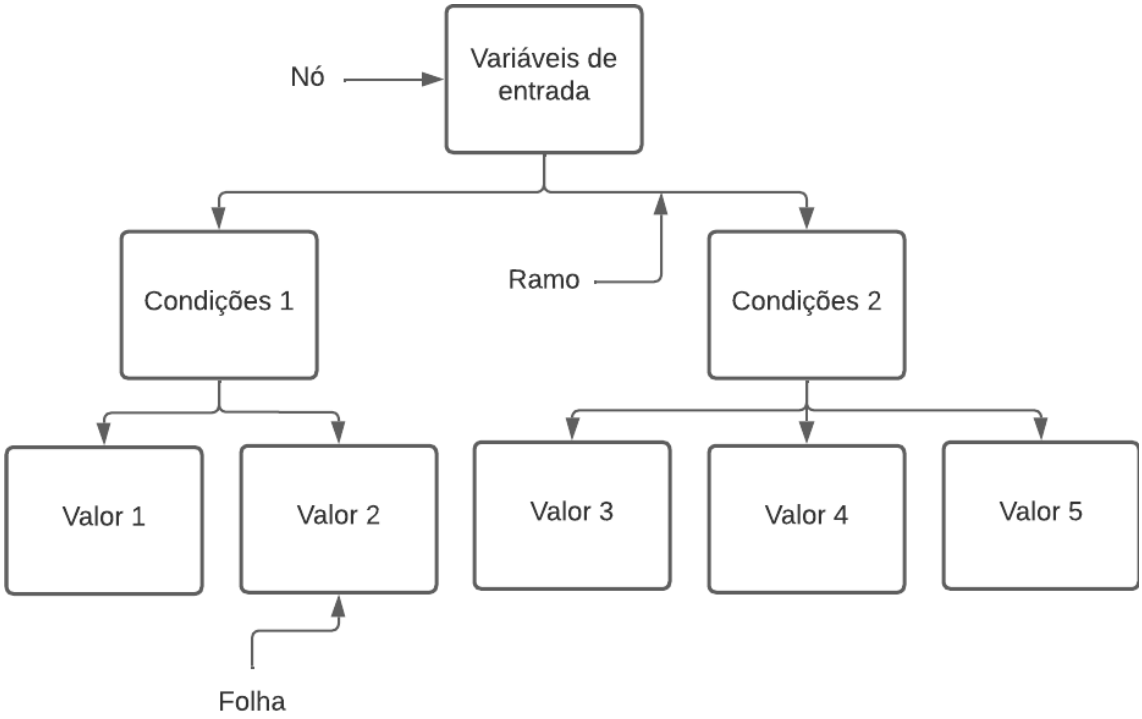
As árvores de decisão são formadas por nós, ramos e folhas. Os nós são os pontos internos à estrutura em que os dados encontram uma bifurcação. Os ramos são os caminhos para onde os dados podem seguir e as folhas representam os valores que a saída pode assumir (Figura 5).

Esses modelos são robustos e simples de implementar. Como as estruturas da árvore são auto-ajustadas durante o treinamento, o desenvolvedor não precisa se preocupar com a sua topologia, diferentemente das RNA. Isso facilita no desenvolvimento na manutenção do modelo, principalmente na adição e remoção das variáveis de entrada. Contudo, essas estruturas são mais sensíveis a alterações nas relações entre as variáveis de entrada e de saída. Dessa forma, pequenas mudanças no sistema físico podem inutilizar um modelo treinado, sendo necessário treiná-lo novamente, o que acarreta numa alta necessidade de manutenção para que esse modelo permaneça assertivo.

Aplicadas à regressão numérica, as árvores de decisão irão encontrar, para cada nó, um intervalo para cada variável de entrada. Cada intervalo levará a um ramo, que, por sua vez, levará a um novo nó ou a uma folha. Dessa forma, podemos considerar a variável de saída como sendo uma classificação (RUSSELL; NORVIG; MACEDO, 2013). Para cada valor que ela pode assumir, está associado um conjunto único de condições (intervalos) ligados a cada variável de entrada, de forma que para a saída assumir um valor específico todas as condições devem ser atendidas simultâneamente. Em outras palavras, cada variável de entrada pertencerá a um intervalo específico simultâneamente ligados ao caminho que leva até um determinado valor que a variável alvo irá assumir.

Árvores de decisão aprendem ajustando esses intervalos para cada nó, a fim de minimizar o erro entre cada amostra da variável alvo e os valores previstos. Elas também estão sujeitas a *overfitting*, quando o modelo se especaliza no conjunto de dados de treinamento, perdendo capacidade de generalização e, com isso, perdendo performance para novos conjuntos de dados. Por isso, é recomendado que ela seja o menor possível. A avaliação dos modelos de árvore de decisão é similar aos modelos de RNA e podem ser utilizadas as mesmas métricas de avaliação.

Figura 5 – Esquema de uma árvore de decisão.




Fonte: Autoria Própria.

### 2.1.2 Seleção de variáveis

A seleção de variáveis tem uma função crucial para aplicações de IA como um todo e, principalmente no treinamento e modelagem das RNA. Ela pode ser feita por diversos métodos, nos quais a maioria atribui uma pontuação às variáveis, indicando a sua similaridade. Isso significa que por meio delas é possível gerar uma classificação ordenada com as variáveis que estão mais relacionadas entre si. Dessa forma, o desenvolvedor pode escolher um subconjunto a partir das variáveis disponíveis e obter resultados similares e por vezes até melhor que se utilizasse todo o conjunto de variáveis disponíveis.

Um conjunto reduzido de variáveis de entrada implica em RNA com menos neurônios por camada, o que se traduz em velocidade de processamento dos dados em operação e tempo de treinamento reduzido. Isso facilita no desenvolvimento do modelo, visto que na grande maioria das vezes é feito por tentativa e erro, ajustando aos poucos os seus parâmetros até que se alcance um resultado satisfatório, atingindo métricas mínimas estipuladas pelo desenvolvedor. Além disso, ao utilizar uma quantidade de variáveis além da necessária, o desempenho da RNA pode ser prejudicado por conta do *overfitting*. Essa é uma situação em que a RNA se ajusta muito bem ao conjunto de dados de treinamento, porém tem um baixo desempenho quando exposta a novos dados, perdendo sua capacidade de generalização do problema.

Há diversos tipos de algoritmos de seleção de variáveis, variando de acordo com

o tipo de variável, numérica ou categórica, linearidade, linear ou não e também com a abordagem, que pode ser estatística, recursiva, inteligente, dentre outras. No presente trabalho, será utilizado o método da máxima relevância e mínima redundância (MRMR), que é um método automatizado para seleção de variáveis e funciona muito bem para relações lineares e não-lineares. Vale mencionar que a biblioteca utilizada para a construção do modelo baseado em árvores de decisão (XGBoost) já utiliza internamente um algoritmo próprio para seleção de variáveis, porém esse não será um ponto focal na monografia. Dessa forma, ambos os modelos serão alimentados pelo mesmo banco de dados, já com as variáveis pré-selecionadas.

#### 2.1.2.1 Máxima relevância e mínima redundância (MRMR)

O algoritmo de máxima relevância e mínima redundância foi proposto pela primeira vez em (DING; PENG, 2005) e aprimorada no trabalho de (ZHAO; ANAND; WANG, 2019). Em suma, os autores estipulam duas condições, uma de máxima relevância e outra de mínima redundância, para aplicar às variáveis. Essas condições são utilizadas em um algoritmo baseado em inteligência artificial que irá atribuir uma pontuação a cada uma das variáveis. As duas principais funções de otimização para variáveis contínuas são a diferença entre a estatística F e a correlação (FCD) (2.3) e o quociente entre a estatística F e a correlação (FCQ) (2.4). Note que a análise é feita entre as variáveis preditivas, ou atributos, e a variável alvo.

$$FCD = \max_{i \in \Omega_s} \left\{ F(i,h) - \frac{1}{|S|} \sum_{j \in S} |c(i,j)| \right\} \tag{2.3}$$

$$FCQ = \max_{i \in \Omega_s} \left\{ \frac{F(i,h)}{\frac{1}{|S|} \sum_{j \in S} |c(i,j)|} \right\} \tag{2.4}$$

### 2.1.3 Métricas de avaliação dos modelos

Para definir se os modelos desenvolvidos são boas representações da realidade, ou seja, se elas de fato aprenderam durante o treinamento, podem ser feitos testes para quantificar esse aprendizado e, posteriormente, comparar seus desempenhos. Assim, calculando o erro entre as previsões feitas pelos modelos com um novo conjunto de dados de entrada, separado especificamente para testes, e a saída real relacionada a esses dados de entrada, podemos analisar algumas medidas estatísticas ligadas a ele.

Algumas dessas medidas e que foram as utilizadas no presente trabalho são a média absoluta do erro (*Mean Absolute Error* - MAE) (2.5), a média do erro ao quadrado (*Mean Squared Error* - MSE) (2.6), a média do erro absoluto percentual (*Mean Absolute Percentage Error* - MAPE) (2.7) e o valor máximo do erro percentual (MEP) 2.8. Onde $y_i$ é o valor real da i-ésima amostra, $\overline{y_i}$ é o valor previsto pelo modelo e $N$ é a quantidade de

amostras de teste. Todas essas são medidas bem conhecidas e amplamente utilizadas na avaliação de modelos de inteligência artificial voltados à regressão numérica.

Note que o MEP não é tão utilizado quanto as outras métricas, porém ele diz muito a repeito do conjunto de dados. Juntamente com o MAPE, ele representa o quão homogêneo é o conjunto de previsões, indicando qual o máximo valor percentual do erro cometido pelo modelo.

$$MAE = \sum_{i=1}^{N} \frac{|y_i - \overline{y_i}|}{N} \tag{2.5}$$

$$MSE = \sum_{i=1}^{N} \frac{(y_i - \overline{y_i})^2}{N} \tag{2.6}$$

$$MAPE = \sum_{i=1}^{N} \frac{|y_i - \overline{y_i}|}{y_i \cdot N} \cdot 100\% \tag{2.7}$$

$$MEP = \max_{i \in [1,N]} \left( \frac{|y_i - \overline{y_i}|}{y_i} \right) \cdot 100\% \tag{2.8}$$

## 2.2 Incertezas de medição

Qualquer instrumento de medição tem um erro de medição associado a ele. Dessa forma, toda medição feita será composta de um valor medido e a incerteza, que será o desvio padrão. Assim, o valor verdadeiro da variável medida deve estar compreendido no intervalo de confiança que é descrito como o valor medido mais ou menos o desvio padrão, como descrito em 2.9. Onde $i_{conf}$ é o intervalo de confiança, $m$ é o valor medido e $V_v$ é o valor verdadeiro da variável medida. (CAMPILHO, 2011).

$$\begin{aligned} i_{conf} &= [m - \sigma, m + \sigma] \\ V_v &\in i_{conf} \end{aligned} \tag{2.9}$$

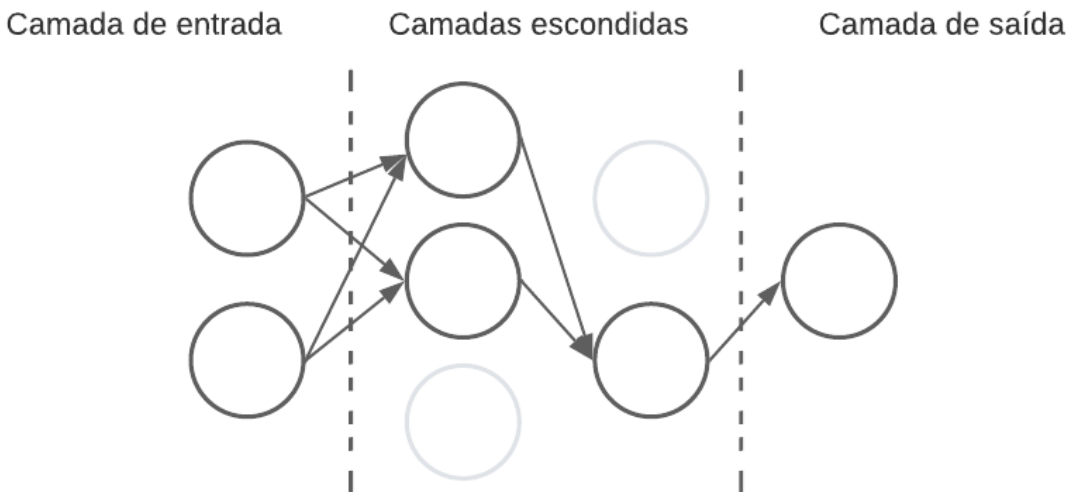
Para um SV, medir a incerteza associada à medição é uma tarefa que depende da abordagem utilizada na construção do modelo matemático. Para SV baseados em RNA, uma abordagem que pode ser utilizada é o Monte Carlo Dropout.

### 2.2.1 Monte Carlo Dropout

O *droupout* é uma técnica de regularização dos modelos de RNA. As técnicas de regularização têm a função de ajustar o modelo a fim de evitar o *overfitting*, deixando a rede o mais generalista possível, sem que ela perca muito com relação à exatidão das suas previsões. O *dropout* consiste em desativar, aleatoriamente, neurônios das camadas de entrada e das camadas escondidas, de acordo com as probabiblidades de desativação $p_{in}$ e $p$,

respectivamente (Figura 6). A técnica Monte Carlo Dropout consiste em aplicar o *dropout* também durante o treinamento da RNA e pode ser utilizada para calcular a incerteza que está intrinsecamente ligada ao modelo, ou incerteza epistêmica (RODRIGUES, 2022).

Figura 6 – Esquema de uma RNA com camadas de *dropout*.




Fonte: Autoria própria.

#### 2.2.1.1 Cálculo de incertezas

Por conta da desativação dos neurônios, os pesos da RNA podem ser tratados como variáveis aleatórias, sendo possível aproximar a RNA por uma distribuição Bayesiana a *posteriori*. Essa aproximação é o que permite a análise de incertezas da RNA, uma vez que ela será a dispersão dessa distribuição. Dessa forma, são feitas $N$ previsões para o conjunto de dados de teste. Para cada previsão é aplicado o *dropout* das camadas, obtendo assim um conjunto de $N$ previsões, cada um com $k$ amostras. Para cada uma das $N$ previsões, haverá uma matriz de pesos da RNA, cada uma representando amostras das variáveis aleatórias em análise. Portanto, é possível calcular, para cada uma das amostras, uma média e uma variância e, calculando o desvio padrão a partir dela, podemos traçar um intervalo de confiança para as medições. A média e a variância são calculadas pelas equações 2.10 e 2.11 (GAL; GHAHRAMANI, 2015).

$$\mathbb{E}(\hat{y}) = \frac{1}{N}\sum_{t=1}^{N}\hat{y}_t \tag{2.10}$$

$$Var(\hat{y}) \approx \tau^{-1} + \frac{1}{N}\sum_{t=1}^{N}\hat{y}_t^N\hat{y}_t - \mathbb{E}(\hat{y})^N\mathbb{E}(\hat{y}) \tag{2.11}$$

Onde:

$$\tau = \frac{pl^2}{2K\lambda} \tag{2.12}$$

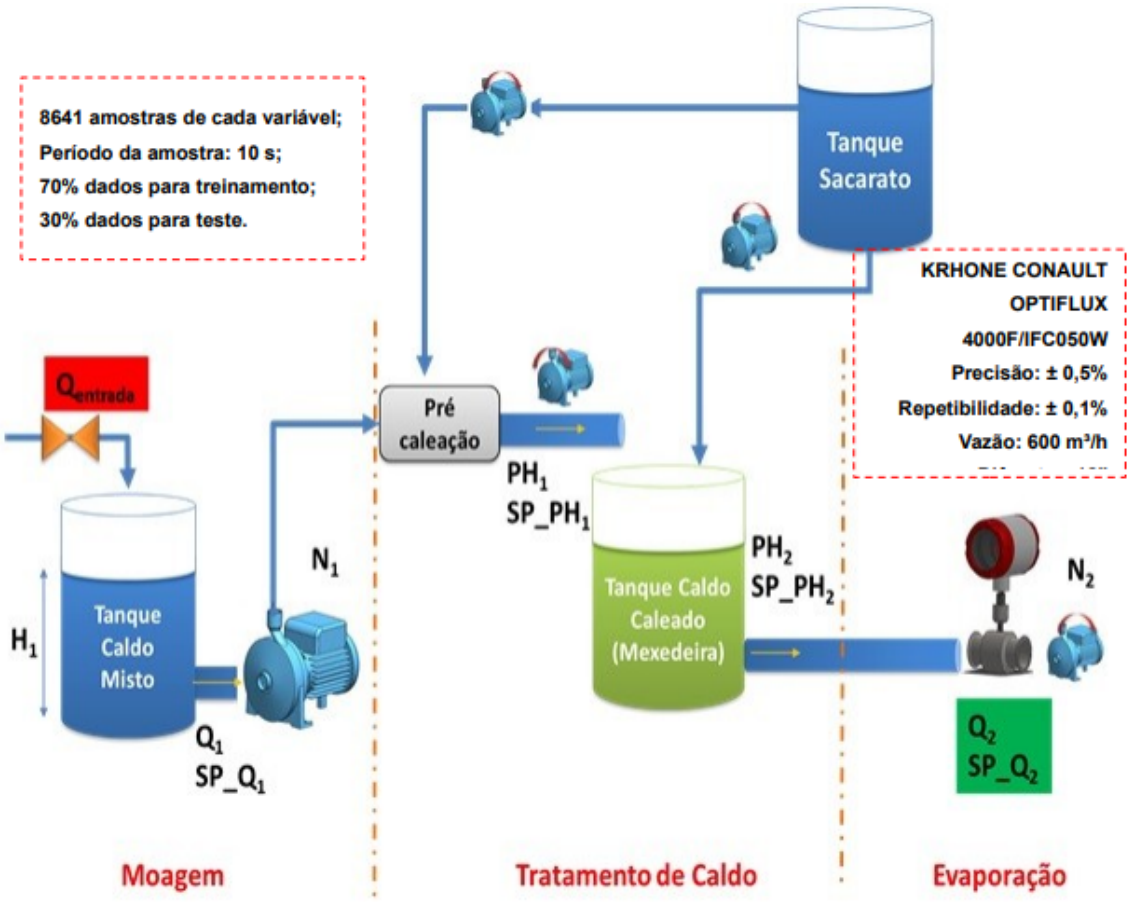
Na equação 2.12, $\tau$ é a precisão do modelo, $p$ é a probabilidade de um neurônio aleatório de uma camada escondida ser desativado, $l$ é o *length scale*, que representa a confiabilidade dos dados de entrada. Além disso, $K$ é o número de amostras de treinamento, $\lambda$ é o peso da regularização L2 e $T$ é o número de previsões. Note que número de previsões é diferente de número de amostras. Por exemplo, podemos ter 100 previsões, cada uma com 500 amostras.

# 3 Materiais e métodos

## 3.1 Estudo de caso

O sensor virtual desenvolvido neste trabalho foi baseado no trabalho de (LIMA, 2022a). O instrumento de medição foi elaborado com base nos dados fornecidos pela autora do trabalho citado e são relativos a uma planta industrial do setor sucroenergético. A planta em questão faz parte da Usina Central Olho D'Água (USICODA), que produz açúcar, álcool e produtos derivados e gera energia elétrica por meio da cogeração de biomassa. Ela fica localizada no município de Camutanga - PE.

Figura 7 – Esquema da planta industrial.




Fonte: (LIMA, 2022a)

### 3.1.1 A planta

O esquema da planta, que está ilustrada na Figura 7, consiste em dois reservatórios, sendo eles o tanque de Sacarato e o tanque de caldo misto. O caldo misto e o sacarato são

misturados na proporção adequada na pré-caleação. O pré-caleado é enviado ao terceiro reservatório, o tanque de caldo caleado, onde será misturado com mais sacarato. As proporções são controladas por bombas nas linhas, nos locais indicados na Figura 7.

A mistura adequada desses caldos é o objetivo do sistema de controle da planta. A qualidade dos produtos gerados e a eficiência da planta dependem diretamente da desse processo. O SV desenvolvido tem o intuito de emular o sensor de vazão do caldo caleado. Esse sensor é crucial para a operação da planta, pois a vazão do caldo caleado está ligada diretamente a parâmetros dele, como cor, acidez e até mesmo a população bacteriana presente nele. Estes parâmetros devem ser controlados com exatidão, a fim de garantir a qualidade do açúcar e otimizar a distribuição de vapor na planta, o que interfere diretamente na exportação de energia feita pela usina. Seu custo de aquisição é consideravelmente alto (R$ 45000,00, em 2022), quando comparado com sensores do mesmo tipo e, além disso, tem a necessidade de ajustes e manutenções periódicas, como qualquer sensor físico, o que encarece sua operação (LIMA, 2022a).

### 3.1.2 Banco de dados

O banco de dados é composto por dez variáveis, sendo elas nove variáveis secundárias e uma variável de interesse, a qual se deseja medir por meio do instrumento virtual. Essas variáveis são:

- **Q1:** Vazão de saída do tanque de caldo misto;
- **SP_Q1:** Set point de vazão de saída do tanque de caldo misto;
- **PH1:** PH do caldo pré caleado;
- **SP_PH1:** Set point do pH do caldo pré-caleado;
- **N1:** Rotação da bomba de caldo misto;
- **N2:** Rotação da bomba de caldo caleado;
- **H1:** Nível do tanque de caldo misto;
- **PH2:** PH do caldo caleado;
- **SP_PH2:** Set point do pH do caldo caleado;
- **Q2:** Vazão de saída do tanque de caldo caleado (mexedeira).

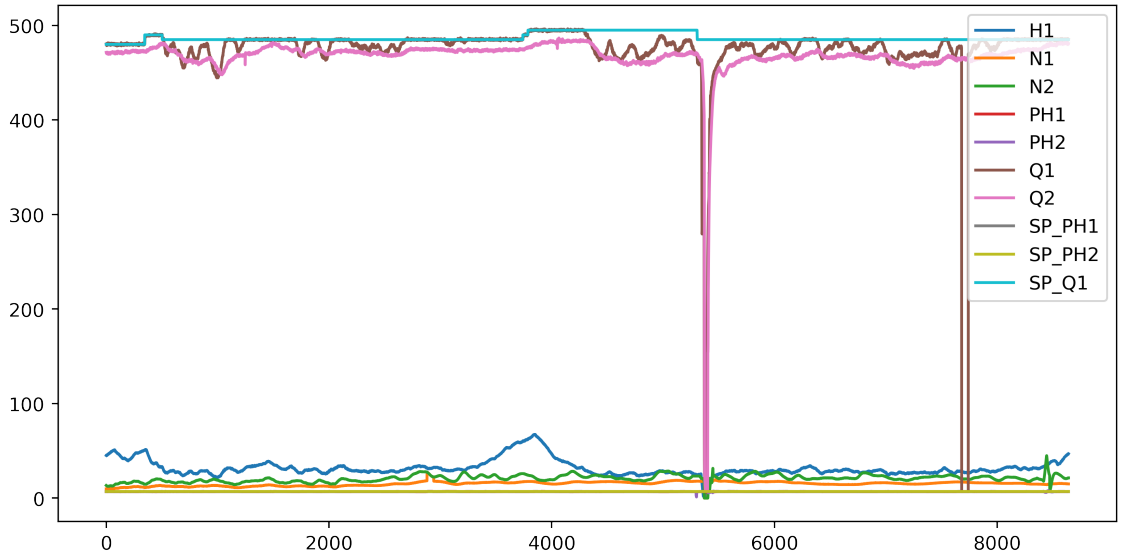
Esses dados correspondem ao funcionamento da planta no dia 20 de fevereiro de 2020, contendo 8641 amostras de cada variável. Note que as variáveis SP_Q1, SP_PH1 e SP_PH2 são *set points* do sistema de controle, portanto, seu valor é praticamente

constante, podendo ser descartadas sem a necessidade de um método matemático para seleção de variáveis, pois não são dados de medições. Dessa forma, o banco de dados utilizado será composto pelas outras variáveis.

### 3.1.3 Tratamento dos dados

Antes de realizar o treinamento dos modelos é necessário fazer o tratamento dos dados, visando remover ruídos e *outliers*. Para isso, foi feita a substituição dos *outliers* pela média e aplicado um filtro de média móvel, duas técnicas bastante usuais. Na Figura 8, estão plotados os dados de todas as variáveis, ainda sem tratamento.

Figura 8 – Banco de dados sem tratamento.



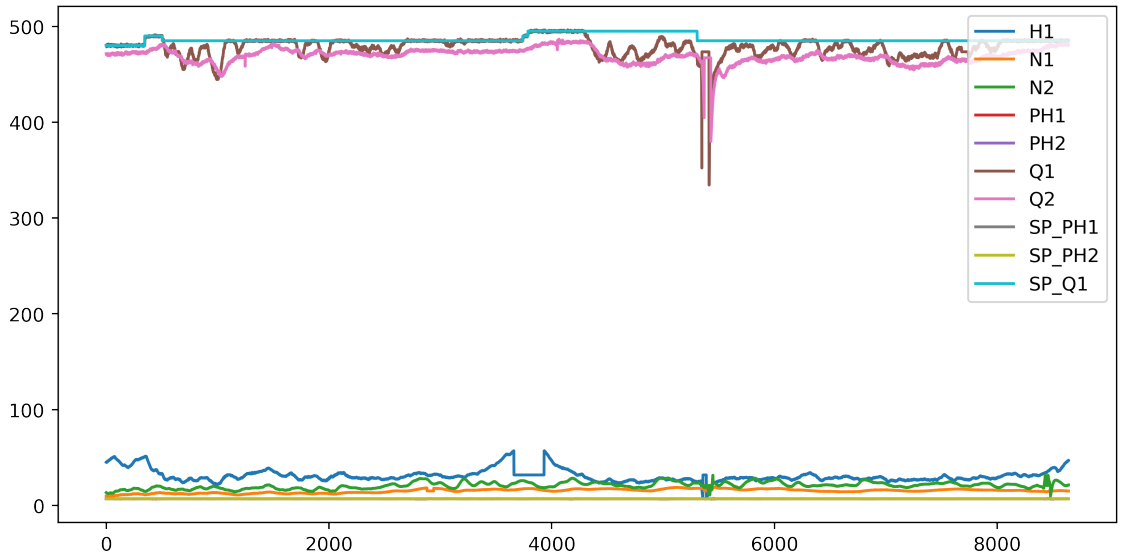

Fonte: Autoria própria.

#### 3.1.3.1 Remoção de *outliers*

Os *outliers* são amostras das variáveis que se comportam de forma distinta das demais, fugindo da tendência do conjunto global de amostras. Esses valores não podem ser utilizados como dados de entrada para métodos de previsão ou análise baseados em dados visto que, por se tratarem de erros de medição ou perda de dados durante a transmissão, são valores que não têm correlação com o fenômeno observado. Uma forma de lidar com essas falhas é substituir esses dados pela média do conjunto de dados da variável em questão.

Para um conjunto de dados com tendência constante, são considerados *outliers* valores que não estejam no intervalo $[\overline{y} - 3\sigma, \overline{y} + 3\sigma]$, onde $\overline{y}$ é a média do conjunto de dados e $\sigma$ é o desvio padrão. Dessa forma, substituímos os valores que estão fora desse intervalo pela média do conjunto de dados, para todas as variáveis. Os resultados estão ilustrados na Figura 9.

Figura 9 – Banco de dados após remoção de *outliers*.




Fonte: Autoria própria.

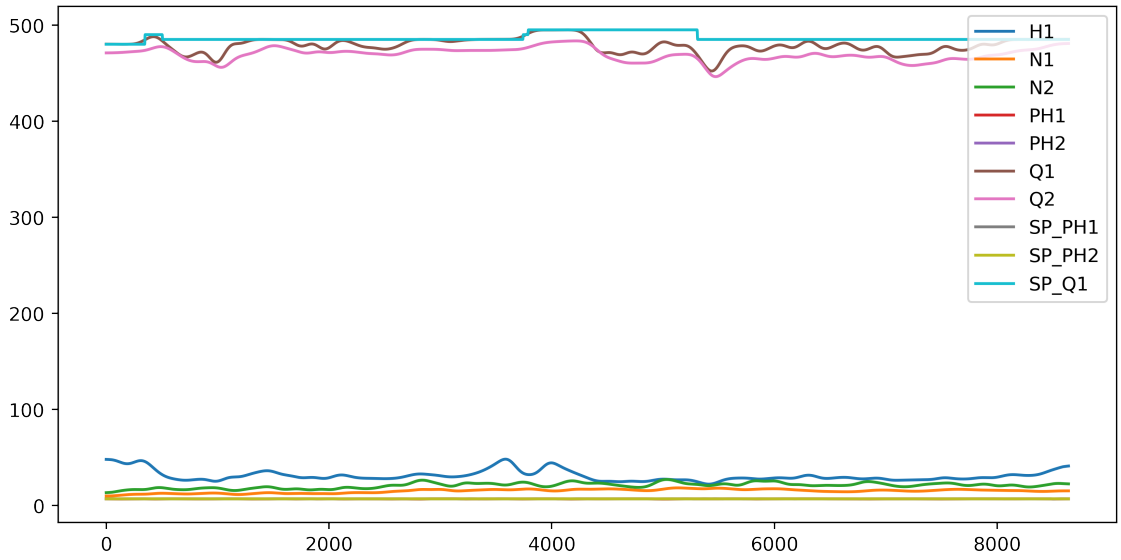
#### 3.1.3.2 Filtro de média móvel

Outra característica geral dos conjuntos de dados não tratados é o ruído associado à medição e transmissão dos dados. Esse ruído pode ser proveniente de interferência eletromagnética dos diversos equipamentos presentes na planta, perturbações externas nos sensores, dentre outros, alterando levemente o valor das variáveis. Geralmente, esse ruído é um sinal aleatório de alta frequência somado ao valor medido. Assim, por meio de filtros é possível remover ou, pelo menos, suprimir esse sinal, restando um conjunto de dados que se aproxima do que seria o equivalente ao sinal puro da variável medida.

Uma das formas de fazer isso, é utilizando um filtro de média móvel. Ele calcula a média das medições em subconjuntos adjacentes a cada uma das amostras da variável e as substitui por essa média. Esse subconjunto é conhecido como janela ou *kernel*. Para o banco de dados fornecido, foi utilizada uma janela contendo 51 amostras. Note que, para uma janela com um número ímpar de amostras, utilizaremos $(N-1/2)$ amostras anteriores e posteriores a ela. Dessa forma, chamaremos este valor de $R$, que, neste caso, assumirá o valor $R = 25$. Portanto, para um vetor com $N$ amostras, temos que cada amostra $x_i$ será calculado de acordo com 3.1. Na Figura 10, está disposto o resultado da aplicação da média móvel no banco de dados.

$$x_i = \begin{cases} \sum_{j=0}^{i+R} x_j/(R+i), & i < R \\ \sum_{j=i-R}^{i+R} x_j/(2R+1), & R \leqslant i \leqslant N-R \\ \sum_{j=i}^{N} x_j/(N+R-i), & N-i < R \end{cases} \tag{3.1}$$

Figura 10 – Banco de dados após tratamento.




Fonte: Autoria própria.

## 3.2 Incertezas da RNA

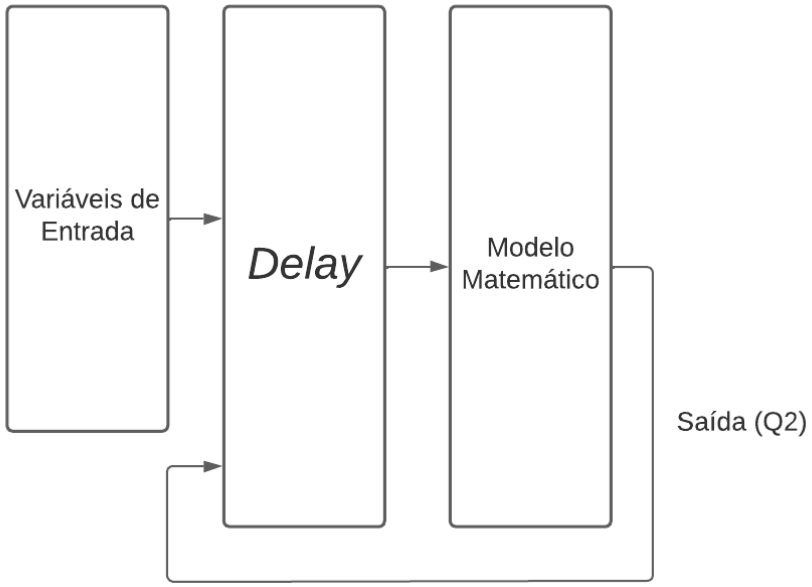
## 3.3 Topologia geral da aplicação

Os dois modelos desenvolvidos são baseados na topologia do modelo 2 (Figura 11), desenvolvido no trabalho de (LIMA, 2022a). O primeiro modelo do presente trabalho também é baseado em RNA e visa obter resultados similares aos alcançados pela autora no estudo citado. Já o segundo modelo desenvolvido tem a mesma estrutura, porém substituindo a RNA por uma árvore de decisão aplicada à regressão.

É ilustrado na Figura 11 o modelo geral, sendo ele composto pelos estágios de tratamento de dados, seleção de variáveis e modelo matemático. Para facilitar o entendimento, chamaremos os modelos desenvolvidos neste trabalho de Modelo RNA e Modelo XGBoost, sendo eles, respectivamente, baseados em RNA e árvores de decisão. Serão analisados os desempenhos dos modelos para três conjuntos de variáveis preditivas, sendo eles o conjunto completo, formado pelas seis variáveis preditivas e o *delay* da saída do modelo, e outros dois conjuntos contendo, respectivamente, as cinco e quatro variáveis mais importantes do conjunto original com relação à variável de saída.

A comparação entre os modelos será feita de acordo com as métricas convencionais de avaliação de modelos de regressão, já citados anteriormente, para cada conjunto de variáveis selecionado. Também será feita a avaliação de incertezas do Modelo RNA, que será calculada por meio da técnica Monte Carlo dropout. Também será analisado o efeito do conjunto de variáveis de entrada na incerteza do Modelo RNA.

Figura 11 – Esquema genérico do modelo do sensor virtual.




Fonte: Autoria Própria.

# 4 Resultados

Em um primeiro momento, foram gerados dois conjuntos de resultados, cada um referente a um conjunto de variáveis preditivas. Cada conjunto é composto por dois resultados, cada um referente ao treinamento dos respectivos modelos com as sete e cinco variáveis preditivas mais significativas. Os resultados dos modelos são comparados por conjunto. Em seguida, é feito o mesmo procedimento, porém apenas com o Modelo RNA, aplicando a técnica Monte Carlo dropout, a fim de analisar o impacto da retirada de variáveis de entrada na incerteza do modelo. Note que, para o Modelo RNA, é necessário ajustar as quantidades de neurônios por camada para cada conjunto de dados. A proporção ideal encontrada foi de, para cada $N$ variáveis de entrada, a primeira e a segunda camadas escondidas terão $2N$ e $N$ neurônios, respectivamente. Esta configuração foi obtida empíricamente, sendo a que teve melhor performance nos treinamentos, para um conjunto de entrada contendo sete ou cinco variáveis. Como há uma variável de saída, só haverá um neurônio na última camada. Dessa forma, a RNA terá a configuração $N-2N-N-1$. Para o Modelo RNA, foi utilizada a função de ativação ReLu. Os dados foram normalizados apenas na etapa da análise de incertezas, uma vez que não houve necessidade da aplicação de tal técnica na primeira parte do desenvolvimento do modelo.

Os parâmetros mínimos para os modelos são: $MSE < 5 \cdot 10^{-3}$, $MAPE < 5 \cdot 10^{-2}\%$, $MAE < 5 \cdot 10^{-2}$ e $MEP < 7 \cdot 10^{-2}$. Os conjuntos de teste e treinamento foram divididos na proporção $70\% : 30\%$, para os dois modelos. Todo o desenvolvimento do trabalho foi feito na linguagem de programação Python. Para o Modelo RNA, foi utilizado o módulo Keras, da biblioteca Tensorflow e, para o Modelo XGBoost, foi utilizado o módulo XGBRegressor, da biblioteca XGBoost. Por fim, a seleção de variáveis foi feita utilizando o módulo mrmr_regression da biblioteca mrmr.

## 4.1 Influência da seleção de variáveis

O conjunto 1 é composto pelos resultados dos testes dos modelos treinados com todo o banco de dados, cujo tratamento foi apresentado no capítulo anterior. Assim, as variáveis de entrada utilizadas no treinamento dos modelos serão Q2_delay, Q1, H1, N1, PH1, PH2, N2 (Figura 12). Note que a ordem em que as variáveis foram dispostas é a ordem de importância que o algoritmo de seleção (MRMR) atribuiu, com relação à sua similiaridade com a variável de saída Q2. Os resultados estão dispostos na Tabela 1.

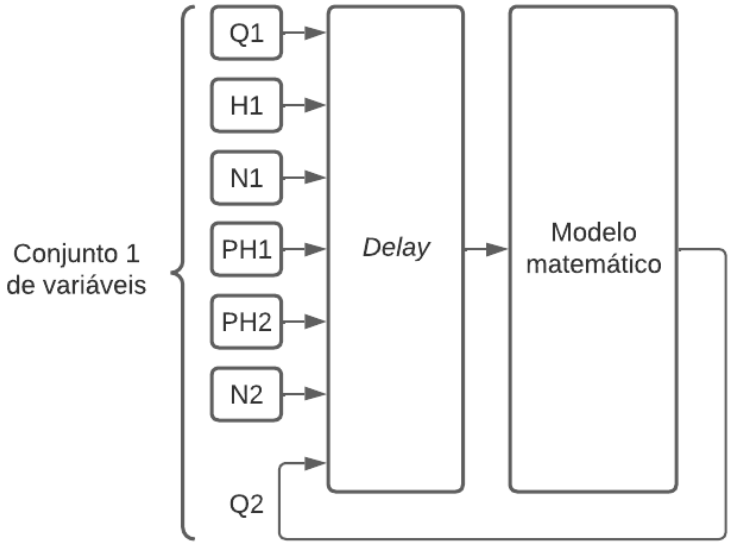
Observando as métricas de avaliação, temos que, embora o MSE e o MAE dos dois modelos sejam os mesmos, o MAPE e o MEP do Modelo RNA são menores, o que indica um desempenho melhor do modelo. A partir desses dados, é possível inferir que o erro do

Tabela 1 – Métricas dos modelos para sete variáveis de entrada.

| Modelo | MSE | MAPE | MAE | MEP |
|---|---|---|---|---|
| RNA | $8,96 \cdot 10^{-4}$ | $5,13 \cdot 10^{-3}\%$ | $2,4 \cdot 10^{-2}$ | $1,92 \cdot 10^{-2}$ |
| XGBoost | $8,96 \cdot 10^{-4}$ | $1,98 \cdot 10^{-2}\%$ | $2,4 \cdot 10^{-2}$ | $1,57 \cdot 10^{-1}$ |

Fonte: Autoria própria.

Figura 12 – Conjunto 1 de variáveis aplicado aos modelos.




Fonte: Autoria própria.

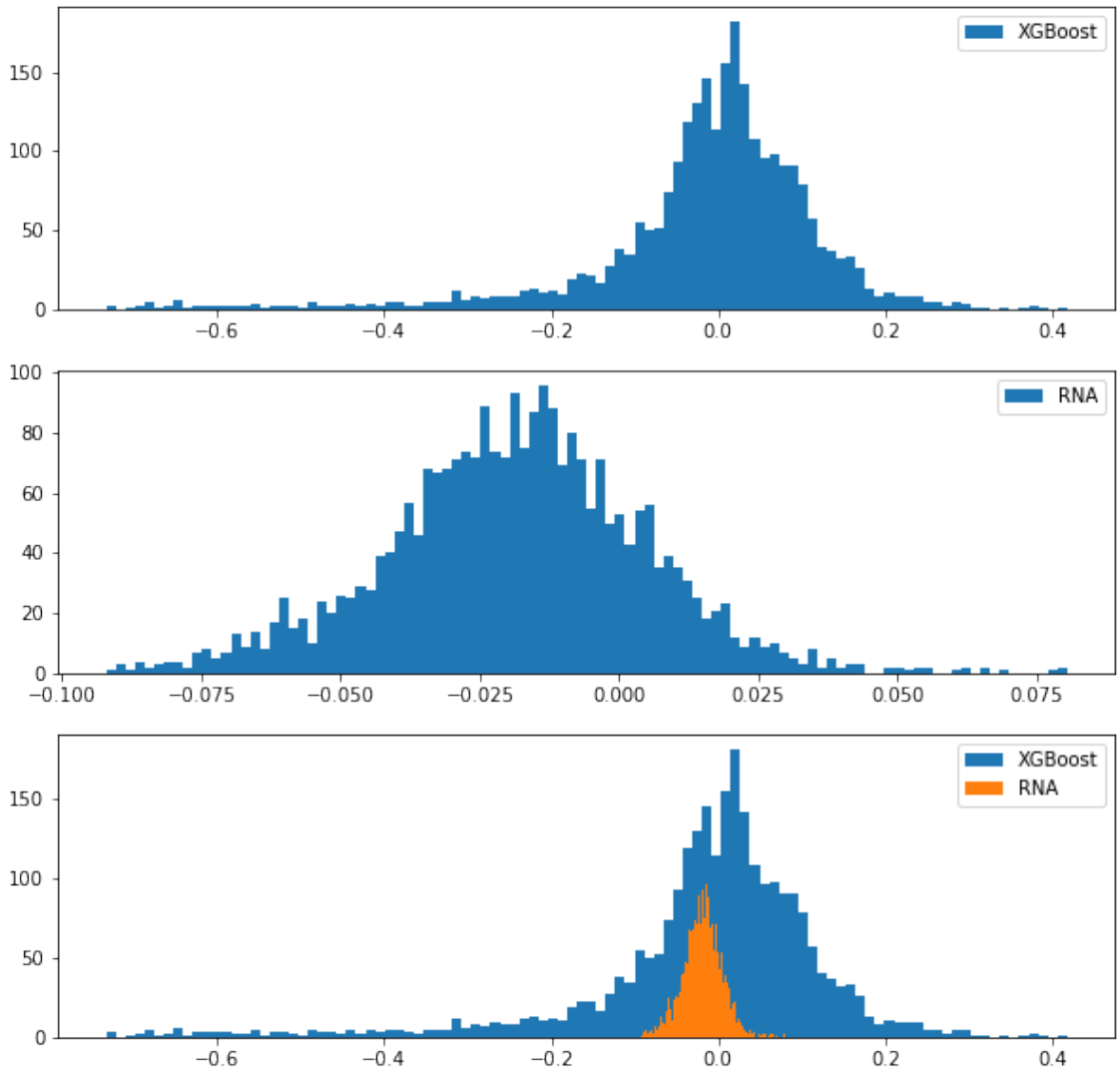
Modelo RNA é mais concentrado que o erro do Modelo XGBoost. Isso significa que, para cada amostra, o erro do Modelo RNA é mais próximo de um valor central (MAE) que o erro do Modelo XGBoost. Isso pode ser verificado pelo histograma do erro, ilustrado na Figura 13.

As previsões dos dois modelos para o conjunto de testes pode ser verificada na Figura 14. É possível perceber que a saída do Modelo XGBoost é formada por patamares. Isso se dá por conta da tentativa do modelo de classificar a variável de saída no menor número de categorias possível. Caso isso não fosse feito, a árvore de decisão resultante teria um número muito grande, teoricamente infinito, de folhas. Dessa forma, a árvore resultante seria impraticável.

O conjunto 2, com cinco variáveis de entrada, é composto por Q2_delay, Q1, H1, N1 e PH1 (Figura 15). Isso significa que, fora Q2_delay, todas as variáveis utilizadas nos modelos são referentes às grandezas do caldo misto e pré-caleado. As métricas para esse conjunto estão dispostas na Tabela 2. Comparando com os dados da Tabela 1, nota-se que todas as métricas foram levemente prejudicadas pela retirada das variáveis. Contudo, essa perda de desempenho não é muito significativa para ambos os modelos.

Os histogramas de erro e gráficos das previsões feitas com o conjunto 2 estão

Figura 13 – Amostras por valor de erro para o conjunto 1.




Fonte: Autoria própria.

Tabela 2 – Métricas dos modelos para cinco variáveis de entrada.

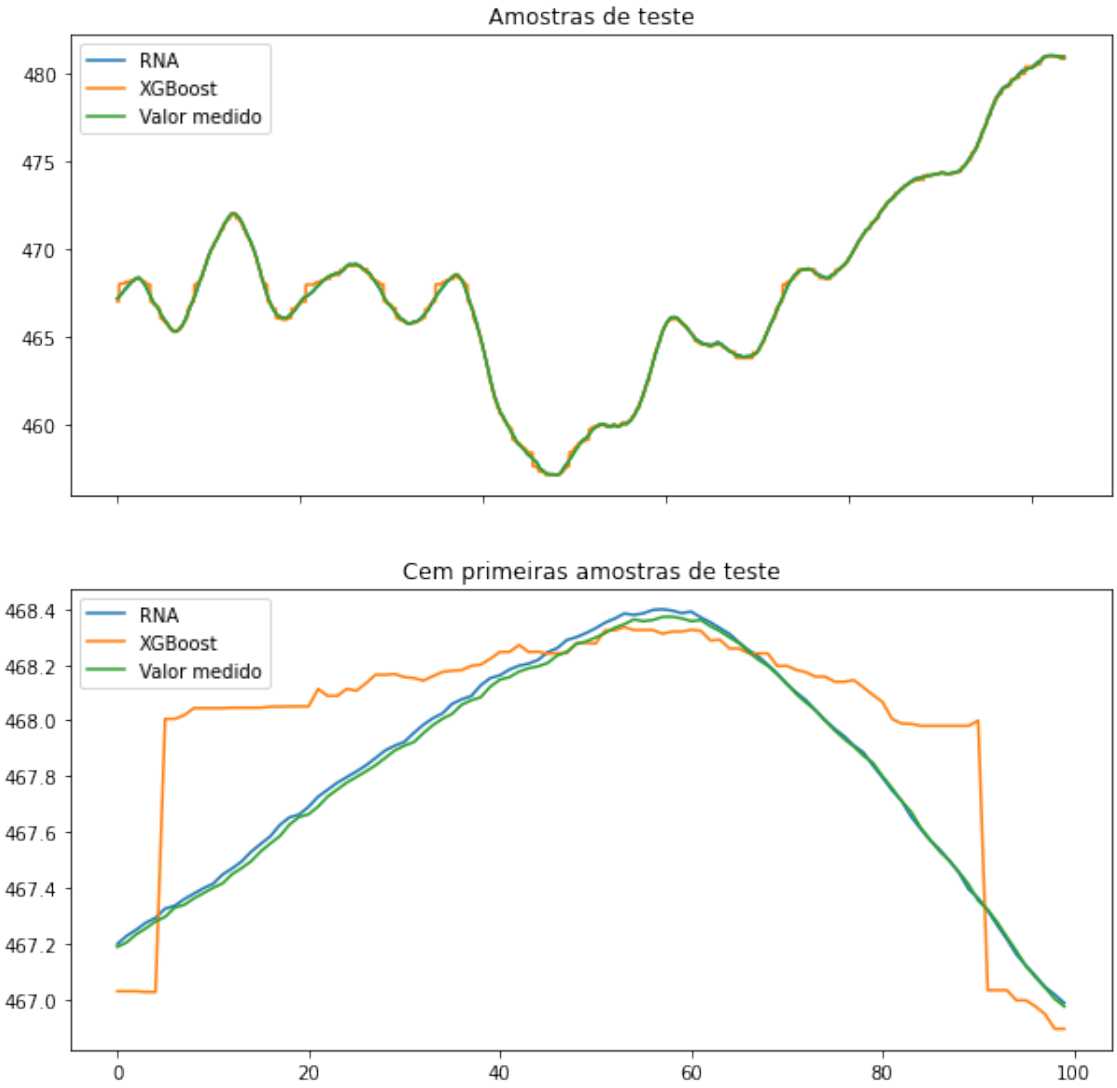
| Modelo | MSE | MAPE | MAE | MEP |
|---|---|---|---|---|
| RNA | $1,02 \cdot 10^{-3}$ | $5,61 \cdot 10^{-3}\%$ | $2,62 \cdot 10^{-2}$ | $2,0 \cdot 10^{-2}$ |
| XGBoost | $1,03 \cdot 10^{-3}$ | $1,78 \cdot 10^{-2}\%$ | $2,62 \cdot 10^{-2}$ | $1,28 \cdot 10^{-1}$ |

Fonte: Autoria própria.

dispostos nas Figuras 16 e 17. Analisando as duas figuras, é possível perceber que, de fato, os resultados pouco sofreram variação. Podemos concluir então que as variáveis medidas que mais interferem na vazão do caldo caleado são aquelas medidas pelos sensores montante do ponto de medição da vazão, com excessão da própria vazão medida e realimentada após aplicado *delay*.

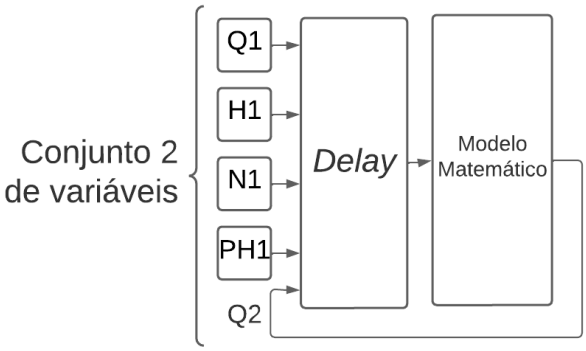
Para todos os conjuntos de variáveis de entrada, o Modelo RNA apresentou resultados que atingiam os critérios mínimos estabelecidos. Já o Modelo XGBoost, para todos

Figura 14 – Previsões dos modelos para o conjunto 1.



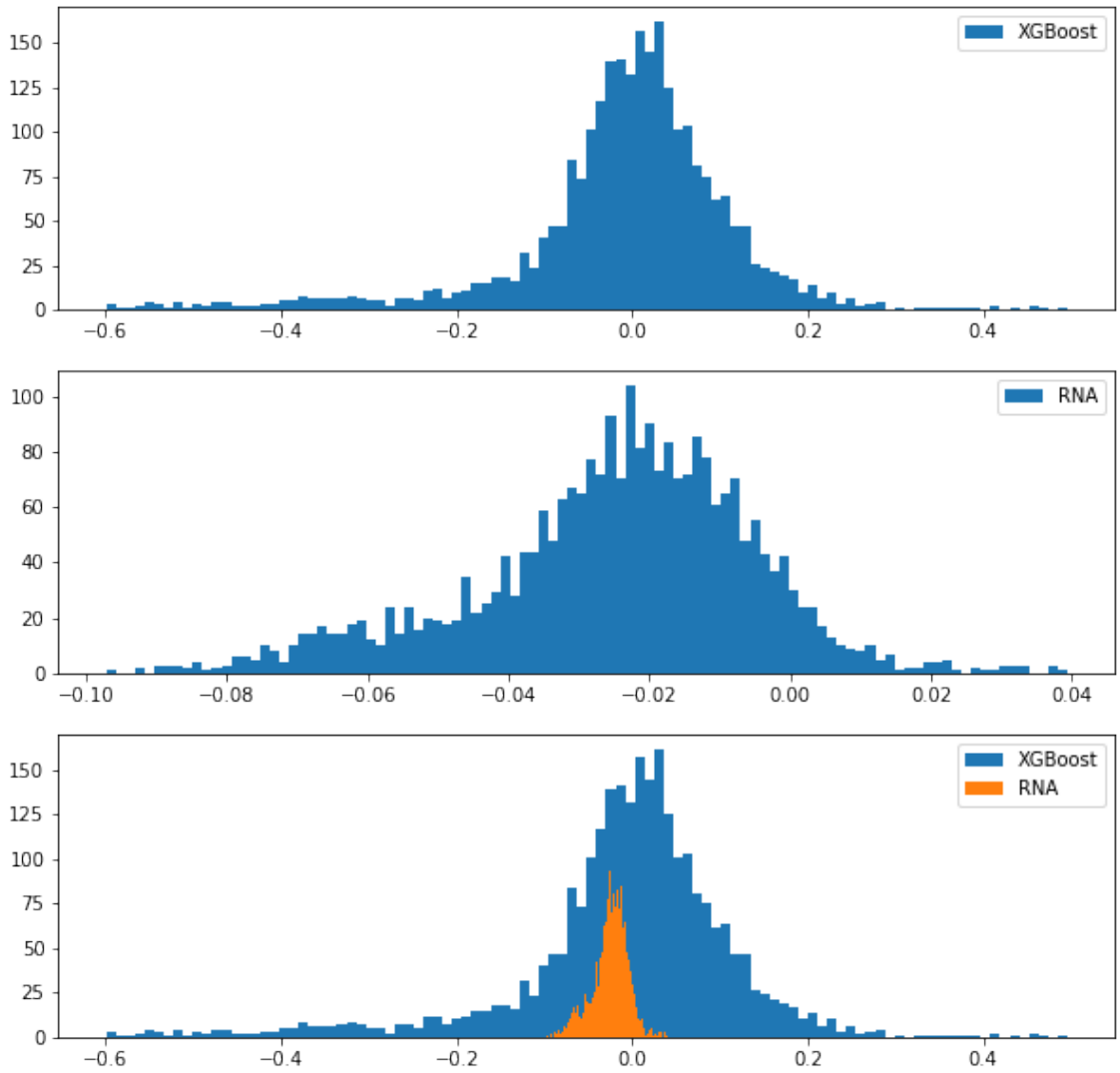

Fonte: Autoria própria.

Figura 15 – Conjunto 2 de variáveis aplicado aos modelos.




Fonte: Autoria própria.

Figura 16 – Amostras por valor de erro para o conjunto 2.



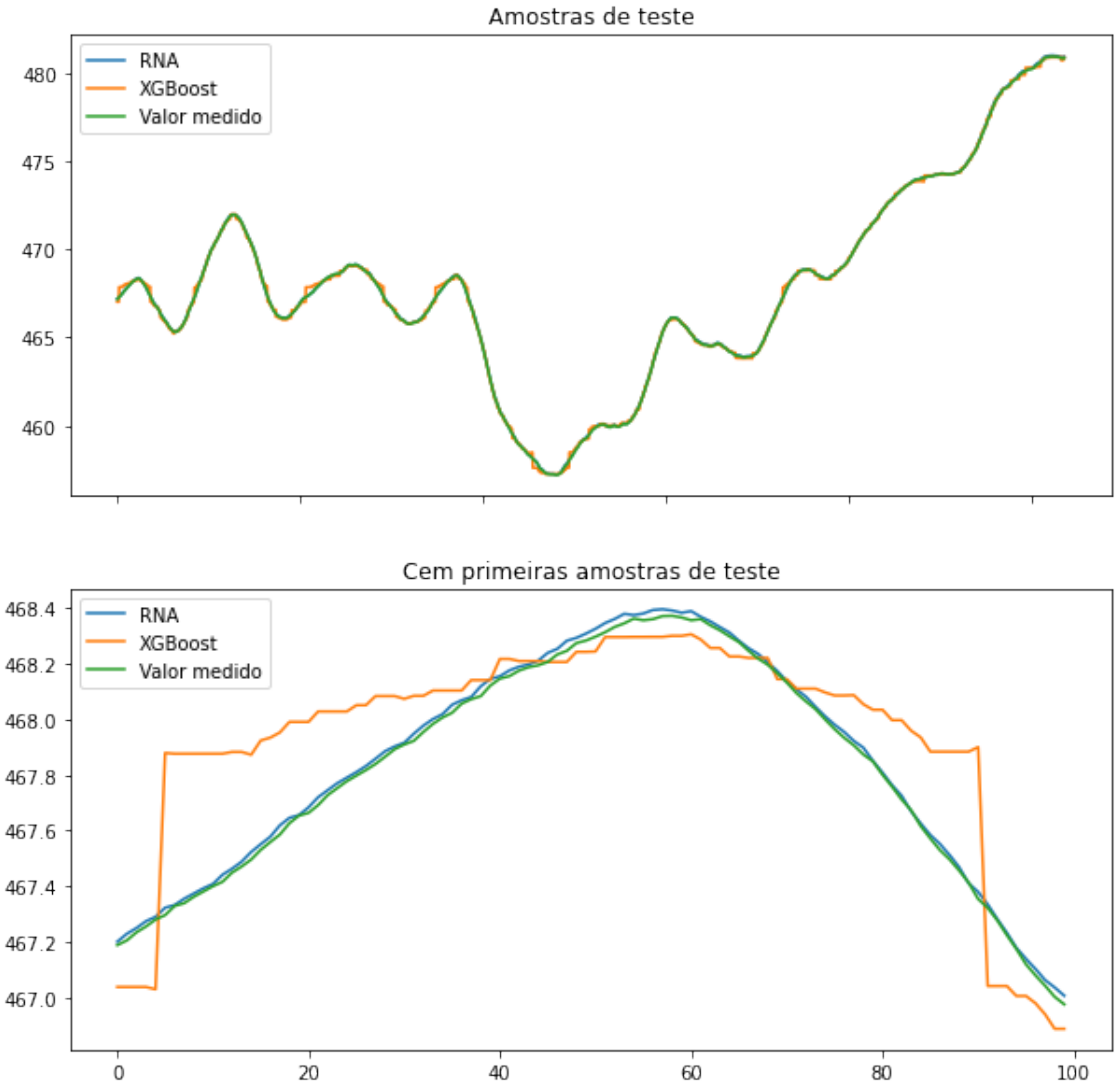

Fonte: Autoria própria.

os conjuntos de variáveis, apresentou valores de $MEP$ acima do estipulado. As outras métricas estão dentro dos limites e, visualmente, os resultados expostos nos gráficos das Figuras 14 e 17 estão aceitáveis, na maior parte de sua extensão. Contudo, há pontos em que as previsões se distanciam demasiadamente do valor real medido. Portanto, o Modelo XGBoost, embora seja muito robusto, apresentando pouca variação de resultados entre os testes, não atende às necessidades do problema por si só, sendo necessária a aplicação de um filtro para suavizar o sinal de saída gerado por ele.

## 4.2 Análise de incertezas do Modelo RNA

A análise de incertezas do Modelo RNA foi feita utilizando a técnica Montecarlo dropout. O cálculo das incrtezas também foi feito de acordo com os conjuntos de variáveis

Figura 17 – Previsões dos modelos para o conjunto 2.




Fonte: Autoria própria.

selecionadas, aplicados na seção anterior. Os parâmetros utilizados, aplicados diretamente em 2.11 e 2.12, estão dispostos na Tabela 3. Esses parâmetros foram escolhidos com base nas observações de Rodrigues (2022) e por experimentação ao longo do desenvolvimento do trabalho.

O valor de $l$ foi escolhido arbitrariamente, estimando uma incerteza global dos sensores utilizados na aquisição de dados da ordem do valor utilizado. Já o valor do regularizador L2 foi escolhido por meio de tentativa e erro. O baixo valor de L2 impacta numa menor incerteza, analisando pelas equações 2.11 e 2.12. Apesar disso, o valor de L2 não pode ser aumentado levianamente, a fim de aumentar o intervalo de confiança, pois ele interfere nos pesos da RNA. A regularização L2 tem o objetivo de evitar *overfitting* penalizando pesos muito discrepantes. Isso significa que, caso uma variável de entrada tenha uma afinidade muito maior que as outras com a variável de saída, os pesos atribuídos às

Tabela 3 – Parâmetros da RNA para o Monte Carlo dropout.

| Parâmetro | Valor | Comentário |
|---|---|---|
| p | 0,25 | Probabilidade de um neurônio aleatório de uma camada escondida ser desativado |
| l | 0,1 | Confiabilidade dos dados de entrada |
| K | 6048 | Número de amostras de treinamento |
| $\lambda$ | $1 \cdot 10^{-}6$ | Peso da regularização l2 |
| T | 1000 | Número de previsões |

Fonte: Autoria própria.

ligações com essa variável serão proporcionalmente maiores que os pesos das ligações com as outras variáveis. Dessa forma, uma regularização que penalize esses pesos desproporcionais irá piorar o resultado geral das predições, afastando-o do valor real. Portanto, embora a regularização L2 aumente o intervalo de confiança, dando uma maior margem de erro para o modelo, ela tende a piorar o resultado geral do modelo à medida em que aumenta.

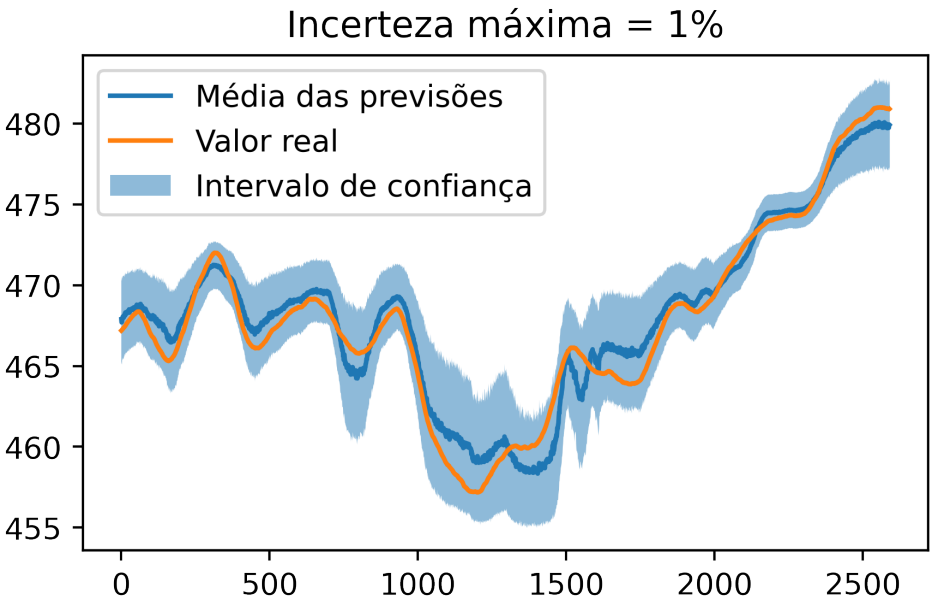
Aplicando a tecnica Monte Carlo dropout, foram calculadas as incertezas percentuais máximas para cada conjunto de vriáveis de entrada utilizado. Para os conjuntos 1 e 2 de variáveis de entrada, a incerteza máxima associada é, respectivamente, $\sigma_1 = 1\%$ e $\sigma_2 = 1,6\%$.

O efeito da ausência das duas variáveis no conjunto 2 tem um impacto mais significativo na análise de incertezas do que no desempenho do Modelo RNA, como pode ser verificado na Figura 18.

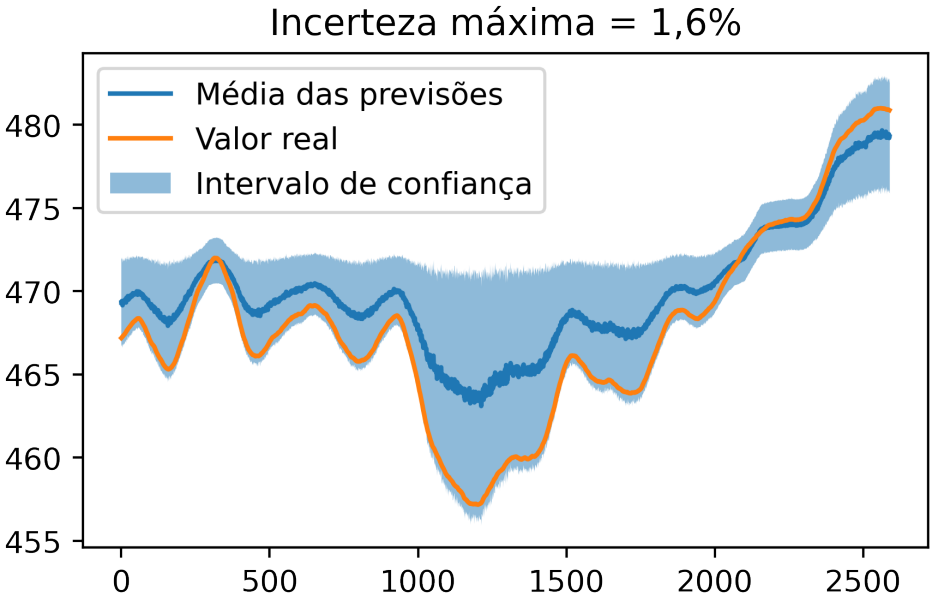
Note que, mesmo a incerteza máxima relativa não tendo aumentado muito, o intervalo de confiança acabou se tornando mais largo para o modelo com menos variáveis de entrada. Além disso, o valor real se aproximou mais do limite inferior do intervalo de confiança para esse mesmo conjunto de dados, enquanto o mesmo não ocorreu para o conjunto de dados maior.

Figura 18 – Incertezas do Modelo RNA.

(a) Intervalo de cofiança para conjunto de sete variáveis.




(b) Intervalo de cofiança para conjunto de cinco variáveis.




Fonte: Autoria própria.

# 5 Conclusões

Foi aplicada a metodologia para desenvolvimento de um sensor virtual, comparando dois modelos matemáticos distintos. O Modelo RNA teve uma performance superior em comparação ao Modelo XGBoost e, portanto, é o mais adequado para a utilização no caso estudado, dadas as condições apresentadas. Contudo, apesar da performance abaixo da meta estabelecida, caso seja feito o tratamento dos dados de saída, o Modelo XGBoost tende a se equiparar e, possivelmente ser preferível ao Modelo RNA. Isso porque as árvores de decisão com *gradient boosting* tendem a performar melhor com uma grande quantidade de variáveis de entrada.

Por isso, embora no estudo de caso essa característica seja irrelevante, dada a natureza sucinta do banco de dados, à medida em que ele cresce, o modelo com árvores de decisão tende a se tornar mais atrativo. Isso é verdade principalmente se o desenvolvimento do modelo for levado em consideração. Modelos baseados em RNA, como observado, necessitam de testes durante seu desenvolvimento para encontrar uma configuração adequada de quantidade de camadas escondidas e neurônios por camada. Dessa forma, grandes bancos de dados e, principalmente, bancos de dados que recebem cada vez mais informações de sensores, impõem dificuldades no desenvolvimento e manutenção desses modelos. Enquanto, por outro lado, modelos baseados em árvores de decisão são mais versáteis durante o desenvolvimento, porém mais sensíveis a mudanças no sistema, necessitando de ajustes mais constantes.

Ponderando todos os argumentos e tendo em vista o caso apresentado, o Modelo RNA é o mais indicado, sendo ideal a utilização de todo o banco de dados (conjunto 1) como dados de entrada. O referido modelo teve uma boa performance, baseado nas métricas de avaliação utilizadas e apresenta uma incerteza relativa máxima de 1%, que é um valor baixo para um instrumento virtual desenvolvido com uma abordagem semelhante. Portanto, para uma aplicação que necessita de exatidão nas medições, esse é um modelo satisfatório.

# Referências

CAMPILHO, A. *Instrumentaçao Electronica: METODOS E TECNICAS DE MEDIÇAO*. [S.l.]: FEUP EDIÇOES, 2011. ISBN 9789727521630. Citado na página 24.

DING, C.; PENG, H. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, v. 03, n. 02, p. 185–205, 2005. Disponível em: <https://doi.org/10.1142/S0219720005001004>. Citado na página 23.

GAL, Y.; GHAHRAMANI, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. 2015. Citado na página 25.

JIANG, Y. et al. A review on soft sensors for monitoring, control, and optimization of industrial processes. *IEEE sensors journal*, IEEE, v. 21, n. 11, p. 12868–12881, 2021. ISSN 1530-437X. Citado 2 vezes nas páginas 16 e 17.

LIMA, J. *CONSTRUÇÃO DE SENSOR VIRTUAL PARA MEDIÇÃO DE VAZÃO EM UMA USINA DO SETOR SUCROENERGETICO BASEADO EM REDES NEURAIS ARTIFICIAIS*. Dissertação (Mestrado) — Universidade Federal da Paraíba, 2022. Citado 4 vezes nas páginas 14, 27, 28 e 31.

LIMA, R. *DESENVOLVIMENTO DE UM SOFT SENSOR PARA ESTIMAÇÃO DA VAZÃO EM SISTEMAS DE ABASTECIMENTO DE ÁGUA UTILIZANDO REDES NEURAIS ARTIFICIAIS*. Tese (Doutorado) — Universidade Federal da Paraíba, 2022. Citado na página 16.

LIU, Z. et al. Adaptive soft sensors for quality prediction under the framework of bayesian network. *Control engineering practice*, Elsevier Ltd, v. 72, p. 19–28, 2018. ISSN 0967-0661. Citado 2 vezes nas páginas 13 e 14.

RODRIGUES, J. *Análise de Incertezas em Modelos de Deep Learning Utilizados em Previsão de Geração de Energia Fotovoltaica*. Dissertação (Mestrado) — Universidade Federal da Paraíba, 2022. Citado 3 vezes nas páginas 19, 25 e 38.

RUSSELL, S.; NORVIG, P.; MACEDO, R. *Inteligência Artificial*. 3. ed. [S.l.]: Rio de Janeiro: Elsevier, 2013. ISBN 9788535237016. Citado 2 vezes nas páginas 19 e 21.

SUN, Q.; GE, Z. A survey on deep learning for data-driven soft sensors. *IEEE transactions on industrial informatics*, IEEE, v. 17, n. 9, p. 5853–5866, 2021. ISSN 1551-3203. Citado 2 vezes nas páginas 16 e 18.

VAIDYA, S.; AMBAD, P.; BHOSLE, S. Industry 4.0 – a glimpse. *Procedia manufacturing*, Elsevier B.V, v. 20, p. 233–238, 2018. ISSN 2351-9789. Citado na página 13.

YIN, S.; ZHU, X.; KARIMI, H. R. Quality evaluation based on multivariate statistical methods. *Mathematical problems in engineering*, Hindawi Publishing Corporation, LONDON, v. 2013, p. 1–10, 2013. ISSN 1024-123X. Citado na página 13.

ZHAO, Z.; ANAND, R.; WANG, M. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. 2019. Citado na página 23.