



Universidade Federal de Campina Grande - UFCG  
Centro de Engenharia Elétrica e Informática - CEEI  
Departamento de Engenharia Elétrica - DEE

# Metodologia para Classificação do Nível de Poluição de Isoladores com Base na Medição da Corrente de Fuga Utilizando Inteligência Artificial

Nathanael Pereira Silva

Campina Grande - PB

Março - 2022

Nathanael Pereira Silva

Metodologia para Classificação do Nível de Poluição de  
Isoladores com Base na Medição da Corrente de Fuga  
Utilizando Inteligência Artificial

Trabalho de Conclusão de Curso submetido à  
Unidade Acadêmica de Engenharia Elétrica da  
Universidade Federal de Campina Grande como  
parte dos requisitos necessários para obtenção  
do título de Bacharel em Engenharia Elétrica.

Orientador: George Rossany Soares de Lira, D. Sc.

Campina Grande - PB

Março - 2022

Nathanael Pereira Silva

Metodologia para Classificação do Nível de Poluição de  
Isoladores com Base na Medição da Corrente de Fuga  
Utilizando Inteligência Artificial

Trabalho de Conclusão de Curso submetido à  
Unidade Acadêmica de Engenharia Elétrica da  
Universidade Federal de Campina Grande como  
parte dos requisitos necessários para obtenção  
do título de Bacharel em Engenharia Elétrica.

Aprovado em 01 de abril de 2022.

---

George Rossany Soares de Lira, D. Sc.  
Orientador

---

Pablo Bezerra Vilar, D. Sc.  
Avaliador

Campina Grande - PB  
Março - 2022

Dedico este trabalho a Deus e  
a meus pais pela paciência,  
compreensão e apoio.

# Agradecimentos

Agradeço primeiramente a Deus, sem a força e fé que Ele me concedeu eu não teria chegado tão longe.

Quero agradecer a meus pais, Ednaldo e Socorro, por todo o apoio durante os meus estudos, gratidão aos meus irmãos Déborah e Nathan por sempre estarem por perto nos diversos momentos desta jornada.

Agradeço a todos os meus colegas e amigos da graduação que tive o prazer de conhecer e conviver, sou grato pelos momentos de descontração, conversas e por todo o conhecimento compartilhado.

Gratidão a todos os professores do Departamento de Engenharia Elétrica (DEE) da Universidade Federal de Campina Grande (UFCG) que fizeram parte da minha formação em especial, o meu professor e orientador, George Lira, pelo conhecimento transmitido e por me fazer despertar para novas áreas de conhecimento. Agradeço a Vandilson pela ajuda e apoio durante todo o desenvolvimento deste trabalho.

*”Os problemas nunca vão desaparecer,  
mesmo na mais bela existência.  
Problemas existem para serem resolvidos,  
e não para perturbar-nos.”*  
*(Augusto Cury)*

# Resumo

Os isoladores são equipamentos responsáveis por fornecer isolamento elétrico e sustentação mecânica aos condutores, sendo amplamente utilizados nos Sistemas Elétricos de Potência. A confiabilidade e continuidade do sistema dependem que esses equipamentos operem de forma correta. Devido à importância dos isoladores para o sistema elétrico, estudos que desenvolvam técnicas capazes de realizar o diagnóstico/prognóstico desses equipamentos, fazem-se necessários. Este trabalho apresenta uma técnica para classificação ou diagnóstico de isoladores a partir de parâmetros da corrente de fuga. Inicialmente, o banco de dados foi construído, por meio de ensaios elétricos para obtenção da corrente de fuga com isoladores de vidro poluídos artificialmente em cinco diferentes níveis: limpo, muito leve, leve, moderado e pesado. Os parâmetros da corrente de fuga que apresentaram maior correlação com a poluição foram: valor de pico, valor eficaz e a componente harmônica fundamental. Os modelos de classificação utilizados para determinar o nível de poluição, basearam-se em técnicas de inteligência artificial, sendo elas: Máquinas de Vetor de Suporte, Mapas Auto-Organizáveis de *Kohonen* e *Naive Bayes*. Indicadores de desempenho foram utilizados para avaliar a confiabilidade de cada uma das técnicas como também realizar uma análise comparativa entre elas. Os indicadores utilizados foram: acurácia, precisão, sensibilidade, *f1-score* e matriz de confusão. Verificou-se que os modelos de classificação apresentaram ótimos resultados, principalmente o algoritmo baseado em aprendizado não supervisionado, Mapas de *Kohonen*. Com isso, obteve-se um modelo de classificação com desempenho satisfatório que pode ser utilizado em um sistema de monitoramento e diagnóstico de isoladores capaz de auxiliar nos processos de tomada de decisão relacionados ao gerenciamento de ativos.

**Palavras-chave:** Corrente de Fuga; Diagnóstico; Inteligência Artificial; Isoladores; Poluição.

# Abstract

Insulators are equipment responsible for providing electrical insulation and mechanical support to conductors, and are widely used in the Electric Power Systems. System reliability and continuity depend on these equipment operating correctly. Due to the importance of insulators for the electrical system, studies that develop techniques capable of performing the diagnosis/prognosis of this equipment are necessary. This work presents a technique for classification or diagnosis of insulators based on leakage current parameters. Initially, the database was built, through electrical tests to obtain the leakage current with artificially polluted glass insulators at five different levels: clean, very light, light, moderate and heavy. The leakage current parameters that presented the greatest correlation with pollution were: peak value, RMS value and the fundamental harmonic component. The classification models used to determine the level of pollution were based on artificial intelligence techniques, namely: Support Vector Machines, Kohonen Self-Organizing Maps and Naive Bayes. Performance indicators were used to assess the reliability of each of the techniques as well as perform a comparative analysis between them. The indicators used were: accuracy, precision, recall, f1-score and confusion matrix. It was verified that the classification models presented excellent results, mainly the algorithm based on unsupervised learning, Kohonen Maps. Thus, we obtained a classification model with satisfactory performance that can be used in an insulator monitoring and diagnosis system capable of assisting in decision-making processes related to asset management.

**Key-words:** Artificial Intelligence; Diagnosis; Insulators; Leakage Current; Pollution.



# Lista de ilustrações

Figura 1 – Isolador de vidro e seus componentes. . . . .	17
Figura 2 – Relação entre a condutividade da superfície ( $\mu\text{S}$ ) e a salinidade ( $\text{kg}/\text{m}^3$ ). . .	19
Figura 3 – Relação entre o valor máximo da corrente de fuga (A) e a salinidade ( $\text{kg}/\text{m}^3$ ). . .	20
Figura 4 – Classificador SVM linear . . . . .	22
Figura 5 – Efeitos da constante de suavização, $C$ , na fronteira de decisão do modelo. . .	24
Figura 6 – Efeitos na fronteira de decisão para diferentes tipos de <i>kernel</i> . . . . .	25
Figura 7 – Mapeamento $\mathbb{R}^n \Rightarrow \mathbb{R}^2$ . . . . .	26
Figura 8 – Mapa topológico. . . . .	28
Figura 9 – Representação de um classificador <i>Naive Bayes</i> . . . . .	28
Figura 10 – Método de validação cruzada usando amostragem aleatória. . . . .	33
Figura 11 – Método de validação cruzada usando $k$ partições. . . . .	33
Figura 12 – Resumo dos procedimentos propostos. . . . .	40
Figura 13 – Níveis de poluição com base nos índices DDSE e DDNS. . . . .	41
Figura 14 – Fotografia dos isoladores utilizados. . . . .	43
Figura 15 – Fotografias das partes a) externa e b) interna da câmara de névoa. . . . .	44
Figura 16 – Diagrama do arranjo experimental utilizado nos ensaios elétricos. . . . .	44
Figura 17 – Sinal de tensão aplicado. . . . .	45
Figura 18 – Sinais da corrente de fuga para diferentes níveis de poluição. . . . .	45
Figura 19 – Resumo do procedimento utilizado para construção do banco de dados. . . . .	46
Figura 20 – Matriz de confusão - modelo SVM considerando dados de teste. . . . .	51
Figura 21 – Evolução dos parâmetros de avaliação do modelo. . . . .	52
Figura 22 – Mapa topológico do classificador do nível de poluição dos isoladores. . . . .	53
Figura 23 – Matriz de confusão - modelo <i>Naive Bayes</i> considerando dados de teste. . . . .	54

# Lista de tabelas

Tabela 1 – Classificação do nível de contaminação . . . . .	20
Tabela 2 – Resumo funções <i>kernel</i> . . . . .	24
Tabela 3 – Classificação do nível de correlação com base no coeficiente de <i>Spearman</i> .	31
Tabela 4 – Exemplo de uma matriz de confusão. . . . .	34
Tabela 5 – Características físicas e elétricas individuais dos isoladores de vidro utilizados	40
Tabela 6 – Relação entre DDSE e condutividade elétrica da solução. . . . .	42
Tabela 7 – Níveis de poluição e valores correspondentes de DDSE e condutividade elétrica.	42
Tabela 8 – Análise estatística dos valores de tensão, temperatura e umidade. . . . .	46
Tabela 9 – Resumo da matriz de correlação de <i>Spearman</i> . . . . .	47
Tabela 10 – Valor da precisão e sensibilidade para os diferentes níveis de poluição - SVM.	51
Tabela 11 – Valor da precisão e sensibilidade para os diferentes níveis de poluição - <i>Naive</i> <i>Bayes</i> . . . . .	54
Tabela 12 – Comparação dos indicadores de desempenho dos modelos de classificação. .	55

# Lista de abreviaturas e siglas

ANFIS	<i>Adaptive Neuro Fuzzy Inference System</i>
BMU	<i>Best-Matching Unit</i>
CCR	<i>Correct Classification Rate</i> (Taxa de Classificação Correta)
DDS	Densidade de Depósito de Sal
DDSE	Densidade de Depósito de Sal Equivalente
DDNS	Densidade de Depósito Não Solúvel
DEE	Departamento de Engenharia Elétrica
DFT	<i>Discrete Fourier Transform</i> (Transformada Discreta de Fourier)
FFT	<i>Fast Fourier Transform</i> (Transformada Rápida de Fourier)
IA	Inteligência Artificial
IEC	<i>International Electrotechnical Commission</i>
KNN	<i>K-Nearest Neighbor Classifier</i>
LAT	Laboratório de Alta Tensão
LS-SVM	<i>Least-Squares Support Vector Machine</i> (Máquinas de Vetores de Suporte - Mínimos Quadrados)
mRMR	<i>Minimum Redundancy – Maximum Relevance</i>
NBR	Norma Brasileira
NRMSE	<i>Normalized Root Mean Square Error</i>
RMS	<i>Root Mean Square</i>
RNA	Rede Neural Artificial
SEP	Sistema Elétrico de Potência

SOM	<i>Self-Organization Maps</i> (Mapas Auto-Organizáveis)
SVM	<i>Support Vector Machine</i> (Máquinas de Vetores de Suporte)
TCC	Trabalho de Conclusão de Curso
THD	<i>Total Harmonic Distortion</i> (Distorção Harmônica Total)
UFCG	Universidade Federal de Campina Grande

# Sumário

<b>1</b>	<b>INTRODUÇÃO</b>	<b>14</b>
<b>1.1</b>	<b>Objetivos</b>	<b>15</b>
1.1.1	Objetivo Geral	15
1.1.2	Objetivos Específicos	15
<b>1.2</b>	<b>Organização do Trabalho</b>	<b>15</b>
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>17</b>
<b>2.1</b>	<b>Isoladores</b>	<b>17</b>
2.1.1	Inspeções e Ensaios em Isoladores	18
2.1.2	Contaminação em Isoladores	19
<b>2.2</b>	<b>Aprendizado de Máquina</b>	<b>21</b>
2.2.1	Máquinas de Vetores de Suporte	21
2.2.2	Mapas Auto-Organizáveis de <i>Kohonen</i>	25
2.2.3	<i>Naive Bayes</i>	28
<b>2.3</b>	<b>Conceitos Estatísticos</b>	<b>29</b>
2.3.1	Estatística Descritiva	30
2.3.2	Coefficiente de Correlação de <i>Spearman</i>	31
2.3.3	Normalização de Dados	32
2.3.4	Validação Cruzada	32
2.3.5	Avaliadores de Desempenho	34
<b>3</b>	<b>REVISÃO BIBLIOGRÁFICA</b>	<b>36</b>
<b>3.1</b>	<b>Levantamento do Estado da Arte</b>	<b>36</b>
<b>4</b>	<b>METODOLOGIA PROPOSTA</b>	<b>40</b>
<b>4.1</b>	<b>Construção do Banco de Dados</b>	<b>40</b>
<b>4.2</b>	<b>Análise Exploratória dos Dados</b>	<b>46</b>
<b>4.3</b>	<b>Implementação dos Modelos</b>	<b>48</b>
4.3.1	Modelo de Classificação Utilizando SVM	48
4.3.2	Modelo para Reconhecimento de Padrões Utilizando Rede SOM	48

4.3.3	Modelo de Classificação Utilizando <i>Naive Bayes</i> . . . . .	49
<b>5</b>	<b>RESULTADOS E ANÁLISES</b> . . . . .	<b>50</b>
5.1	Modelo SVM . . . . .	50
5.2	Modelo - Rede SOM . . . . .	52
5.3	Modelo - <i>Naive Bayes</i> . . . . .	53
5.4	Análise Comparativa entre os Modelos de Classificação . . . . .	54
<b>6</b>	<b>CONCLUSÃO</b> . . . . .	<b>56</b>
6.1	Trabalhos Futuros . . . . .	56
	<b>REFERÊNCIAS</b> . . . . .	<b>58</b>
<b>A</b>	<b>APÊNDICE - RESUMO BANCO DE DADOS</b> . . . . .	<b>61</b>
<b>B</b>	<b>APÊNDICE - MATRIZ DE CORRELAÇÃO DE <i>SPEARMAN</i></b> . . . . .	<b>62</b>

# 1 Introdução

Os isoladores são um dos equipamentos mais presentes e importantes nos Sistemas Elétricos de Potência (SEP), possuem propriedades mecânicas capazes de suportar os esforços produzidos pelos condutores, e propriedades elétricas que exercem a função de isolar os condutores submetidos a uma diferença de potencial em relação à terra ou a um outro condutor fase (MAMEDE FILHO, 2013). Esses equipamentos são uma das bases do SEP, sendo fundamental o seu correto funcionamento. Falhas em isoladores, podem provocar interrupções no fornecimento de energia elétrica. Com isso, deve-se estar atento as diferentes causas que resultam na degradação desses equipamentos, a fim de prever e diagnosticar tais falhas.

Os isoladores estão sujeitos ao acúmulo de poluição sobre a superfície, alto grau de umidade, intempéries climáticas, fontes de poluição local e vandalismo. Todos esses fatores, podem levar a uma diminuição no desempenho das suas funções (ROCHA, 2019). O diagnóstico de isoladores não é uma tarefa fácil, visto que o estado operacional e vida útil do equipamento dependem das condições ambientais a qual estará exposto e alguns defeitos não são visíveis (STEFENON et al., 2018), tornando a inspeção visual subjetiva, pois está condicionada a experiência do profissional que está realizando a inspeção (DIAS, 2021).

Além da inspeção visual outras técnicas não invasivas têm sido desenvolvidas como a análise de imagens coletadas por câmeras infravermelhas, ultravioleta e de luz visível (YANG et al., 2018), ou a detecção de sinais de ultrassom emitidos pelos isoladores poluídos (STEFENON et al., 2019). Além disso, ensaios elétricos de diagnóstico como detecção de descargas parciais, tensão suportável na frequência de 60 Hz e medições de corrente de fuga têm sido utilizadas (GERMANO, 2010). A corrente de fuga tem sido um dos parâmetros mais eficazes para avaliação de desempenho (LI et al., 2010). A partir de medições realizadas em campo da corrente de fuga é possível correlacionar esse parâmetro com o nível de poluição do isolador.

Entretanto mesmo a corrente de fuga sendo um parâmetro importante para fins de monitoramento e diagnóstico de isoladores, sua interpretação e ponderação não é simples. Portanto faz-se necessário o desenvolvimento de aplicações de técnicas que auxiliem na interpretação dos sinais de corrente de fuga de isoladores, visando seu diagnóstico. Dentre as abordagens mais utilizadas hoje em dia, destaca-se a aplicação de algoritmos de Inteligência Artificial (IA). A IA consiste no estudo de agentes que recebem percepções do ambiente e executam ações (RUSSEL;

NORVIG, 2013).

A literatura mostra que a IA tem fornecido resultados consistentes em diversas aplicações tais como: diagnósticos médicos, recomendações de vendas, transporte, automatização de processos, apoio a tomada de decisão na gestão de pessoas, entre outros. Portanto a utilização de técnicas inteligentes que auxiliem no diagnóstico de equipamentos e tomada de decisão são uma ótima alternativa às empresas do setor elétrico, pois permitem que essas empresas atendam a determinados padrões de qualidade, minimizem custos, acelerem processos e evitem prejuízos em relação a possíveis falhas do sistema.

## 1.1 Objetivos

### 1.1.1 Objetivo Geral

Este trabalho tem como objetivo geral propor uma metodologia para a classificação do nível de poluição superficial de isoladores a partir de medições da corrente de fuga utilizando modelos de inteligência artificial, sendo capaz de realizar o diagnóstico/prognóstico de isoladores.

### 1.1.2 Objetivos Específicos

Os objetivos específicos são apresentados a seguir.

- Verificar a capacidade dos modelos para determinar o nível de poluição frente as variações dos parâmetros da corrente de fuga, avaliando os três algoritmos utilizados, sendo eles: Máquinas de Vetor de Suporte, Mapas Auto-organizáveis de *Kohonen* e *Naive Bayes*.
- Avaliar a correlação entre a poluição dos isoladores e os parâmetros extraídos da corrente de fuga.
- Analisar as vantagens e viabilidade do uso de técnicas de IA que auxiliem na tomada de decisão, realizando o diagnóstico de equipamentos como isoladores.

## 1.2 Organização do Trabalho

O presente Trabalho de Conclusão de Curso (TCC) está organizado em seis capítulos, considerando este capítulo, em que está apresentado a introdução e objetivos do trabalho.



No capítulo 2, apresenta-se a fundamentação teórica com os conceitos teóricos necessários para o entendimento do trabalho.

No capítulo 3, tem-se o levantamento do estado da arte do uso da corrente de fuga para classificação do nível de poluição em isoladores.

No capítulo 4, apresentam-se os procedimentos práticos e computacionais utilizados para construção do banco de dados, a análise exploratória e estatística dos dados e implementação dos modelos de classificação.

No capítulo 5, tem-se a análise dos resultados obtidos por cada um dos modelos como também uma análise comparativa entre os mesmos.

No capítulo 6, apresentam-se as considerações finais e as contribuições do trabalho ao gerenciamento de equipamentos de alta tensão.

O apêndice A apresenta um resumo do banco de dados utilizado para implementação dos modelos de classificação.

O apêndice B apresenta a matriz de correlação de *Spearman*, utilizada para análise da relação entre as variáveis.

## 2 Fundamentação Teórica

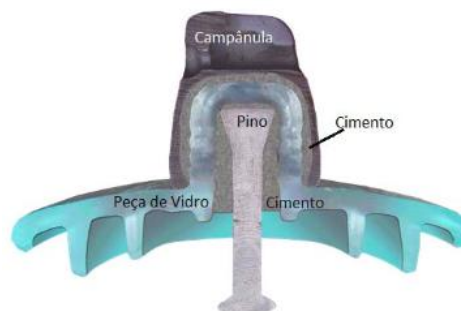
Neste capítulo, serão apresentados os conceitos teóricos necessários para o entendimento do trabalho. Na seção 2.1 será apresentado uma breve fundamentação teórica sobre isoladores e os parâmetros utilizados para caracterizar o nível de contaminação. Na seção 2.2 é apresentado a base teórica dos algoritmos de IA utilizados. Na seção 2.3 serão apresentados os conceitos básicos estatísticos e o coeficiente de correlação de *Spearman* utilizados durante a análise exploratória e estatística dos dados, também será apresentado a técnica de validação cruzada para determinação dos parâmetros dos modelos.

### 2.1 Isoladores

O isolador é um componente elétrico projetado para suportar condutores e separá-los fisicamente e eletricamente de outros condutores, objetos ou a terra (GORUR; CHERNEY; BURNHAM, 1999). Os três principais materiais dielétricos utilizados na fabricação de isoladores são: porcelana, vidro e polímeros.

Os isoladores de vidro são utilizados desde o final da década de 1880 com o surgimento das primeiras linhas de transmissão no mundo. A dificuldade inicial no uso desse material era a sua fragilidade, mas com as melhorias desenvolvidas por técnicas como tratamento térmico de têmpera ou endurecimento térmico os dielétricos de vidro apresentaram uma maior resistência à tração e ao impacto, entre outras melhorias (GORUR; CHERNEY; BURNHAM, 1999). Os componentes de um isolador de vidro basicamente são: campânula, cimento, dielétrico de vidro e pino. Um isolador de vidro é ilustrado na Figura 1.

Figura 1 – Isolador de vidro e seus componentes.



Fonte: Ferreira (2017).

### 2.1.1 Inspeções e Ensaios em Isoladores

Os isoladores de vidro podem apresentar uma vida útil considerada longa, apresentando vantagens consideráveis como: resistência a danos por radiação ultravioleta, alta rigidez dielétrica, alta resistência à compressão e estilhaçamento quando danificado (DIAS, 2017). Como desvantagens, eles são componentes vulneráveis a vários fenômenos de envelhecimento causados por agentes ambientais e estresse elétrico, como também ações de vandalismo. Portanto, o desempenho e a vida útil dos isoladores são influenciados pelas condições operacionais e ambientais a que estarão expostos.

Defeitos em isoladores impactam diretamente na continuidade e confiabilidade do SEP. Para evitar falhas e aumentar a confiabilidade do sistema faz-se necessário a rápida detecção de isoladores defeituosos por meio de inspeções ou ensaios. As concessionárias, utilizam-se de algumas técnicas para a análise do estado físico e dielétrico dos isoladores (SOUZA, 2016), algumas delas são: inspeções visuais por meio de binóculos ou câmeras de luz visível, câmeras de radiação ultravioleta, análise termográfica (YANG et al., 2018), radiografia, detecção de rádio interferência, entre outros (STEFENON et al., 2019). Os ensaios elétricos se sobressaem em relação aos físicos-químicos pois a maioria dos testes preservam a integridade do equipamento permitindo-o ser devolvido à operação sem modificação das suas propriedades (GERMANO, 2010). Os principais parâmetros elétricos utilizados para realização de diagnósticos são apresentados a seguir.

- Descargas parciais: de acordo com IEEE (2000) são descargas elétricas localizadas, que são ocasionadas pelo aumento da intensidade de campo elétrico local acima da rigidez dielétrica do isolamento.
- Tensão suportável: segundo Mamede Filho (2013) é o valor eficaz de tensão à frequência nominal que um isolador pode suportar durante 60 segundos.
- Corrente de fuga: a corrente que atravessa pela superfície externa do isolador do condutor vivo para a terra ou para a parte não condutora (AMIN; AMIN; ALI, 2009).

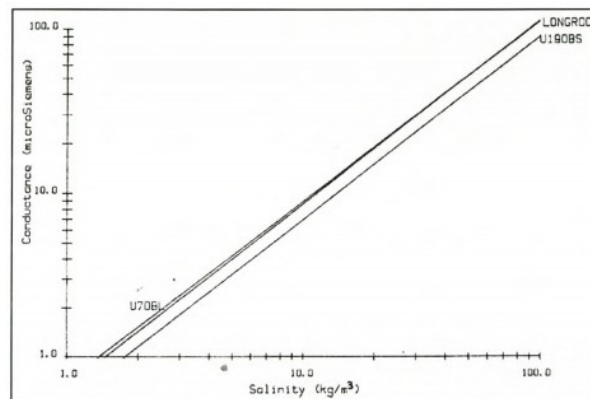
Com a elevação da umidade, a poluição existente no isolador como poeira, sal, areia e etc., forma uma camada condutora na superfície que leva ao aumento da corrente de fuga (ROCHA, 2019). Se a corrente de fuga for alta o suficiente ocorre uma descarga disruptiva (*flashover*) sobre toda a superfície do isolador resultando em danos ao equipamento e a falha do sistema.

A corrente de fuga é afetada pela tensão de operação, temperatura, umidade e poluição, fornecendo assim informações sobre o estado operacional dos isoladores. Parâmetros da corrente de fuga, tais como: componente fundamental, terceira e quinta componentes harmônicas, valor de pico, número de pulsos e valor RMS são utilizados na avaliação do estado operacional de isoladores (LI et al., 2010).

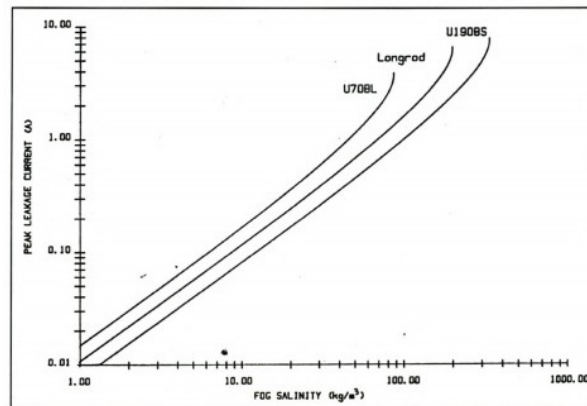
### 2.1.2 Contaminação em Isoladores

Podem existir dois tipos de contaminantes, os solúveis e não solúveis, os efeitos de um contaminante são avaliados por dois fatores: (1) o grau de condutividade em condições úmidas determinado pela solubilidade do contaminante em água e (2) a capacidade de reter umidade. A magnitude da corrente de fuga é proporcional à condutividade da superfície. Quanto maior a quantidade de contaminação, menor é a resistência apresentada na superfície do isolador (GORUR; CHERNEY; BURNHAM, 1999). Essas relações podem ser observadas nas Figuras 2 e 3 em que três tipos de isoladores (um de vidro padrão, um isolador de haste longa e um isolador anti poluição) foram poluídos artificialmente e foi verificada a relação entre o nível de salinidade e a condutividade da superfície e o valor máximo da corrente de fuga.

Figura 2 – Relação entre a condutividade da superfície ( $\mu\text{S}$ ) e a salinidade ( $\text{kg}/\text{m}^3$ ).



Fonte: Schwardt (2005).

Figura 3 – Relação entre o valor máximo da corrente de fuga (A) e a salinidade (kg/m<sup>3</sup>).

Fonte: Schwardt (2005).

A salinidade pode ser representada pelos índices, Densidade de Depósito de Sal Equivalente (DDSE) e Densidade de Depósito Não Solúvel (DDNS). Esses índices são parâmetros utilizados para caracterizar o nível de contaminação na superfície de isoladores. Segundo a Norma Brasileira (NBR) 10621, a salinidade é definida como a concentração de sal em água, expressa pela quantidade de sal dividida pelo volume da solução, geralmente expressa em quilogramas por metro cúbico (ABNT, 2005). O parâmetro DDSE é definido pela *International Electrotechnical Commission* (IEC) 60815 como a quantidade de sal que, quando dissolvido em água, dá a mesma condutividade volumétrica que a do depósito natural retirado de uma dada superfície do isolador dividido pela área desta superfície, geralmente expresso em miligrama por centímetros quadrados (IEC, 2008). Essa relação está expressa em (2.1).

$$DDSE = S_a \times \frac{V}{A} [mg/cm^2], \quad (2.1)$$

onde,  $S_a$  é a salinidade em Kg/cm<sup>3</sup>,  $V$  é o volume total da água (cm<sup>3</sup>) e  $A$  é a área da superfície (cm<sup>2</sup>) do isolador utilizada para coleta da poluição.

O DDNS, segundo a IEC 60815 (2008), é a quantidade de resíduo insolúvel removido de uma determinada superfície do isolador dividido pela área dessa superfície, geralmente expresso em mg/cm<sup>2</sup>. Na Tabela 1 é apresentado uma classificação do nível de contaminação em função das faixas de valores de DDSE baseado em experiências em campo, fornecido pela IEC 60815.

Tabela 1 – Classificação do nível de contaminação

DDSE (mg/cm <sup>2</sup> )	Classificação do Nível de Contaminação
0-0,005	Limpo ou poluição muito leve
0,006-0,02	Poluição leve
0,03-0,1	Poluição moderada
0,15-1	Poluição alta ou muito alta

Fonte: Adaptado da IEC 60815 (2008).

## 2.2 Aprendizado de Máquina

O aprendizado de máquina é uma das subáreas da IA, segundo Marsland (2014), aprendizado de máquina, consiste em fazer com que os computadores modifiquem ou adaptem suas ações para que essas ações sejam mais precisas, onde a precisão é medida pelo quão bem as ações escolhidas refletem as corretas. Existem dois tipos principais de aprendizado: o supervisionado e não supervisionado. O primeiro envolve fornecer ao algoritmo um conjunto de dados de treinamento que contém os alvos ou as soluções desejadas. A classificação é uma tarefa típica deste tipo de aprendizado. O segundo, envolve identificar semelhanças entre as entradas sem que os alvos ou as respostas sejam fornecidas ao algoritmo (GÉRON, 2019). A clusterização, detecção de anomalias e reconhecimento de padrões são casos típicos do aprendizado não supervisionado.

Nesta seção será apresentada uma breve teoria sobre as técnicas de inteligência artificial utilizadas para classificação: Máquina de Vetores de Suporte (SVM, do inglês - *Support Vector Machine*), Mapas Auto-Organizáveis (SOM, do inglês - *Self-Organizing Map*) e *Naive Bayes*. Essas técnicas foram escolhidas devido ao bom desempenho em problemas de classificação segundo a literatura, além de aspectos construtivos diferentes permitindo a análise comparativa entre os modelos.

### 2.2.1 Máquinas de Vetores de Suporte

A máquina de vetores de suporte é um modelo de aprendizado de máquina capaz de realizar classificações lineares, não lineares, regressão e detecção de *outliers*<sup>1</sup>. Esse modelo é adequado para a classificação de dados complexos de pequeno ou médio porte (GÉRON, 2019). A ideia principal de uma SVM é que dado uma amostra de dados de treinamento, é construído um hiperplano como superfície de decisão de tal forma que a margem de separação entre as classes seja maximizado (HAYKIN, 2009), aumentando a capacidade de generalização evitando problemas de *overfitting*<sup>2</sup>.

Os primeiros algoritmos desenvolvidos eram capazes apenas de classificar duas classes

<sup>1</sup> Um *outlier* é qualquer valor que seja muito distante dos outros valores em um conjunto de dados, ou seja, são pontos de dados não representativos (BRUCE; BRUCE, 2017).

<sup>2</sup> O *overfitting* ou memorização excessiva, ocorre quando o modelo de aprendizado acaba decorando as suas respostas frente as suas entradas, durante a fase de aprendizado o erro tende a ser muito baixo, contudo, durante a fase de generalização frente aos subconjuntos de teste, o erro tende a assumir valores bem elevados (SILVA; SPATTI; FLAUZINO, 2010).

com dados linearmente separáveis, conhecidos como SVMs de margem rígida (*hard margin*). Após aprimoramentos, o algoritmo tornou-se capaz de realizar a classificação de dados não linearmente separáveis, utilizando os conceitos de margem flexível (*soft margin*) e de função *kernel* (DIAS, 2019).

## SVM Linear

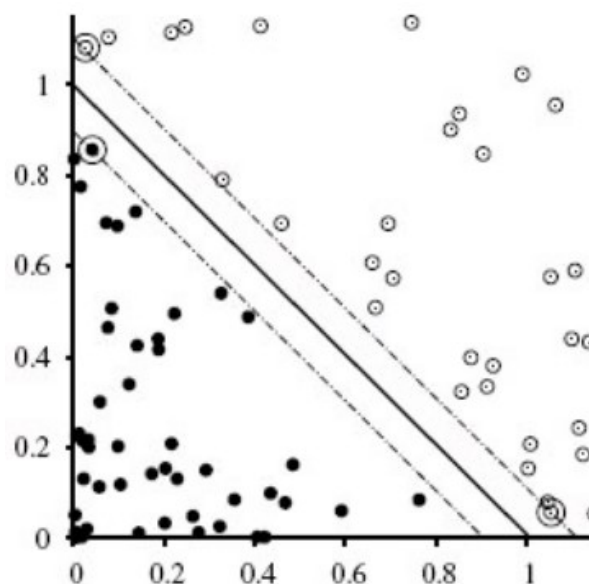
A forma mais básica de uma máquina de vetor de suporte é dada quando se tem duas classes linearmente separáveis, por exemplo, elementos positivos e negativos, essa forma é denominada SVM linear, cuja equação geral do hiperplano é apresentada em (2.2).

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b, \quad (2.2)$$

em que  $\mathbf{x}$  é um vetor com componentes  $x_i$ ,  $i$  é o  $i$ -ésimo vetor em um conjunto de dados composto por  $n$  exemplos,  $\mathbf{w}$  é o vetor de pesos e  $b$  é chamado de *bias*. O *bias* é um elemento que translada o hiperplano para longe da origem.

O hiperplano divide o espaço de acordo com a função discriminante (2.2), como pode ser observado na Figura 4, duas regiões foram classificadas, uma como positiva e a outra como negativa. Os pontos de dados mais próximos, em destaque, são chamados de margem de separação. A SVM encontra o hiperplano particular na qual essa margem de separação é maximizada.

Figura 4 – Classificador SVM linear



Fonte: Russel e Norvig (2013).

O classificador apresentado na Figura 4 também é chamado de separador de margem máxima ou rígida. A margem é a largura delimitada pela linha tracejada e é igual a duas vezes a distância do separador até o ponto de exemplo mais próximo (RUSSEL; NORVIG, 2013). Dado um conjunto de treinamento é encontrado os valores ótimos do vetor de pesos ( $\mathbf{w}$ ) e *bias* ( $b$ ), de acordo com as restrições e função de custo, apresentadas em (2.3) e (2.4), respectivamente.

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \text{ para } i = 1, 2, \dots, N \quad (2.3)$$

$$\Phi(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} \quad (2.4)$$

### SVM com margens flexíveis

A imposição de que todas as instâncias estejam fora da via de separação entre as classes indica a aplicação de uma SVM de margem rígida, que pode apresentar dois problemas. Primeiramente, só é aplicável para dados linearmente separáveis e segundo é muito sensível a *outliers* (GÉRON, 2019). Para solucionar esses problemas pode ser utilizado o conceito de margem flexível em que é inserido um valor  $\epsilon_i \geq 0$ , que suaviza as restrições impostas pelo algoritmo SVM com margem rígida, temos nas Equações (2.5) e (2.6) uma adaptação das restrições e função de custo, respectivamente, apresentados em (2.3) e (2.4), considerando agora uma margem flexível.

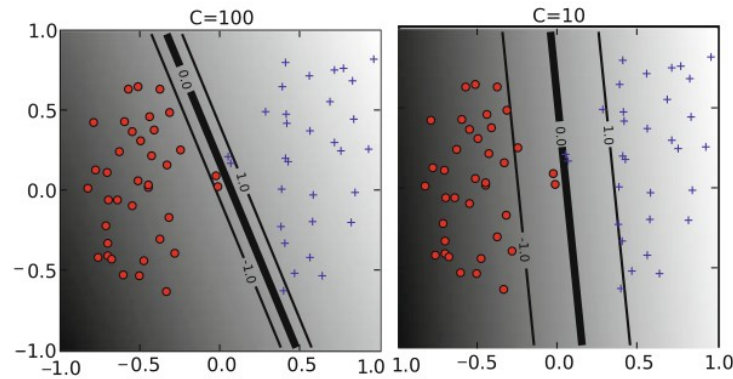
$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \epsilon_i, \text{ onde } \epsilon_i > 0, \text{ para } i = 1, 2, \dots, N \quad (2.5)$$

$$\Phi(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \epsilon_i \quad (2.6)$$

O parâmetro  $C$  é o fator de suavização, uma maneira de controlar o *overfitting*, ou seja, controlar a importância de maximizar a margem a fim de se ajustar os dados de treinamento à margem suave (DIAS, 2019). Quanto menor o valor de  $C$ , mais pontos próximos a margem são ignorados, aumentando assim a margem de separação entre as classes, para valores altos de  $C$  a margem de separação é reduzida aumentando a capacidade de generalização do modelo, contudo deve-se atentar a possíveis sobreajustes. Os efeitos da aplicação de diferentes valores de  $C$  para o problema de classificação linear de duas classes pode ser observado na Figura 5.



Figura 5 – Efeitos da constante de suavização,  $C$ , na fronteira de decisão do modelo.



Fonte: Ben-Hur e Weston (2010).

### Função *kernel*

A fronteira de decisão é dita não linear quando os dados não são linearmente separáveis. Esses dados de entrada são mapeados em um novo espaço de dimensão maior (HAYKIN, 2009), tornando o problema linearmente separável, conforme (2.7).

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{2.7}$$

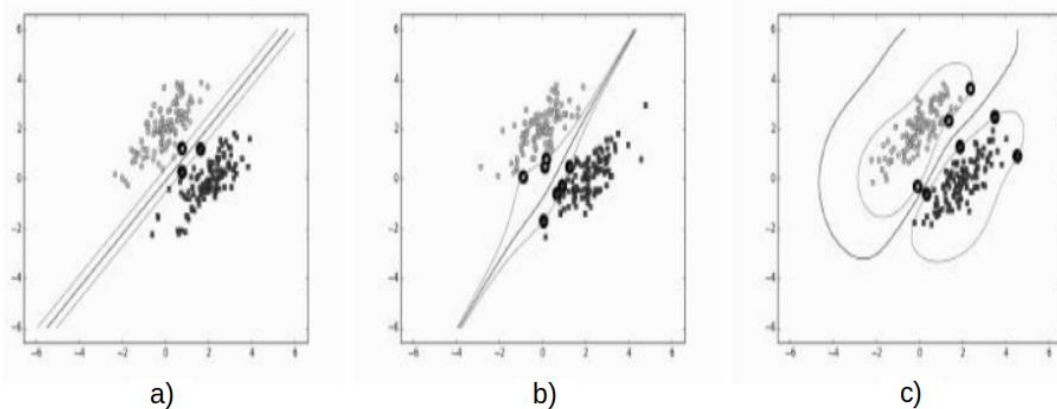
A transformação do espaço vetorial dos dados de entrada é realizado utilizando uma função *kernel* ( $\phi$ ). Os tipos de funções *Kernel* mais utilizados são: Linear, Polinomial, Gaussiano (RBF) e Sigmóide. A Tabela 2 apresenta um resumo com as funções e parâmetros a serem determinados.

Tabela 2 – Resumo funções *kernel*

Kernel	Função	Parâmetros
Linear	$\mathbf{x}^T \mathbf{x}_i$	-
Polinomial	$(\gamma(\mathbf{x}^T \mathbf{x}_i) + coef0)^d$	$d, coef0, \gamma$
Gaussiano	$exp(-\gamma \ \mathbf{x} - \mathbf{x}_i\ ^2)$	$\gamma$
Sigmóide	$tanh(\gamma(\mathbf{x}^T \mathbf{x}_i) + coef0)$	$\gamma, coef0$

Fonte: Autoria Própria (2022).

Os parâmetros  $d$ ,  $coef0$  e  $\gamma$  são chamados hiperparâmetros de ajuste, em que  $d$  é o grau do polinômio,  $coef0$  controla o quanto o modelo é influenciado por polinômios de alto grau versus polinômios de baixo grau (GÉRON, 2019) e  $\gamma$  é o parâmetro de regularização, sabendo que valores baixos tornam o classificador restrito, valores maiores podem levar a *overfitting*. A Figura 6 apresenta a influência de diferentes tipos de *kernel* na classificação da base de dados *Iris*, em que foi considerado três tipos de *kernel*, linear (a), polinomial com grau três (b) e RBF (c), em todos os casos o valor de  $C$  foi igual a 0,1.

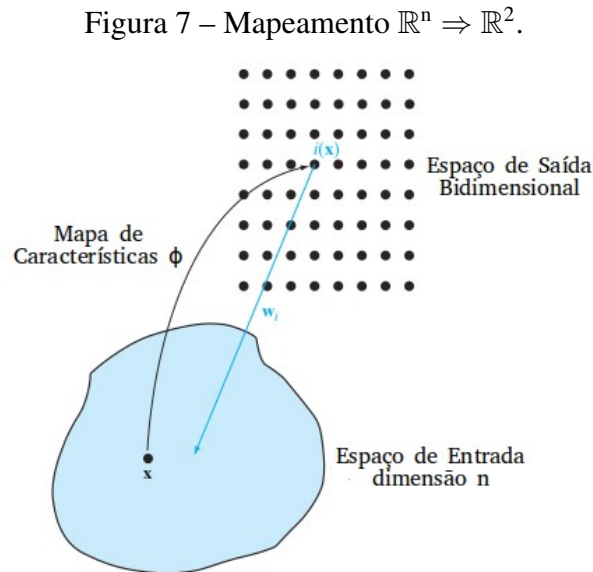
Figura 6 – Efeitos na fronteira de decisão para diferentes tipos de *kernel*.

Fonte: Marsland (2014).

## 2.2.2 Mapas Auto-Organizáveis de *Kohonen*

As redes neurais que utilizam aprendizado não supervisionado com métodos de treinamento competitivos são uma ferramenta valiosa, principalmente nos casos em que se possui apenas o conjunto de padrões de entrada, sem as respectivas saídas (SILVA; SPATTI; FLAUZINO, 2010). Os Mapas Auto-Organizáveis também conhecidos como Rede SOM (do inglês - *Self-Organizing Map*) desenvolvidos por Teuvo Kohonen são uma das técnicas de aprendizado não supervisionado mais utilizadas para clusterização e reconhecimento de padrões, permitindo a análise e visualização de dados (ASAN; ERCAN, 2012) como também o agrupamento dos dados de entrada em subconjuntos, em que cada um representa uma classe discreta de padrões (KOHONEN, 1990).

O principal objetivo da rede SOM é transformar um sinal de entrada de alta dimensão, complexo, em um mapa discreto de baixa dimensão, geralmente unidimensional ou bidimensional (HAYKIN, 2009), preservando as relações, ou seja, a topologia dos dados mas não as distâncias reais (ASAN; ERCAN, 2012).



Fonte: Haykin (2009).

Na Figura 7, temos  $x$  que é o sinal de entrada pertencente ao espaço de dimensão  $n$  representado no espaço de saída bidimensional pelo neurônio vencedor  $i(x)$  cuja localização está em função de  $x$  e  $w_i$  é o vetor de pesos do neurônio vencedor.

Segundo Asan e Ercan (2012) o treinamento de uma rede SOM envolve três fases: *i*) Competição, *ii*) Cooperação e *iii*) Adaptação.

## Competição

Na fase competitiva, os neurônios no mapa competem entre si para melhor representar a amostra de entrada. A melhor representação é determinada usando uma função discriminante, em que o vetor de entrada é comparado com o vetor de peso de cada nó. O neurônio com seus pesos de conexão semelhante à amostra é declarado vencedor, esse neurônio também é chamado de *Best-Matching Unit* (BMU), a função discriminante mais utilizada é a distância euclidiana, apresentada em (2.8).

$$dist_j^{(k)} = \sqrt{\sum_{i=1}^n (x_i^{(k)} - w_i^{(j)})^2}, \text{ onde } j = 1, \dots, n_1, \quad (2.8)$$

sendo  $dist_j^{(k)}$  a norma euclidiana entre o vetor  $x^{(k)}$  referente a  $k$ -ésima amostra e o vetor de pesos  $w^{(j)}$  do  $j$ -ésimo neurônio,  $n$  é o número de amostras de entrada e  $n_1$  o número de neurônios da grade neural (FARIA, 2012).

## Cooperação

Nesta segunda fase o BMU determina a localização espacial de uma vizinhança de nós cooperantes que compartilham características comuns. O BMU tem influência decrescente sobre a vizinhança, ou seja, à medida que a distância entre um neurônio e o BMU aumenta, diminui a influência do mesmo naquele neurônio (LIRA, 2012). Normalmente, usa-se uma curva gaussiana semelhante à (2.9) como função de vizinhança.

$$h_{ic}(t) = \exp\left(\frac{-d_{ic}^2}{2\sigma^2(t)}\right) \quad (2.9)$$

Em que  $d_{ic}$  é a distância euclidiana entre o neurônio vencedor  $i$  e um determinado neurônio  $c$ ,  $\sigma(t)$  representa a largura da vizinhança topológica que deve diminuir ao longo do treinamento a fim de restringir a vizinhança.

## Adaptação

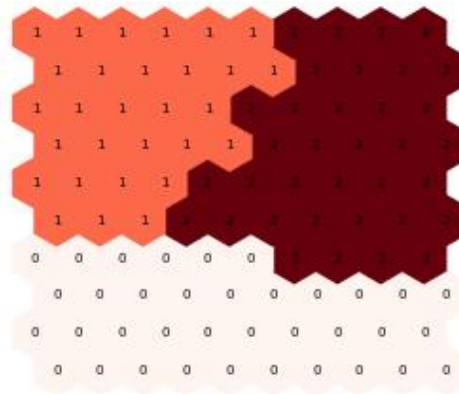
Na fase adaptativa, os vetores dos pesos do neurônio vencedor e da vizinhança são ajustados. De tal forma que os nós relevantes se tornam mais semelhantes à amostra de entrada, tendo uma chance maior de responder a dados de entrada semelhantes no futuro (ASAN; ERCAN, 2012). O ajuste dos pesos sinápticos é dado por (2.10):

$$\Delta w_j = \eta(t)h(t)(\mathbf{x}_k - \mathbf{w}_j), \quad (2.10)$$

em que  $\eta(t)$  é a taxa de aprendizagem da rede SOM, recebe valores que variam entre 0 e 1, partindo de um valor inicial e diminuindo gradualmente ao longo do tempo de treinamento a fim de evitar que novos dados desorganizem as informações já estabelecidas (LIRA, 2012).

O processo de aprendizagem da rede SOM pode ser comparado com o comportamento do cérebro humano em que diferentes entradas sensoriais são mapeadas em diferentes áreas do córtex cerebral de forma ordenada. De forma semelhante, após o treinamento da rede SOM é possível obter um mapa topológico dividido em regiões que definem classes representativas, um exemplo de mapa topológico é apresentado na Figura 8.

Figura 8 – Mapa topológico.

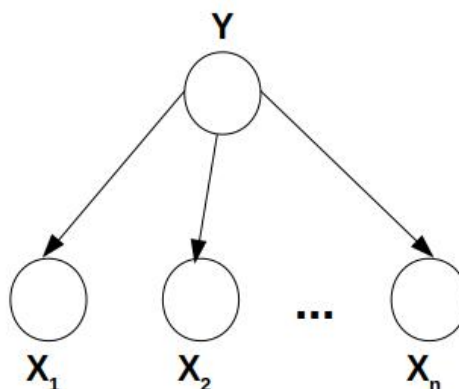


Fonte: Autoria Própria (2022).

Neste exemplo a rede SOM dividiu o mapa em três classes distintas numeradas de 0 a 2 que correspondem as três classes presentes no banco de dados *Iris*, cujo parâmetros foram utilizados para o treinamento da rede.

### 2.2.3 *Naive Bayes*

As redes *Bayesianas* são algoritmos de aprendizagem de máquina baseados na teoria de *Bayes*. O classificador *Naive Bayes* está entre os métodos de aprendizado *Bayesiano* em que as classes são descritas a partir da distribuição probabilística sobre os atributos dos exemplos (PADILHA, 1999). Esse classificador aprende a partir dos dados de treinamento a probabilidade condicional de cada atributo, dado o valor da classe (WANKE et al., 2014). Na Figura 9 temos um grafo representando um classificador *Naive Bayes* em que os vértices representam variáveis aleatórias e as arestas representam a relação de dependência estatística entre as variáveis (SANTOS, 2010).

Figura 9 – Representação de um classificador *Naive Bayes*.

Fonte: Autoria Própria (2022).

Para compreender o funcionamento do algoritmo é necessário definir o Teorema de *Bayes*. Thomas Bayes tratou nos anos de 1700 sobre a probabilidade de um evento ocorrer baseado em um conhecimento a priori (MONTGOMERY; RUNGER, 2008), conforme (2.11).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}, \quad (2.11)$$

em que,  $P(A)$  é a probabilidade a priori, ou seja, a probabilidade de um evento ter ocorrido ou não,  $P(B|A)$ , probabilidade condicional que é a probabilidade de um evento  $B$  ocorrer quando um evento  $A$  ocorrer,  $P(B)$ , probabilidade do evento  $B$  ocorrer e  $P(A|B)$ , probabilidade à posteriori baseado nas novas informações.

O classificador *Naive Bayes* se baseia na regra de *Bayes* para classificar um novo exemplo associando-o a classe mais provável,  $y$ , a partir dos valores de seu atributo  $(x_1, x_2, \dots, x_n)$  (PADILHA, 1999), conforme (2.12).

$$y = \operatorname{argmax}_y P(y|x_1, x_2, \dots, x_n), \quad (2.12)$$

reescrevendo (2.12) utilizando a Regra de *Bayes*, obtemos (2.13).

$$y = \operatorname{argmax}_y \frac{P(x_1, x_2, \dots, x_n|y)P(y)}{P(x_1, x_2, \dots, x_n)}, \quad (2.13)$$

considerando que o denominador,  $P(x_1, x_2, \dots, x_n)$ , é independente da classe,  $y$ , temos (2.14).

$$y = \operatorname{argmax}_y P(x_1, x_2, \dots, x_n|y)P(y) \quad (2.14)$$

Esse cálculo é possível devido a suposição de que todos os atributos são independentes em relação a classe, ou seja, dada uma classe, a probabilidade de um atributo não depende da probabilidade de nenhum outro (WANKE et al., 2014). Após essas considerações obteve-se (2.15), denominada de função de decisão do classificador *Naive Bayes*.

$$y = \operatorname{argmax}_y P(y) \prod_i P(x_i|y) \quad (2.15)$$

Apesar da forte suposição de independência entre os atributos, esta suposição nem sempre é verdade, mesmo assim o classificador pode aprender a função de classificação de modo satisfatório e apresentar um bom desempenho, mesmo quando não se verifica independência (SANTOS, 2010; RUSSEL; NORVIG, 2013).

## 2.3 Conceitos Estatísticos

Nesta seção, serão apresentados alguns conceitos relacionados a estatística descritiva, normalização de dados, uma breve fundamentação teórica sobre o coeficiente de correlação de

*Spearman*, apresentação da técnica de validação cruzada e dos avaliadores de desempenho. Esses conceitos foram utilizados durante a análise exploratória e estatística dos dados e implementação dos modelos.

### 2.3.1 Estatística Descritiva

Alguns conceitos básicos são úteis para se analisar um conjunto de dados obtido por meio de medições, inicialmente é importante determinar uma estimativa de onde a maioria dos dados está localizada, ou seja, sua tendência central (BRUCE; BRUCE, 2017). Para isso duas métricas se destacam, a média e a mediana. A média é a soma de todas as amostras dividido pelo número total de amostras, conforme (2.16).

$$\mu = \frac{\sum_i^n x_i}{n}, \quad (2.16)$$

em que  $\mu$  é a média,  $x_i$  corresponde à  $i$ -ésima amostra e  $n$  o número total de amostras.

A mediana é o valor central em um conjunto de dados e é considerada uma estimativa robusta de localização da tendência central, pois não é influenciada por *outliers*. Um *outlier* é qualquer valor que seja muito distante dos outros valores em um conjunto de dados e podem enviesar os resultados (BRUCE; BRUCE, 2017).

A variância e desvio padrão são medidas de dispersão ou variabilidade na distribuição dos dados (MONTGOMERY; RUNGER, 2008). A variância é a soma dos quadrados dos desvios da média, divididos por  $n - 1$ , em que  $n$  é o número de valores de dados, conforme (2.17). O desvio padrão é dado pela raiz quadrada da variância, conforme (2.18).

$$\sigma^2 = \frac{\sum_i^n (x_i - \mu)^2}{n - 1} \quad (2.17)$$

$$\sigma = \sqrt{\sigma^2} \quad (2.18)$$

Na análise dos dados é comum lidar com grau de incerteza, limites que representam um intervalo de valores plausíveis para um parâmetro de uma população, por exemplo, uma média é chamado intervalo de confiança. A definição de um intervalo de confiança é útil pois delimita observações futuras provenientes da população (MONTGOMERY; RUNGER, 2008). A equação (2.19) determina um intervalo de confiança para um valor médio.

$$\mu \pm Z \frac{\sigma}{\sqrt{n}}, \quad (2.19)$$

em que  $\mu$  é a media amostral,  $Z$  é o valor associado ao intervalo de confiança, valores típicos são 90%, 95% e 99% e  $\sigma/\sqrt{n}$  é o desvio padrão amostral da média.

### 2.3.2 Coeficiente de Correlação de *Spearman*

A correlação é um meio para analisar a relação entre duas ou mais variáveis definidas em um espaço probabilístico (MONTGOMERY; RUNGER, 2008). Para avaliação dessa relação podem ser aplicados diferentes critérios ou coeficientes a depender da distribuição dos dados. É comum a utilização do coeficiente de *Pearson*, entretanto outras alternativas são: coeficiente de *Spearman*, coeficiente de determinação, coeficiente de contingência, entre outros (GUIMARÃES, 2007).

Neste trabalho foi utilizado o coeficiente de correlação de *Spearman* destinado para dados não paramétricos, ou seja, dados que não possuem uma distribuição normal (MONTGOMERY; RUNGER, 2008), a saber que a definição matemática do coeficiente é dado por (2.20).

$$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}, \quad (2.20)$$

$rg_X, rg_Y$  são as variáveis convertidas em posto, uma sequência de números inteiros e positivos, obtidos por meio de um ordenamento numérico,  $r_s$  é o coeficiente de *Spearman*,  $\rho$  é o coeficiente de *Pearson* aplicado às variáveis em posto,  $cov(rg_X, rg_Y)$  é a covariância das variáveis,  $\sigma_{rg_X}$  e  $\sigma_{rg_Y}$  são os desvios padrão das variáveis.

Um sinal positivo do coeficiente de correlação indica que quando a variável  $X$  aumenta,  $Y$  aumenta. O sinal negativo indica que quando  $X$  aumenta,  $Y$  tende a diminuir. Um coeficiente de valor igual a zero indica que não existe relação entre as variáveis. Coeficientes com magnitudes maiores próximos a um indicam uma correlação mais forte entre as variáveis. Na Tabela 3 é apresentado uma classificação da correlação em função dos coeficientes de *Spearman*.

Tabela 3 – Classificação do nível de correlação com base no coeficiente de *Spearman*

Correlação	Coeficiente
Perfeita Negativa	$r_s = -1$
Negativa	$-1 < r_s < 0$
Nula	$r_s = 0$
Positiva	$0 < r_s < 1$
Perfeita Positiva	$r_s = 1$
Muito Fraca*	$0 < r_s < 0,19$
Fraca*	$0,2 < r_s < 0,39$
Moderada*	$0,4 < r_s < 0,69$
Forte*	$0,7 < r_s < 0,89$
Muito forte*	$0,9 < r_s < 1$

\*Possui o mesmo significado para os casos negativos ou positivos.

Fonte. Adaptado de Guimarães (2007).



### 2.3.3 Normalização de Dados

A normalização ou padronização dos dados é uma etapa importante para o desenvolvimento e aplicação dos algoritmos de IA, pois a maioria dos algoritmos não funcionam de forma satisfatória quando os parâmetros possuem intervalos diferentes entre si (GÉRON, 2019). A normalização pode ser realizada de duas formas, primeiro por meio da criação de uma nova variável a partir de (2.21).

$$Z = \frac{X - \mu}{\sigma}, \quad (2.21)$$

no qual  $Z$  representa a distância de  $X$  a partir de sua média,  $\mu$ , em termos dos desvios padrão,  $\sigma$  (MONTGOMERY; RUNGER, 2008).

A segunda forma é o escalonamento min-max em que os valores são deslocados e redimensionados para que variem em uma determinada faixa de valores, por exemplo, entre 0 e 1 (GÉRON, 2019). Para isso a variável é subtraída do valor mínimo e dividido pelo máximo menos o mínimo conforme (2.22).

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}, \quad (2.22)$$

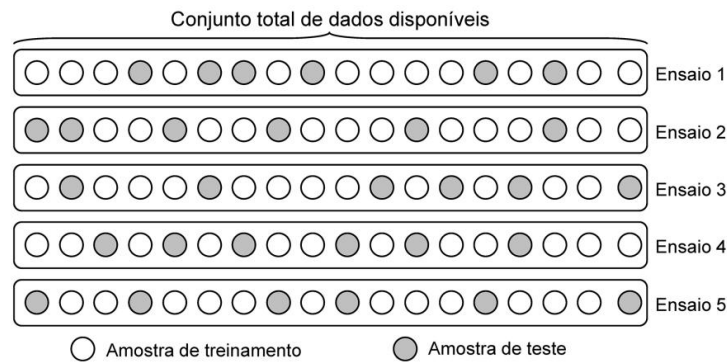
em que  $X_{new}$  é o valor normalizado,  $X$  a variável antes do escalonamento,  $X_{min}$  é o menor valor atribuído e  $X_{max}$  é o valor máximo atribuído.

### 2.3.4 Validação Cruzada

A validação cruzada (do inglês - *cross-validation*) é uma das técnicas estatísticas mais utilizadas para seleção de topologias de redes neurais e determinação de hiperparâmetros (SILVA; SPATTI; FLAUZINO, 2010), o objetivo é avaliar a capacidade de generalização das topologias diante de um conjunto de dados diferente daquele utilizado no ajuste dos parâmetros. A seguir, serão apresentados três métodos tipicamente utilizados.

O primeiro método é denominado validação cruzada por amostragem aleatória, consiste na divisão aleatória do conjunto de dados em duas partes, o subconjunto de treinamento utilizado para treinar todas as topologias candidatas, e o subconjunto de teste aplicado para selecionar a topologia que apresenta maior capacidade de generalização (SILVA; SPATTI; FLAUZINO, 2010). A Figura 10 apresenta a seleção de amostras de treinamento e testes escolhidas aleatoriamente para cada um dos ensaios.

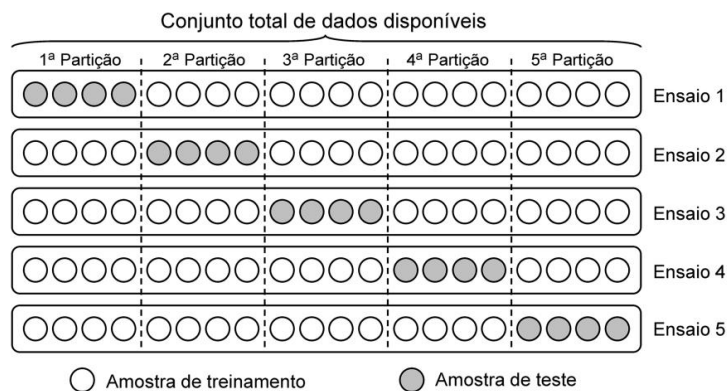
Figura 10 – Método de validação cruzada usando amostragem aleatória.



Fonte: Silva, Spatti e Flauzino (2010).

O segundo método é a validação cruzada com  $k$ -partições a ideia é que cada exemplo sirva duplamente como dados de treinamento e dados de teste. Os dados são divididos em  $k$  subconjuntos e realiza  $k$  rodadas de aprendizagem, em cada rodada  $1/k$  dos dados é utilizado como conjunto de teste (RUSSEL; NORVIG, 2013). Na Figura 11 é apresentado um esquema da seleção de amostras considerando  $k$  igual a 5 e um total de 20 amostras.

Figura 11 – Método de validação cruzada usando  $k$  partições.



Fonte: Silva, Spatti e Flauzino (2010).

O terceiro método é o caso extremo do anterior quando  $k = n$ , em que  $n$  é o número total de amostras (RUSSEL; NORVIG, 2013). Conhecido como validação cruzada por unidade em que uma única amostra constitui o subconjunto de teste e as demais o subconjunto de treinamento. O processo de aprendizado se repete até que todas as amostras sejam utilizadas como subconjunto de teste, portanto esse método é menos usual (SILVA; SPATTI; FLAUZINO, 2010).

### 2.3.5 Avaliadores de Desempenho

Durante a implementação das técnicas de inteligência artificial algumas métricas foram utilizadas tanto nas fases de treinamento quanto na fase de validação dos modelos. As métricas são utilizadas com a finalidade de escolher os melhores parâmetros e avaliar o desempenho dos modelos.

Para os modelos SVM e *Naive Bayes* foram utilizados como avaliadores de desempenho a matriz de confusão, acurácia, precisão e sensibilidade. Uma matriz de confusão de tamanho  $N \times N$  esta associada a um classificador, em que cada linha da matriz representa uma classe real, enquanto cada coluna representa uma classe prevista (GÉRON, 2019),  $N$  é a quantidade de classes. Na Tabela 4 tem-se um exemplo de uma matriz de confusão de 2 classes.

Tabela 4 – Exemplo de uma matriz de confusão.

Matriz de Confusão		Classe predita	
		Positiva	Negativa
Classe Original	Positiva	VP	FN
	Negativa	FP	VN

Fonte: Mariano (2021)

Em que VP são os verdadeiros positivos, VN os verdadeiros negativos, FP os falsos positivos e FN os falsos negativos.

A acurácia (ACC, do inglês *accuracy*) uma das métricas mais utilizadas, avalia o percentual de acertos, ou seja, é obtida pela razão entre a quantidade de acertos e o total de entradas, conforme (2.23), tomando como base a matriz de confusão (Tabela 4).

$$acurácia = \frac{VP + VN}{VP + FN + VN + FP} \quad (2.23)$$

O problema com a acurácia é que essa métrica não nos informa tudo sobre os resultados obtidos (MARSLAND, 2014), ela apresenta apenas uma informação mais geral sobre o desempenho do classificador. Portanto, além da acurácia outras métricas foram utilizadas.

A precisão é uma métrica que avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos (MARIANO, 2021) conforme (2.24) geralmente é utilizada em conjunto com outra métrica chamada revocação também conhecida como sensibilidade (GÉRON, 2019).

$$precisão = \frac{VP}{VP + FP} \quad (2.24)$$

A sensibilidade (do inglês, *recall*) é utilizada com a finalidade de avaliar a capacidade do modelo de detectar resultados classificados como positivos. A sensibilidade é útil quando

se deseja um sistema com taxa de negativos próximo a zero (MARIANO, 2021). Valores altos de sensibilidade indicam alto valores de verdadeiros positivos frente a falsos negativos, conforme (2.25). Essa métrica foi utilizada a fim de avaliar principalmente, como o classificador se comporta nos níveis mais altos de poluição.

$$\text{sensibilidade} = \frac{VP}{VP + FN} \quad (2.25)$$

Juntos, esses pares de medidas fornecem mais informações do que apenas a acurácia. Eles podem ser combinados para se obter uma única medida, o  $f1 - score$ , que pode ser obtido em termos de precisão e sensibilidade conforme (2.26) (MARSLAND, 2014). O  $f1 - score$  é a média harmônica da precisão e revocação, desta forma, valores mais baixos terão mais peso. Como resultado, o classificador só obterá um  $f1 - score$  alto se a sensibilidade e a precisão forem altas (GÉRON, 2019). Essa métrica foi utilizada para comparar o desempenho entre os diferentes classificadores.

$$f1 - score = 2 * \frac{\text{precisão} * \text{sensibilidade}}{\text{precisão} + \text{sensibilidade}} \quad (2.26)$$

As métricas utilizadas para avaliação dos Mapas Auto-Organizáveis foram: o erro de quantização vetorial ( $E_q$ ) e a taxa de classificação correta. O  $E_q$  corresponde a média do erro entre o vetor de entrada  $X_i$  e o vetor de pesos sinápticos correspondente ao neurônio vencedor,  $w_{BMU}$ , essa relação é apresentada em (2.27).

$$E_q = \frac{\sum_{i=1}^n \|x_i - w_{BMU}\|}{n} \quad (2.27)$$

A CCR também é utilizada para avaliar a exatidão do sistema (LIRA, 2012), e é dado por (2.28).

$$CCR = \frac{c}{d} * 100, \quad (2.28)$$

em que  $c$  é o número de padrões classificados corretamente e  $d$  é o número total de padrões.

## 3 Revisão Bibliográfica

Uma revisão bibliográfica e levantamento do estado da arte sobre determinada área é importante para verificarmos aquilo que foi ou está sendo desenvolvido. Na seção 3.1 são apresentados algumas pesquisas desenvolvidas sobre o diagnóstico de isoladores de alta tensão, especificamente utilizando a corrente de fuga como parâmetro de classificação e a aplicação de modelos de inteligência artificial.

### 3.1 Levantamento do Estado da Arte

O trabalho de Suda (2001) apresenta um método para o monitoramento do nível de poluição dos isoladores baseado na forma de onda e nas componentes de frequência da corrente de fuga. Um isolador do tipo pino foi poluído artificialmente, considerando o índice DDSE e as formas de onda e espectro da frequência foram obtidas. Com base na forma de onda e espectro da frequência da corrente de fuga, determinam-se seis estágios até a ocorrência da tensão de ruptura. Estes estágios compreendem a ausência de descargas, o aparecimento de descargas fracas sem sinal visível, aparecimento de pontos visíveis, arcos locais fracos, arcos locais fortes e intermitentes e arcos locais muito fortes logo antes do *flashover*. Além disso, é determinada a evolução temporal do valor de pico da corrente de fuga e das 1<sup>a</sup>, 5<sup>a</sup>, 7<sup>a</sup> e 9<sup>a</sup> componentes harmônicas. Verifica-se que a possibilidade da ocorrência de *flashover* aumenta quando esses parâmetros atingem determinados níveis. O trabalho apresenta uma análise superficial entre os diferentes níveis de poluição e os parâmetros analisados, ou seja, não é feita uma correlação entre os níveis atingidos pelos parâmetros da corrente de fuga e os diferentes valores de DDSE, também não é apresentado formas de prever a ocorrência de *flashover*, tornando o método refém da análise de especialistas no estudo das formas de onda e componentes de frequência da corrente de fuga.

Li et al. (2010) apresenta um método para predição do nível de contaminação na superfície de isoladores utilizando o valor médio, valor máximo e o desvio padrão do valor *Root Mean Square* (RMS) da corrente de fuga. Foram realizados testes de poluição artificial em duas cadeias formadas por três isoladores cada uma delas, com diferentes aspectos construtivos. Após os testes, foram determinadas as relações entre as características da corrente de fuga e o índice DDSE, essa relação foi obtida por meio de equações de regressão linear, os coeficientes

de determinação ( $R^2$ ) obtidos foram maiores que 0,95. Com o objetivo de prever o nível de contaminação dos isoladores os autores utilizaram uma Rede Neural Artificial (RNA) com algoritmo *backpropagation*. O erro absoluto entre os valores de DDSE obtidos pela rede neural e os determinados durante os ensaios foram utilizados para analisar a eficiência do modelo de previsão. Pode-se concluir que o valor máximo da corrente de fuga apresentou maior relação com o nível de contaminação. Verificou-se também que a estrutura e tipo do isolador influencia no valor da corrente de fuga, mas neste caso as análises dos autores foram superficiais pois foi considerado apenas a distância de escoamento. Um dos problemas deste método é que foram considerados durante os testes apenas cinco valores de DDSE o que pode influenciar na capacidade de generalização do modelo.

Muniraj e Chandrasekar (2011) apresentam um sistema de previsão do estado da poluição de isoladores poliméricos de 11 kV, a partir das seguintes características da corrente de fuga: valor médio, valor máximo, desvio padrão e Distorção Harmônica Total (THD, do inglês *Total Harmonic Distortion*). Essas características foram utilizadas como entrada de um modelo *Adaptive Neuro Fuzzy Inference System* (ANFIS). Os isoladores foram poluídos artificialmente e o índice DDSE foi utilizado como saída do modelo. Uma RNA com algoritmo *backpropagation* foi implementada para fins de comparação com o modelo ANFIS. Portanto, o modelo apresentou um maior coeficiente de determinação (próximo a 1) em relação à RNA. Novamente, neste caso, foram considerados apenas cinco valores de DDSE durante o processo de poluição, para o treinamento do modelo.

No trabalho de Pylarinos et al. (2012) é apresentado um método de monitoramento e classificação das formas de onda da corrente de fuga em isoladores tanto no domínio do tempo como no da frequência. Os isoladores pertencentes a subestações de 150 kV foram monitorados durante seis anos, portanto a poluição nos isoladores foi resultado da exposição intensa a poluição marinha. Com um arranjo de 18 isoladores, foi construído um banco de dados com 80.000 formas de onda e selecionadas 1540 para as análises. Com o intuito de classificar as formas de onda da corrente de fuga foi selecionado vinte parâmetros da corrente, sendo dez no domínio do tempo e dez no domínio da frequência. Dois algoritmos de seleção de recursos foram utilizados teste-t e *Minimum Redundancy–Maximum Relevance* (mRMR). Algoritmos de *machine learning* como: *K-Nearest Neighbor Classifier* (KNN), *Naive Bayes* e SVM foram utilizados para classificação. Pode-se observar que o teste-t é ineficaz neste caso, pois ofereceu os piores resultados em comparação ao mRMR. Observou-se que os testes de classificação com os algoritmos de *machine learning* tiveram um bom desempenho com uma taxa de acerto de até

90%. Os autores concluíram que os algoritmos de seleção de recursos podem ser usados para aprimorar o desempenho de classificação e reduzir o número de parâmetros utilizados. Por fim, observou-se que o modelo SVM apresentou um desempenho superior com o conjunto de recursos fornecidos pelo algoritmo mRMR. O trabalho apesar da ótima contribuição em relação as análises da corrente de fuga não apresenta, de forma clara, uma relação com o estado de degradação dos isoladores devido a ausência de indicadores do nível de poluição não contribuindo para os processos de tomada de decisão em relação aos equipamentos.

Em Zhao et al. (2013) é desenvolvido um método para previsão da tensão de ruptura (*flashover*) em isoladores, que ocorre devido ao depósito de poluição na superfície dos equipamentos. Para isso são analisadas as relações entre o índice de Densidade de Depósito de Sal (DDS) e da hidrofobicidade na corrente de fuga e na tensão de ruptura. Foram realizados testes de medição da corrente de fuga e tensão de ruptura em isoladores poliméricos de 35 kV, poluídos artificialmente. Inicialmente foram determinadas as relações entre a tensão disruptiva e o índice DDS e a hidrofobicidade através de equações de regressão linear. Os parâmetros da corrente de fuga como: a entropia da amplitude do pulso, amplitude máxima do pulso, a razão da energia e energia foram extraídos. As relações entre os parâmetros da corrente de fuga e o índice DDS e a hidrofobicidade também foram determinados. Para prever o valor da tensão disruptiva foi utilizado uma técnica de aprendizagem de máquina denominado Máquinas de Vetores de Suporte de Mínimos Quadrados (LS-SVM, do inglês Least-Squares Support Vector Machine), como entrada do modelo, foram utilizados os parâmetros extraídos da corrente de fuga. O erro relativo foi empregado para verificar a performance do algoritmo, para isso foi calculado o erro entre o valor da tensão de ruptura prevista pelo modelo e o valor medido durante os testes, como também o erro entre a tensão de ruptura prevista e os valores calculados pelas equações de regressão linear, nos dois casos foi apresentado erros abaixo de 7%. O método não utiliza o nível de poluição como entrada ou saída do modelo, tornando as análises da relação entre o nível de poluição e tensão de ruptura ou da corrente de fuga superficiais, pois não é apresentado uma previsão do estado de degradação do isolador apenas uma indicação da ocorrência de *flashover*.

Em Khaled, El-Hag e Assaleh (2016) é apresentado uma técnica de monitoramento do nível de contaminação da superfície de isoladores poliméricos a partir da corrente de fuga. Os isoladores foram poluídos artificialmente com determinado nível de DDSE e ensaiados em uma câmara de névoa. As correntes de fuga foram medidas durante o ensaio e os valores de pico, componente fundamental e terceira harmônica foram extraídos a fim de serem usados para prever o nível atual de DDSE. Para criar o modelo, três técnicas de classificação foram utilizadas:

KNN, Classificador Polinomial e Classificador *Neuro-Fuzzy*. Os autores verificaram que das três técnicas, a que apresentou uma maior acurácia geral foi o KNN, menos para o nível de poluição média, em que o classificador Polinomial apresentou desempenho maior. Verificou-se que o problema de uma menor taxa de acerto para níveis de poluição intermediárias ocorreu devido ao banco de dados desbalanceado que apresenta um menor número de amostras nesse caso, influenciando na capacidade de classificação dos modelos.

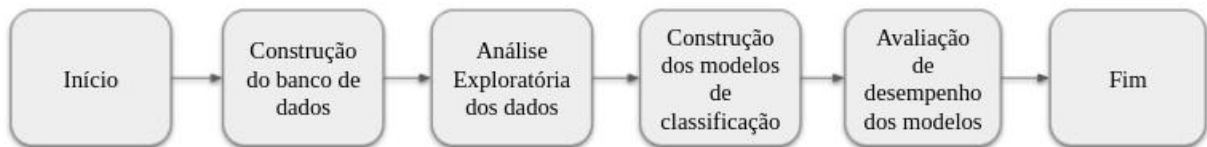
Khafaf e El-Hag (2018) utiliza RNA com regularização *Bayesiana* para prever valores futuros da componente fundamental e terceira harmônica da corrente de fuga em isoladores poliméricos poluídos artificialmente. Três tipos de RNA foram utilizadas e o *Normalized Root Mean Square Error* (NRMSE) foi utilizado como avaliador de desempenho das redes neurais. Foi considerado uma janela de tempo de 30 a 240 minutos, verificando que o NRMSE diminui a medida que a janela de previsão aumenta, isso pode ser atribuído as frequentes mudanças dos parâmetros da corrente de fuga a medida que o envelhecimento do equipamento avança. Portanto, o modelo implementado por Khafaf e El-Hag (2018), pode ser configurado para prever a série temporal da corrente de fuga por um longo período alcançando melhores resultados. Apesar das grandes contribuições relacionadas a previsão da corrente de fuga, o modelo não apresenta uma classificação do nível do estado de poluição dos isoladores, tornando o método refém da análise de especialistas no estudo da previsão dos parâmetros da corrente de fuga.



## 4 Metodologia Proposta

Neste capítulo são apresentados os procedimentos práticos e computacionais utilizados para a classificação dos níveis de poluição de isoladores de vidro utilizando medições de corrente de fuga. As etapas apresentadas a seguir são: construção do banco de dados, análise exploratória e estatística dos dados e a implementação dos modelos de classificação. Na Figura 12 é apresentado de forma esquemática um resumo dos procedimentos realizados.

Figura 12 – Resumo dos procedimentos propostos.



Fonte: Autoria Própria (2022).

### 4.1 Construção do Banco de Dados

A simulação dos níveis de poluição se deu por meio de ensaios em laboratório em uma cadeia de isoladores poluída artificialmente. A cadeia de isoladores foi composta por cinco isoladores de vidro, tipo suspensão, compondo um arranjo para um sistema de 69 kV. Os isoladores selecionados são do mesmo fabricante e possuem as mesmas características físicas e elétricas, estas apresentadas na Tabela 5. Esses isoladores pertencem ao Laboratório de Alta Tensão (LAT) da UFCG.

Tabela 5 – Características físicas e elétricas individuais dos isoladores de vidro utilizados

Características Físicas		Características Elétricas	
Resistência eletromecânica combinada	70 kN	Tensão disruptiva em frequência industrial a seco	80 kV
Resistência ao impacto	45 kN	Tensão disruptiva em frequência industrial sob chuva	50 kV
Distância de escoamento	320 mm	Tensão crítica de impulso atmosférico - Positivo	125 kV
-	-	Tensão crítica de impulso atmosférico - Negativo	130 kV
-	-	Tensão de perfuração na frequência industrial	130kV

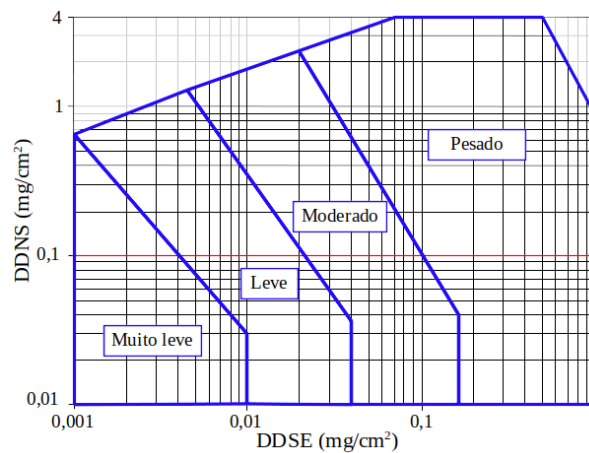
Fonte: Autoria Própria (2022).

Os isoladores foram lavados e ensaiados antes de serem poluídos com o objetivo de

se obter valores de referência para o estado do isolador limpo. Esses dados foram usados posteriormente nos modelos de classificação. A higienização dos isoladores foi feita com a aplicação de detergente líquido e água, a fim de garantir a remoção de impurezas presentes na superfície dos equipamentos, esse procedimento foi repetido antes de cada aplicação da solução de poluição.

A realização da poluição artificial foi feita seguindo as instruções presentes nas normas IEC 60815-1 (IEC, 2008) e IEC 6057 (IEC, 2013). Para a determinação do nível de poluição, considerou-se a combinação dos índices de DDSE e DDNS bem como a condutividade associada. A norma IEC 60815-1 (IEC, 2008) indica cinco níveis de poluição: muito leve, leve, média, pesada e muito pesada. Devido a aspectos práticos, relacionados a escala do condutivímetro utilizado para medição da condutividade da solução, os níveis de poluição pesado e muito pesado foram classificados como pesado. A Figura 13 apresenta graficamente, os níveis de poluição relacionados ao índices DDSE e DDNS.

Figura 13 – Níveis de poluição com base nos índices DDSE e DDNS.



Fonte: Adaptado de IEC 60815-1 (2008)

Neste trabalho além dos níveis de poluição considerados da Figura 13 foi incluído os valores de referência para os isoladores limpos, totalizando cinco níveis a serem classificados: limpo, muito leve, leve, moderado e pesado.

O procedimento de poluição iniciou-se com a preparação da solução salina, composta por água, material inerte e sal de cozinha. O material inerte utilizado foi caulim, na proporção de 40 g para cada 1 litro de água e a quantidade de sal variou conforme o nível de condutividade elétrica da solução. O nível de poluição é determinado pelos índices DDSE e DDNS, mas o parâmetro utilizado para o preparo da solução foi a condutividade elétrica. A norma IEC 6057 (IEC, 2013) apresenta um conjunto de dados que relaciona de forma aproximada esses

parâmetros, conforme reproduzido na Tabela 6.

Tabela 6 – Relação entre DDSE e condutividade elétrica da solução.

DDSE (mg/cm <sup>2</sup> )	Condutividade Elétrica (mS/cm)
0,012	5
0,018	7
0,025	10
0,035	14
0,05	20
0,07	28
0,1	40
0,14	56
0,2	80
0,28	112
0,4	160

Fonte: Adaptado da IEC 6057 (2013).

Para se obter soluções que simulassem os níveis de poluição, utilizaram-se em conjunto as informações presentes na Figura 13 e Tabela 6 a fim de se obter uma relação entre o DDSE e a condutividade, a relação apresentada é linear, conforme (4.1).

$$DDSE = 0,0025 \times \sigma, \quad (4.1)$$

em que  $\sigma$  é a condutividade da solução e  $DDSE$  é o valor correspondente a  $\sigma$ . O índice DDNS foi mantido constante já que a proporção do material inerte e a água foi mantida constante.

Determinadas essas relações, os valores de condutividade elétrica da solução são apresentados na Tabela 7, como também o valor de DDSE e o nível de poluição.

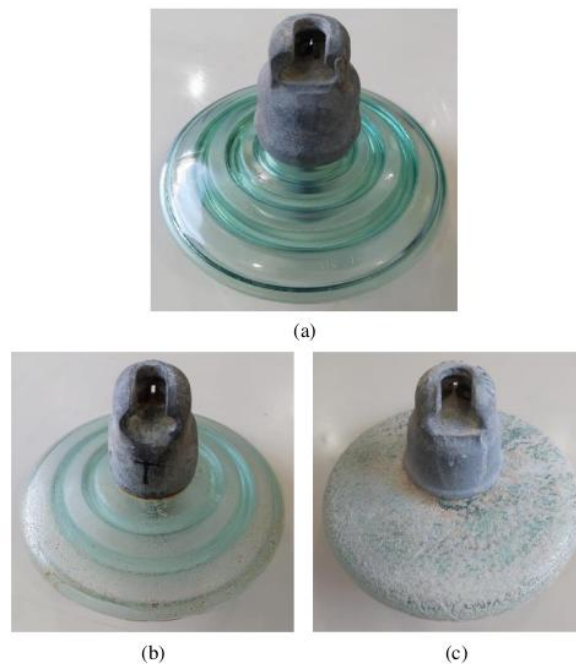
Tabela 7 – Níveis de poluição e valores correspondentes de DDSE e condutividade elétrica.

Níveis de Poluição	DDSE (mg/cm <sup>2</sup> )	Condutividade Elétrica da Solução (mS/cm)
Limpo	$5,10 \times 10^{-4}$	0,21
Limpo	$8,64 \times 10^{-4}$	0,36
Muito Leve	$2,16 \times 10^{-3}$	0,98
Muito Leve	$3,45 \times 10^{-3}$	1,44
Muito Leve	$4,60 \times 10^{-3}$	1,90
Leve	$8,04 \times 10^{-3}$	3,35
Leve	$1,53 \times 10^{-2}$	6,38
Leve	$1,90 \times 10^{-2}$	8,21
Moderado	$3,82 \times 10^{-2}$	15,91
Moderado	$6,22 \times 10^{-2}$	25,90
Moderado	$9,90 \times 10^{-2}$	41,50
Pesado	$1,69 \times 10^{-1}$	70,60
Pesado	$3,15 \times 10^{-1}$	131,2
Pesado	$3,80 \times 10^{-1}$	160,5

Fonte: Autoria Própria (2022).

Com a solução preparada, após a higienização dos isoladores, os mesmos foram imersos na solução com o objetivo de garantir uma contaminação ou depósito de poluição uniforme, depois passaram por um processo de secagem de 24 horas antes do início da realização dos ensaios, com a finalidade de garantir a secagem da superfície de todos os isoladores ensaiados. Na Figura 14 são apresentadas fotografias do isolador a) higienizado, b) logo após a aplicação da solução salina e c) após o processo de secagem.

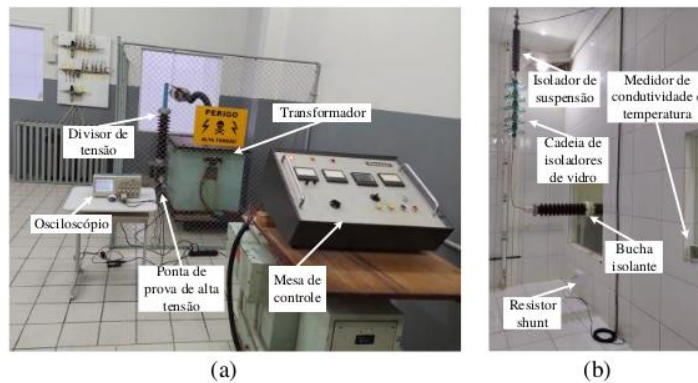
Figura 14 – Fotografia dos isoladores utilizados.



Fonte: Aatoria Própria (2022).

Após os procedimentos de poluição, os isoladores foram instalados no interior de uma câmara de névoa capaz de isolar o objeto de teste da temperatura e umidade ambiente, com a finalidade de analisar os níveis de poluição depositados na superfície dos isoladores de vidro. As dimensões da câmara de névoa são 6,5 m x 2,0 m x 4,0 m. Na Figura 15 são apresentadas as fotografias dos ambientes externo e interno da câmara de névoa, como também dos componentes utilizados para realização dos ensaios elétricos.

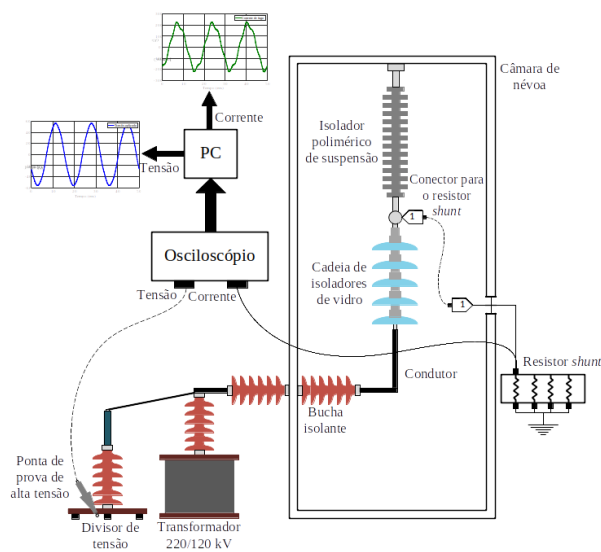
Figura 15 – Fotografias das partes a) externa e b) interna da câmara de névoa.



Fonte: Autoria Própria (2022).

O arranjo experimental é composto por uma mesa de controle, um transformador elevador (220/120 kV), um divisor de tensão capacitivo, uma ponta de prova de alta tensão Tektronix P66015A, objeto de teste (cadeia de isoladores para um sistema de 69 kV) em série com um resistor *shunt* de 1 k $\Omega$  e osciloscópio digital. Isoladores poliméricos de suporte foram utilizados a fim de proporcionarem sustentação mecânica e atuarem como uma alta impedância, permitindo que a corrente de fuga escoe para o resistor *shunt*. Além disso, um termo-higrômetro digital foi utilizado para realizações de medições de temperatura e umidade. Um diagrama do arranjo experimental completo utilizado para aquisição dos sinais de corrente de fuga se encontra na Figura 16.

Figura 16 – Diagrama do arranjo experimental utilizado nos ensaios elétricos.

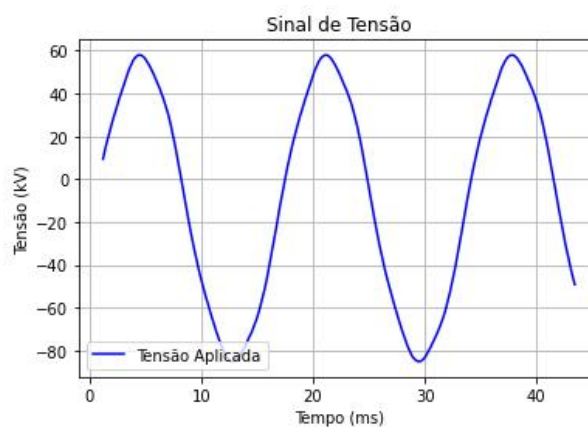


Fonte: Autoria Própria (2022).

Os sinais de corrente de fuga foram adquiridos de forma indireta pelo princípio da Lei de Ohm, levando em consideração a queda de tensão sobre o resistor *shunt*.

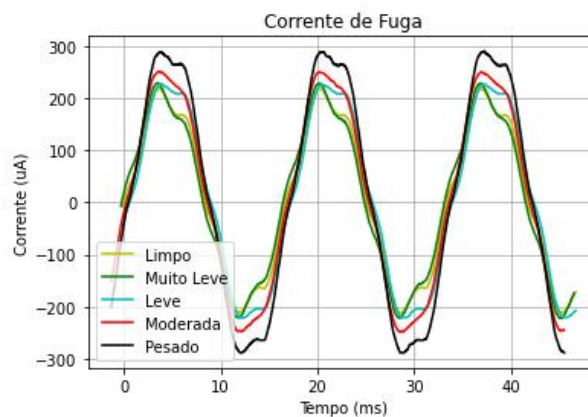
Cada ensaio foi realizado durante duas horas e trinta minutos, considerando o tempo para a montagem do arranjo, o tempo de espera de uma hora para que o isolador entre em estado de operação semelhante à campo em termos de estabilidade térmica e o tempo para aquisição dos dados. Para cada ensaio, foram realizadas dez medições dos sinais de corrente de fuga. Os sinais foram digitalizados e armazenados em um cartão de memória a fim de serem processados em um computador. Nas Figuras 17 e 18 podem ser observados sinais de tensão e corrente de fuga, respectivamente, para diferentes estados de poluição.

Figura 17 – Sinal de tensão aplicado.



Fonte: Autoria Própria (2022).

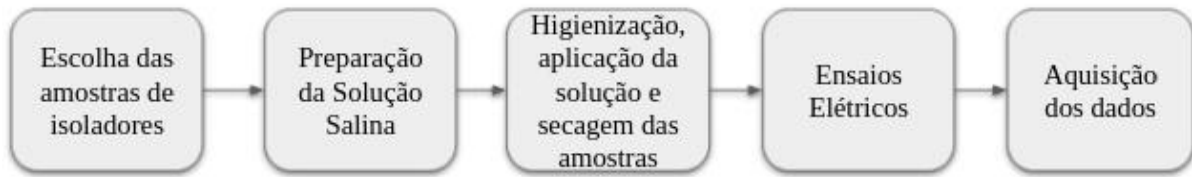
Figura 18 – Sinais da corrente de fuga para diferentes níveis de poluição.



Fonte: Autoria Própria (2022).

A Figura 19 apresenta um resumo do processo para construção do banco de dados apresentado nesta seção.

Figura 19 – Resumo do procedimento utilizado para construção do banco de dados.



Fonte: Autoria Própria (2022).

## 4.2 Análise Exploratória dos Dados

A análise exploratória dos dados envolveu uma análise descritiva dos sinais de tensão aplicado ao longo dos ensaios, dos parâmetros ambientais como temperatura e umidade e a extração dos parâmetros da corrente de fuga, como também uma análise da correlação entre os mesmos e o nível de poluição. O sinal de tensão cuja amostra pode ser vista na Figura 17, permaneceu praticamente constante ao longo dos ensaios com um desvio padrão de 0,26. A média da temperatura durante os ensaios foi de 27,02°C e da umidade 63,71%, também apresentaram baixa variação. A Tabela 8 apresenta um resumo das estatísticas descritivas obtidas para os valores de tensão, temperatura e umidade.

Tabela 8 – Análise estatística dos valores de tensão, temperatura e umidade.

Parâmetros Estatísticos	Tensão (kV)	Temperatura (°C)	Umidade (%)
Número de Amostras	140	140	140
Média	40,23	27,02	63,71
Mediana	40,33	27,0	65,0
Desvio Padrão	0,26	0,65	2,70
Mínimo	39,82	25,8	59,0
Máximo	40,55	28,0	67,0

Fonte: Autoria Própria (2022).

Os parâmetros da corrente de fuga foram extraídos utilizando ferramentas computacionais. Os parâmetros obtidos foram: valor de pico, valor eficaz, THD, nível CC, componente fundamental, 3ª componente harmônica, 5ª componente harmônica, 7ª componente harmônica, 9ª componente harmônica e 11ª componente harmônica. O valor de pico foi obtido a partir do valor máximo do sinal de corrente de fuga e o valor eficaz foi determinado por (4.2).

$$i_{eficaz} = \sqrt{\frac{1}{N} \sum_{j=1}^N i_j^2}, \quad (4.2)$$

em que  $N$  é o número equivalente a múltiplos de um período e  $i_j$  é o valor do sinal de corrente no  $j$ -ésimo ponto.

A THD também foi obtida para análise da relação com o grau de poluição, esse valor foi calculado utilizando (4.3).

$$THD\% = \frac{\sqrt{\sum_{j>1}^{j_{max}} i_j^2}}{i_1} \times 100, \quad (4.3)$$

em que  $i_1$  é a amplitude da componente harmônica fundamental e  $i_j$  é amplitude da  $j$ -ésima componente harmônica.

As componentes harmônicas da corrente de fuga e o nível CC foram determinados utilizando a Transformada Rápida de *Fourier* (FFT, do inglês *Fast Fourier Transform*). A FFT é um algoritmo desenvolvido para calcular a Transforma Discreta de *Fourier* (DFT, do inglês *Discrete Fourier Transform*) de forma prática e rápida. Sua velocidade vem do fato que ela utiliza os cálculos anteriores para reduzir o número de operações e explora a periodicidade e simetria das funções trigonométricas (CHAPRA; CANALE, 2011). O resumo do banco de dados obtido após a extração dos atributos se encontra no Apêndice A, em que tem-se o grau de poluição, o índice DDSE, a condutividade e os parâmetros da corrente de fuga obtidos.

Os dados obtidos foram utilizados para construir a matriz de correlação, baseado no coeficiente de correlação de *Spearman*. O resumo da matriz de correlação é apresentada na Tabela 9 e contém a correlação entre o índice DDSE e a condutividade, variáveis de saída, e os parâmetros da corrente de fuga.

Tabela 9 – Resumo da matriz de correlação de *Spearman*.

	$I_{pico}$	$I_{rms}$	$I_{cc}$	$I_{1har}$	$I_{3har}$	$I_{5har}$	$I_{7har}$	$I_{9har}$	$I_{11har}$	$I_{THD}$
Cond.	0,874	0,981	-0,273	0,981	0,228	0,040	0,218	0,211	0,238	-0,771
DDSE	0,874	0,981	-0,273	0,981	0,228	0,040	0,218	0,211	0,238	-0,771

Fonte: Autoria Própria (2022).

Os parâmetros que mais se correlacionam com o índice DDSE e a condutividade são: valor eficaz da corrente de fuga, componente harmônica fundamental e o valor de pico. Esses parâmetros apresentaram coeficientes de correlação próximos a 1 indicando, segundo a Tabela 3 uma forte ou muito forte correlação positiva. Com isso os três parâmetros foram escolhidos para modelagem dos modelos de classificação. No Apêndice B é apresentado a matriz de correlação de *Spearman* completa contendo a relação entre todas as variáveis analisadas.



## 4.3 Implementação dos Modelos

Os algoritmos para classificação do nível de poluição dos isoladores foram implementados utilizando linguagem de programação *Python*. As três técnicas de inteligência artificial utilizadas foram: SVM, Rede SOM e *Naive Bayes*. A seguir, será apresentado, a descrição da implementação dos modelos.

### 4.3.1 Modelo de Classificação Utilizando SVM

A implementação deste modelo foi realizada utilizando a biblioteca *scikit-learn* do *Python*. O desenvolvimento do modelo envolveu duas fases: treinamento e teste, como sugerido pela literatura 80% dos dados foram reservados para treinamento e 20% para teste. Os dados de entrada foram normalizados utilizando (2.21) a fim de evitar erros durante a implementação devido a variação da escala entre as variáveis.

Os parâmetros do modelo SVM escolhidos foram aqueles que apresentaram um melhor desempenho durante a fase de treinamento. A função *kernel* utilizada foi a ‘linear’ devido ao comportamento linear da corrente de fuga em relação ao nível de poluição. Os parâmetros  $C$  e  $\gamma$  (*gamma*) foram determinados a partir do processo de validação cruzada com  $k$  partições utilizando apenas os dados de treinamento e considerando  $k$  igual a 5. Foram escolhidas faixas de valores contendo potências de base dez para os valores de  $C$  e  $\gamma$ , normalmente utilizados em problemas de classificação. A média da precisão obtida após o processo de validação cruzada para cada um dos parâmetros testados foi utilizada como indicador para a escolha dos mesmos. Com isso, os parâmetros que apresentaram melhores resultados após o processo de validação foram  $C = 1000$  e  $\gamma = 1$ .

Após a fase de treinamento o modelo foi testado, utilizando o conjunto de dados reservado. Com isso, foi analisado, o comportamento do modelo quando apresentado novos dados, com o objetivo de verificar a capacidade de generalização.

### 4.3.2 Modelo para Reconhecimento de Padrões Utilizando Rede SOM

Para implementação da Rede SOM, foram utilizadas as bibliotecas *MiniSOM* e *SOMPY* do *Python*. As características da topologia da rede neural desenvolvida são: bidimensional com vizinhança hexagonal, a função de vizinhança escolhida foi a *gaussiana*, a função de ativação utilizada a distância euclidiana e a taxa de aprendizagem  $\eta = 0,5$ . Durante o treinamento foi

utilizado 80% do banco de dados e os dados de entrada foram normalizados utilizando (2.21). A determinação da dimensão da grade neural se deu por meio de mecanismos para avaliação da qualidade de classificação como o erro de quantização vetorial ( $E_q$ ) e a Taxa de Classificação Correta (CCR, do inglês, *Correct Classification Rate*). O procedimento para determinar a melhor dimensão do mapa envolveu avaliar a evolução dos parâmetros CCR e  $E_q$  para diferentes dimensões. A dimensão escolhida foi aquela em que o valor de CCR se estabeleceu e o erro de quantização atingiu uma região de patamar média. Esse procedimento para determinação do mapa neural foi adaptado de Lira (2012).

Com a arquitetura definida, a rede SOM é treinada e os dados são agrupados em *clusters* definindo assim cinco classes cada uma correspondente a um nível de poluição. Novos dados são apresentados à rede, com o objetivo de verificar a capacidade de generalização da mesma, a rede neural calcula a distância entre os novos dados de entrada e o neurônio vencedor pertencente a determinada classe, a menor distância indica a qual nível de poluição os novos dados de entrada pertencem.

### 4.3.3 Modelo de Classificação Utilizando *Naive Bayes*

O modelo baseado em aprendizado *Bayesiano*, *Naive Bayes*, foi implementado utilizando a biblioteca *scikit-learn* do *Python*. O banco de dados foi separado em duas partes, uma para fase de treinamento que corresponde a 80% dos dados, a outra parte, 20%, para fase de teste. Os dados de entrada foram normalizados utilizando (2.21) a fim de evitar erros durante o treinamento do modelo devido a variação da escala entre as variáveis.

Durante a fase de treinamento os dados de entrada reservados foram utilizados para treinar o algoritmo *Naive Bayes* para classificar cinco níveis de poluição dos isoladores. A acurácia e o respectivo intervalo de confiança foram determinados utilizando o processo de validação cruzada com  $k$  partições, considerando  $k = 5$ . Não se faz necessário a determinação de parâmetros para implementação do algoritmo *Naive Bayes* simples.

A fase de teste foi realizada após o treinamento do modelo, utilizando os dados de entrada reservados para teste. O modelo teve sua capacidade de generalização analisado quando apresentado novos dados a partir da acurácia e sensibilidade.

## 5 Resultados e Análises

A aplicação das técnicas utilizadas neste trabalho para classificação dos níveis de poluição de isoladores foram avaliadas utilizando os indicadores apresentados no capítulo 2. Os indicadores foram utilizados com a finalidade de avaliar o grau de confiabilidade dos modelos.

Espera-se que os valores de acurácia, sensibilidade, precisão, CCR e  $f1 - score$  sejam maiores possíveis tendo cuidado de evitar o sobreajuste dos modelos. A matriz de confusão em conjunto com a precisão e sensibilidade foram utilizados com a finalidade de se obter uma análise mais detalhada para as diferentes classes. O erro de quantização usado para avaliação do modelo da rede SOM deve ser o menor possível sem que haja sobreajuste.

Após a implementação dos modelos de classificação utilizando os valores RMS, pico e componente fundamental da corrente fuga é possível determinar o nível de poluição para novos isoladores e assim gerenciar manutenções preventivas destes equipamentos. Para implementação das três técnicas foi utilizado a mesma base de dados composta por valor de pico, valor RMS e componente harmônica fundamental da corrente de fuga para cinco diferentes níveis. Os cinco níveis são: limpo, muito leve, leve, moderada e pesada.

Nas próximas seções serão apresentados os resultados obtidos por meio dos três modelos implementados e a comparação entre os indicadores de desempenho de cada uma das técnicas.

### 5.1 Modelo SVM

O desempenho do modelo baseado na aplicação de SVM apresentou uma acurácia de 97,31% após o treinamento da rede. Esse valor é utilizado como uma média para avaliação do modelo quando apresentado novos dados. O intervalo de confiança considerado, meio importante para indicar o grau de incerteza de um cálculo, foi de 99%, com isso o valor da acurácia deve estar entre 91,65% e 99,99%. A precisão e sensibilidade (*recall*) para os diferentes níveis de poluição se encontra na Tabela 10.

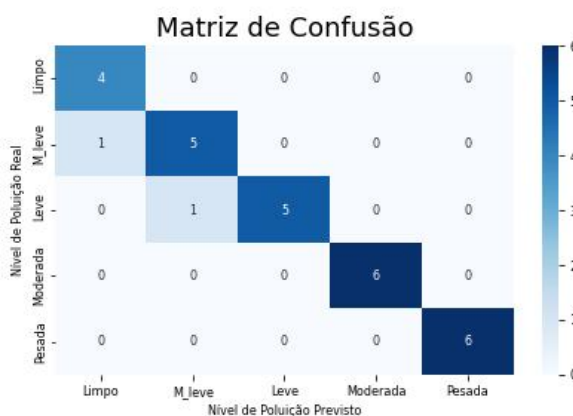
Tabela 10 – Valor da precisão e sensibilidade para os diferentes níveis de poluição - SVM.

Nível de Poluição	Precisão (%)	Recall (%)
Limpo	83,33	93,75
Muito Leve	91,30	87,50
Leve	100,0	91,67
Moderado	96,0	100,0
Pesado	100,0	100,0

Fonte: Aatoria Própria (2022).

Verifica-se na Tabela 10 que após o treinamento todos os níveis apresentaram valores maiores que 80% tanto para a sensibilidade quanto para a precisão, chegando a apresentar valores de sensibilidade máximo nos casos moderado e pesado. O modelo apresentou uma menor precisão na classe limpo, e menor sensibilidade na classe muito leve. Alguns fatores podem influenciar nos valores de sensibilidade e precisão do modelo como o desbalanceamento dos dados em relação a classe de isoladores limpos, a baixa variância dos parâmetros da corrente de fuga para os casos limpo, muito leve e leve inclusive faixas muito próximas do índice DDSE, o que pode levar a erros na definição das fronteiras de decisão dos hiperplanos. Na Figura 20 é apresentado a Matriz de Confusão do modelo para um novo conjunto de dados, em que temos a relação entre os dados reais e os dados previstos pelo modelo.

Figura 20 – Matriz de confusão - modelo SVM considerando dados de teste.



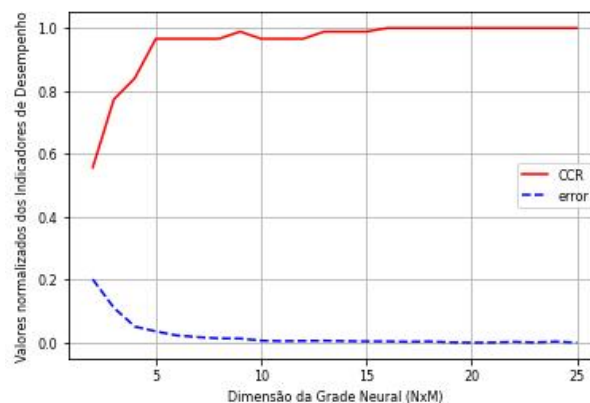
Fonte: Aatoria Própria (2022).

Verifica-se que para o conjunto de dados de teste a acurácia foi de 92,86%, ou seja, esta dentro da faixa do intervalo de confiança. A sensibilidade apresentada para os casos muito leve e leve foi de 83,33%. O nível de poluição pesada e moderada apresentou uma boa precisão e sensibilidade tanto na fase de treinamento quanto na fase de validação. É importante destacar o bom desempenho do modelo em classificar os casos de poluição mais grave.

## 5.2 Modelo - Rede SOM

Os resultados obtidos pelo modelo utilizando Mapas Auto-Organizáveis foram adquiridos a partir de análises do CCR e do erro de quantização. Para determinação da melhor dimensão da grade neural a ser utilizada foi analisado a região de patamar média do erro e a estabilização da curva da taxa de classificação correta a partir da Figura 21.

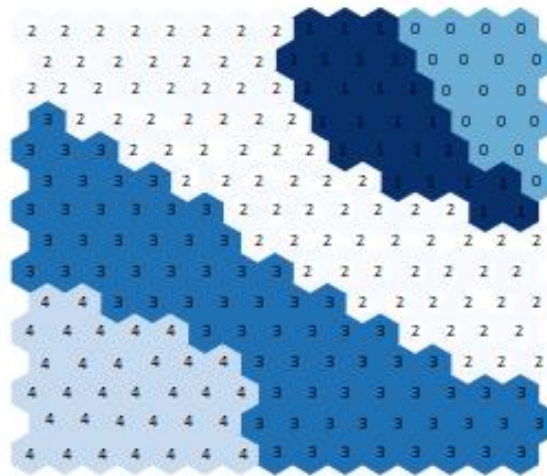
Figura 21 – Evolução dos parâmetros de avaliação do modelo.



Fonte: Autoria Própria (2022).

Verifica-se que o  $E_q$  varia de valores próximos a 0,40 para dimensões menores até valores próximos a 0,015 para grades de dimensões maiores. Desta forma o erro de quantização apresenta uma redução média de 96%. A taxa de acerto partiu de valores próximos a 65% até valores próximos a 100%, casos em que o sistema apresenta sobreajuste. A dimensão do classificador utilizado foi de 15x15, o valor do CCR neste caso foi de 99,11%, valor próximo a estabilização da curva. A redução do erro para essa dimensão foi cerca de 95%. Ao apresentar novos dados para realização do reconhecimento dos padrões por parte da rede neural constatou-se um valor de CCR igual a 96,43%. A Figura 22 apresenta um mapa topológico, esse mapa apresenta a distribuição dos padrões relacionados ao nível de poluição presente na superfície de isoladores.

Figura 22 – Mapa topológico do classificador do nível de poluição dos isoladores.



Legenda: 0-Limpo, 1-Muito Leve, 2-Leve, 3-Moderado, 4-Pesado

Fonte: Autoria Própria (2022).

Verifica-se que cada estado de poluição ativa regiões diferentes do mapa, calculando a distância de cada amostra em relação ao neurônio vencedor, verificando a qual *cluster* a amostra pertence e comparando com os rótulos obtidos durante a construção do banco de dados foi possível obter indicadores de desempenho para os dados de treinamento, assim verifica-se que os níveis leve e moderado apresentaram uma taxa de classificação menor, cerca de 96,00%, em relação a classe limpo e o caso mais grave para o sistema que é a poluição pesada que apresentaram uma taxa de acerto de 100%. A matriz de confusão não foi utilizada para avaliação do treinamento da Rede SOM devido à natureza não supervisionada do algoritmo. Entretanto é possível adaptar o algoritmo a fim de solucionar de forma satisfatória problemas de classificação como apresentado neste trabalho.

### 5.3 Modelo - *Naive Bayes*

A acurácia média do modelo *Naive Bayes* após o processo de treinamento foi de 72,33% com um intervalo de confiança de 99%, o valor da acurácia deve estar entre 58,97% e 85,70%. Na Tabela 11 é apresentado a precisão e sensibilidade para os diferentes níveis de poluição.

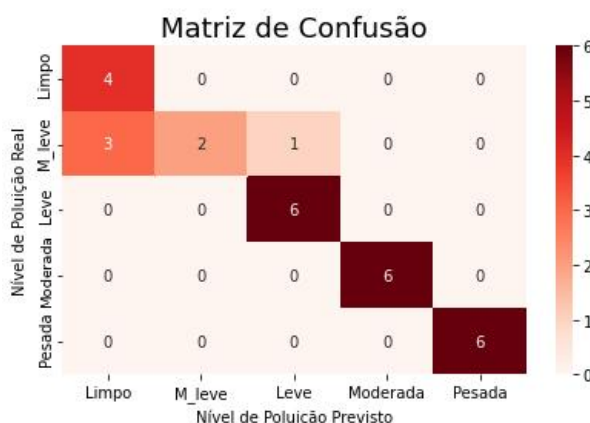
Tabela 11 – Valor da precisão e sensibilidade para os diferentes níveis de poluição - *Naive Bayes*.

Nível de Poluição	Precisão (%)	Recall (%)
Limpo	68,18	93,75
Muito Leve	88,89	33,33
Leve	59,26	66,67
Moderado	73,33	91,67
Pesado	100,0	100,0

Fonte: Aatoria Própria (2022).

Os valores de sensibilidade apresentados na Tabela 11 indicam uma boa classificação para os níveis de poluição pesado, moderado e limpo, com sensibilidade maior que 90%. Os casos muito leve e leve apresentaram um erro maior durante o treinamento. Na Figura 23 é apresentado a matriz de confusão do modelo para novos dados. Na matriz tem-se a relação entre os dados reais e os dados previstos pelo modelo.

Figura 23 – Matriz de confusão - modelo *Naive Bayes* considerando dados de teste.



Fonte: Aatoria Própria (2022).

A acurácia para os novos dados foi de 85,71%, o modelo apresentou um baixo desempenho apenas para o caso muito leve, sensibilidade de 33,33%. O modelo apresentou um bom desempenho para classificação dos casos mais graves tanto para os dados de treinamento quanto para os novos dados.

## 5.4 Análise Comparativa entre os Modelos de Classificação

Os resultados obtidos pelos três modelos de classificação tiveram seus desempenhos comparados utilizando a acurácia média e o *f1 – score* para cada uma das cinco classes definidas.

A Tabela 12 apresenta um resumo desses indicadores para cada um dos modelos.

Tabela 12 – Comparação dos indicadores de desempenho dos modelos de classificação.

Modelo	F1-score (%)					Acurácia (%)
	Limpo	Muito Leve	Leve	Moderado	Pesado	
SVM	88,89	83,33	90,91	100,0	100,0	92,86
Rede SOM	100,0	92,31	100,0	90,91	100,0	96,43
Naive Bayes	72,73	50,00	92,31	100,0	100,0	85,71

Fonte: Autoria Própria (2022).

O  $f1 - score$  apresenta uma média entre a precisão e sensibilidade, assim pode-se ter uma perspectiva mais geral do desempenho dos modelos. Verifica-se pela Tabela 12 que, considerando a acurácia o modelo que apresentou um maior desempenho foi aquele em que se utilizou aprendizado não supervisionado, Mapas Auto-Organizáveis de *Kohonen* e o que apresentou menor desempenho foi o algoritmo *Naive Bayes*. Entretanto, ao analisar o  $f1 - score$  individual para cada uma das classes, verificou-se que tanto o modelo utilizando *Naive Bayes* e SVM apresentaram um desempenho melhor na classificação do nível moderado em relação a Rede SOM. A capacidade de classificar de forma mais precisa o nível de poluição moderado pode ser considerado uma boa vantagem, principalmente para programação de manutenções preventivas em locais em que o aumento da poluição na superfície de isoladores se dá de forma rápida.

Comparando o modelo SVM e *Naive Bayes* verificou-se que o primeiro apresenta melhor desempenho em todos os níveis exceto no nível leve, mas não é apresentado uma diferença significativa. Por fim tem-se que os três modelos apresentaram um ótimo desempenho ao classificar o caso mais grave de poluição que é o nível pesado. Com isso, o modelo SOM se sobressai em relação aos outros pois além de apresentar um bom desempenho individual na maioria dos casos, têm a vantagem da aprendizagem não supervisionada, pois em aplicações práticas nem sempre se tem a presença de um especialista para instruir o modelo.



## 6 Conclusão

Este trabalho, cumpre com o objetivo principal proposto, ao criar um modelo de classificação capaz de determinar o grau de poluição de isoladores por meio de medições da corrente de fuga. Verificou-se a partir dos indicadores de desempenho como: acurácia e  $f1 - score$  que, os modelos utilizados baseados em inteligência artificial apresentaram resultados satisfatórios em determinar os casos mais críticos para o funcionamento do SEP, quando os isoladores apresentam um nível de poluição alto ou muito alto. Com destaque para a técnica que se utiliza de aprendizado não supervisionado, Rede SOM, que apresentou um bom desempenho no reconhecimento de novos dados para a maioria dos níveis de poluição e apresenta a vantagem de não necessitar de um “professor” para o reconhecimento de novos padrões.

Além disso, os objetivos específicos também foram atingidos, com o levantamento do estado da arte, do uso da corrente de fuga para determinar o estado de poluição de isoladores, com foco na utilização de modelos de classificação. Verificando que não há um consenso na literatura de quais parâmetros estão mais correlacionados com o nível de poluição na superfície de isoladores. Portanto, os parâmetros da corrente de fuga foram analisados e foi determinado a partir do coeficiente de correlação de *Spearman* quais estão mais relacionados com o nível de poluição. A partir das análises dos resultados obtidos pelos modelos e o estudo dos indicadores de desempenho, verifica-se a importância do uso de técnicas baseadas em IA para o diagnóstico de isoladores devido ao bom nível de assertividade e capacidade de generalização.

Portanto, é possível por meio de medições da corrente de fuga em campo alimentar um sistema de diagnóstico baseado em técnicas de inteligência artificial, capaz de auxiliar tomadores de decisão e contribuir para o gerenciamento de ativos, especificamente isoladores, do sistema elétrico de potência elevando a confiabilidade do sistema.

### 6.1 Trabalhos Futuros

O estudo de aplicações de inteligência artificial relacionadas ao gerenciamento de ativos do sistema elétrico de potência é amplo. Algumas sugestões para melhorias e trabalhos futuros são:

- 
- Verificar se o banco de dados construído e utilizado durante o aprendizado dos modelos é estatisticamente significativo a fim de comprovar a capacidade de generalização e aplicabilidade dos modelos.
  - Analisar e validar o modelo baseado nos Mapas Auto-organizáveis de *Kohonen* utilizando dados não rotulados, respeitando o princípio de aprendizado não supervisionado do algoritmo.

# Referências

- ABNT. *ABNT NBR 10621: Isoladores utilizados em sistemas de alta tensão em corrente alternada — ensaios de poluição artificial*. 2. ed. Rio de Janeiro, 2005. Citado na página 20.
- AMIN, M.; AMIN, S.; ALI, M. Monitoring of leakage current for composite insulators and electrical devices. *Advanced Study Center Co. Ltd.*, p. 75–89, 2009. Citado na página 18.
- ASAN, U.; ERCAN, S. An introduction to self-organizing maps. In: \_\_\_\_\_. [S.l.: s.n.], 2012. p. 299–319. ISBN 978-94-91216-76-3. Citado 3 vezes nas páginas 25, 26 e 27.
- BEN-HUR, A.; WESTON, J. A user’s guide to support vector machines. *Methods in molecular biology (Clifton, N.J.)*, v. 609, p. 223–39, 01 2010. Citado na página 24.
- BRUCE, P.; BRUCE, A. *Practical Statistics for Data Scientists: 50 Essential Concepts*. 1. ed. [S.l.]: O’Reilly Media, 2017. Citado 2 vezes nas páginas 21 e 30.
- CHAPRA, S. C.; CANALE, R. P. *Métodos Numéricos para Engenharia*. 5. ed. Porto Alegre: AMGH Editora Ltda, 2011. Citado na página 47.
- DIAS, A. L. *Diagnóstico de anomalias em aplicações de acionamento de motores elétricos a partir de dados de processo de rede PROFINET e aprendizagem de máquinas*. Tese (Doutorado) — Escola de Engenharia de São Carlos da Universidade de São Paulo, 2019. Citado 2 vezes nas páginas 22 e 23.
- DIAS, B. A. *Simulação computacional da poluição biológica em isoladores de vidro*. Dissertação (Mestrado) — Universidade Federal de Campina Grande, 2017. Citado na página 18.
- DIAS, B. A. *Gestão Técnica de Isoladores Poliméricos Utilizando Aprendizado de Máquina*. Tese (Doutorado) — Universidade Federal de Campina Grande, 2021. Citado na página 14.
- FARIA, L. T. de. *Sistema Inteligente Híbrido Intercomunicativo para Detecção de Perdas Comerciais*. Dissertação (Mestrado) — Universidade Estadual Paulista, 2012. Citado na página 26.
- FERREIRA, L. F. P. *Curso de Isoladores: Processo de Fabricação de Isoladores de Vidro Temperado*. [S.l.], 2017. Citado na página 17.
- GERMANO, A. D. *Análise de Degradação de Isolamentos Poliméricos*. Dissertação (Mestrado) — Universidade Federal de Campina Grande, 2010. Citado 2 vezes nas páginas 14 e 18.
- GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow*. [S.l.]: O’Reilly Media, 2019. v. 2. Citado 6 vezes nas páginas 21, 23, 24, 32, 34 e 35.
- GORUR, R. S.; CHERNEY, E. A.; BURNHAM, J. T. *Outdoor Insulators*. [S.l.]: Ravi S. Gorur, Inc., Phoenix, Arizona 85044, USA, 1999. Citado 2 vezes nas páginas 17 e 19.
- GUIMARÃES, P. R. B. *Métodos Quantitativos Estatísticos*. 1. ed. Curitiba: IESDE Brasil S.A, 2007. Citado na página 31.

HAYKIN, S. *Neural Networks and Learning Machines*. 3. ed. Ontario, Canada: Pearson, 2009. Citado 4 vezes nas páginas 21, 24, 25 e 26.

IEC. *Selection and dimensioning of high-voltage insulators intended for use in polluted conditions – Part 1: Definitions, information and general principles*. [S.l.], 2008. (IEC/TS 60815-1). Citado 2 vezes nas páginas 20 e 41.

IEC. *Artificial pollution tests on high-voltage ceramic and glass insulators to be used on a.c. systems*. [S.l.], 2013. (IEC 6057). Citado 2 vezes nas páginas 41 e 42.

IEEE. *High-Voltage test techniques - Partial discharge measurements*. [S.l.], 2000. (IEC 60.270). Citado na página 18.

KHAFAN, N. A.; EL-HAG, A. Bayesian regularization of neural network to predict leakage current in a salt fog environment. *IEEE Transactions on Dielectrics and Electrical Insulation*, v. 25, n. 2, p. 686–693, 2018. Citado na página 39.

KHALED, A.; EL-HAG, A.; ASSALEH, K. Equivalent salt deposit density prediction of outdoor polymer insulators during salt fog test. In: *2016 IEEE Conference on Electrical Insulation and Dielectric Phenomena (CEIDP)*. [S.l.: s.n.], 2016. p. 786–789. Citado na página 38.

KOHONEN, T. The self-organizing map. *Proceedings of the IEEE*, v. 78, n. 9, p. 1464–1480, 1990. Citado na página 25.

LI, J.; SUN, C.; SIMA, W.; YANG, Q.; HU, J. Contamination level prediction of insulators based on the characteristics of leakage current. *IEEE Transactions on Power Delivery*, v. 25, n. 1, p. 417–424, 2010. Citado 3 vezes nas páginas 14, 19 e 36.

LIRA, G. R. S. de. *Monitoramento de Para-raios de Óxido de Zinco com Base na Medição da Corrente de Fuga Total*. Tese (Doutorado) — Universidade Federal de Campina Grande, 2012. Citado 3 vezes nas páginas 27, 35 e 49.

MAMEDE FILHO, J. *Manual de Equipamentos Elétricos*. Rio de Janeiro: LTC, 2013. v. 4. Citado 2 vezes nas páginas 14 e 18.

MARIANO, D. Métricas de avaliação em machine learning: acurácia, sensibilidade, precisão, especificidade e f-score. In: \_\_\_\_\_. [S.l.: s.n.], 2021. ISBN 9786599275326. Citado 2 vezes nas páginas 34 e 35.

MARSLAND, S. *Machine Learning - An Algorithmic Perspective*. 2. ed. [S.l.]: Chapman and Hall/CRC, 2014. Citado 4 vezes nas páginas 21, 25, 34 e 35.

MONTGOMERY, D. C.; RUNGER, G. C. *Applied Statistics and Probability for Engineers*. 4. ed. [S.l.]: LTC, 2008. Citado 4 vezes nas páginas 29, 30, 31 e 32.

MUNIRAJ, C.; CHANDRASEKAR, S. Adaptive neurofuzzy inference system-based pollution severity prediction of polymeric insulators in power transmission lines. *Advances in Artificial Neural Systems*, v. 2011, 08 2011. Citado na página 37.

PADILHA, T. P. P. *Investigação de Algoritmos de Aprendizado de Máquina Pertencentes ao Paradigma Estatístico para Aquisição de Conhecimento*. Dissertação (Mestrado) — Universidade de São Paulo - São Carlos, 1999. Citado 2 vezes nas páginas 28 e 29.

- PYLARINOS, D.; THEOFILATOS, K.; SIDERAKIS, K.; THALASSINAKIS, E.; VITELLAS, I.; ALEXANDRIDIS, A. T.; PYRGIOTI, E. Investigation and classification of field leakage current waveforms. *IEEE Transactions on Dielectrics and Electrical Insulation*, v. 19, n. 6, p. 2111–2118, 2012. Citado na página 37.
- ROCHA, P. H. V. da. *Estimação da Poluição de Isoladores em Linhas de Transmissão a partir da Irradiação de RF*. Tese (Doutorado) — Universidade Federal de Campina Grande, 2019. Citado 2 vezes nas páginas 14 e 18.
- RUSSEL, S.; NORVIG, P. *Inteligência Artificial*. 3. ed. Rio de Janeiro, Brasil: Elsevier, 2013. Citado 5 vezes nas páginas 15, 22, 23, 29 e 33.
- SANTOS, C. da Silva dos. *Construção de atributos binários baseada em análise de interações*. Tese (Doutorado) — Universidade de São Paulo, 2010. Citado 2 vezes nas páginas 28 e 29.
- SCHWARDT, W. H. *Insulator Pollution Monitoring Device: Development, Calibration and Field Evaluation*. Dissertação (Mestrado) — University of Stellenbosch, 2005. Citado 2 vezes nas páginas 19 e 20.
- SILVA, I. N. da; SPATTI, D. H.; FLAUZINO, R. A. *Redes Neurais Artificiais para engenharia e ciências aplicadas*. São Paulo, Brasil: Artliber Editora Ltda, 2010. Citado 4 vezes nas páginas 21, 25, 32 e 33.
- SOUZA, B. A. de. *Monitoramento e Diagnóstico de Isoladores Poliméricos Utilizando Processamento Digital de Imagens Termográficas*. Dissertação (Mestrado) — Universidade Federal de Campina Grande, 2016. Citado na página 18.
- STEFENON, S. F.; AMERICO, J. P.; MEYER, L. H.; GREBOGI, R. B.; NIED, A. Analysis of the electric field in porcelain pin-type insulators via finite elements software. *IEEE Latin America Transactions*, v. 16, n. 10, p. 2505–2512, 2018. Citado na página 14.
- STEFENON, S. F.; SILVA, M. C.; BERTOL, D.; MEYER, L.; NIED, A. Fault diagnosis of insulators from ultrasound detection using neural networks. *Journal of Intelligent and Fuzzy Systems*, v. 37, p. 6655–6664, 12 2019. Citado 2 vezes nas páginas 14 e 18.
- SUDA, T. Frequency characteristics of leakage current waveforms of an artificially polluted suspension insulator. *IEEE Transactions on Dielectrics and Electrical Insulation*, v. 8, n. 4, p. 705–709, 2001. Citado na página 36.
- WANKE, B. dos S. L.; COSTA, V. O.; PINA, A. C. de; FILHO, A. C. de P. Aplicação do classificador naive bayes para identificação de falhas de um manipulador robótico. *ABCMSymposium Series in Mechatronics*, v. 6, p. 888–895, 2014. Citado 2 vezes nas páginas 28 e 29.
- YANG, Y.; WANG, L.; WANG, Y.; MEI, X. Insulator self-shattering detection: a deep convolutional neural network approach. *Springer Science+Business Media, LLC, part of Springer Nature*, 2018. Citado 2 vezes nas páginas 14 e 18.
- ZHAO, S.; JIANG, X.; ZHANG, Z.; HU, J.; SHU, L. Flashover voltage prediction of composite insulators based on the characteristics of leakage current. *IEEE Transactions on Power Delivery*, v. 28, n. 3, p. 1699–1708, 2013. Citado na página 38.

# A Apêndice - Resumo Banco de Dados

Graus de Poluição	DDSE (mg/cm <sub>2</sub> )	Cond. (mS/m)	$I_{rms}$ ( $\mu$ A)	$I_{pico}$ ( $\mu$ A)	$I_{cc}$ ( $\mu$ A)	$I_{DHT}$ (%)	$I_{t_1}$ ( $\mu$ A)	$I_{t_3}$ ( $\mu$ A)	$I_{t_5}$ ( $\mu$ A)	$I_{t_7}$ ( $\mu$ A)	$I_{t_9}$ ( $\mu$ A)	$I_{t_{11}}$ ( $\mu$ A)
Limpo	$5,10 \times 10^{-4}$	0,21	144	205	1,89	10,59	203	6,36	20,4	2,50	0,46	0,49
Limpo	$8,64 \times 10^{-4}$	0,36	151	232	1,16	9,26	213	6,23	18,6	1,77	0,51	0,70
Muito Leve	$2,16 \times 10^{-3}$	0,98	148	233	2,45	9,62	208	8,27	18,1	0,57	0,40	0,14
Muito Leve	$3,45 \times 10^{-3}$	1,44	158	241	2,00	10,88	223	5,62	23,5	2,70	0,26	0,22
Muito Leve	$4,60 \times 10^{-3}$	1,90	164	249	0,63	9,13	230	7,28	19,5	13,8	0,80	0,78
Leve	$8,04 \times 10^{-3}$	3,35	162	231	2,09	8,31	228	5,19	18,1	2,49	0,45	0,45
Leve	$1,53 \times 10^{-2}$	6,38	164	243	0,74	9,09	230	6,27	19,7	2,32	0,82	0,67
Leve	$1,90 \times 10^{-2}$	8,21	172	253	0,20	7,40	243	6,26	16,5	2,49	0,46	0,49
Moderado	$3,82 \times 10^{-2}$	15,91	174	250	0,47	7,38	244	7,49	16,0	2,99	0,78	0,54
Moderado	$6,22 \times 10^{-2}$	25,90	179	251	1,27	7,80	253	6,25	18,6	1,68	0,47	0,64
Moderado	$9,90 \times 10^{-2}$	41,50	182	245	1,79	7,84	259	5,30	19,3	1,66	0,48	0,58
Pesado	$1,69 \times 10^{-1}$	70,60	187	264	2,69	8,61	263	6,52	21,6	2,04	0,43	0,25
Pesado	$3,15 \times 10^{-1}$	131,20	205	287	1,24	7,29	289	9,48	18,5	2,95	0,58	0,68
Pesado	$3,80 \times 10^{-1}$	160,50	206	291	0,12	7,85	291	7,79	21,2	2,35	0,83	0,97

## B Apêndice - Matriz de Correlação de Spearman

	Cond.	DDSE	$I_{pico}$	$I_{rms}$	$I_{cc}$	$I_{1har}$	$I_{3har}$	$I_{5har}$	$I_{7har}$	$I_{9har}$	$I_{11har}$	$I_{THD}$
Cond.	1	1	0,874	0,981	-0,273	0,981	0,228	0,040	0,218	0,212	0,238	-0,771
DDSE	1	1	0,874	0,981	-0,273	0,981	0,228	0,040	0,218	0,212	0,238	-0,771
$I_{pico}$	0,874	0,874	1	0,913	-0,430	0,912	0,431	0,064	0,198	0,245	0,332	-0,674
$I_{rms}$	0,981	0,981	0,913	1	-0,328	0,999	0,254	0,055	0,198	0,236	0,301	-0,769
$I_{cc}$	-0,273	-0,273	-0,430	-0,328	1	-0,330	-0,280	0,257	-0,235	-0,568	-0,649	0,462
$I_{1har}$	0,981	0,981	0,912	0,999	-0,330	1	0,249	0,054	0,196	0,229	0,300	-0,768
$I_{3har}$	0,228	0,228	0,431	0,254	-0,280	0,249	1	-0,106	0,046	0,388	0,223	-0,213
$I_{5har}$	0,040	0,040	0,065	0,055	0,257	0,054	-0,106	1	-0,056	-0,047	0,001	0,520
$I_{7har}$	0,218	0,218	0,198	0,198	-0,235	0,196	0,046	-0,056	1	0,026	0,009	-0,349
$I_{9har}$	0,211	0,211	0,245	0,236	-0,568	0,229	0,388	-0,047	0,026	1	0,594	-0,237
$I_{11har}$	0,238	0,238	0,332	0,301	-0,649	0,300	0,223	0,001	0,009	0,594	1	-0,285
$I_{THD}$	-0,771	-0,771	-0,674	-0,769	0,462	-0,768	-0,213	0,520	-0,349	-0,237	-0,285	1