UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS

BRUNO AUGUSTO TREVISAM

Aplicação de ferramentas inteligentes de classificação de dados não estruturados como suporte a gestão de ativos em sistemas elétricos de potência

São Carlos
2023

BRUNO AUGUSTO TREVISAM

Aplicação de ferramentas inteligentes de classificação de dados não estruturados como suporte a gestão de ativos em sistemas elétricos de potência

Dissertação de mestrado apresentada à Escola de Engenharia de São Carlos da Universidade de São Paulo, como requisito para a obtenção do Título de Mestre em Ciências, Programa de Engenharia Elétrica, Área de Concentração em Sistemas Elétricos de Potência.

Orientador: Prof. Dr. Rogério Andrade Flauzino

São Carlos
2023

FOLHA DE JULGAMENTO

Candidato: Bacharel **BRUNO AUGUSTO TREVISAM.**

Título da dissertação: "Aplicação de ferramentas inteligentes de classificação de dados não estruturados como suporte a gestão de ativos em sistemas elétricos de potência".

Data da defesa: 24/02/2023.

Comissão Julgadora	**Resultado**
Prof. Associado **Rogério Andrade Flauzino (Orientador)** (Escola de Engenharia de São Carlos – EESC/USP)	Aprovado
Prof. Dr. **Eduardo Coelho Marques da Costa** (Escola Politécnica/EP-USP)	Aprovado
Prof. Dr. **Pablo Torrez Caballero** (Universidade Federal do Acre/UFAC)	Aprovado

Coordenador do Programa de Pós-Graduação em Engenharia Elétrica:
Prof. Associado **João Bosco Augusto London Junior**

Presidente da Comissão de Pós-Graduação:
Prof. Titular **Murilo Araujo Romero**

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Trevisam, Bruno Augusto
T814a Aplicação de ferramentas inteligentes de classificação de dados não estruturados como suporte a gestão de ativos em sistemas elétricos de potência / Bruno Augusto Trevisam; orientador Rogério Andrade Flauzino. São Carlos, 2023.

Dissertação (Mestrado) - Programa de Pós-Graduação em Engenharia Elétrica e Área de Concentração em Sistemas Elétricos de Potência -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2023.

1. Gestão de Ativos. 2. Árvore de Decisão. 3. Rede Bayesiana. 4. Máquinas de Vetores de Suporte. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

DEDICATÓRIA

Aos meus pais por tudo o que eu sou hoje e pela constante dedicação para o alcance dos meus sonhos.

A minha esposa pela compreensão, carinho e apoio incansável.

AGRADECIMENTOS

A Deus pela vida e por guiar toda minha trajetória.

Ao Dr. Rogério Andrade Flauzino pela disponibilidade e apoio para realização desse trabalho, bem como por toda contribuição em meu crescimento cientifico.

Ao Mestre Narco Afonso Ravazzoli Maciejewski pelo pronto atendimento e colaboração em apoio as buscas de informações contidas no trabalho.

A USP e EESC pelo acesso a infraestrutura e informação.

A toda equipe da ISA CTEEP, especialmente da Análise da Manutenção, pela oportunidade oferecida e aprendizados constantes.

A todos os meus familiares e amigos pelo incentivo e compreensão em momentos de dificuldades. Principalmente meus pais e esposa que sempre me apoiaram e estiveram ao meu lado.

EPÍGRAFE

"*In God We Trust, All Others Must Bring Data*"

William Edwards Deming (1988)

RESUMO

TREVISAM, Bruno Augusto. **Aplicação de ferramentas inteligentes de classificação de dados não estruturados como suporte a gestão de ativos em sistemas elétricos de potência**. 2023. Dissertação de Mestrado em Sistema Elétrico de Potência – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2023.

Este trabalho apresenta a aplicação de ferramentas inteligentes com a finalidade de tratar e melhorar a qualidade de dados não estruturados. A aplicação é realizada no contexto de gestão de ativos do Sistema Elétrico de Potência, com ênfase no transformador de potência. A abordagem nesse seguimento de negócio visa solucionar lacunas de informações que possam subsidiar tomadas de decisões mais eficazes, além de aumentar a visibilidade e integração na gestão de riscos em ativos. Pelo SIN – Sistema Interligado Nacional ser regulado, qualquer iniciativa em prol de reduzir falhas ou manutenções não planejadas possui relevância na confiabilidade e disponibilidade de seus ativos. Na mineração de dados, destaca-se a metodologia KDD - *Knowledge Discovery in Database* com etapas pré-estabelecidas sequenciais para desenvolvimento do trabalho, abordando desde a compreensão inicial dos objetivos propostos, bem como a escolha dos métodos mais aderentes a cada aplicação e também avaliação sobre os resultados obtidos. Em suma, verifica-se a aplicabilidade das metodologias de Árvore de Decisão, Rede Bayesiana e Máquinas de Vetores de Suporte na finalidade de classificação de textos oriundos dos registros de histórico de manutenções em transformadores de potência.

Palavras-chave: Gestão de Ativos, Árvore de Decisão, Rede Bayesiana, Máquinas de Vetores de Suporte.

ABSTRACT

TREVISAM, Bruno Augusto. ***Application of intelligent tools for classification of unstructured data to support asset management in electrical power systems*.** 2023. Master's Dissertation in Electrical Power System – São Carlos School of Engineering, University of São Paulo, 2023.

This work presents the application of intelligent tools in order to treat and improve the quality of unstructured data. The application is carried out in the context of asset management of the Electric Power System, with an emphasis on the power transformer. The approach in this business segment aims to solve information gaps that can support more effective decision-making, in addition to increasing visibility and integration in asset risk management. As SIN - National Interconnected System is regulated, any initiative to reduce failures or unplanned maintenance has relevance in the reliability and availability of its assets. In data mining, the KDD - Knowledge Discovery in Database methodology stands out with pre-established sequential steps for the development of the work, addressing from the initial understanding of the proposed objectives, as well as the choice of the most adherent methods to each application and also evaluation about the results obtained. In summary, the applicability of the Decision Tree, Bayesian Network and Support Vector Machines methodologies is applied in order to classify texts from the history of maintenance records in power transformers.

Keywords: Asset Management, Decision Tree. Bayesian Network. Support Vector Machines.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANEEL	-	Agência Nacional de Energia Elétrica
ANSI	-	*American National Standards Institute*
FT	-	Função Transmissão
IEEE	-	*Institute of Electrical and Electronics Engineers*
LTC	-	*Load Tap Changer*
ONS	-	Operador Nacional do Sistema Elétrico
PV	-	Parcela Variável
RAP	-	Receita Anual Permitida
REN	-	Resolução Normativa
SEP	-	Sistema Elétrico de Potência
SIN	-	Sistema Interligado Nacional
KDD	-	*Knowledge Discovery in Database*
NLP	-	*Natural Language Processing*

SUMÁRIO

1 INTRODUÇÃO

No atual cenário mundial, é cada vez mais comum deparar-se com a integração humana e a virtual na busca de soluções que atendam atividades ou rotinas que exigem muito esforço e demandam grande tempo para serem executadas. As ferramentas inteligentes então, ganham muito espaço nos mais diversos setores, tornando-os mais competitivos no que tange ao seu posicionamento de mercado, devido maior disponibilidade de tempo dedicado em análises estratégicas para uma melhor tomada de decisão, e consequentemente agregando mais valor ao negócio.

Em termos de manutenção, em um passado próximo, as decisões eram tomadas através de fatos e evidencias em sua maioria física, obtidos por uma investigação presencial que se demandava muito tempo e dependia também do nível de conhecimento do profissional sobre o assunto em questão. Tanto a subjetividade quanto a padronização no procedimento realizado nesse método geram uma incerteza no resultado devido a não ser algo sistemático.

Atualmente, com todo o avanço tecnológico e facilidade na obtenção de dados proporcionada, as análises e diagnósticos são baseados em informações provenientes diretamente do ativo. A atuação do fator humano então passa a ser mais voltada ao perfil analítico do que operacional e assim soluções integradas de ferramentas inteligentes vem se tornando essenciais na busca de excelência operacional.

1.1 Gestão de ativos do setor elétrico brasileiro

O Brasil é um país de proporções continentais, o SIN – Sistema Interligado Nacional é constituído por quatro subsistemas: Sul, Sudeste/Centro-Oeste, Nordeste e a maior parte da região Norte. Tanto a coordenação e controle das operações de geração e transmissão de energia elétrica é realizada pelo ONS – Operador Nacional do Sistema Elétrico, bem como também o planejamento da operação dos sistemas isolados do país. O órgão responsável pela fiscalização e regulação é a ANEEL – Agência Nacional de Energia Elétrica [1].

Como o SIN é bastante complexo, em termos de magnitude pela alta quantidade de ativos envolvidos de diferentes agentes, as intervenções ao sistema

devem ser coordenadas com sinergia. De modo que todas as partes afetadas sejam atendidas e nenhuma delas prejudicada. Então, o processo de manutenção dos ativos do setor elétrico segue especificações geridas pela ANEEL que disponibiliza através de REN - Resoluções Normativas as regras direcionadoras aplicadas aos agentes.

Em termos de manutenção preventiva e preditiva, a REN 669 de 2015 [2] regulamenta os Requisitos Mínimos de Manutenção e o monitoramento da manutenção de instalações de transmissão de Rede Básica, isto é, ativos com nível de tensão superior a 230kV. Em 2019, teve alterações realizadas pela REN 853 [3], e posteriormente em 2020 também pela REN 906 [4], conforme *Tabela 1*.

Tabela 1: Resumo das atividades de manutenção

Atividade	Equipamento	Periodicidades máximas (meses)	Tolerância (meses)
Inspeções Termográficas	Equipamentos de Subestações	6	1
Análise de gases dissolvidos no óleo isolante	Transformadores de Potência ou Autotransformadores	6	1
	Reatores de Potência		
Ensaio físico-químico do óleo isolante	Transformadores de Potência ou Autotransformadores	24	4
	Reatores de Potência		
Manutenção Preventiva Periódica	Transformadores de Potência ou Autotransformadores	72	12
	Reatores de Potência		
	Disjuntores		
	Chave Seccionadora		
	Chave de Alta Velocidade		
	Medidores de Tensão e Corrente em CCAT		
	Transformadores para Instrumento		
	Para-raios		
Manutenção Preventiva Periódica	Banco de Capacitores Paralelos	36	6
Manutenção Preventiva Periódica	Filtros	48	8
Manutenção Preventiva Periódica	Válvulas	24	4
Inspeção de Rotina	Linha de Transmissão	12	2

Fonte: REN 906 [4]

No que se refere as transmissoras, nem sempre as intervenções ao sistema são realizadas de maneira planejada, a indisponibilidade portanto é contabilizada para

o agente, resultando em um desconto de PV – Parcela Variável da respectiva RAP – Receita Anual Permitida associada a FT – Função Transmissão. O cálculo é realizado de acordo com as diretrizes da REN 729 de 2016 [5], e complementos e revisões conforme as REN 782 de 2017, REN 853 de 2019 [6] [7] e REN 906 de 2020 [4]. Na *Tabela 2* são elencados os critérios aplicados de acordo com o tipo de FT envolvida no evento, onde utiliza-se fatores distintos, Ko – Indisponibilidade de Urgência e Kp – Indisponibilidade Programada.

Tabela 2 - Padrão de Frequência de Outros Desligamentos e Fatores Ko e Kp

FT	Família de FT		Padrão de Frequência de Outros Desligamentos (desl./ano)	Fator Ko	Fator Kp
MG	(*)		não possui	150	10
LT	≤ 5km(*)		1	150	10
	>5km e ≤50Km(*)		1		
	>50km - 230kV		3		
	345kV		2		
	440kV		2		
	500kV		2		
	750kV		3		
	Cabo Isolado(*)		não possui	50	2,5
	CCAT(*)		3	50	10
TR	Trifásico (*)		1	50	5,0
	≤345kV		1	150	10
	>345kV		1		
CR	REA	≤345kV	1	150	10
		>345kV	1		
	CRE	(*)	3	150	7,5
	CSI	(*)	3	50	2,5
	BC	(*)	3	100	5,0
	CSE	(*)	3	150	7,5

Fonte: REN 906 [4]

Legenda:

(*) Qualquer nível de tensão

Ko - Fator multiplicador para Outros Desligamentos

Kp - Fator multiplicador para Desligamento Programado

LT - Linha de Transmissão

TR - Transformação

CR - Controle de Reativo
REA - Reator
CRE - Compensador Estático
CSI - Compensador Síncrono
BC - Banco de Capacitor
CSE - Compensação Série
CCAT - Corrente Contínua em Alta Tensão

É notório a diferença de valor de desconto da RAP em casos onde o evento é caracterizado de maneira urgente ao invés de planejada. Isso reitera a ideia de excelência operacional pelos agentes, estimulando-os na busca por novas práticas ou soluções preventivas e preditivas que mitiguem o risco de desligamentos indesejados. Desse modo, quanto mais informação a respeito do ativo que possibilite a tomada da melhor decisão é extremamente bem vinda e benéfica, pois possibilita aos agentes realizar predições e diagnósticos antevendo uma possível falha catastrófica.

1.2 Objetivo

O propósito do trabalho é realizar a aplicação de ferramentas inteligentes com finalidade de melhorar a qualidade de banco de dados de registros de manutenções e atividades em transformadores de potência. A utilização de um banco de dados não estruturados demanda em uma investigação analítica a fim de identificar as técnicas mais aderentes ao modelo, que possui várias etapas de tratamento até que se atinge um nível aceitável no qual se possa subsidiar em alguma informação significativa.

A utilização desses tipos de dados não estruturados são muito dificultosas e demandam muito tempo de análise para se conseguir algo que seja realmente significativo e confiável, visto que a variabilidade dos dados implica em um aumento de erro de categorizações e classificação. A possibilidade de utilizar essas informações em tomadas de decisões mais assertivas, principalmente no que se refere ao histórico do ativo, agrega muito no âmbito de ampliar a visão que se tem sobre o problema ou evento em si.

1.3 Organização

O trabalho de pesquisa da dissertação de mestrado foi estruturado em seções, e organizados de tal forma que denota um sequenciamento lógico para o tema estudado. Na seção 1, a Introdução do trabalho conta com subseção dedicada a abordar o tema de gestão de ativos no âmbito dos SEP, citando modelos de regras aplicados aos agentes de transmissão e suas particularidades. Na mesma seção de Introdução, encontra-se também subseções referentes objetivo e organização do trabalho.

A Seção 2 aborda os principais aspectos do transformador de potência, com estratificação de seus componentes em sistemas dedicados a cada funcionalidade do conjunto que o compõe.

Na seção 3 do trabalho são elencados aspectos introdutórios de tratamento de dados não estruturados, onde conta com subseção dedicada ao tema de mineração de dados no âmbito de como esse tema é incorporado em um processo geral de nove etapas de descoberta de conhecimento em banco de dados. Inclui-se também subseções relacionadas aos principais classificadores de texto com possibilidade de aplicação em grande volume de dados.

A metodologia aplicada no projeto é relatada na seção 4, onde há o detalhamento das atividades realizadas envolvendo as fases do KDD - *Knowledge Discovery in Database*. Em cada uma das subseções são evidenciadas as etapas de do KDD, iniciando no entendimento do objetivo e criação do conjunto de dados até a obtenção dos resultados e avaliações após as aplicações dos diferentes algoritmos de classificação.

Na seção 5 serão apresentadas as considerações finais a respeito deste trabalho, as conclusões e os ganhos desta pesquisa para gestão de ativos de transformadores de potência.

2 PRINCIPAIS ASPECTOS DO TRANSFORMADOR DE POTÊNCIA

Um transformador é definido pela ANSI/IEEE como um dispositivo elétrico estático, que não envolve nenhuma peça em movimento contínuo, utilizado em sistemas de energia elétrica para transferir energia entre circuitos através do uso de indução eletromagnética [8]. É um equipamento muito relevante no SEP - Sistema Elétrico de Potência, uma vez que tem responsabilidade da transformação de tensão, ora como função de elevação, utilizados em aplicações de subestações elevadoras de geração, ora como função de rebaixamento, muito comumente aplicados em sistemas de transmissão e distribuição.

Tem como objetivo principal a conversão de diferentes níveis de tensão, entre o circuito primário, e os circuitos secundário e terciário, onde tem-se mantida a mesma frequência e diferem-se as tensões e correntes. É caracterizado por não possuir partes internas moveis, onde a transferência de energia ocorre pela indução eletromagnética entre um circuito e outro [9].

A construção de um transformador depende da aplicação. Os transformadores destinados ao uso interno são principalmente do tipo seco, porém também podem ser imersos em líquido. Para uso ao ar livre, os transformadores geralmente são imersos em líquido [8], conforme ilustrado na Figura *1* [9].

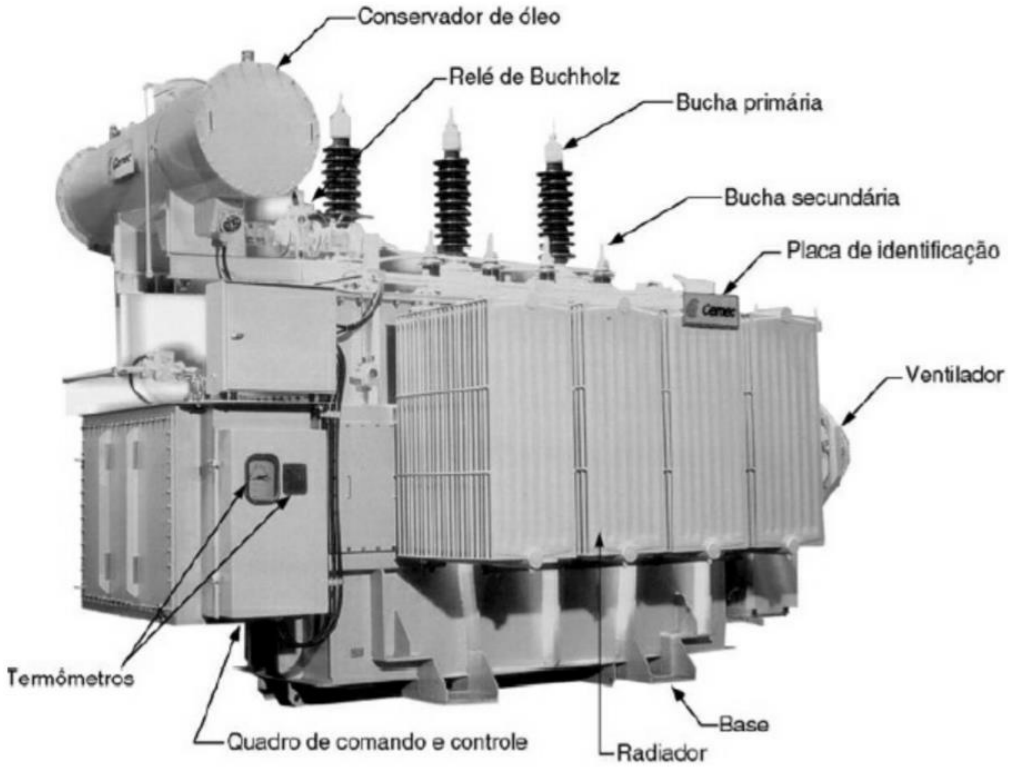


Figura 1: Transformador trifásico com conservador de óleo *[9]*

A composição de um transformador de potência imerso em liquido isolante se dá não somente pelo próprio transformador, como também por equipamentos associados. Devido à alta quantidade de componentes envolvidos, geralmente são associados a sistemas funcionais, nos quais são concentrados os componentes envolvidos respectivamente [10].

2.1 Sistema de Comutação

É definido como o conjunto de dispositivos eletromecânicos e eletrônicos com função de alterar as relações de tensões do transformador, e podem ser caracterizados por dois tipos, comutadores de derivações em carga e os comutadores de derivações sem tensão [10];

Figura 2: Comutador de derivação em carga *[9]*

2.2 Sistema de Proteção

São conjuntos de três tipos de proteção utilizados para detectar falhas, evitando danos extensos e e/ou preservar a estabilidade e a qualidade do sistema de energia. Normalmente, são definidos pela proteção sobrecorrente para correntes de fase, proteção diferencial para correntes diferenciais e acumulador de gás ou proteção de aumento de pressão para falhas de arco [8].

Figura 3: Relé Buchholz

2.3 Sistema de Conexão

É constituído pelo conjunto de componentes responsáveis pela conexão dos cabos ou barramentos elétricos, como por exemplo a bucha, que ao mesmo tempo se conecta e elementos externos ao transformador, mas mantem a estanqueidade do equipamento, bem como sua isolação.

2.4 Sistema de Resfriamento

Tem em sua composição componente ou elementos que auxiliam na refrigeração do transformador, como os radiadores. Esses que atuam na dissipação térmica do óleo, pela troca de calor com o meio externo através das aletas que possuem maior contato com o ar. Normalmente, o sistema de resfriamento conta também com ventilação forçada, onde se obtém maior desempenho e melhor eficiência na refrigeração do transformador.

2.5 Sistema de Estrutural

É constituído principalmente pelo tanque principal, uma vez que é responsável por abrigar a parte ativa do transformador como núcleo e o óleo isolante, possuindo conexões aos radiadores para circulação dele com meio externo. Deve ser um componente extremamente robusto a fim de suportar esforços mecânicos e altas pressões de acordo com a respectiva especificação de potência aplicada.

2.6 Sistema de Conservação do Óleo

É o conjunto de componentes dedicados a receber o óleo do tanque quando este se expande, devido aos efeitos do aquecimento por perdas internas. O conservador de óleo é tido como requisito para o uso do relé Buchholz, necessário para detecção de pequenas falhas internas [9].

2.7 Sistema Ativo

Constitui-se pelos componentes elétricos de alta tensão, como núcleo magnético e isolamento dielétrico do transformador, ao qual é destinado na transformação de tensão e corrente de alta tensão [10].

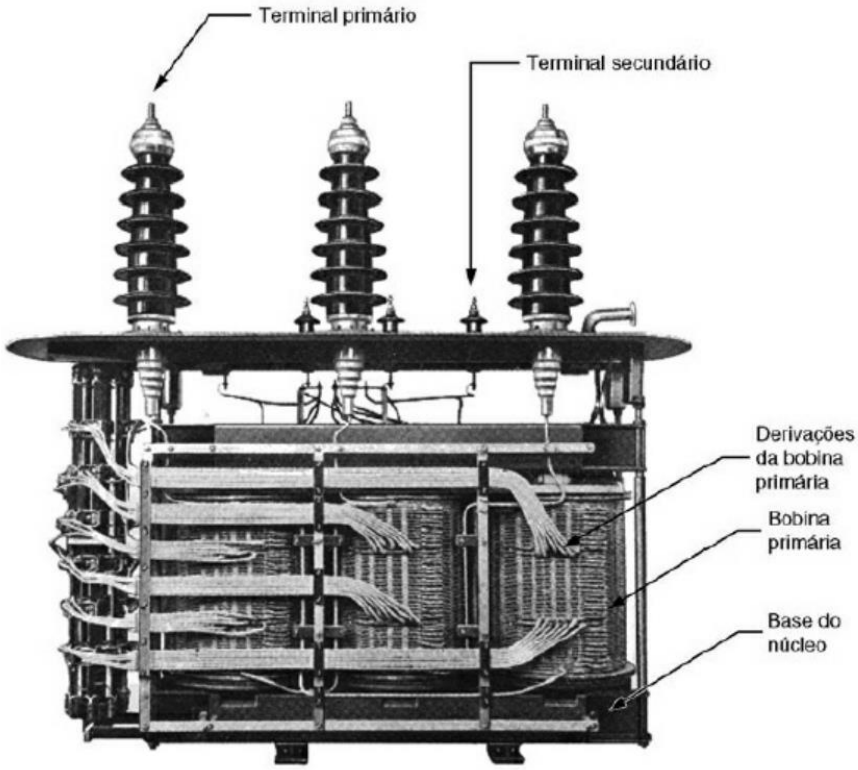


Figura 4: Vista do núcleo de um transformador trifásico *[9]*

2.8 Sistema de Controle, Supervisão e Monitoramento

É constituído pelo conjunto de dispositivos relacionados a estados e grandezas do transformador. São componentes de extrema importância no que tange ao regime de operação, logo que é através de periféricos como termômetro de óleo, indicador de nível de óleo, válvula de alívio de pressão ou relé de súbita pressão que atuação de controle pode ser realizadas no objetivo de evitar uma falha catastrófica. Talvez, atualmente esse sistema seja um dos que mais evoluiu ao longo do tempo, com aplicações de monitoramento online para transformadores modernos com sensores instalados.

3 ASPECTOS INTRODUTÓRIOS DE TRATAMENTO DE DADOS NÃO ESTRUTURADOS

Os ativos são tidos como máquinas, edifícios, veículos, as informações associadas e sistemas de *software* que são utilizados para servir uma função empresarial ou organizacional [11]. Desse modo, os dados não deixam de ser ativos, pois são através deles que se geram informações. Ao longo da evolução tecnológica ao qual o planeta vem passando, principalmente nos últimos anos, onde essa transformação está se dando mais rápida, os dados estão ganhando cada vez mais espaço. Muito disso se deve pela facilidade de se obter diagnósticos mais rápidos [12].

Por tratar-se de um ativo de grande importância no SEP, o transformador de potência tem alto nível de controle e gestão por parte dos agentes, como em atividades de manutenções, nas quais este equipamento é envolvido, sejam elas preventivas, corretivas ou preditivas. Na busca do prolongamento do ciclo de vida do transformador, há priorização por soluções preventivas e preditivas, logo que os benefícios quantificados pelo menor custo de reparo, diminuição do risco de falha e aumento de desempenho possibilitem assegurar uma maior confiabilidade ao sistema elétrico.

A avaliação relacionada ao estado atual do transformador pode ser obtida através do índice de saúde [13]. Para obtenção deste índice, são consideradas algumas variáveis, como: degradação do papel isolante, geração de gases combustíveis, taxa de falha, etc. De maneira geral, é tomada a maior quantidade de dados históricos, como registros de manutenções, resultados de ensaios físico-químicos, cromatográficos, elétricos, etc. E nem sempre esses dados estão estruturados, principalmente pelo transformador de potência possuir registros bastante antigos que comumente não eram realizados de maneira sistemática, isto é, em banco de dados como em sistemas digitais atuais.

A base histórica de dados é de grande valia para avaliações e diagnósticos de um determinado ativo, por exemplo na estimativa de condição de degradação do transformador de potência e estimativa do índice de isolamento com base no histórico de dados de óleo [14]. Nesse artigo, é enfatizada a urgência em se estabelecer um arranjo adequado de análise histórica de óleo de transformador de potência desde de

quando teve sua entrada em operação pela primeira vez até a data atual. Desse modo, é proporcionada uma identificação que auxilia a equipe de gestão de ativos a tomar uma ação de manutenção oportuna.

3.1 Mineração de dados

No que se refere a análise de banco de dados, é inevitável não abordar o termo de Data Mining ou Mineração de Dados na tradução livre, que se trata do processo de mudança, através de grandes bancos de dados em busca de padrões interessantes e anteriormente desconhecidos, segundo [15]. Sua grande importância e necessidade se devem a abundancia e acessibilidade proporcionada aos dados atualmente. Data Mining faz parte de um processo geral de *Knowledge Discovery in Database* – KDD, ou Descoberta de Conhecimento em Banco de Dados, na tradução livre, e pode ser considerado como o passo central de todo o processo.

O processo de KDD foi definido por Fayaad et al [16] como “o processo não trivial de identificar validos, novos, potencialmente úteis e, em última análise, compreensíveis padrões em dados”. Basicamente, todo o processo conta com nove passos, conforme ilustrado na Figura *5*.

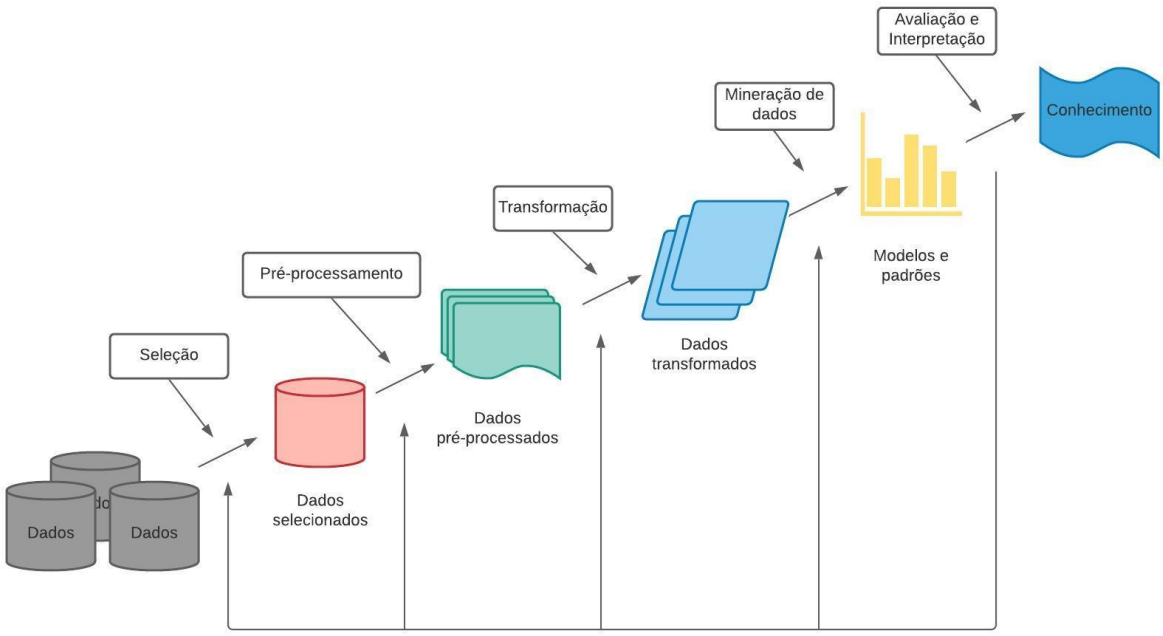


Figura 5: Representação do Processo KDD adaptado de *[16]*.

1. Compreensão de domínio e objetivos do KDD trata-se da etapa preparatória inicial que visa entender o que deve ser feito com as muitas

decisões relacionadas a transformação ou algoritmos, por exemplo. É nesse passo também que se necessita entender os objetivos do trabalho. No decorrer no processo, podem haver revisões e ajustes desta etapa.

2. Criar um conjunto de dados no qual a descoberta será realizada. Com os objetivos já definidos, deve-se determinar quais dados serão utilizados para a descoberta do conhecimento. É nesse passo que se verifica a disponibilidade dos dados, bem como a necessidade de inclusão de dados adicionais, integrando-os em um conjunto de dados com seus respectivos atributos.
3. Pré-processamento e limpeza. É a etapa que melhora a confiabilidade dos dados, pois inclui limpeza de dados, remoção de ruídos ou outliers e tratamento de valores ausentes. Podendo envolver métodos estatísticos complexos ou até mesmo algoritmo de mineração de dados específicos.
4. Transformação de dados. Trata-se da fase onde ocorre a preparação e desenvolvimento dos melhores dados para mineração. Tem como possibilidades de utilização a redução de dimensão, como seleção de recursos e extração, assim como amostragem de registro. Ou então, a transformação de atributos, como discretizar atributos numéricos e transformação funcional.
5. Escolha da tarefa de mineração de dados apropriada. Nesse passo é onde ocorre a decisão de qual tarefa se demonstra mais aderente ao proposito, por exemplo: classificação, regressão ou agrupamento. Há dois principais objetivos na mineração de dados: previsão e descrição, sendo que a previsão se refere muitas vezes a mineração de dados supervisionada, em contrapartida a mineração de dados descritiva inclui classificação não supervisionada e aspectos de visualização de mineração de dados.
6. Escolha do algoritmo de mineração de dados. É nessa etapa que ocorre a seleção do método especifico a ser utilizado para pesquisar padrões. Por exemplo, ao comparar precisão com compressibilidade, o primeiro é melhor com redes neurais, logo o segundo é melhor com árvores de decisão.
7. Empregando o algoritmo de mineração de dados. Nessa etapa, é onde se aplica o algoritmo várias vezes até que se possa obter um resultado satisfatório. Nesse momento também, podem ocorrer ajustes de controle

de parâmetros do algoritmo, bem como números de neurônios de uma rede neural.

8. Avaliação. Fase onde ocorre a interpretação de padrões com relação aos objetivos definidos na primeira etapa. Tem como finalidade a compreensão verificação de utilidade do modelo induzido. É também o momento onde se registram as descobertas de conhecimento para uso posterior.
9. Usando a descoberta de conhecimento. É o momento onde o conhecimento se torna dinâmico, com implicações em alterações no sistema com possibilidade de medir seus efeitos. As estruturas de dados podem mudar conforme tornam-se indisponíveis e o domínio de dados pode ser modificado, por exemplo um atributo ter um valor que não assumido antes.

3.2 Classificadores de texto

A aplicação de regras fixas em larga escala de maneira manual é possível, porém depende de alto esforço e tempo. Características estas que são extremamente importantes nos dias atuais onde se buscam o máximo de eficiência. Com a classificação de texto, torna-se fácil a classificação de artigos em diferentes categorias, como esportes, política ou mercado de ações. Este processo pode ser realizado com apoio de NLP – *Natural Language Processing*, processamento de linguagem natural na tradução livre, através de aprendizagem de máquina que proporciona a possibilidade de viabilizar a classificação e ordenação de um grande volume de dados.

Há inúmeras maneiras de classificação textuais, dentre as quais pode-se citar:

- Árvore de decisão
- Rede Bayesiana
- Máquina de Vetores de Suporte

3.2.1 Árvore de decisão

A representação por meio de regras mapeadas em Árvores de Decisão tem como objetivo a criação de um modelo viável que irá prever o valor de uma variável de destino com base no conjunto de variáveis de entrada [17]. Na mineração de dados, uma árvore de decisão é um modelo preditivo que pode ser utilizado para representar

classificadores e modelos de regressão, sendo também chamada de árvores de classificação e árvores de regressão, respectivamente, de acordo com o propósito de aplicação [15].

O funcionamento desta técnica ocorre de maneira estruturada hierarquicamente através de conjuntos de nós interconectados. Onde os atributos de entrada são testados nos nós internos, de modo que se tome uma decisão determinando qual será o nó descendente, já a classificação das instâncias é realizada de acordo com seu respectivo rótulo associado [18].

Entre as principais características tidas como vantagem e que viabilizam o método, tem-se as seguintes [18]:

- Precisão: habilidade do modelo para avaliar ou prever corretamente classes, agrupamentos, regras;
- Velocidade: uma vez construída uma árvore de decisão, seu uso é imediato e sua execução é computacionalmente muito rápida;
- Robustez: habilidade do modelo para avaliar ou prever corretamente, utilizando dados ruidosos ou com valores ausentes;
- Escalabilidade: capacidade de construir modelos eficientemente a partir de grandes volumes de dados;
- Interpretabilidade: alto nível de compreensão fornecido pelo modelo;
- Flexibilidade: o espaço das instâncias é particionado em subespaços e cada subespaço é adaptado a diferentes modelos.

Além das vantagens supracitadas, nesse método pode-se citar algumas limitações, como por exemplo a criação de modelos excessivamente complexos, dependendo dos dados apresentados no conjunto de treinamento. Para evitar que o algoritmo de aprendizado de máquina superajuste (*Over-fitting*) os dados, é importante revisar os dados de treinamento e podar os valores para categorias e assim produza um modelo mais refinado e melhor ajustado. Outro ponto é de que alguns dos conceitos da árvore de decisão podem ser difíceis de aprender porque o modelo não pode expressá-los facilmente. Essa deficiência às vezes resulta em um modelo maior do que o normal. Desse modo, pode ser necessário alterar o modelo ou olhar para métodos diferentes de aprendizado de máquina [17].

Há diversos algoritmos desenvolvidos para aplicação das Árvores de Decisão e classificação, dentre os quais os mais relevantes são CART [19] de Breiman *et al.*, o ID3 [20] e C4.5 [21], de Quinlan:

- CART: é uma técnica aplicada tanto pra árvores de classificação (atributo nominal) quanto pra árvores de regressão (atributo contínuo) [22]. Caracteriza-se pelo fato de construir árvores binárias, ou seja, cada nó interno tem exatamente duas arestas de saída. As divisões são selecionadas usando os dois critérios e a árvore obtida é podada por meio da redução do fator custo-complexidade [15]. Permite também além do tratamento diferenciado para atributos ordenados, a possibilidade de combinações lineares entre atributos (agrupamento de valores em vários conjuntos) [22].
- ID3: é considerado um algoritmo de árvore de decisão muito simples. Utiliza o ganho de informação como critério de divisão, o algoritmo para de crescer quando todas as instâncias pertencem a um único valor de um recurso de destino ou quando o melhor ganho de informação não é maior que zero. ID3 não aplica nenhum procedimento de poda nem lida com atributos numéricos ou valores ausentes. Tem como principal vantagem sua simplicidade, no entanto, possui várias desvantagens, como ser necessário a conversão de dados contínuos devido ser projetado para atributos nominais, medidas de contorno devem ser aplicadas para evitar superajuste (*over-fitting)*, escolhendo as menores árvores em relação as maiores, mesmo que esse algoritmo produza árvores pequenas, porém não a menor árvores possível [15].
- C4.5: é uma evolução do ID3 e também usa a taxa de ganho como critério de divisão. A divisão cessa quando o número de instâncias a serem divididas está abaixo de um certo limite. Remoção baseada em erros é realizada após a fase de crescimento. C4.5 pode lidar com atributos numéricos. Também pode induzir a partir de um conjunto de treinamento que incorpora valores ausentes usando critérios de razão de ganho corrigidos. Dentre os aprimoramentos em relação ao ID3, o C4.5 usa um procedimento de poda que remove ramos que não contribuir para a precisão e substituí-los por nós folha, também permite que os valores dos atributos estejam ausentes, e lida com atributos contínuos dividindo o valor

do atributo variando em dois subconjuntos (divisão binária). Especificamente, ele procura o melhor limite que maximiza o critério de razão de ganho. Todos os valores acima do limite constituem o primeiro subconjunto e todos os outros valores constituem o segundo subconjunto [15].

3.2.2 Rede Bayesiana

A rede Bayesiana como o próprio nome faz referência, leva como um de seus principais fatores, a aplicação do teorema de Bayes. O teorema de Bayes é representado de acordo como a seguinte equação.

$$p(c_j|d) = \frac{p(d|c_j)p(c_j)}{p(d)}$$

Onde:

$p(c_j|d)$ representa a probabilidade da instância *d* estar na classe c_j,

$p(d|c_j)$ indica a probabilidade de gerar instância *d* dada a classe c_j.

$p(c_j)$ representa a probabilidade de ocorrência da classe c_j.

A $p(d)$ é a probabilidade de ocorrência da instância *d.*

O teorema de Bayes pode ser definido como uma suposição de independência condicional dentre todas as variáveis $A_1,...A_n$ em uma determinada categoria *C*, conforme ilustrado na Figura *6***.**

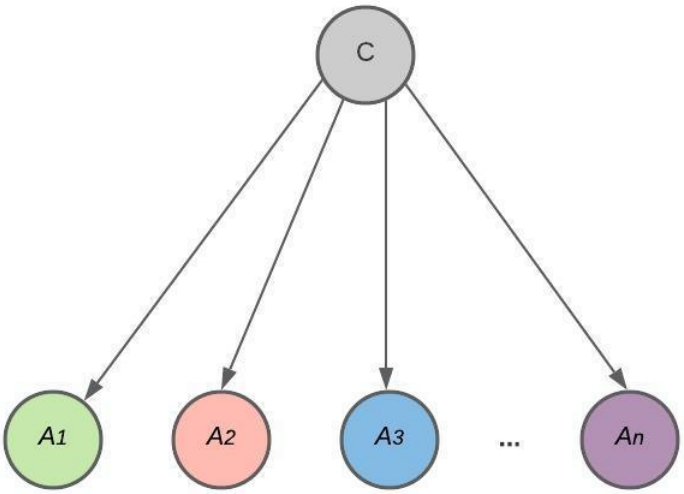


Figura 6: Exemplo típico de estrutura de rede Bayesiana, adaptado de Pavlenko e Chernyak *[23]*

A rede Bayesiana, também conhecida como modelos gráficos direcionados acíclicos, é a técnica que combina a teoria da probabilidade com a teoria dos grafos. Com base em um conjunto de variáveis ou parâmetros, é possível prever resultados com base em probabilidades. Essas variáveis estão conectadas de tal maneira que o valor resultante de uma variável irá influenciar a probabilidade de saída de outro, daí o uso de nós em rede [17].

Pavlenko e Chernyak [23] definem uma rede bayesiana como um grafo acíclico dirigido que codifica a distribuição de probabilidade conjunta sobre um conjunto de variáveis aleatórias $X= \{X_1...X_d\}$. Formalmente, podendo ser definida uma rede bayesiana para X pelo par $<G,P>$. Onde G representa um grafo acíclico dirigido cujos nós correspondem as variáveis aleatórias $X= \{X_1...X_d\}$ e P é representado pela fórmula: $P = \{P(x_1|\Pi_{[1]}), ..., P(x_d|\Pi_{[d]})\}$, indica o conjunto de distribuições de probabilidade condicionais de d, dado o conjunto de nós pais desses vértices, como: $\Pi_{[i]}$, para cada $Xi, i = 1, ..., d$.

A aplicação de Naive Bayes demonstra algumas vantagens, como possuir um tempo curto para treinamento, devido a varredura única, obtendo dessa maneira uma rápida classificação. Também não se mostra com sensibilidade a recursos irrelevantes, e por fim, lida bem com dados reais e discretos, assim como dados de *streaming*. Outro ponto bastante relevante é o retorno da probabilidade sobre o modelo, algo importante para analisar a confiança no algoritmo em questão.

3.2.3 Máquina de Vetores de Suporte

Support Vector Machines (SVMs), na tradução livre, Máquinas de Vetores de Suporte, são tidas como um conjunto de aprendizagem supervisionada relacionada a métodos de classificação e regressão. Faz uso da teoria de aprendizagem de máquina para maximizar a precisão preditiva enquanto evita automaticamente o ajuste excessivo dos dados [24].

Os fundamentos das Máquinas de Vetores de Suporte foram introduzidos por Vapnik em 1995, na primeira edição de seu livro, ocasião na qual foi trazida para discussão o tema se contrapondo aos métodos clássicos de estatística, onde para se controlar o desempenho, diminui-se a dimensionalidade de um espaço de recursos, logo o SVM aumenta drasticamente a dimensionalidade e depende do chamado de fator de margem grande [25].

O SVM tem como objetivo principal a segregação dos dados. Desse modo, à medida que a segregação é realizada, a mínima distância entre os pontos é denominada de margem, conforme ilustrado na Figura 7. Assim se busca a máxima segregação dos pontos de dados mais próximos, isto é, a máxima margem.

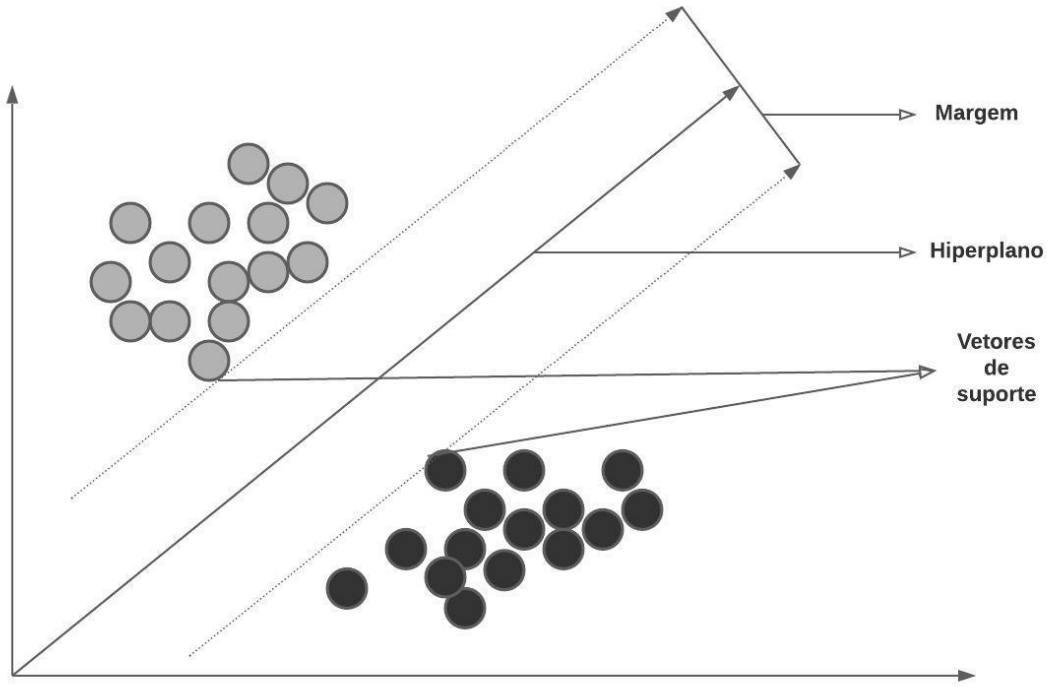


Figura 7: Exemplo de SVM linear. Elaboração própria.

Nem sempre os hiperplanos são a melhor solução para segregação de pontos de dados, uma vez que não estão distribuídos linearmente. Para esses casos, se utiliza um truque de Kernel, que transforma a entrada de dados em um espaço de dimensão superior. Dentre os tipos comumente utilizados em SVM, pode-se citar o kernel de função base radial que tem a possibilidade de mapeamento espaço em dimensões infinitas, ilustrado na equação seguinte [26].

$$K(x, y) = e^{-||x-y||^2/2\sigma^2}$$

Outro método bastante popular para modelos não lineares é o Kernel polinomial que possui a capacidade de distinção de espaço de entrada curvo ou não linear, expresso pela equação a seguir [26].

$$K(x, y) = (\mathrm{x} \cdot \mathrm{y} + 1)^{\mathrm{p}}$$

A utilização da metodologia do SVM possui vantagens por se tratar de uma ferramenta útil para análise de insolvência, por exemplo quando os dados não são regularmente distribuídos ou tem uma distribuição desconhecida, desse modo se torna eficaz em espaços dimensionais elevados. Além disso, os SVMs oferecem uma solução única, uma vez que o problema de otimalidade é convexo. Esta é uma vantagem comparada para Redes Neurais, que têm várias soluções associadas a mínimos locais e, por este motivo, podem não ser robusto em diferentes amostras [27].

A aplicação de SVM inclui a necessidade de uma boa função Kernel [26], e isso pode ser entendido como desvantagem. Outro ponto que se pode citar, comum de técnicas não paramétricas, como SVMs, é a falta de transparência nos resultados, principalmente no que tange a dados financeiros, utilizados no estudo de Deutsche Bundesbank como alternativa para classificação de empresas de Auria e Moro em 2008 [27].

4 METODOLOGIA

De maneira a sistematizar o processo metodológico nesse trabalho, foi utilizado KDD - *Knowledge Discovery in Database* como referência, e deste modo, foram seguidas as etapas conforme definição do item 3.1. Na sequência, há o detalhamento das atividades realizadas em cada fase do KDD voltadas para o trabalho proposto.

4.1 Compreensão de domínio e objetivos do KDD

A primeira fase da metodologia corresponde ao entendimento do negócio na tradução livre. É nessa etapa onde se inclui a determinação dos objetivos, baseando-se na situação atual e então se desenvolve um plano de projeto.

A principal finalidade deste trabalho atua na busca de soluções em ferramentas inteligentes para classificação de dados não estruturados na gestão de ativos do setor elétrico. Diante da alta quantidade de ativos que o compõem, foi priorizado o transformador de potência, visto que se trata de um equipamento de grande importância ao sistema elétrico, além de também ser um dos tipos de equipamentos que mais impactam os agentes em caso de indisponibilidades indesejadas.

Como fonte de dados para o estudo, foi escolhida a base de registros de manutenção em transformadores de potência de uma agente de Transmissão de Energia Elétrica ao longo de 10 anos, tendo início em janeiro de 2009. Nesse banco de dados há inúmeras informações relacionadas ao problema e/ou defeito relatado. Dentre os dados mais relevantes, pode-se citar além de detalhes como modelo e fabricante do respectivo ativo transformador de potência, como também o componente envolvido, os modos de falha e as ações realizadas para reparar o problema e/ou defeito.

4.2 Criação de conjunto de dados

No entendimento dos dados é onde se incluem a coleta inicial de dados, exploração, descrição, bem como também a constatação da qualidade dos dados.

Para apoio na análise dos dados foi utilizada o Python, linguagem de programação criada por Guido van Rossum e publicada em 1991, que incorpora vários módulos, pacotes e bibliotecas. A linguagem traz uma grande variedade de *libraries*, possibilitando a utilização das mais variadas aplicações, como *web development*, computação numérica, desenvolvimento de jogos e *softwares*, *machine* e *deep learning* [28].

A linguagem Python demanda a utilização de uma IDE - *Integrated Development Enviroment*, ambientes computacionais para desenvolvimento e compilação de códigos. Dentre as várias IDE's para utilização com o Python, foi escolhida para o desenvolvimento do trabalho Jupyter Notebook, devido a ter como uma de suas principais vantagens, a compilação em partes, gerando assim resultados preliminares, e consequentemente facilitando a interpretação, algo de extrema relevância pois minimiza o risco de erros no código e melhora a interatividade.

Os dados utilizados para esse trabalho são relacionados a registros de manutenção, totalizando 6192 eventos caracterizados em 40 colunas. Para a análise dos dados crus desse banco de dados, foi utilizada a biblioteca pandas-profiling. De início, já se constata uma baixa qualidade de dados, como também falta de homogeneidade e uma quantidade de células faltantes de 5,5%.

Na realização de uma análise geral sobre o banco de dados, é possível perceber campos que agregam maior valor ao objetivo do trabalho quando comparados a outros. Por exemplo, o campo descrição que possui preenchimento com texto livre, sendo um campo com dados de alta cardinalidade com 78,8% de valores distintos. Nesse caso em especifico, se constata variações textuais que implicam no mesmo significado, possibilitando sua categorização, conforme Figura 8.

Common Values | Overview

Value	Count	Frequency (%)
AA.Vazamento de óleo isolante	42	0.7%
AA. Vazamento de óleo isolante	41	0.7%
AA.Anormalidade nos acessórios	30	0.5%
AA.Anormalidade silica-gel	28	0.5%
AA. Anormalidade silica-gel	28	0.5%
Silica gel saturada	23	0.4%
AA. Anormalidade Silica-Gel	23	0.4%
Silica Gel Saturada	17	0.3%
AA. Aquecimento detectado por termovisão	17	0.3%
Defeito resistência de aquecimento cx ac	16	0.3%
Other values (4868)	5927	95.7%

Figura 8: Base de dados original - Campo descrição

Outro campo relevante que demonstra não muita variação é “Txt.code prob” com informações valiosas a respeito do problema em questão. Um dado interessante constatado, é a existência de um registro genérico “Outro – qual?” fazendo com que a qualidade seja ainda menor, simplesmente pelo fato de não caracterizar corretamente o problema.

Common Values | Overview

Value	Count	Frequency (%)
Outro - qual?	1622	26.2%
Componentes danificados ou desgastados	903	14.6%
Vazamento de óleo	888	14.3%
Dano ventiladores	499	8.1%
Sílica-Gel Defeituosa	425	6.9%
Ponto quente	383	6.2%
nan	172	2.8%
Ventilador(es) Defeito(s)	139	2.2%
Nível de óleo anormal	137	2.2%
Vazamento de Óleo	102	1.6%
Other values (74)	922	14.9%

Figura 9: Base de dados original - Campo "Txt.code prob"

Até mesmo campos de cadastro de transformadores como respectivo fabricante é apresentado de forma não padronizada, com evidencias de diferentes registros ao se referir sobre o mesmo fabricante, como por exemplo: “ABB” e “ASEA BROWN BOVERI”, conforme ilustrado na Figura *10*.

```
In [15]: np.sort(notas_tr['Fabricante'].unique())

Out[15]: array(['ABB', 'AEG', 'ALGE', 'ALSTHOM', 'ALSTHOM ATLANTIQUE', 'ALSTOM',
                'ANSALDO', 'AREVA', 'ASEA', 'ASEA  BROWN BOVERI',
                'ASEA BROWN BOVERI', 'ASEA ELÉTRICA', 'ASEA ELÉTRICA S.A',
                'ASEA ELÉTRICA SA', 'Alstom Grid Energia Ltda', 'Areva', 'BBC',
                'BELIMA', 'BROWN BOVERI', 'BROWN BOVERI / WEG', 'Brown Boveri',
                'CEMEC', 'CHARLEROY-INDUSELET', 'COEMSA', 'COEMSA-ANSALDO',
                'COENS.A.', 'COENSA', 'DEDINI', 'EASA', 'GE', 'GENERAL ELECTRIC',
                'GENERAL ELETRIC', 'GORDON', 'GORDON S/A', 'INDUSELET',
                'INDUSELET S.A.', 'ITAIPU', 'ITEL', 'Itaipu', 'LEGNANO', 'MITRUS',
                'MITSUBISHI', 'MITSUBISHI ELECTRIC CO.', 'NATIVA', 'ORTENG LTDA.',
                'ROMAGNOLE', 'SADE', 'SIEMENS', 'SIEMENS LTDA', 'SUPERKAVEA',
                'Siemens', 'T. EQUIP. ELETRONICOS S/A', 'TOSHIBA',
                'TOSHIBA DO BRASIL', 'TRAFO', 'TRAFO (ASEA)', 'TRAFO (COEMSA)',
                'TRAFO EQUIP.ELETRIC.', 'TRAFO T', 'TRAFOMIL',
                'TRANSFORMADORES JUNDIAI', 'TRANSFORMADORES UNIÃO',
                'TRANSFORMADORES UNIÃO S.A.', 'TRANSFORMADORES UNIÃO S/A',
                'TRANSFORMADORES UNIÃO(TUSA)', 'TUSA', 'TUSA / SIEMENS',
                'TUSA TRANSFORMADORES UNIÃO S.A', 'TUSA.TRANSF. UNIÃO LTDA',
                'Toshiba', 'Trafo', 'Trafo Equip. Elétricos S.A.',
                'Transformadores União', 'ULTRA TRAFO', 'UNIÃO', 'União', 'WEG',
                'WEG (ASEA)', 'WEG (BROWN BOVERI)', 'WEG (TRAFO)', 'WEG / ITEL',
                'WEG/MR/TREE TECH', 'WESTINGHOUSE', 'WTW', 'ZILMER', 'nan'],
               dtype=object)
```

Figura 10: Base de dados original - Campo "Fabricante"

De maneira similar ao caso anterior, a não padronização de informação relacionada ao modelo também tem registro com divergências cadastrais, como por exemplo “TMY 33”, “TMY-33” e “TMY.33”, ilustrado na Figura *11*. Observa-se a utilização de caracteres especiais no registro do respectivo modelo em inúmeros casos, algo que dificulta a categorização de dados para qualquer estratificação simples. Portanto, são itens que necessitam de padronização preliminar para proporcionar um tratamento dos dados mais assertivos.

```
In [8]: np.sort(notas_tr['Denomin.tipo'].unique())

Out[8]: array(['10953734', '112,5KVA', '11254362', '11300457', '11468179',
               '11554982', '11881450', '11931767', '11932181', '12471160',
               '13847349', '13847352', '13847356', '13859901', '225KVA',
               '3005.9999', '3007 0342', '3007.0417', '30070542',
               '345/88/34,5 kV', '3PO15/1-75P', '45 KVA', '50025213 (TE3LF)',
               '50042907', '500kVA', '5356/07', '615/009-010', '75KVA',
               '843274262', '843274645', '843274819', 'A', 'A NORMA NBR-535681',
               'AD3LF', 'AE3RA', 'AKOUS-402/15', 'AMO-CH', 'AMOC-FFA', 'AMOC-NF',
               'AMOV-NF', 'ATOC-NF', 'ATTOE', 'ATTOE/REG', 'ATTOE/TERRA', 'B',
               'BID-15001/138', 'BID-46/15', 'BM 400/2/15', 'BUC304050M24592,4',
               'CR', 'CRB', 'CTF', 'CTR3-YDDA', 'EFPN 8157/440',
               'ELCN 8156/345-B2', 'ELCN8156345B2HLAK544', 'ELCN8356+HLAK5646',
               'ELCN8356-HLAK5646', 'ELGN 7854', 'ELGN-7954',
               'ELMN 8157+HLAK 5742', 'ELMN8356+HLAK5646', 'ELPN 8057/440',
               'ELPN-8057/440', 'ELPN8157+HLAK5742', 'ELUN 7354', 'ELUN 7854',
               'ELUN7354', 'FOA', 'H-8', 'H8', 'HC / O P R - D', 'HC/OP/APLR-D',
               'HC/OP/OPLAR-D', 'HC/OP/OPLR-D', 'HC/OP/OPR-D', 'HC/OP/QPLR-D',
               'HC/OPTLR-D', 'HCTLR-D', 'HCTR-D', 'HO/OPTLAR-D', 'HTL',
               'ITS 75/1.2', 'KLRM-1545T/138', 'KLUM 1194/138', 'KLUM1194/138',
               'KOHM 1154/138B', 'KOHM1154/138', 'KOPHM-1154/138', 'LCF-VF',
               'LN-T', 'LN-VF', 'LN-VF-VF', 'LN-VF-VF MONOFASICO', 'LN/LVF',
               'LN/VF', 'LNVF', 'MA-4', 'MLPN 8054', 'MLPN-7954', 'MOC - NF',
               'MOC-FFA', 'MOC-NF', 'MOC37,5-50M/242/92,2', 'MOV-NF', 'MOVC-NF',
               'MUC 35-50M/242/92,4', 'MUC35-50M/212/92,4', 'NBR 5440 / 75 Kva',
               'NLPN 8157', 'NLPN7957+HLAK6342', 'NLUN 7856', 'NOS / ONP', 'NPN',
               'OA-T', 'ONAN', 'ONP', 'PDOE', 'PTO-PS', 'PTOC', 'PTOE',
               'PTOE 30071-0672', 'PTOE-REG', 'PTOO', 'PTOP8', 'PTOPS',
               'PTOPS-LN', 'RTI', 'SECO', 'SECO TDS-102T', 'SRB', 'T 1000/146',
               'T-0075/120', 'T-05 00/34', 'T-0500/34A', 'T-0500/34A/NAT/CLASB',
               'T24000/132-6', 'T5000/1326', 'T6250/696', 'T75K38', 'TA-1',
               'TAC 25-30M /145/72,5', 'TAC-15M/36,2/24,2', 'TAC-15M/38/25,8',
               'TAI', 'TAO', 'TC 500/15', 'TCO', 'TCX-500', 'TCY-75', 'TCZ-1000',
               'TCZ225', 'TD3LF', 'TDO-105E', 'TE 2 LF', 'TE2AF', 'TE2LF',
               'TE3LF', 'TE3RF', 'TE3RF/843274707', 'TEC-1000/15/1,2', 'TED 9009',
               'TEEY 820-9', 'TEEY 9309', 'TEEY-820-9', 'TET 656', 'TET 659',
               'TET-400-9', 'TET-7009', 'TET-739', 'TEZRF', 'TFDY 1180', 'TJ',
               'TL', 'TL-1000/15', 'TL-10000/138', 'TL-45/15', 'TL-5000/69',
               'TL-75/15', 'TL45/15', 'TLGN-8154/230-B1', 'TLLN-7251.138-8',
               'TLLN-7251/138B', 'TLMN 7352/138 B', 'TLMN-7852', 'TLPN 7752',
               'TLSN 6949', 'TLSN 7349', 'TLSN 7652', 'TLSN 7752',
               'TLSN 7752-2010', 'TLSN 8256', 'TLSN8356', 'TLTF-17500/138B',
               'TLTR 15000/1388', 'TM-46', 'TM-56', 'TM56', 'TMY 33', 'TMY 43',
               'TMY-33', 'TMY.33', 'TMZ 55', 'TMZ-44', 'TMZ-46', 'TMZ-53',
               'TMZ-55', 'TNBA', 'TNSYA', 'TNYA', 'TOC - NF', 'TOC-NF',
               'TODN-500/15A', 'TOFF-12,5/138J', 'TOT 1000', 'TOT 4708',
               'TOT-6509', 'TOT6509', 'TOTAL 300', 'TOVC - NF', 'TR-92',
               'TRANSFORMADOR CONVER', 'TS', 'TSA', 'TSAE 7250', 'TSAE-7049',
               'TSSE7250', 'TSUN6946', 'TT- 23', 'TT54', 'TTS45/1,2', 'TTSN-8156',
               'TTSNB 156', 'TUC 10-12,5M/145/15', 'TUC 10/12,5/138/15',
               'TUC 10/12,5M/145/15', 'TUC 15-18,5M/145/15', 'TUC 20-25M/145/15',
               'TUC 21800/92,4/38', 'TUC 21800/92,4/3B', 'TUC 25-33,3M/145/15/',
               'TUC 5/6,25M/38/15', 'TUC 500/15/1,2', 'TUC 500/15/1.2',
               'TUC 6250/72,5/15', 'TUC-10-12,5/138/15', 'TUC-25/33,3M-145/15/',
               'TUC-25/33,3M-145/15R', 'TUC-500/15/1,2', 'TUC25-33,3MVA/145',
               'TUC3040M/145/72,5/15', 'TUDFKOW', 'TUDFKWZ', 'TWZ 55', 'TYA',
               'Tnba', 'Tnya', 'UTO 15', 'WLHM-1734T/230-B1', 'WLHRM-1855/230-B1',
               'nan'], dtype=object)
```

Figura 11: Base de dados original - Campo de modelo "Denomin.tipo"

4.3 Pré-processamento e Limpeza

Diante do exposto de alguns exemplos de variáveis na etapa anterior, demonstrando a não padronização categórica devido aos dados se encontrarem de

modo não estruturado, foi realizada uma análise global sobre o *Dataset* – conjunto de dados, conforme ilustrado na Tabela *3*. Através dessa consolidação, nota-se outros pontos importantes do conjunto de dados com necessidade de ações de melhoria de qualidade, por exemplo com linhas duplicadas e variáveis com categorias não suportadas.

É observado também uma lacuna de 5,5% do conjunto de dados que se encontra com células faltantes. É notório que essa lacuna reflete na qualidade de dados resultantes em qualquer análise, no entanto se torna um potencial de aprimoramento do conjunto de dados com aplicação de técnicas visando o preenchimento destas células e dessa forma melhorando sua qualidade.

Tabela 3 - Estatísticas gerais do conjunto de dados de registros de manutenção de 2009 a 2020.

Estatística do conjunto de dados	
Número de variáveis	40
Número de observações	6192
Células faltantes	13643
Células faltantes (%)	5,5%
Linhas duplicadas	48
Linhas duplicadas (%)	0,8%

Tipos de variáveis	
Categoria	28
Data	6
Numero	4
Não suportado	2

Fonte: Elaboração própria.

Com o emprego do processo de limpeza da base de dados foram selecionadas apenas 9 variáveis das 40 iniciais, isso ocorreu devido muitas delas não oferecerem nenhum tipo de informação válida sobre os transformadores registrados. Basicamente, para a metodologia de escolha das variáveis candidatas ao projeto, foi levado em consideração o maior número de registros categóricos distintos possível, a fim de possibilitar uma análise de melhor qualidade. Na *Figura 12*, temos um exemplo da realização dessa etapa, onde foram verificadas as contagens de frequências de variáveis categóricas.

```
tr['Txt.code prob.'].value_counts()

Outro - qual?                                1620
Componentes danificados ou desgastados        903
Vazamento de óleo                             888
Dano ventiladores                             499
Sílica-Gel Defeituosa                         425
                                              ...
Alarme Ausência de Tensão VAC                   1
Defeito corpo indutivo                          1
Isolamento externo quebrado                     1
Relé alarmado / bloqueado                       1
Isolamento gasto ou quebrado                    1
Name: Txt.code prob., Length: 84, dtype: int64
```

Figura 12: Contagem de valores distintos por variável da base de dados.

Durante essa etapa notou-se que as variáveis possuíam categorias distintas e ao mesmo tempo subjetivas com relação ao propósito, portanto também foram eliminados os seguintes registros.

- 'Outra - ¿quál?';
- 'Outra (especificar)';
- 'Outro - qual?'.

Como o projeto se destina a classificação de palavras, houve a necessidade de realizar outras etapas envolvendo o pré-processamento dos dados. Em resumo, após a remoção das linhas em branco nos dados, a base amostral obtida foi de 2741 registros e 9 variáveis. Com isso, a conclui-se as etapas de pré-processamento e limpeza para aplicação dos algoritmos de árvore de decisão.

Contudo, tanto Rede Bayesiana quanto SVM necessitaram de um pré-processamento com etapas especificas, em virtude da característica intrínseca a esses dois modelos classificadores. Basicamente, foi necessário a criação de uma coluna agrupando todas as variáveis, denominada “Texto total”, além da realização das 5 etapas listadas na sequência:

1. Alteração de texto para minúsculas.
2. Divisão de fluxo de texto em palavras, frases, símbolos ou outros elementos significativos.
3. Remoção de palavras de parada.
4. Remoção de texto não alfa
5. Redução de formas inflexionais de cada palavra para uma base ou raiz comum.

Para possibilitar execução das etapas 2. Divisão de fluxo de texto em palavras, frases, símbolos ou outros elementos significativos e 5. Redução de formas inflexionais de cada palavra para uma base ou raiz comum, houve a necessidade de realizar previamente a categorização das palavras em substantivo, verbo ou adjetivo. Na sequência, verificou-se a existência de palavras de parada, considerando apenas o alfabeto da língua portuguesa, conforme ilustrado na *Figura 13*. E então, somente após todas essas etapas, o conjunto de palavras processadas para cada iteração foram armazenados na coluna “texto_final”, conforme ilustrado na Figura *14*.

```
tag_map = defaultdict(lambda : wn.NOUN)
tag_map['J'] = wn.ADJ
tag_map['V'] = wn.VERB
tag_map['R'] = wn.ADV

for index,entry in enumerate(nbsvm['Texto total']):
    Final_words = []
    word_Lemmatized = WordNetLemmatizer()
    for word, tag in pos_tag(entry):
    if word not in stopwords.words('portuguese') and word.isalpha():
            word_Final = word_Lemmatized.lemmatize(word,tag_map[tag[0]])
            Final_words.append(word_Final)
    nbsvm.loc[index,'texto_final'] = str(Final_words)
```

Figura 13: Etapas de criação da variável "texto total"

```
print(nbsvm['texto_final'].head())
```

```
24          ['baixo', 'nivel', 'óleo', 'nível', 'óleo']
38           ['nivel', 'óleo', 'óleo', 'nível', 'óleo']
39    ['oleo', 'isolante', 'bucha', 'óleo', 'nível',...
40    ['oleo', 'isolante', 'bucha', 'cuba', 'nível',...
41    ['falha', 'comando', 'baixa', 'mecanismo', 'co...
Name: texto_final, dtype: object
```

Figura 14: Conjunto final de palavras processadas para cada iteração - "texto_final".

A etapa seguinte foi de declarar o vetor de recurso e variável de destino. Nesse momento, com a finalidade de padronizar as simulações com os diferentes algoritmos, tomou-se como premissa a utilização da variável categórica “Tipo”, como ilustrado na Figura *15*. A variável “Tipo” faz menção ao tipo característico e/ou aplicação do transformador, sendo uma informação relevante ao modelo por possuir um número reduzido de categorias e consequentemente um menor ruído proporcionado aos modelos quando em momento de simulações.

```
tr['Tipo'].value_counts()
```

```
TRF     2233
ATR      340
TRSA     134
TRRG      28
TRAT       6
Name: Tipo, dtype: int64
```

Figura 15: Variável categórica de transformadores "Tipo"

Legenda:

TRF: Transformador de Força, de característica construtiva por bancos de unidades monofásicas

ATR – Autotransformador, de característica construtiva por unidade trifásica

TRSA: Transformador de Serviço Auxiliar

TRRG: Transformador Regulador

TRAT: Transformador de Aterramento

A partir dessa definição foi possível dividir os dados em conjuntos de treinamento e teste, utilizando a proporção de 33% para o conjunto de teste e 67% para o conjunto de treinamento, como mostrado na *Figura 16*.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.33, random_state = 42)
```

Figura 16: Divisão de dados em conjuntos de treinamento e teste

4.4 Transformação de dados

Mesmo com os dados terem passado pela etapa de pré-processamento e limpeza, há necessidade transformação dos dados brutos em recursos uteis que ajudam a entender melhor o modelo com o aumento de seu poder preditivo. Esse processo também é conhecido pelo termo inglês como *Feature Engineering*, na tradução livre, Engenharia de Recursos. Basicamente, essa é uma etapa de transformação de dados categóricos em valores numéricos através da codificação das variáveis com codificação ordinal. Tem-se o objetivo com isso, viabilizar utilização de

métodos vetorizados. Através da vetorização, é possível descobrir a importância de uma palavra no documento em comparação com a base de dados.

Para esse processo de transformação de dados, foi utilizado o método TF-IDF – *Term Frequency – Inverse Document Frequency*, na tradução livre, Frequência de Termo - Frequência de Documentos Inversos [29]. São pontuações de frequência de palavras buscando destaque para as palavras mais interessantes em um documento, onde ocorre a atribuição de um número exclusivo para cada palavra. Como resultado desse processo, têm-se então o número da linha analisada, uma associação numérica exclusiva para cada palavra e uma pontuação calculada pela vetorização, remetendo a importância associada, conforme exemplo na *Figura 17*.

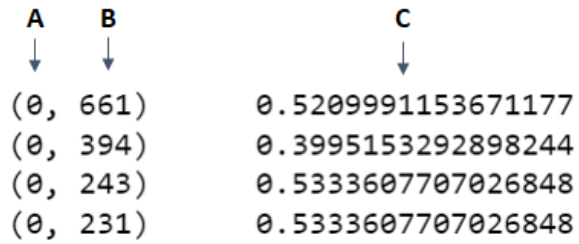


Figura 17: Exemplo de resultados de vetorização por TF-IDF

Legenda:
A - Número da linha da vetorização do conjunto de treinamento X
B - Número exclusivo de cada palavra da primeira linha
C – Pontuação calculada pela vetorização TF-IDF

4.5 Escolha da tarefa e algoritmos de mineração de dados

A tarefa mais aderente ao trabalho foi de classificação, com objetivos de previsão, ou seja, mineração de dados supervisionada. Dentre as inúmeras opções de algoritmos de classificação, foram escolhidas a Árvore de Decisão [17], Rede Bayesiana [17] e SVM – Máquina de Vetores de Suporte [24].

Com o objetivo de testar não apenas uma opção, mas também outras soluções de mineração de dados com objetivo de classificação, os três algoritmos contaram com as mesmas etapas que antecedem a escolha e definição do algoritmo. Salvo

algumas exceções de adequações preliminares devido a característica do algoritmo, com por exemplo a necessidade da criação de uma coluna agrupando todas as variáveis, denominada “Texto total”, além da realização das 5 etapas listadas no 4.3 Pré-processamento e Limpeza.

4.6 Empregando o algoritmo de mineração de dados

Como foram escolhidos três diferentes algoritmos de mineração de dados, houve a aplicação de cada um deles de modo diferenciado a partir dessa etapa. Isso se deve pelo fato deles serem distintos entre si quanto ao seu funcionamento, apesar de todos terem possibilidade de utilização similares.

4.7 Avaliação

Nas seções seguintes serão apresentadas em detalhes o emprego das três metodologias propostas: Árvore de decisão, rede Bayesiana e SVM – Máquina de Vetores de Suporte.

4.7.1 Aplicação de Árvore de Decisão

Para as simulações do algoritmo de Árvore de Decisão foram utilizados dois critérios para seleção de atributos, a entropia e o índice Gini, proveniente de CART - *Categorical and Regression Trees*, na tradução livre, árvores categóricas e de regressão. Com isso, foi possível obter resultados para ambos e compará-los não somente com os outros algoritmos, mas entre si, como soluções intrínsecas a Árvore de Decisão.

Adicionalmente, se verificou a possível existência de superajuste (*Over-fitting*) nos conjuntos de teste e treinamento. Conforme ilustrado na *Figura 18*, a pontuação

do conjunto de treinamento foi de 0,9069 e a pontuação do conjunto de teste de 0,8873. Como esses dois valores são muito próximos e comparáveis, se nota a inexistência de superajuste que poderia estar enviesando o trabalho.

```
print('Training set score: {:.4f}'.format(clf_gini.score(X_train, y_train)))

print('Test set score: {:.4f}'.format(clf_gini.score(X_test, y_test)))
```

```
Training set score: 0.9069
Test set score: 0.8873
```

Figura 18: Verificação de superajuste nos conjuntos de treinamento e teste - índice Gini

Na *Figura 19*, é demonstrada a árvore de decisão utilizando critério de índice Gini de maneira ilustrativa, facilitando a visualização do trabalho desse algoritmo.

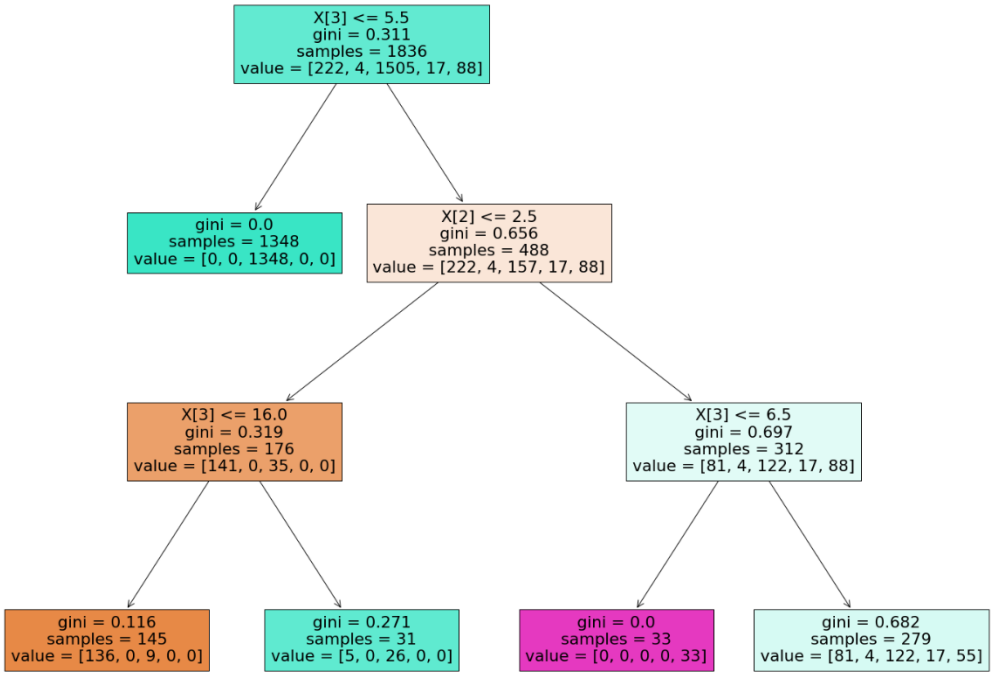


Figura 19: Árvore de decisão utilizando critério índice Gini

O mesmo processo se repetiu para aplicação do algoritmo de árvore de decisão com critério de entropia. Na verificação de possível existência de superajuste (*Over-fitting*) nos conjuntos de treinamento e teste foram obtidas as pontuações de 0,9259 e 0,9182, respectivamente, como mostrado na *Figura 20*. De modo similar a verificação

com critério de índice Gini, os valores se demonstraram próximos, e, portanto, se nota também a inexistência de superajuste que poderia estar enviesando o trabalho.

```
print('Training set score: {:.4f}'.format(clf_en.score(X_train, y_train)))

print('Test set score: {:.4f}'.format(clf_en.score(X_test, y_test)))
```

```
Training set score: 0.9259
Test set score: 0.9182
```

Figura 20: Verificação de superajuste nos conjuntos de treinamento e teste – Entropia

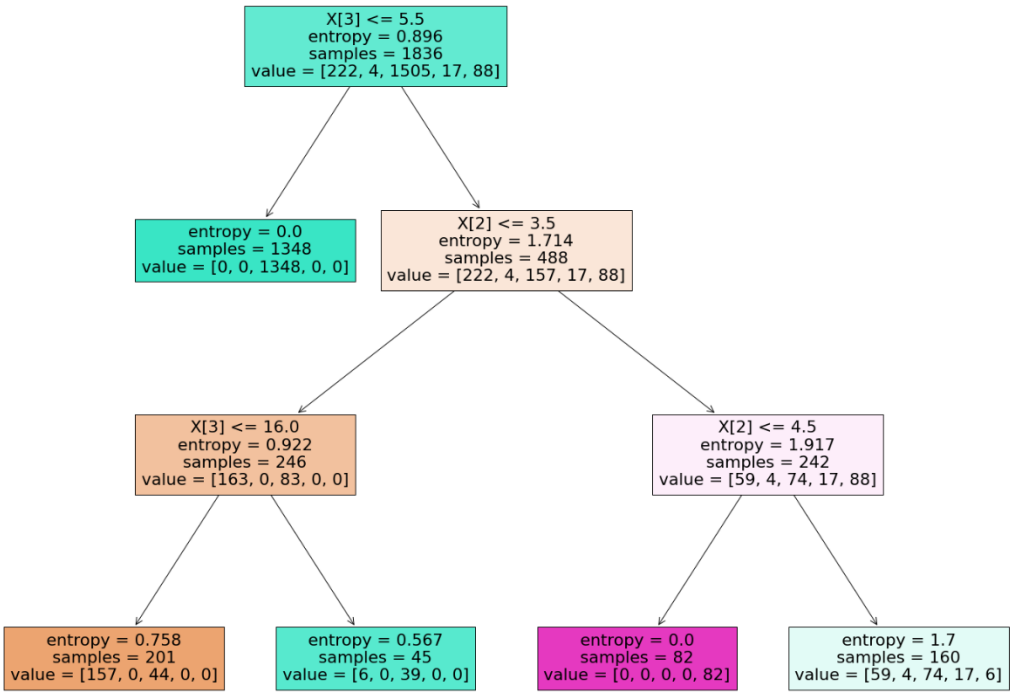


Figura 21: Árvore de decisão utilizando critério de entropia

Na Figura 21, é ilustrada a árvore de decisão utilizando o critério de entropia. Esse tipo de visualização é bastante útil para entender o funcionamento desse algoritmo, porém não apresenta os erros que o classificador possa estar cometendo. Nesse sentido, utilizou-se também outra forma de verificar o desempenho do classificador, inclusive com os erros produzidos pelo modelo. Para tal, a matriz confusão é uma boa opção de identificação de resultados em quatro tipos: verdadeiros

positivos (TP), verdadeiros negativos (TN), falsos positivos (FP) e falsos negativos (FN), sendo que os dois últimos, são tidos como erros mais críticos como mostrada nas *Tabela 4: Matriz confusão utilizando critério índice Gini* e *Tabela 5: Matriz confusão utilizando critério entropia*.

Tabela 4: Matriz confusão utilizando critério índice Gini

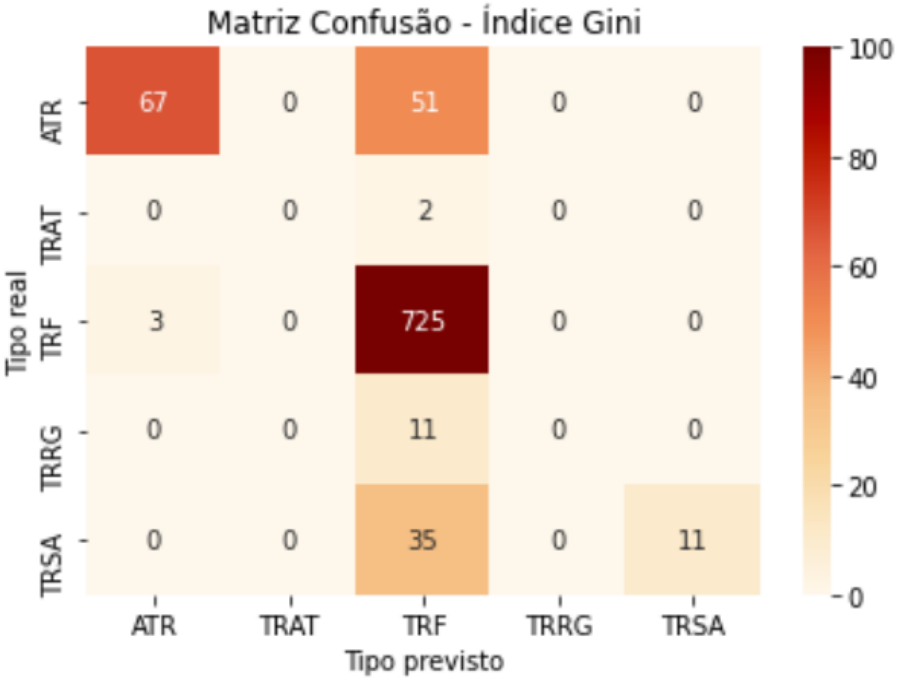
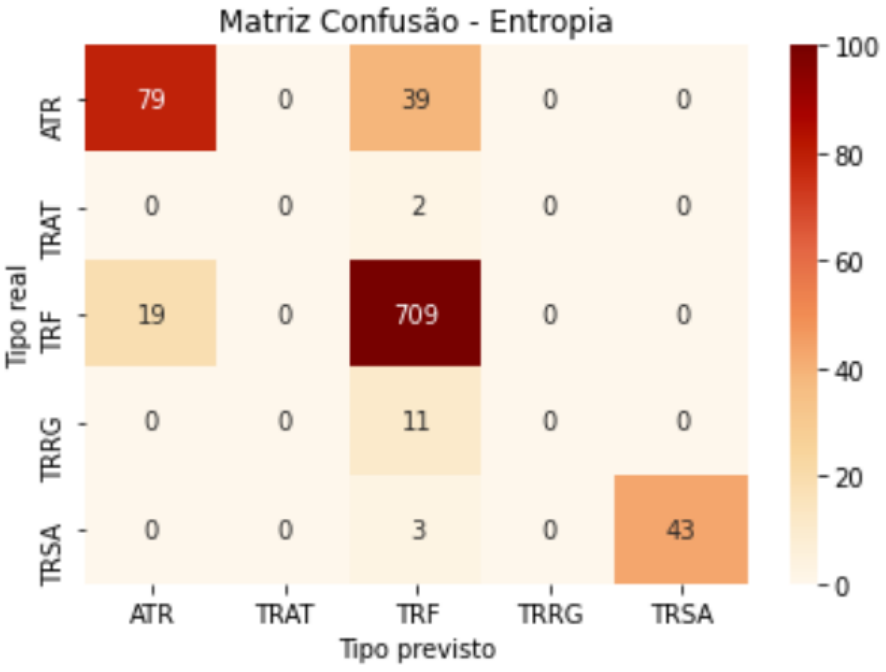


Tabela 5: Matriz confusão utilizando critério entropia



Em um primeiro olhar sobre os resultados apresentados nas matrizes se identifica o porquê de o desempenho não ter sido de 100%. Nota-se erros tanto em aprendizado utilizando critérios de índice Gini, como também entropia. Com uma vantagem para o critério de entropia que apresentou melhor desempenho na classificação do tipo TRSA quando comparado ao índice Gini.

Ressalta-se que na matriz confusão de critério de entropia também é possível verificar que houve erros na previsão do sistema. Dentre eles, a predição de 39 equipamentos tipo ATR que foram classificados como TRF. E também ao não conseguir prever corretamente os tipos TRAT e TRRG, classificando-os como outro tipo na predição.

Um ponto a se observar é de os resultados apresentados na matriz confusão terem um melhor desempenho proporcionalmente a quantidade de valores utilizadas no modelo para essa variável. E que isso teria influência nesses acertos e erros de aprendizado. De fato, como mostrado na Figura *15*: Variável categórica de transformadores "Tipo", as quantidades são distribuídas de forma não uniforme, e percebe-se também o mesmo comportamento da matriz confusão, isto é, quanto maior a quantidade, maior o número de acertos e consequentemente menor número de erros.

Na busca de um método padronizado de comparação de desempenho de algoritmos, e que possibilitasse a utilização também com as avaliações de Rede Bayesiana e SVM - Máquina de Vetores de Suporte, foi identificado o relatório de classificação. É um outro meio de realizar avaliações o desempenho do modelo de classificação, exibindo a precisão, recall, na tradução livre, chamar novamente, pontuação f1 e suporte.

Onde:

Precisão é a relação entre a quantidade de resultados corretos dividido pelo total de todos os resultados retornados. Dada pela seguinte expressão [30]:

$$Precisão = \frac{TP}{TP + FP}$$

Recall é a relação entre a quantidade de resultados corretos sobre a quantidade de resultados que deveriam ser retornados. Também conhecida como sensibilidade. Dada pela seguinte expressão *[30]*:

$$Recall = \frac{TP}{TP + FN}$$

Pontuação F1 é a média harmônica da precisão e recall. É dada pela seguinte expressão [30]:

$$Pontuação\ F1 = \frac{2}{\frac{1}{Precisão} + \frac{1}{Recall}}$$

A acurácia geral de um classificador é estimada dividindo-se o total de positivos e negativos corretamente classificados pelo número total de amostras [30]. É dada pela seguinte expressão:

$$Acurácia = \frac{TP + TN}{TP + FN + FP + FN}$$

E por fim, o termo suporte remete ao número de ocorrências de cada rótulo.

Através da Tabela *6* e Tabela *7* é possível visualizar os resultados obtidos do modelo de aprendizado e constar que o desempenho de ambos os critérios, índice Gini e entropia foram muito bons.

Tabela 6: Relatório de classificação - Árvore de Decisão com critério índice Gini

Tipo	**Precisão**	**recall**	**Pontuação F1**	**Suporte**
ATR	0,96	0,57	0,71	118
TRAT	0,00	0,00	0,00	2
TRF	0,88	1,00	0,93	728
TRRG	0,00	0,00	0,00	11
TRSA	1,00	0,24	0,39	46
Acurácia			0,89	905

Tabela 7: Relatório de classificação - Árvore de Decisão com critério entropia

Tipo	Precisão	recall	Pontuação F1	Suporte
ATR	0,81	0,67	0,73	118
TRAT	0,00	0,00	0,00	2
TRF	0,93	0,97	0,95	728
TRRG	0,00	0,00	0,00	11
TRSA	1,00	0,93	0,97	46
Acurácia			0,92	905

4.7.2 Aplicação de Rede Bayesiana

Na simulação utilizando Rede Bayesiana foram realizados os devidos ajustes no conjunto de dados de treinamento do classificador, conforme segue na *Figura 22*. e na sequencia realizou-se a previsão dos rótulos no conjunto de dados de validação expressa pela linha de código da *Figura 23*. E por fim, obteve-se a acurácia de 0,8104, como mostrado na *Figura 24*.

```
Naive = naive_bayes.MultinomialNB()
Naive.fit(Train_X_Tfidf,Train_Y)
```

```
▾ MultinomialNB
MultinomialNB()
```

Figura 22: Ajuste no conjunto de dados para o classificador Naive Bayes

```
predictions_NB = Naive.predict(Test_X_Tfidf)
```

Figura 23: Previsão dos rótulos por Naive Bayes

```
print("Naive Bayes Accuracy Score -> ",accuracy_score(predictions_NB, Test_Y))
```

```
Naive Bayes Accuracy Score ->  0.8104166666666667
```

Figura 24: Acurácia do modelo com Naive Bayes

Com a finalidade de entender maiores detalhes e seguir as mesmas análises empregadas na simulação com árvore de decisão, para Rede Bayesiana também foram geradas a matriz confusão e o relatório de classificação, conforme ilustrado na *Tabela 8* e *Tabela 9*.

Tabela 8: Matriz confusão Rede Bayesiana

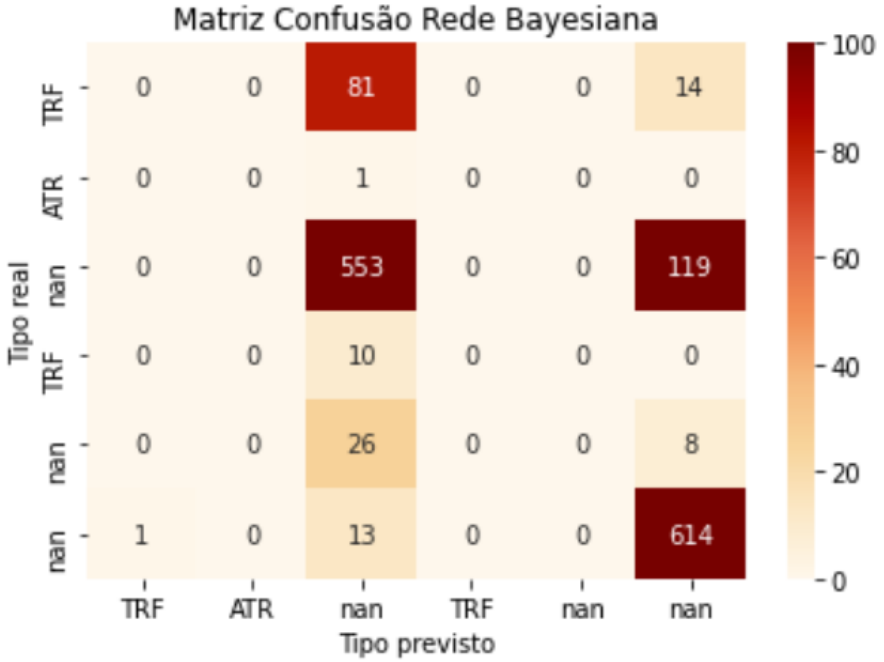


Tabela 9: Relatório de classificação - Rede Bayesiana

Tipo	Precisão	recall	Pontuação F1	Suporte
TRF	0,00	0,00	0,00	1
ATR	0,00	0,00	0,00	0
nan	0,82	0,81	0,82	684
TRF	0,00	0,00	0,00	0
nan	0,00	0,00	0,00	0
nan	0,98	0,81	0,89	755
Acurácia			0,81	1440

Através da matriz confusão é possível notar que o alto valor de acurácia não se refere a uma classificação válida. Pelo contrário, fica evidente na *Tabela 9: Relatório de classificação - Rede Bayesiana* que a alta precisão do modelo teve como suporte uma

grande quantidade de dados nulos, chamados pela terminologia “nan” na linguagem em python.

4.7.3 Aplicação de SVM – Máquina de Vetores de Suporte

Para o algoritmo SVM foi realizada a mesma simulação, iniciando pelo ajuste do conjunto de dados de treinamento do classificador, como demonstrado na *Figura 25*. Após isso, se realizou a previsão dos rótulos no conjunto de dados de validação, expressa pela linha de código da *Figura 26*. E então, obteve-se a acurácia de 0,8187, como mostrado na Figura *27*.

```
SVM = svm.SVC(C=1.0, kernel='linear', degree=3, gamma='auto')
SVM.fit(Train_X_Tfidf,Train_Y)
```

```
SVC
SVC(gamma='auto', kernel='linear')
```

Figura 25: Ajuste no conjunto de dados para o classificador SVM

```
predictions_SVM = SVM.predict(Test_X_Tfidf)
```

Figura 26: Previsão dos rótulos por SVM

```
print("SVM Accuracy Score -> ",accuracy_score(Test_Y, predictions_SVM))
```

```
SVM Accuracy Score ->  0.81875
```

Figura 27: Acurácia do modelo com SVM

De maneira similar, gerou-se também a matriz confusão e o relatório de classificação, a fim de entender melhor os detalhes a respeito da aplicação do algoritmo SVM ao trabalho, de acordo como segue nas *Tabela 10* e *Tabela 11*.

Tabela 10: Matriz confusão SVM

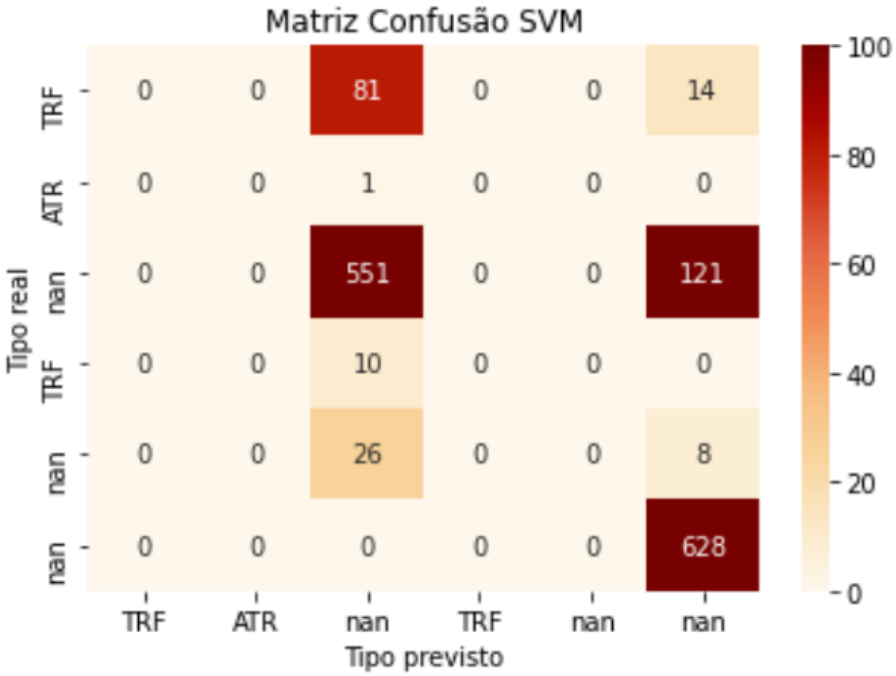


Tabela 11: Relatório de classificação - SVM

Tipo	Precisão	recall	Pontuação F1	Suporte
TRF	0,00	0,00	0,00	0
ATR	0,00	0,00	0,00	0
nan	0,82	0,82	0,82	669
TRF	0,00	0,00	0,00	0
nan	0,00	0,00	0,00	0
nan	1,00	0,81	0,90	771
Acurácia			0,82	1440

Com resultado bastante similar a aplicação do algoritmo rede Bayesiana, o alto valor de acurácia geral não remete a uma boa classificação, pois do mesmo modo, foram classificados rótulos nulos. Cabe destacar a importância de utilizar mais de um método de avaliação do classificador, logo que, se assumir apenas os resultados de acurácia geral de cada um deles, todos os algoritmos tiveram desempenhos satisfatórios. E na realidade, analisando detalhadamente o funcionamento discretizado por rótulos, nota-se que o desempenho de Rede Bayesiana e SVM não foram bons.

4.8 Resultados experimentais

Os algoritmos utilizados no trabalho com objetivo de classificação de dados não estruturados desempenharam um comportamento bastante interessante e distinto. Nas aplicações de Árvore de Decisão, tanto com critério de índice Gini quanto com entropia demonstraram um resultado muito bom, com geração de erros sim, mas em quantidades reduzidas quando comparada com os acertos. A matriz confusão e o relatório de classificação apresentaram um desempenho de modelo muito bom e bastante aderente ao objetivo do trabalho.

Já os algoritmos de Rede Bayesiana e SVM – Máquina de Vetores de Suporte geraram uma classificação aquém do esperado, principalmente após a obtenção dos resultados do algoritmo árvore de decisão. É possível notar através das respectivas matrizes confusão e dos relatórios de classificação, onde os eixos que demonstraram acertos remetem a rótulos nulos, isto é, não válidos ao propósito do trabalho.

Na *Tabela 12*, se verifica que mesmo possuindo um número de menor de dados de suporte, os algoritmos de árvore de decisão apresentaram melhores desempenho para rótulos validos e não nulos. Já Rede Bayesiana e SVM realizaram a classificação apenas para rótulos nulos, mesmo tendo retornado um número maior de dados de suporte.

Tabela 12: Comparação de resultados dos algoritmos

	Pontuação F1	**Suporte**	**Percentual de acerto de rótulos válidos**
Árvore de Decisão - Índice Gini	0,89	905	89%
Árvore de Decisão - Entropia	0,92	905	92%
Rede Bayesiana	0,81	1440	-
SVM - Máquina de Vetores de Suporte	0,82	1440	-

Uma informação relevante também, é sobre a classificação obtida pelos algoritmos conseguir apresentar resultados somente para uma parcela reduzida da base de dados bruta a qual se iniciou todo o trabalho. Em resumo, o conjunto de dados inicial era de 6192 registros e 40 variáveis e que ao passar pelas etapas de pré-processamento e limpeza, teve seu tamanho reduzido para 2741 registros e 9

variáveis. E por fim, como retorno do aprendizado dos algoritmos obteve-se apenas 905 com classificação válida, incluindo os erros.

4.9 Considerações parciais

Ao longo desse capitulo, foi mostrado todas as etapas do trabalho baseadas no processo metodológico de *Knowledge Discovery in Database* – KDD, que teve como resultado a classificação de dados não estruturados de uma base de dados histórica de atividades de manutenção de transformadores de potência.

Ao partir da base de dado bruta, se realizou a compreensão do domínio do trabalho em questão e somente após isso é que foi criado o conjunto de dados com a finalidade de conter o máximo de informações relevantes ao equipamento em questão. Como etapa seguinte, houve o pré-processamento e limpeza de tudo que poderia interferir na busca de resultados satisfatórios. Nessa etapa também, foi realizada a definição da variável categórica Tipo como padrão nos classificadores. Houve também a divisão dos dados em conjuntos de treinamento e teste.

Na sequência, realizou-se a transformação dos dados, uma adequação necessária para que possibilitassem os diferentes algoritmos de serem utilizados. Dessa forma, houve a transformação dos dados categóricos em valores numéricos através da codificação das variáveis com codificação ordinal.

Como a escolha da tarefa a ser executada já estava estabelecida como a classificação dos dados, assim como a definição dos algoritmos a serem utilizados estava baseada na aplicação de Árvore de Decisão, Rede Bayesiana e SVM – Maquina de Vetores de Suporte, partiu-se então para o emprego dos três algoritmos nos conjuntos de dados.

Através da utilização desses três algoritmos foi possível a realização de simulações e testes com cada um deles, na busca de capturar suas particularidades de funcionamento e exigências de adequações, caso fossem necessários. Após a realização de vários testes de maneira empírica com os atributos de cada algoritmo com objetivo de obter o valor mais alto de acurácia, no primeiro momento todos os

algoritmos apresentaram resultados satisfatórios no que tange ao objetivo de classificação, com valores de acurácia geral partindo de 0,81 à 0,92.

No entanto, ao realizar outros métodos de avaliação dos modelos utilizados como matriz de confusão e relatório de classificação, foi notória a discrepância em relação a qualidade do que foi classificado por cada um dos modelos. Os algoritmos de Rede Bayesiana e SVM tiveram um bom desempenho apenas classificando dados nulos. Logo, cabe um destaque para o algoritmo de árvore de decisão que obteve resultados elevados tanto para simulações com critério de índice gini quanto para critério de entropia.

5 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho de mestrado mostrou a importância de Gestão de Ativos no Sistema Elétrico de Potência, trazendo uma visão do agente de transmissão quanto as suas regras de operação e manutenção, geridas pelas Resoluções Normativas. Regras estas que ora podem se caracterizar em requisitos mínimos de manutenção com periodicidade definida, ora em penalizações das receitas associadas a Função Transmissão.

Mostrou que o transformador de potência, equipamento de grande porte, possui papel fundamental no Sistema Elétrico de Potência, tido como um dos mais importantes. Desse modo, toda e qualquer informação relacionada a esse equipamento pode contribuir muito para um melhor diagnóstico ou tomada de decisão sobre uma ação que transmita mais confiabilidade.

Diante de uma lacuna de registros históricos de atividades de manutenção em transformadores de potência, como ausência de dados, variáveis não padronizadas e categorias genéricas, a proposta desse trabalho de mestrado buscou a aplicação de ferramentas inteligentes na classificação de dados não estruturados com o objetivo de possibilitar a utilização desses registros de forma que gerassem uma informação valida e mais precisa.

Após a realização de todos os testes e simulações considerando os métodos de Árvore de Decisão, Rede Bayesiana e SVM - Máquina de Vetores de Suporte, foi possível atestar que os três algoritmos são aderentes a classificação de dados não estruturados, ressalta-se então, as seguintes constatações.

A base de dados bruta reduziu de tamanho de maneira significativa após as etapas de pré-processamento e limpeza. Devido a isso, o emprego dos algoritmos de mineração de dados foi realizado apenas em uma amostra da base de dados total que contempla o parque todo de transformadores de potência do agente de transmissão.

As três metodologias demonstraram resultados de alta acurácia geral, com valores próximos e destaque para Árvore de Decisões que alcançou os maiores valores. Desse modo, o algoritmo mais aderente a proposta de classificação ao qual o trabalho se propôs a realizar é a Árvore de decisão, principalmente pelo fato de conseguir classificar rótulos válidos e não nulos, diferentemente dos algoritmos de rede Bayesiana e SVM.

Para os trabalhos futuros, sugere-se aprofundar na melhoria de qualidade do preenchimento dos dados de registros de manutenção, criando alternativas e soluções sistêmicas como medidas de contorno, a fim de garantir o correto e completo registro de atividade de manutenção de transformadores de potência e também uma cobertura maior do parque instalado, uma vez que com a etapa de pré-processamento e limpeza, muitos dados faltantes e/ou não padronizados foram descartados das simulações por enviesar os resultados. Uma sugestão que vem a esse encontro é a utilização dos classificadores para outras variáveis dos conjuntos de dados, logo que a aplicação foi validada de maneira satisfatória, há possibilidade de realizar o preenchimento da lacuna de informações da base de dados bruta com a retroalimentação obtida através dos resultados dos classificadores.

Também se sugere o aperfeiçoamento nos ensaios experimentais envolvendo os diferentes algoritmos com a otimização de hiper parâmetros, principalmente com Árvore de Decisão. Enquanto os parâmetros são ajustados dentro do processo de aprendizagem, os hiper parâmetros se diferenciam por serem definidos antes do treinamento atuando como atributos de controle e identificando qual a combinação pode produzir os melhores valores dos respectivos critérios de precisão. Espera-se então que com a implantação dos hiper parâmetros seja possível obter resultados ainda melhores de desempenho.

REFERÊNCIAS

[1] ONS - Operador Nacional do Sistema Elétrico, “Sobre o ONS: O que é o ONS,” [Online]. Available: http://www.ons.org.br/paginas/sobre-o-ons/o-que-e-ons. [Acesso em 24 Setembro 2020].

[2] ANEEL, “Resolução Normativa nº669,” 14 Julho 2015. [Online]. Available: http://www2.aneel.gov.br/cedoc/ren2015669.pdf. [Acesso em 17 Novembro 2020].

[3] ANEEL, “Resolução Normativa nº853,” 13 Agosto 2019. [Online]. Available: http://www2.aneel.gov.br/cedoc/ren2019853.pdf. [Acesso em 17 Novembro 2020].

[4] ANEEL, “Resolução Normativa Nº 906,” 8 Dezembro 2020. [Online]. Available: https://www.in.gov.br/en/web/dou/-/resolucao-normativa-aneel-n-906-de-8-de-dezembro-de-2020-294354729. [Acesso em 20 Janeiro 2023].

[5] ANEEL, “Resolução Normativa nº729,” 28 Junho 2016. [Online]. Available: http://www2.aneel.gov.br/cedoc/ren2016729.pdf. [Acesso em 25 Setembro 2020].

[6] ANEEL, “Resolução Normativa nº782,” 19 Setembro 2017. [Online]. Available: http://www2.aneel.gov.br/cedoc/ren2017782.pdf. [Acesso em 25 Setembro 2020].

[7] ANEEL, “Resolução Normativa nº853,” 13 Agosto 2019. [Online]. Available: http://www2.aneel.gov.br/cedoc/ren2019853.pdf. [Acesso em 25 Setembro 2020].

[8] J. H. HARLOW, Electric Power Transformer Engineering, Boca Raton, Florida: CRC Press, 2012.

[9] J. MAMEDE FILHO, Manual de Equipamentos Elétricos, Rio de Janeiro: LTC, 2013.

[10] Grupo de Trabalho B3/B5/A2.01 - CIGRÉ Brasil, “Transformadores Imersos em Liquidos Isolantes - Guia de Manutenção Centrada na Confiabilidade,” 2007.

[11] N. A. J. HASTINGS, Physical Asset Management, London: Springer, 2010.

[12] J. R. R. M. J. M. F. A. G. C. B. J. M. J. J. G. SARMIENTO, “A predictive model for the maintenance of industrial machinery in the context of industry 4.0,” *Engineering Applications of Artificial Intelligence,* vol. 87, 2020.

[13] U.S. Department of the Interior Bureau of Reclamation, Transformers: Basics, Maintenance, and Diagnostics, Denver, 2005.

[14] M. F. P. R. A. S. a. A.-S. A. AL HAMDANI, “Power Transformer Degradation Condition and Insulation Index Estimation Based on Historical Oil Dat,” em *2nd International Conference on High Voltage Engineering and Power Systems (ICHVEPS)*, Bali, 2019.

[15] L. M. O. ROKACH, Data Mining with Decision Trees - Theory and Applications, Singapore: World Scientific Publishing Co. Pte. Ltd., 2015.

[16] U. P.-S. G. a. S. P. FAYYAD, “From Data Mining to Knowledge Discovery in Databases.,” *AI Magazine,* vol. 17, nº 3, pp. 37-54, 1996.

[17] J. BELL, Machine Learning: Hands-on for developers and techinical professionals, Indianapolis: John Wiley & Sons, Inc., 2015.

[18] F. A. S. BORGES, “Extração de Caracteristicas Combinadas com Árvores de Decisão para Detecção e Classificação dos Distúrbios de Qualidade da Energia Elétrica.,” São Carlos, 2013.

[19] L. F. J. H. O. R. A. &. C. J. BREIMAN, Classification and Regression Trees, 1st ed., Chapman and Hall/CRC, 1984.

[20] J. R. QUINLAN, “Induction of decison Trees,” *Machine Learning,* vol. 1, nº 1, pp. 81-106, 1986.

[21] J. R. QUINLAN, “C4.5: Programs for Machine Learning,” *Machine Learning,* vol. 16, pp. 245-240, 1994.

[22] M. A. ARAUJO, “Tese de Doutorado: Metodologia Baseada em Medidas Dispersas de Tensão e Árvores de Decisão para Localização de Faltas em Sistemas de Distribuição Modernos,” São Carlos, 2017.

[23] T. C. O. PAVLENKO, “Credit risk modeling using bayesian networks,” *International Journal of Intelligent Systems,* vol. 25, pp. 326-344, 2010.

[24] V. JAKKULA, “Tutorial on Support Vector Machine (SVM),” Pullman, 2011.

[25] V. N. VAPNIK, The Nature of Statistical Learning Theory, New York: Springer, 2000.

[26] C. J. BURGES, “A Tutorial on Support Vector Machines for Pattern Recognition,” Kluwer Academic Publishers, Boston, 1998.

[27] L. AURIA e R. A. MORO, Support Vector Machines (SVM) as a technique for solvency analysis, vol. No. 811, Berlin: Deutsches Institut für Wirtschaftsforschung (DIW), 2008.

[28] Python, 2020. [Online]. Available: https://python.org/. [Acesso em 29 Setembro 2020].

[29] O. S. D. S. I. Yahav, “Comments Mining With TF-IDF: The Inherent Bias and Its Removal,” *IEEE Transactions on Knowledge and Data Engineering,* pp. 437-450, 1 Março 2019.

[30] D. L. D. D. Olson, Advanced Data Mining Techniques, 1ª ed., Springer, 2008, p. 138.

[31] M. WALTON, The Deming Management Method, Perigee Books, 1988.