UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

AVALIAÇÃO DA DISTORÇÃO HARMÔNICA TOTAL DE TENSÃO NO PONTO DE ACOPLAMENTO COMUM INDUSTRIAL USANDO O PROCESSO KDD BASEADO EM MEDIÇÃO

EDSON FARIAS DE OLIVEIRA

TD 05/2018

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2018

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

EDSON FARIAS DE OLIVEIRA

AVALIAÇÃO DA DISTORÇÃO HARMÔNICA TOTAL DE TENSÃO NO PONTO DE ACOPLAMENTO COMUM INDUSTRIAL USANDO O PROCESSO KDD BASEADO EM MEDIÇÃO

TD 05/2018

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2018

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

EDSON FARIAS DE OLIVEIRA

AVALIAÇÃO DA DISTORÇÃO HARMÔNICA TOTAL DE TENSÃO NO PONTO DE ACOPLAMENTO COMUM INDUSTRIAL USANDO O PROCESSO KDD BASEADO EM MEDIÇÃO

Tese submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará para a obtenção do Grau de Doutor em Engenharia Elétrica na área de Sistemas de Energia Elétrica.

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2018

Dados Internacionais de Catalogação – na – Publicação (CIP)
Sistema de Bibliotecas da UFPA

Oliveira, Edson Farias de, 1966 –

Avaliação da distorção harmônica total de tensão no ponto de acoplamento comum industrial usando o processo KDD baseado em medição / Edson Farias de Oliveira – 2018.

Orientadora: Maria Emília de Lima Tostes.

Tese (Doutorado) – Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2018.

1. Energia elétrica – distribuição. 2. Sistemas de energia elétrica – controle de qualidade. 3. Harmônicos (ondas elétricas). 4. Inteligência computacional. I. Título.

CDD 23. ed. 621.319

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"AVALIAÇÃO DA DISTORÇÃO HARMÔNICA TOTAL DE TENSÃO NO PONTO DE ACOPLAMENTO COMUM INDUSTRIAL USANDO O PROCESSO KDD BASEADO EM MEDIÇÃO"

AUTOR: **EDSON FARIAS DE OLIVEIRA**

TESE DE DOUTORADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE DOUTOR EM ENGENHARIA ELÉTRICA NA ÁREA DE SISTEMAS DE ENERGIA ELÉTRICA.

APROVADA EM: 27/03/2018

BANCA EXAMINADORA:

_____
Presidente: Profa. Dra. Maria Emília de Lima Tostes - Orientadora

_____
Membro Interno ao PPGEE: Prof. Dr. Ubiratan Holanda Bezerra

_____
Membro Interno ao PPGEE: Profa. Dra. Adriana Rosa Garcez Castro

_____
Membro Externo ao PPGEE - UFOPA: Prof. Dr. Anderson Menezes

_____
Membro Externo ao PPGEE - FEE: Prof. Dr. Edson Ortiz de Matos

_____
Membro Externo ao PPGEE - UNIFAP: Prof. Dr. Werbeston Douglas de Oliveira

DEDICATÓRIA

A DEUS, Inteligência Suprema e Causa Primária de todas as coisas.

AGRADECIMENTOS

A Universidade Federal do Pará – UFPA

O Instituto de Tecnologia e Educação Galileo da Amazônia – ITEGAM

Ao Instituto de Tecnologia José Rocha Sérgio Cardoso – ITJRSC

Aos professores Dra. Maria Emília Tostes e Dr. Jandecy Cabral Leite
pela orientação e condução durante todo o curso

Aos amigos Carlos Freitas e Haroldo Melo
pelo incentivo para conclusão do curso.

Aos colegas Rildo de Mendonça Nogueira e Waterloo Ferreira da Silva que
contribuíram e participaram nesse projeto de pesquisa.

Aos colegas e todos os professores que contribuíram com o curso

Á minha mãe Maria Auxiliadora Farias de Oliveira,
que sempre me incentivou.

À minha querida esposa Lucilene e minhas filhas Luciana e Ellen,
pelo amor, apoio e compreensão.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PIM	Polo Industrial de Manaus
PC	*Personal Computer* (Computador Pessoal)
ITEGAM	Instituto de Tecnologia e Educação Galileo da Amazônia
ITJRSC	Instituto de Tecnologia José Rocha Sérgio Cardoso
UFPA	Universidade Federal do Pará
SUFRAMA	Superintendência da Zona Franca de Manaus
ZFM	Zona Franca de Manaus
PAC	Ponto de Acoplamento Comum
THD	*Total Harmonic Distortion* (Distorção Harmônica Total)
QEE	Qualidade da Energia Elétrica
KDD	*Knowledge Discovery in DataBase* (Descoberta do Conhecimento em Banco de Dados
SEP	Sistema Elétrico de Potência
ONS	Operador Nacional do Sistema
PRODIST	Procedimentos de Distribuição
ANEEL	Agência Nacional de Energia Elétrica
SRD	Superintendência de Regulação dos Serviços de Distribuição
IC	Inteligência Computacional
SVM	*Support Vector Machine*
MARS	*Multivariate Adaptive Regression Splines*
EPRI	*Electric Power Research Institute*
LPQI	*Leonardo Power Quality Initiative*
IEEE	*Institute of Electrical and Electronics Engineers*
IEC	*International Electrotechnical Comission*
PECI	*Power and Energy Conference at Illinois*

RESUMO

Nas últimas décadas, a indústria de transformação, tem proporcionado a introdução de produtos cada vez mais rápidos e energeticamente mais eficientes para utilização residencial, comercial e industrial, no entanto essas cargas devido à sua não linearidade têm contribuído significativamente para o aumento dos níveis de distorção harmônica de tensão em decorrência da corrente conforme indicadores de Qualidade de Energia Elétrica do sistema brasileiro de distribuição de energia elétrico. O constante aumento dos níveis das distorções, principalmente no ponto de acoplamento comum, tem gerado nos dias atuais muita preocupação nas concessionárias e nos consumidores de energia elétrica, devido aos problemas que causam como perdas da qualidade de energia elétrica no fornecimento e nas instalações dos consumidores e isso têm proporcionado diversos estudos sobre o assunto. Com o intuito de contribuir com o assunto, a presente tese propõe um procedimento com base no processo *Knowledge Discovery in Database - KDD* para identificação das cargas impactantes das distorções harmônicas de tensão no ponto de acoplamento comum. A metodologia proposta utiliza técnicas de Inteligência computacional e mineração de dados para análise dos dados coletados por medidores de qualidade de energia instalados nas cargas principais e no ponto de acoplamento comum do consumidor e consequentemente estabelecer a correlação entre as correntes harmônicas das cargas não lineares com a distorção harmônica no ponto de acoplamento comum. O processo proposto consiste na análise das cargas e do *layout* do local onde a metodologia será aplicada, na escolha e na instalação dos medidores de QEE e na aplicação do processo *KDD* completo, incluindo os procedimentos de coleta, seleção, limpeza, integração, transformação e redução, mineração, interpretação, e avaliação dos dados. Com o propósito de contribuição foram aplicadas as técnicas de mineração de dados Árvore de Decisão e *Naïve Bayes* e foram testados diversos algoritmos em busca do algoritmo com resultados mais significativos para esse tipo de análise conforme apresentado nos resultados. Os resultados obtidos evidenciaram que o processo *KDD* possui aplicabilidade na análise da Distorção Harmônica Total de Tensão no Ponto de Acoplamento Comum e deixa como contribuição a descrição completa de cada etapa desse processo, e para isso foram comparados com diferentes índices de balanceamento de dados, treinamento e teste e diferentes cenários em diferentes turnos de análise e apresentaram bom desempenho possibilitando sua aplicação em outros tipos de consumidores e empresas de distribuição de energia. Evidencia também, na aplicação escolhida e utilizando diferentes cenários, que a carga mais impactante foi a sétima harmônica de corrente das centrais de ar condicionado para o conjunto de dados coletados.

Palavras-chave: distorção harmônica total, mineração de dados, *KDD*, inteligência computacional, árvore de decisão, *Naïve Bayes*, rede neural, qualidade de energia.

ABSTRACT

In the last decades, the transformation industry has provided the introduction of increasingly faster and more energy efficient products for residential, commercial and industrial use, however these loads due to their non-linearity have contributed significantly to the increase of distortion levels harmonic of voltage as a result of the current according to the Power Quality indicators of the Brazilian electricity distribution system. The constant increase in the levels of distortions, especially at the point of common coupling, has generated in the current day a lot of concern in the concessionaires and in the consumers of electric power, due to the problems that cause like losses of the quality of electric power in the supply and in the installations of the consumers and this has provided several studies on the subject. In order to contribute to the subject, this thesis proposes a procedure based on the Knowledge Discovery in Database - KDD process to identify the impact loads of harmonic distortions of voltage at the common coupling point. The proposed methodology uses computational intelligence and data mining techniques to analyze the data collected by energy quality meters installed in the main loads and the common coupling point of the consumer and consequently establish the correlation between the harmonic currents of the nonlinear loads with the harmonic distortion at the common coupling point. The proposed process consists in analyzing the loads and the layout of the location where the methodology will be applied, in the choice and installation of the QEE meters and in the application of the complete KDD process, including the procedures for collection, selection, cleaning, integration, transformation and reduction, mining, interpretation, and evaluation of data. In order to contribute, the data mining techniques of Decision Tree and Naïve Bayes were applied and several algorithms were tested for the algorithm with the most significant results for this type of analysis as presented in the results. The results obtained evidenced that the KDD process has applicability in the analysis of the Voltage Total Harmonic Distortion at the Point of Common Coupling and leaves as contribution the complete description of each step of this process, and for this it was compared with different indices of data balancing, training and test and different scenarios in different shifts of analysis and presented good performance allowing their application in other types of consumers and energy distribution companies. It also shows, in the chosen application and using different scenarios, that the most impacting load was the seventh current harmonic of the air conditioning units for the collected data set.

Keywords: harmonic distortion, data mining, KDD, computational intelligence, decision tree, *Naïve Bayes*, neural network, power quality.

SUMÁRIO

1 INTRODUÇÃO

O aumento de eventos e distúrbios relacionados à Qualidade de Energia Elétrica (QEE) no Sistema de Energia Elétrica (SEE) tem tornado cada vez mais relevante à análise da QEE conforme afirmam Rodriguez-Guerrero *et al.* (2017). Para Kahle (2016), esse aumento está relacionado ao constante incremento da quantidade de equipamentos eletrônicos sensíveis nos sistemas de energia modernos. A Grande maioria dos trabalhos nessa área considera a afirmação de Arrillaga e Watson (2004), Bollen e Gu (2006) que os distúrbios harmônicos são os que mais degradam e os mais preocupantes, pois com o passar do tempo tendem a aumentar devido ao uso crescente de equipamentos baseados em conversores eletrônicos. É com base nesses acontecimentos que Iagar, Popa e Dinis (2014) consideram que a proliferação de cargas eletrônicas não lineares está aumentando dia a dia, e que pode antecipar no futuro um influxo de novas tecnologias e afirmam que as fontes chaveadas possuem alta distorção harmônica de corrente e estão presentes em quase todas as cargas não lineares domésticas e comerciais, tais como computadores, monitores, laptops, lâmpadas eletrônicas, etc.

De Araújo *et al.* (2015) e Granados-Lieberman *et al.* (2011) (citados por Rodriguez-Guerrero *et al.*, 2017) afirmam que durante as últimas décadas, aumentou significativamente a análise da qualidade de energia elétrica em sistemas de potência, principalmente devido ao interesse especial no desenvolvimento de novos métodos para medir com precisão o valor de alguns índices de qualidade de energia.

Em integração a todo esse quadro de avanços e impactos na QEE, tanto as empresas de transmissão como as de distribuição de energia elétrica possuem grande interesse em estudos associados à questão de distorções harmônicas presentes em seus respectivos sistemas, uma vez que com a vigência de normas relacionadas à qualidade da energia, como os procedimentos de rede do Operador Nacional do Sistema (ONS) para as transmissoras de energia o módulo 8 dos Procedimentos de Distribuição (PRODIST) desenvolvido pela Agência Nacional de

Energia Elétrica (ANEEL) para as concessionárias de distribuição, na qual a atual revisão entrou em vigência em primeiro de janeiro de 2017 (2016). Dentre os principais objetivos do PRODIST encontra-se o de garantir que os sistemas de distribuição operem com segurança, eficiência, qualidade e confiabilidade.

Empresas transmissoras e distribuidoras de energia se veem obrigadas a conhecerem melhor seus sistemas com relação a estes fenômenos para assim mantê-los em níveis adequados de operação e assim evitar multas pesadas em virtude do não cumprimento das normas vigentes. No entanto com relação à questão da atribuição de responsabilidades entre consumidores e concessionárias, pode-se dizer que ainda não há metodologias consolidadas que sejam capazes de estimar, com segurança, a contribuição harmônica das cargas geradoras de distorções harmônicas nos sistemas elétricos.

Apesar da legislação vigente não penalizar as indústrias pela geração de correntes harmônicas no sistema de energia, já há sinalização de que em breve a ANEEL divulgará nova legislação para avaliar os impactos oriundos das indústrias, conforme citação da nota técnica 0083/2012-SRD/ANEEL (2012), onde a mesma faz referência a norma australiana que impõe valores de referência para a distorção harmônica da tensão (a serem observados pela distribuidora) e outros à corrente harmônica de equipamentos conectados na rede (a serem observados pelo consumidor).

Matos (2016) afirma que é necessária a criação de meio legais para aplicações de sanções aos responsáveis pela geração de níveis de distorções harmônicas acima do permitido e que atualmente inexistem procedimentos para identificação de forma exata das principais fontes dessas distorções no SEE e consequentemente tomar ações para mitigar as distorções.

1.1 Objetivos

1.1.1 Objetivo geral

O objetivo geral da tese é propor uma metodologia utilizando o processo *KDD* para determinar os impactos na distorção harmônica (*THDv)* no ponto de acoplamento comum proporcionada pelas principais cargas instaladas na rede elétrica de um consumidor.

1.1.2 Objetivos Específicos

A presente tese tem como objetivos específicos:

- Descrever o processo *KDD* completo em cada etapa em um consumidor industrial, para deixar essa metodologia como contribuição nos estudos relacionados à análise de distorções harmônicas em outros tipos de consumidores;
- Aplicar e comparar os classificadores Árvore de Decisão e *Naïve Bayes* na etapa de mineração de dados para propor a técnica com as métricas mais significativas para a análise de distorções harmônicas em trabalhos futuros;
- Avaliar quais as principais cargas impactantes no *THDv* do PAC como antecipação à nova legislação para mitigações futuras e evitando assim possíveis penalidades e problemas relacionados a distorções harmônicas no consumidor.

1.2 Revisão bibliográfica

Com o propósito de oferecer uma visão geral sobre o tema em estudo, foi realizado uma pesquisa bibliográfica relacionada à QEE e a relação com metodologia KDD e as técnicas de mineração de dados, mais precisamente na abordagem em relação a distorção harmônica.

Nas seções seguintes são abordadas publicações que serviram base e justificativas para o desenvolvimento dessa tese.

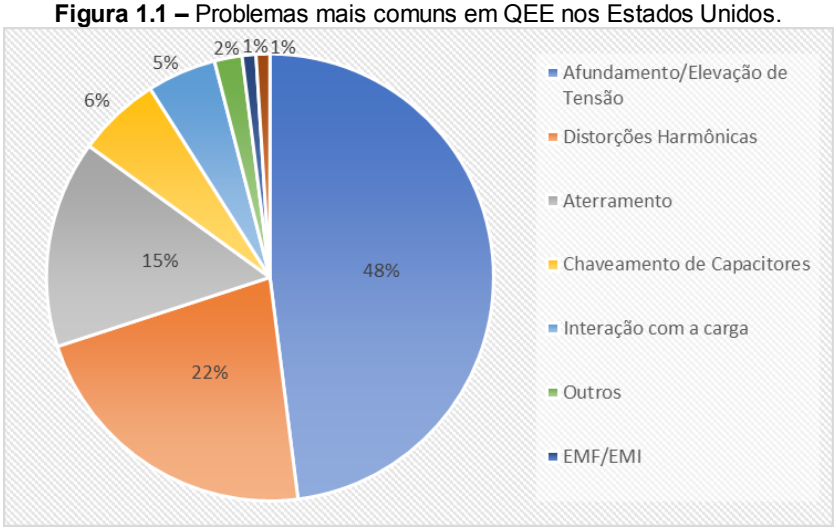
1.2.1 Qualidade de Energia Elétrica e Distorções Harmônicas.

Matos (2016) apresenta uma abordagem com diferentes publicações relacionadas a área de QEE com o propósito de analisar, prever e mitigar os vários problemas que envolvem as distorções harmônicas e a identificação das fontes dessas distorções no SEE e afirma que parece não haver na literatura registros de estudos relacionados a utilização de modelos de regressão não paramétrica para a identificação e quantificação dos impactos gerados por cargas não lineares na distorção harmônica de tensão no PAC .

A preocupação com a QEE é cada vez mais evidente e é decorrente, segundo Deckmann e Pomilio (2010), das ações para viabilizar o setor elétrico sugerem o monitoramento da QEE por meio de 5 passos específicos.

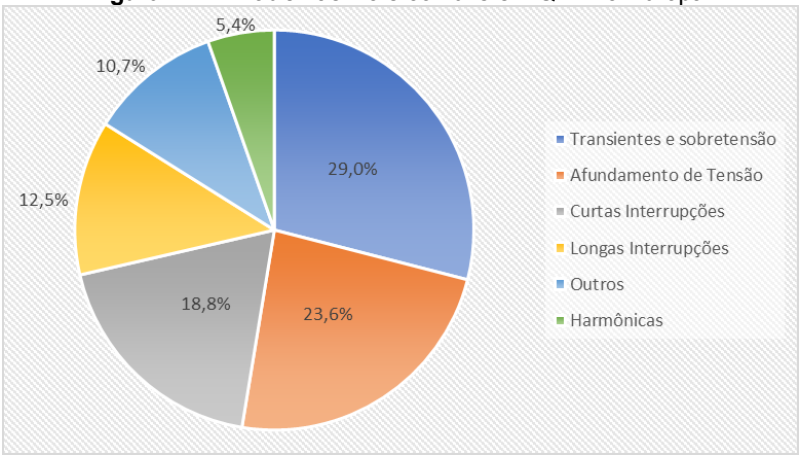
Ignatova *et al.* (2015) afirma que a clássica expressão “Aquilo que você não mede, não consegue gerenciar” é totalmente verdadeira com relação a QEE e apresenta os resultados de uma série de estudos de QEE nos Estados Unidos e na Europa coletados de uma variedade de usuários finais conforme mostrado nas Figuras 1.1 e 1.2.

Figura 1.1 – Problemas mais comuns em QEE nos Estados Unidos.



Fonte: Ignatova *et al.* (2015).

Figura 1.2 – Problemas mais comuns em QEE na Europa.



Fonte: Ignatova *et al.* (2015).

Esses resultados demonstram que nos Estados Unidos os distúrbios de QEE mais preocupantes são afundamento e elevação de tensão e as distorções harmônicas e enquanto na Europa são os transientes, sobretensão, afundamento de tensão e as interrupções, ficando as distorções harmônicas em último indicador. Desta forma Ignatova *et al.* (2015) com base nos dados dos estudos e nas centenas de auditorias realizadas anualmente recomenda monitoramento sistemático dos seguintes problemas de QEE:

- Distorções Harmônicas;
- Fator de Potência;
- Afundamento de tensão e interrupções;
- Transientes;
- Desbalanceamento.

De Almeida *et al.* (2003) apresenta diversos estudos realizados ao longo do tempo evidenciando os custos relacionados a problemas de QEE mostrado na Tabela 1.1.

Tabela 1.1 – Custos anuais estimados relacionados a problemas de QEE.

Entidade	Ano	Custo Anual	Local
Business Week	1991	26 bilhões de dólares	Estados Unidos
EPRI	1994	400 bilhões de dólares	Estados Unidos
US Department of Energy	1995	150 bilhões de dólares	Estados Unidos
Fortune Magazine	1998	10 bilhões de dólares	Estados Unidos
E Source	2001	Entre 60 e 80 bilhões de dólares	Estados Unidos
PQ costs in EU	2001	10 bilhões de euros	Europa

Fonte: De Almeida (2003).

Segundo dados do *Leonardo Power Quality Iniciativa (LPQI)*, o custo anual estimado em decorrência de perdas por distúrbios de QEE na Europa gira em torno de 150 bilhões de euros e os dados do *Electric Power Research Institute (EPRI)* estima as perdas nos Estados Unidos entre 119 bilhões e 188 bilhões de dólares, no entanto entre os dados encontrados pela *EPRI*, afirma Ignatova *et al*. (2015), o dado mais importante é o fato de 80% dos distúrbios de QEE estarem sendo gerados dentro das instalações dos consumidores. Evidenciando assim a necessidade de avançarmos as pesquisas para dentro das empresas.

1.2.2 Qualidade de Energia Elétrica e Inteligência Computacional.

Devido ao grande volume de dados obtidos através de medidores de QEE, é extremamente importante a utilização de técnicas de Inteligência Computacional (IC) nos estudos do SEE. Devido à aplicabilidade das técnicas de mineração de dados, muitos pesquisadores da área de SEE estão dedicando muita atenção nessa área de pesquisa conforme sinalizam Kazerooni *et al.* (2014) e que para Xu *et al.* (2014), isso é devido ao grande volume de dados e afirmam que por meio do processo KDD podem ser obtidos conhecimentos úteis. Tecnicamente, afirma Roiger (2017), o *KDD* é a aplicação do método científico para a técnica de mineração de dados, onde o processo completo inclui (em adição à mineração dos dados), a extração, a preparação dos dados e a tomada de decisão com os dados minerados e Prass (2009) acrescenta que as principais etapas do processo *KDD* necessitam da presença constante de pelo menos um especialista da área de interesse.

Rahmatian *et al.* (2017), utiliza-se de árvore de decisão e o algoritmo *Multivariate Adaptive Regression Splines (MARS)* como ferramentas de avaliação de estabilidade transitória e Seera *et al.* (2016) apresentam estudo sobre análise de QEE usando um modelo híbrido que compreende redes neurais com modelo *fuzzy* e Árvore de Decisão. Nessa mesma linha de pesquisa, Ferreira *et al.* (2016) apresentaram pesquisa citando diversas técnicas computacionais para diagnóstico em linhas de transmissão do SEE. Em trabalho utilizando metodologia de detecção e classificação de distúrbios relacionados à QEE, Borges *et al.* (2016 e 2013) utilizaram Árvore de Decisão como método de classificação e afirmaram com base

em diversas pesquisas correlatas que os métodos mais utilizados para classificação são: lógica fuzzy, redes neurais artificiais, *SVM* e Árvore de Decisão.

Guerrero (2017) afirma que os problemas de medição, análise e compensação da QEE são de grande importância e apresenta 14 trabalhos de alto nível cobrindo os 4 seguintes tópicos:

- Análise de problema de QEE;
- Controle para melhorar a QEE;
- Métodos avançados para avaliação da QEE;
- Técnicas de sincronização.

Dentre os trabalhos apresentados destacam-se o de Chakravorty *et al.* (2017) que fornece uma metodologia para visualizar a variação da impedância dentro de um ciclo devido a fontes chaveadas e propõe três índices para quantificar essa variação de impedância. Guerrero (2017) afirma, ainda, que para os métodos avançados de análise da QEE, Albu *et al.* (2017) propõe um algoritmo alternativo de agregação moldado à QEE a ser implementado em medidores inteligentes de QEE facilitando a análise em tempo real, enquanto propõe um método paramétrico para medição rápida para estimar fasores harmônicos aplicados à norma *IEEE* C37.118.1.

Uma nova técnica de classificação de perturbações de QEE é proposta por Oubrahim *et al.* (2017) para problemas de elevação e afundamento de tensão. Afirma que usando os valores de componentes simétricos estimados, pode classificar as perturbações e para resolver este problema, propuseram dois pré-classificadores com base em Critérios Teóricos da Informação.

A utilização de redes neurais com algoritmo Perceptron Multicamadas (do inglês *MultiLayer Perceptron-MLP*) é apresentada por Da Silva *et al.* (2017) para reconhecimento dos distúrbios relacionados à QEE e afirmam que a avaliação técnica e científica da QEE teve um grande avanço após o uso de sistemas inteligentes para identificação de diversos distúrbios.

Khokhar *et al.* (2017) afirmaram que é uma preocupação desafiadora tanto para a distribuidora como a indústria a classificação automática dos distúrbios de

QEE e propõe um novo algoritmo para classificação automática que consiste a seleção ótima através de Transformada Discreta *Wavelet* (*DWT*) e redes neurais probabilística baseada em colônia de abelha com dados simulados do sistema de distribuição da Malásia.

Entre os diversos trabalhos pesquisados sobre QEE, Matos (2016) aborda o problema de determinar quais cargas não lineares podem ser consideradas fontes harmônicas potenciais para as distorções harmônicas de tensão observadas nos valores de tensão da rede elétrica, para isso utiliza modelos de regressão linear e não paramétrica.

Naik *et al.* (2016) propuseram uma nova metodologia de classificação de eventos de QEE usando *WPT (Wavelet Packet Transform)* e *ELM (Extreme Learning Machine)* e afirmam que a *WPT* pode obter várias características matemáticas e que a nova metodologia foi comparada com o classificador baseado em redes neurais e a precisão alcançada foi maior que 99%.

Uma análise detalhada em artigos publicados entre os anos 1970 e 2013 na base de dados *Scopus* da Elsevier é apresentada por Montoya *et al.* (2016). Dentre os principais resultados na análise destacam-se que as palavras-chave mais encontradas além de QEE, foram harmônicos, filtro ativo, afundamento de tensão, geração distribuída e transformada *Wavelet*, os principais países que publicaram foram a China, Estados Unidos e Índia, sendo as principais instituições contribuintes o *IEEE*, o *Indian Institute of Technology* e a Universidade de Energia Elétrica do Norte da China e as principais contribuições nos campos da otimização, IC e processamento de sinais foram respectivamente os Algoritmos Genéticos, Otimização por Enxame de Partículas, Redes Neurais, Lógica *Fuzzy*, Transformada *Wavelet* e análise de Fourier.

Unsar *et. al*. (2014) publicaram dois artigos sobre a identificação da contribuição da distorção harmônica em uma planta de ferro e aço, onde a parte primeira trata da análise no PAC da indústria e a segunda o trabalho é desenvolvido dentro da planta. Nos referidos trabalhados os autores afirmam que as cargas internas são as grandes geradoras de distorções harmônicas devido às suas

características não lineares e citam com principais geradores os fornos a arco utilizados pela planta estudada. Afirmam, ainda, que as normas internacionais que tratam do assunto como a *IEEE* e *IEC* não fornecem informações para avaliações de contribuições de corrente harmônica individual da carga alimentadas pelo PAC para efeito da distorção total observada no PAC e que parte dessa contribuição é originária das cargas que não estão sendo monitoradas (*background*) e por outro lado afirmam que devido a utilização de filtros de distorções harmônicas, as medições podem ser incorretas. Os autores apresentam o método baseado no circuito equivalente Norton na enésima corrente harmônica.

Kazerooni *et al.* (2014) apresentaram um trabalho de relevante importância para pesquisas envolvendo SEE e Mineração de Dados. Afirmaram que devido ao grande volume de dados coletados através de medidores de QEE é extremamente importante a utilização de técnicas de IC nos estudos de SEE. Citaram que entre técnicas de Mineração de Dados mais utilizados estão as Redes Neurais, *SVM*, agrupamentos K-means, Árvores de Decisão e Técnicas de Visualização. Concluíram que devido à aplicabilidade das técnicas de Mineração de Dados muitos pesquisadores da área de SEE estão dedicando muita atenção nessa área de pesquisa.

Lourenço (2012) apresenta um estudo referente à avaliação da QEE no centro de tecnologia da Universidade Federal do Ceará, onde o autor descreve o processo de coleta utilizando o medidor MARH-21 da RMS e justifica com trabalhos que tratam do monitoramento da QEE como Marques (2011) e Dugan *et al.* (2004) e os que investigaram o uso de medidores como Melo (2010) e Solletto *et al.* (2008) e menciona o módulo 5 do PRODIST (2011). Cita também trabalhos recentes sobre o assunto como Moura (2010), Barros (2010), Vale (2011), Oliveira (2008) e Lopes (2011). Para a análise e diagnóstico o autor utilizou diversas tabelas e gráficos em Excel e propõe um plano de ações corretivas. Excelente material que evidencia a utilização de medidores de QEE, desenvolve análise com base nos dados coletados, no entanto não utilizam técnicas de IC.

Ribeiro e Rocha (2012) apresentaram um trabalho onde teve como objetivo analisar os efeitos das correntes harmônicas em uma fábrica de celulose na situação

de inserção de banco de capacitores para correção do fator de potência na planta e para isso utilizaram o programa *HarmZs* desenvolvido pelo CEPEL. No referido trabalho a ênfase foram as cargas instaladas no secundário do transformador do PAC da indústria e afirmam que esse tipo de estudo tem a finalidade de antecipar a mitigação dos impactos nos indicadores de qualidade de energia elétrica que serão em breve implantados pela ANEEL.

Em Ozgonenel *et al.* (2012), foi proposto um trabalho de diagnóstico de falhas em SEE, com principal objetivo de desenvolver um algoritmo de classificação de sinal para diferentes tipos de distúrbios de QEE (falhas), tendo como base nas mais recentes melhorias incluindo as técnicas de processamento de sinais e reconhecimento de padrões.

Marques (2011) em sua tese de doutorado afirmou que os harmônicos são os principais impactantes no SEE e em seu trabalho desenvolveu técnicas de processamento de sinais para a estimação da frequência e harmônicos de um SEE utilizando o conceito de modulação. Marques (2011) afirmou, também, que pelos resultados obtidos, e pela flexibilidade de projeto, é possível sugerir que as técnicas propostas possam ser utilizadas em diversas aplicações de SEE e *smart grids*, tais como medição, proteção, controle e monitoração da QEE.

Yin *et al.* (2011) contribuem com a pesquisa apresentado trabalho onde é proposto um novo método para identificação de fonte de distorções harmônicas. No referido trabalho os autores fazem uma busca em trabalhos correlatos procurando encontrar ferramentas que tratam do assunto e propuseram o método da regressão linear múltipla e o método de análise de correlação parcial para a identificação das fontes de distorções harmônicas. Também afirmam que recentemente tem-se utilizado Redes Neurais Artificias na busca pelas fontes de distorções harmônicas.

No trabalho apresentado por Santos e De Oliveira (2011) teve por objetivo propor e explorar a superposição das tensões com vistas a fornecer uma avaliação de desempenho, experimental em laboratório, da metodologia, fundamentada no método da superposição, para a atribuição das responsabilidades quanto à geração das distorções harmônicas no PAC entre a rede da concessionária e os

consumidores. Afirmaram, ainda, que os resultados obtidos evidenciaram que o emprego da superposição de correntes para o compartilhamento das tensões harmônicas se mostrou inadequados.

Já em Souza (2010) é utilizada a IC aplicada na identificação e classificação de problemas de medição de energia elétrica e no seu desenvolvimento apresentaram a integração de elementos da IC com ensaios experimentais, visando identificar os efeitos e as causas das interferências nas medições de energia. No referido trabalho os autores utilizaram Redes Neurais Artificiais como técnica de análise e no mapeamento dos erros de medição em ambientes repletos de conteúdo harmônico e também desenvolveram uma interface gráfica aplicada ao *MATLAB* e afirmaram que a metodologia proposta atingiu os seus objetivos e recomendaram o aprofundamento da mesma.

Ferreira (2010) em sua tese de doutorado também aplicou técnicas de Processamento de Sinais, Estatísticas e de IC para análise, detecção e classificação dos distúrbios elétricos no SEE. Para essa classificação usou a técnica estatística Curvas Principais e para o pré-processamento utilizou a técnica Estatísticas de Ordem Superior (EOS). O autor também utilizou técnica de IC no desenvolvimento de sistema automático de classificação de distúrbios baseado em Redes Neurais Especialistas.

Fernandes *et al.* (2010) fizeram uma análise significativa sobre técnicas para identificação de cargas lineares e não lineares aplicadas em um sistema elétrico residencial e afirmaram que atualmente existem muitos trabalhados utilizando diversas técnicas para estudos em ambientes industriais e que o ambiente residencial começa a ser objeto de estudos. Para isso utilizaram técnicas para seleção de atributo e Redes Neurais Artificiais em um experimento em laboratório simulando as cargas residenciais como lâmpadas incandescentes, lâmpada fluorescente, lâmpada fluorescente compacta, ventilador, computador pessoal e monitor. Esses ensaios, onde os dados foram coletados por medidores de qualidade de energia, forneceram dados reais de cargas normalmente utilizadas em residências, os quais possibilitaram o treinamento, validação das RNAs e finalizando com resultados satisfatórios.

Kandev e Chénard (2010) apresentaram um método para identificar a fonte das emissões harmônicas e para determinar a sua contribuição na distorção harmônica em diversos pontos de uma rede de energia elétrica. Teve como objetivo localizar em tempo real o consumidor responsável pelo distúrbio harmônico na rede e para isso propuseram um método baseado na análise do vetor de corrente harmônica e usando dados de medições sincronizadas em múltiplos pontos da rede elétrica. O método do experimento mostrou-se eficaz conforme afirmaram os autores, no entanto indicaram que seria necessário um aprofundamento para múltiplos consumidores.

1.2.3 Qualidade de Energia Elétrica e Árvore de Decisão.

Os classificadores transformada de *Stockwell,* agrupamento *Fuzzy C-Means (FCM) e* Árvore de Decisão são propostos por Mahela *et al.* (2017) para detecção e classificação de distúrbio de QEE. O método utiliza os dados simulados *do IEEE* 1159 no *MATLAB* e o desempenho é comparado com a técnica Árvore de Decisão baseada em transformada S. Os distúrbios de QEE investigados por Mahela *et al.* (2017) incluem elevação de tensão, afundamento de tensão, interrupção, ruído, distorções harmônicas, pico, cintilação, transiente impulsivo e transitório oscilatório.

Da Silva Pessoa e Oleskovicz (2017) propõe a Árvore fe Decisão com o algoritmo J48 para localizar as situações de curto-circuito monofásicas. Os dados utilizados foram compilados por simulações no barramento *IEEE* 34 usando o programa *ATP (Alternative Transients Program)* e utilizado o *software WEKA (Waikato Environment for Knowledge Analysis)* para aplicação do classificador Árvore de Decisão.

Em Ray *et al.*(2014), o estudo abordou a influência sofrida pela penetração dos sistemas destruídos nos sistemas convencionais de energia que causam distúrbios de QEE. Desenvolveram uma melhor classificação dos distúrbios de QEE que está associada a vários fatores como as mudanças de cargas e fatores ambientais. Várias formas de distúrbios são levadas em consideração incluindo afundamentos de tensão, elevação de tensão, cortes de tensão e distorções harmônicas, foram utilizadas várias técnicas como algoritmos genéticos, *SVM* e

Árvore de Decisão, para apoiar na classificação de distúrbios. O estudo foi apoiado por três diferentes estudos de caso, um protótipo de montagem experimental a energia elétrica utilizando energia eólica, um sistema fotovoltaico e um sistema Nórdico de 32 barras.

Borges (2013) apresentou trabalho utilizando metodologia de detecção e classificação de distúrbios relacionados à QEE, no qual utilizou Árvore de Decisão como método de classificação e afirmou com base em diversos trabalhos correlatos que os métodos mais utilizados para classificação para tal são Lógica *Fuzzy*, Redes Neurais Artificiais, *SVM* e Árvore de Decisão. Apresentou, também, em forma de tabela, um resumo relacionando 24 trabalhos publicados, as técnicas de classificação utilizadas e os distúrbios de QEE classificados. O método apresentado no trabalho evidenciou a capacidade de diagnosticar a presença ou não do distúrbio em uma janela e possuindo de forma vantajosa rapidez e tamanho reduzido. A Árvore de Decisão apresentada foi implementada usando a ferramenta computacional *WEKA* e o algoritmo C4.5. O autor afirmou que utilizou metodologia de cálculos simples divergente das metodologias convencionais que utilizam pré-processamento, no entanto afirmou que os 15 distúrbios analisados foram com base em sinais modelados por meio de equações paramétricas e inseridos no banco de dados analisados e propõe para trabalhos futuros que a metodologia seja aplicada em dados reais.

Em Rodríguez *et. al.* (2013), é apresentado uma classificação automática com base em *S-transform* como ferramenta de extração de características e Árvore de Decisão como algoritmo classificador. Os sinais gerados de acordo com os modelos matemáticos, incluindo distúrbios complexos, têm sido utilizados para conceber e testar esta abordagem, em que o ruído é adicionado aos sinais de 40dB a 20dB.

Um estudo para identificação automática da origem de fontes das distorções harmônicas, através da injeção de correntes harmônicas utilizando um diodo retificador trifásico é apresentado por Vaid *et al.* (2011), onde o método de impedância crítica é utilizado para detecção de distúrbios causadores de distorções

harmônicas e classificador Árvore de Decisão foi empregado para identificar automaticamente a fonte das distorções harmônicas.

Silva (2010) desenvolveu trabalho propondo uma nova metodologia para determinação de falhas em medidores eletromecânicos de energia elétrica utilizando duas técnicas de Mineração de Dados: regressão *stepwise* e Árvore de Decisão. No referido trabalho utilizou a base de dados de uma concessionária de energia elétrica e utilizou o processo *KDD*. O trabalho é significativo para este trabalho, pois o autor traz embasamento sobre o processo *KDD* e comenta que comumente tem-se confundido o *KDD* com a etapa de Mineração de Dados.

Já Samantaray (2010), justificou o desenvolvimento de um estudo utilizando Árvore de Decisão com uso de regras *fuzzy* para classificar os diferentes eventos de QEE. Tais eventos são componentes de alta e baixa frequência, sendo difícil classificar através de métodos tradicionais, a Árvore de Decisão foi utilizada para classificar os eventos com uso de regras *fuzzy*.

1.2.4 Qualidade de Energia Elétrica e Naïve Bayes.

A classificação de falhas em micro redes por meio da combinação entre funções *Wavelet* e Aprendizagem de Máquina (*Machine Learning*) é apresentado por Abdelgayed *et al.* (2017), onde a combinação ideal é identificada por meio de otimização por enxame de partículas. Nesse trabalhou os autores automatizaram o processo por meio de quatro técnicas diferentes (Árvore de Decisão, vizinho mais próximo, *SVM* e *Naïve Bayes*).

Hasan *et al.* (2017) investigaram o uso de quatro poderosos classificadores de aprendizado de máquina (Bagging, Boosting, Funções de Base Radial e *Naïve Bayes*) para detectar e prever falhas em uma linha de transmissão de energia de 750kV e afirmam que as descobertas mostraram que o uso da técnica de aprendizado da máquina pode ser viável para essa tarefa e pode representar uma ótima oportunidade para aumentar a proteção e a eficiência do SEE.

Gomes (2011) em sua dissertação propôs um modelo que projeta em um espaço bidimensional o comportamento nominal dos sinais de tensão e corrente e trabalhou na aplicação da metodologia para detectar o momento da falta e sua classificação em um SEE utilizando análise funcional e IC. Embasando a sua análise, o autor afirmou que nos últimos anos vários métodos têm sido propostos para detecção e classificação de faltas como: Redes Neurais Artificiais, Transformada de Fourier, Transformada *Wavelet* ou uma combinação dessas técnicas. Afirmou, ainda, que atualmente métodos inteligentes utilizando IC como: Redes Neurais Artificias, Lógica *Fuzzy* e Redes *Neurofuzzy* estão sendo propostos. O autor também utilizou para o desenvolvimento do seu trabalho o *software WEKA (Waikato Environment for Knowledge Analysis)* com o classificador *Bayes Net* e também o modelo híbrido *DNTB* (*Decision Table and Naïve Bayes*).

1.2.5 Contribuições da tese.

A pesquisa realizada entre 2010 até a presente data parece evidenciar que não há, na literatura, registros de estudos que abordem a análise dos impactos na distorção harmônica de tensão no PAC de indústrias proporcionada por cargas não lineares, por meio do processo *KDD*. Com o propósito de contribuir com essa oportunidade de estudo, a tese apresenta a identificação e quantificação dos impactos na distorção harmônica de tensão no PAC por meio do processo *KDD*, definidos por Xu *et al.* (2017) e Roiger (2017) que tem como base Fayyad (1996), aplicado aos dados reais coletados por medidores de QEE instalados em cargas de uma indústria de transformação de computadores.

A tese também contribui com a interpretação e avaliação dos resultados que foi delimitado com base nas normas presentes no módulo 8 Revisão 8 (PRODIST/ANEEL, 2016), a *IEEE Std.* 519 – 92 (*IEEE*, 2004), *IEC 61000-3-6* (*IEC/TR*, 2008-02), curva *ITIC* e informações da fabricante de processadores INTEL com diversas comparações por meio da utilização das técnicas de mineração de dados Árvore de Decisão e *Naïve Bayes*.

A análise com diferentes cenários considerando em diversos turnos, a aplicação das técnicas de treinamento *cross-validation* e *training split* em diferente

condições, a utilização da técnica de balanceamento de classe SMOTE com diferentes percentuais em diferentes análise e a utilização dos programas de mineração de dados *WEKA* e *Rapid Miner* para a validação dos dados, também fazem parte da contribuição dessa tese.

1.3 Estrutura da Tese

A presente tese é dividida em capítulos, estruturada conforme descrição a seguir.

No primeiro capítulo são apresentados o tema, os objetivos a serem alcançados, o referencial teórico além das contribuições que o presente trabalho apresenta ao meio científico. Este capítulo visa evidenciar os conceitos tratados nessa tese e por fim apresenta a estruturação da mesma.

"Qualidade de Energia Elétrica – QEE" é o título do segundo capítulo, onde são abordados os conceitos básicos de QEE, as distorções harmônicas com causas e consequências e legislação e normas que regulamentam os distúrbios proporcionando as bases para o desenvolvimento da tese.

No terceiro capítulo são apresentados os principais conceitos de Inteligência Computacional, onde é apresentado os aspectos gerais do processo *KDD* etapa por etapa desde a seleção dos dados até a interpretação e avaliação. Na etapa de pré-processamento são apresentados os conceitos e a importância do balanceamento de classes. São consideradas, também, diversas técnicas de mineração de dados e são detalhados os classificadores Árvore de Decisão e Naïve Bayes para entendimento das técnicas utilizadas nessa tese. As principais métricas encontradas na literatura para avaliação do desempenho são apresentados no terceiro capítulo.

A metodologia aplicada utilizada na pesquisa é apresentada no quarto capítulo, descrevendo o meio onde foi desenvolvida a mesma com o detalhamento das cargas, layout e localização dos pontos, a seleção dos equipamentos de coleta e todas as fases do processo de descoberta do conhecimento especificamente no

processo *KDD* e nas técnicas de mineração e todos os algoritmos que foram utilizados nos classificadores Árvore de Decisão e Naïve Bayes.

Os resultados encontrados na referida pesquisa são apresentados no quinto capítulo seguindo as etapas do quarto capítulo e demonstrado os resultados com diferentes cenários para a validação dos mesmos.

Finalmente, no sexto capítulo, são apresentados os resultados gerais encontrados durantes a análise dos resultados de cada etapa e também são apresentadas sugestões para os trabalhos futuros que foram observados durante o desenvolvimento dessa tese.

2 QUALIDADE DE ENERGIA ELÉTRICA - QEE

2.1 Considerações Iniciais

Nesse capítulo fará abordagem sobre os conceitos gerais da QEE. O interesse pela QEE, segundo Bollen (2000), origina-se em 1968 com a publicação de um trabalho que detalhava um estudo da marinha dos Estados Unidos para a especificação da energia para um equipamento eletrônico. Em outra publicação Bollen (2009) afirma que o termo QEE, em geral, compreende a qualidade das formas de onda da tensão e corrente elétricas em um SEE.

Para Sarmanho (2005), o conceito de QEE tem se modificado ao longo dos últimos anos. No decorrer deste período, foram incorporados novos parâmetros para a sua determinação. Estes parâmetros foram originados por meio de normas estabelecidas no âmbito nacional e internacional, exigindo que as empresas do setor elétrico estejam constantemente adequando-se a estas novas regras.

Nas últimas duas décadas, afirma Ferreira (2010), a QEE tem interessado aos pesquisadores devido aos seguintes fatores:

• O uso generalizado de equipamentos computadorizados e também sensíveis às perturbações do sistema elétrico;

• Os consumidores estão cada vez mais exigentes demandando energia de melhor qualidade;

• As instalações elétricas residenciais, comerciais e industriais aumentaram o número de cargas de natureza não linear como sistemas microprocessados, fontes chaveadas e inversores de frequência que estão provocando deformidades nas formas de onda da tensão e corrente nos seus respectivos sistemas de distribuição;

• Finalmente a regulamentação do setor elétrico, no que diz respeito à QEE, em fase de implantação.

Para o *IEEE Std.* 1159-1995 (1995), a QEE se refere a uma grande variedade de fenômenos eletromagnéticos, que caracterizam a tensão e a corrente

em um determinado momento. Estas deformidades são definidas pelo *IEC* (*International Electrotechnical Comission*, IEC-91)(1991) como fenômenos eletromagnéticos ou simplesmente distúrbios, dentre os quais se destacam os harmônicos, Inter harmônicos, flutuações de tensão, afundamentos, transitórios, entre outros conforme visto na Tabela 2.1.

Tabela 2.1 - Principais fenômenos causados por distúrbios eletromagnéticos classificados pelo IEC-91.

Fenômenos Conduzidos de Baixa Frequência	Harmônicos, inter-harmônicos, Sinais de sistemas (portadoras), Flutuações na tensão, Afundamentos de tensão e interrupções, Desbalanceamento de tensão, Variação da frequência da rede elétrica, Tensões induzidas de baixa frequência e CC em rede CA
Fenômenos Irradiados de Baixa Frequência	Campos Magnéticos e Campos Elétricos
Fenômenos Conduzidos de Alta Frequência	Ondas contínuas conduzidas de tensão e corrente, Transitórios unidirecionais e Transitórios oscilatórios
Fenômenos Irradiados de Alta Frequência	Campos Magnéticos, Campos Elétricos, Campos Eletromagnéticos, Ondas Contínuas e Transitórios
Fenômenos de Descargas Eletrostática	
Pulsos Eletromagnéticos Nucleares	

Fonte: Elaborada pelo autor (2017).

Outra definição mais atual é a de Leão (2014) que afirma “que a QEE é a condição do sinal elétrico, de corrente e tensão, que permite que equipamentos, processos, instalações e sistemas elétricos operem de forma satisfatória, sem prejuízo de desempenho e de vida útil”. Os principais distúrbios presentes em um sistema elétrico conforme as normas *IEEE Std* 1159, 1995 e *IEEE Std* 446,1999 são mostrados na Tabela 2.2.

Tabela 2.2 - – Categorias e características dos distúrbios presentes na rede elétrica.

CATEGORIA	ESPECTRO	DURAÇÃO	TENSÃO
1.1 Transitórios Impulsivos			
1.1.1 Nano segundos	5 ns pico	< 50 ns	
1.1.2 Microssegundos	1 us pico	50 s – 1ms	
1.1.3 Milissegundos	0,1 ms pico	> 1ms	
1.2 Transitórios Oscilatórios			
1.2.1 Baixa Frequência	< 5 kHz	0,3 – 50 ms	0 – 4 pu
1.2.2 Média Frequência	5 – 500 kHz	20 us	0 – 8 pu
1.2.3 Alta Frequência	0,5 – 5 MHz	5 us	0 – 4 pu
2 Variações de Tensão de Curta Duração			
2.1 Instantâneo			
2.1.1 Afundamento (*Sag*)		0,5 – 30 ciclos	0,1 – 0,9 pu
2.1.2 Elevação (*Swell*)		0,5 – 30 ciclos	1,1 – 1,4 pu
2.2 Momentâneo			
2.2.1 Interrupção		0,5 ciclo – 3 s	< 0,1 pu
2.2.2 Afundamento (*Sag*)		30 ciclos – 3 s	0,1 – 0,9 pu
2.2.3 Elevação (*Swell*)		30 ciclos – 3 s	1,1 – 1,4 pu
2.3 Temporário			
2.3.1 Interrupção		3 s – 1 min	< 0,1 pu
2.3.2 Afundamento (*Sag*)		3 s – 1 min	0,1 – 0,9 pu
2.3.3 Elevação (*Swell*)		3 s – 1 min	1,1 – 1,4 pu
3 Variações de Tensão de Longa Duração			
3.1 Interrupção Permanente		> 1 min	0,0 pu
3.2 Subtensão		> 1 min	0,8 – 0,9 pu
3.3 Sobretensão		> 1 min	1,1 – 1,2 pu
4 Distorção da Forma de Onda			
4.1 Harmônicos	0 – 100 kHz	Est. Permanente	0 – 20 %
4.2 Inter-harmônicos	0 – 6 kHz	Est. Permanente	0 – 2 %
4.3 Recortes de comutação (*notches*)		Est. Permanente	
4.4 Ruído	Banda Larga	Est. Permanente	0 – 1 %
5 Flutuação de Tensão (*Flicker*)	< 25 Hz	Intermitente	0,1 – 7 %

Fonte: Normas *IEEE* 1159 (1995) e *IEEE* 446 (1999).

2.2 Harmônicos

Dugan *et al.* (2004) afirmam que segundo Fourier, uma função qualquer contínua e periódica, em um intervalo qualquer, pode ser representada por um somatório de componentes senoidais e uma componente constante. Segundo Kamenka (2014), a noção de harmônicos no sentido elétrico tornou-se conhecida na segunda metade do século XX, quando a componente senoidal é da mesma

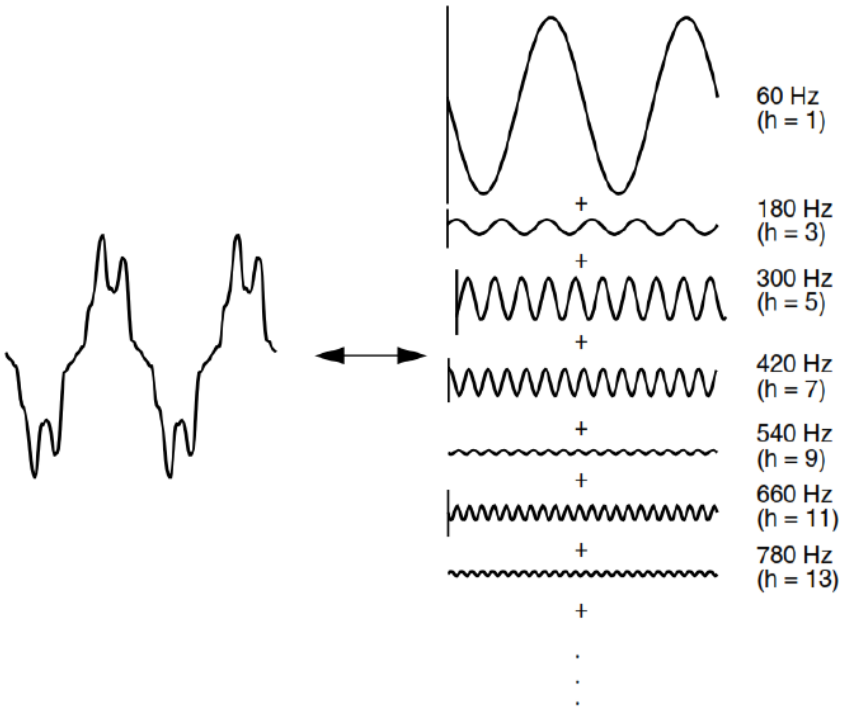
frequência do sinal original denomina-se fundamental e para as demais componentes senoidais, cujas frequências são múltiplas inteiras da frequência fundamental, são denominadas frequências harmônicas conforme mostrado na Tabela 2.3 e na Figura 2.1.

Tabela 2.3 – Frequências harmônicas.

Ordem Harmônica f_n	Frequência em rede de 50 Hz	Frequência em rede de 60 Hz
1ª	50	60
2ª	100	120
3ª	150	180
4ª	200	240
5ª	250	300
...	...	...
Nª	50 x n	60 x n

Fonte: Kamenka (2014).

Figura 2.1 – Série de Fourier representando uma forma de onda distorcida.

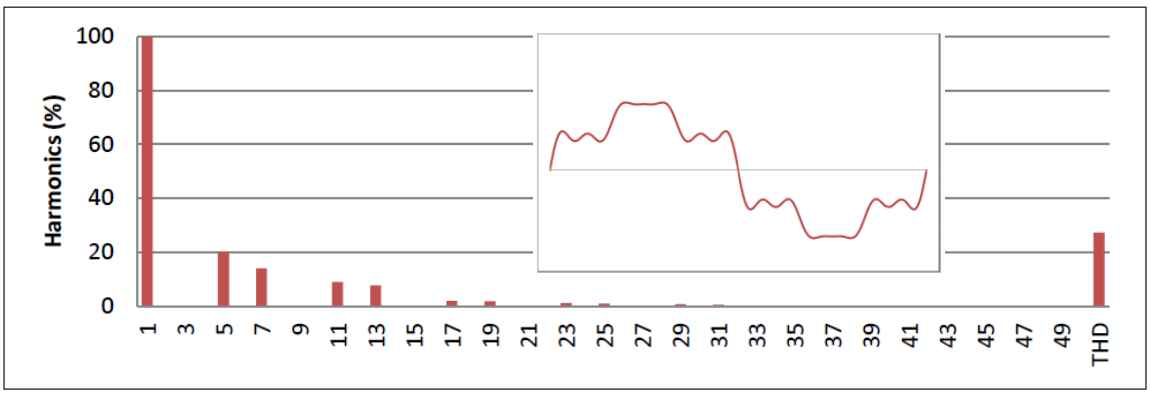


Fonte: Dugan *et al.* (2004)

Para Kamenka (2014), uma forma de onda distorcida sempre pode ser representada como a superposição de uma forma de onda da frequência fundamental com outras formas de onda de diferentes frequências e amplitudes harmônicas. O espectro harmônico mostrado na figura 2.2 é uma boa maneira de

visualizar a decomposição da forma de onda da figura 2.1. Esse tipo de espectro também é usado por quase todos os dispositivos de medição de qualidade de energia.

Figura 2.2 – Onda distorcida e o espectro da onda distorcida



Fonte: Kamenka (2014).

2.2.1 Classificação dos Harmônicos

Os harmônicos podem existir como harmônicos de tensão ou corrente. Existem duas noções comumente usadas para descrevê-las: a noção de componentes simétricos e as ordens harmônicas. A Tabela 2.4 mostra as ordens harmônicas (Peek, 1921; Hearn, 2010; Anar, 2009).

Tabela 2.4 – Ordens dos harmônicos.

	Ímpares	**Pares**	**Triplas**
Ordem dos Harmônicos	5^{th},7^{th}, 11^{th}, 13^{th}, 17^{th}	2^{nd},4^{th}, 6^{th}, 8^{th}, 10^{th}, 12^{th}	3^{rd},9^{th}, 15^{th}, 21^{st}

Fonte: Peek (1921), Hearn (2010) e Anar (2009).

Em um sistema trifásico balanceado mostra a relação simples entre a ordem da harmônica e a sequência de fase, assim sendo podem ser classificadas em harmônicos de sequência positiva, harmônicos de sequência negativa e harmônicos de sequência zero. A Tabela 2.5 mostra as relações entre componentes simétricos e ordens harmônicas.

Tabela 2.5 – Componentes simétricos.

Componentes Simétricos	Sequência Positiva (+)	Sequência Negativa (+)	Sequência Zero (homopolar)
Ordem do Harmônico	1	2	3
	4	5	6
	7	8	9
	10	11	12
	...	...	...
	3k+1	3k+2	3k+3

K=0,1,2,3...

Fonte: Peek (1921), Hearn (2010) e Anar (2009).

Os harmônicos de sequência positiva (4ª, 7ª, 10ª, ...) possuem a mesma rotação de fase que a frequência fundamental. Esses harmônicos circulam entre as fases. Os harmônicos de sequência negativa (2ª, 5ª, 8ª, ...) têm a rotação de fase oposta em relação frequência fundamental. Esses harmônicos circulam entre as fases. Os harmônicos de sequência zero (3ª, 6ª, 9ª, ...) não produzem um campo rotativo. Esses harmônicos circulam entre a fase e o neutro ou o terra. Ao contrário das correntes harmônicas de sequência positiva e negativa, os harmônicos de terceira ordem ou de sequência zero não cancelam, mas somam aritmeticamente no barramento neutro.

2.2.2 Relação entre corrente harmônica, tensão e impedância

Muitos dispositivos podem ser fontes de corrente harmônica, entretanto, além das correntes harmônicas individuais, geralmente a distorção harmônica total no PAC é o que precisa ser analisado. Assim, a interação entre corrente e tensão é um tópico importante que precisa ser entendido quando se trata de problemas de harmônicos. A distorção da tensão resultante da propagação de correntes harmônicas em um sistema de energia depende das características de fontes harmônicas e das características de todos os dispositivos conectados à rede elétrica que precisam ser analisados. A influência da impedância é bem explicada pela lei de Ohm (U = Z * I). A soma das correntes harmônicas dos dispositivos individuais não é direta, pois a distorção harmônica total da corrente também depende da mudança de fase entre as diferentes correntes da mesma frequência. Esta mudança de fase pode ser próxima de zero ou perto de 180 graus, causando uma enorme diferença pelo resultado da soma.

2.2.3 Efeito das Distorções Harmônicas

Conforme a norma *IEEE Std* 519 (1992), a circulação de correntes harmônicas nas redes elétricas deteriora a QEE causando inúmeros prejuízos, tais como:

- Aumento da corrente eficaz nas redes de distribuição causando sobrecarga e operação inadequada de fusíveis e relés de proteção;
- Sobreaquecimento e vibrações nas máquinas rotativas (motores e geradores), devido às perdas no ferro e cobre, afetando o torque e eficiência da máquina, reduzindo sua vida útil;
- Aquecimento dos cabos de alimentação devido ao aumento da frequência de correntes;
- Ruídos e interferências nas linhas telefônicas ou redes de comunicação;
- Mau funcionamento de aparelhos eletrônicos que possuem retificadores, pois resultam em uma tensão de saída menor e pior fator de potência;
- Mau funcionamento ou ineficiência de aparelhos eletrônicos formados por inversores, pois procedem em falhas de operação por curto-circuito interno oriundo de erros de comutação;
- Aumento da temperatura do filamento de lâmpadas incandescentes, devido ao acréscimo no valor eficaz de tensão, reduzindo sua vida útil.

2.2.4 Distorção Harmônica Total

De uma forma geral, qualquer componente harmônico pode ser representado como uma porcentagem da frequência fundamental (% fundamenta) ou uma porcentagem do valor *rms* (% *rms*) da corrente total com a seguinte equação 2.1:

$$I_h = \frac{I_n}{I_1} \times 100\% \qquad (2.1)$$

Onde:

I_n... amplitude da corrente harmônica n

I_1... amplitude da corrente fundamental (ou o valor *rms* da corrente total)

Para a tensão é a mesma abordagem.

Segundo o PRODIST, a distorção harmônica total (*THD*) é a composição das distorções harmônicas individuais que expressa o grau de desvio da onda em relação ao padrão ideal, normalmente referenciada ao valor da componente fundamental.

Este valor *THD* é usado para sistemas de baixa, média e alta tensão. Normalmente, a distorção de corrente é definida como *THDi* e a distorção de tensão como *THDv*, sendo esses índices calculas pelas expressões (2.2) e (2.3):

$$THDv = \frac{\sqrt{\sum_{h=2}^{\infty} V_h^2}}{V_n} x\ 100\% \qquad (2.2)$$

Onde:

V_h = valor *rms* da componente harmônica de tensão de ordem h;

h = ordem harmônica;

V_n = valor *rms* da tensão fundamental.

$$THDi = \frac{\sqrt{\sum_{h=2}^{\infty} I_h^2}}{I_n} x\ 100\% \qquad (2.3)$$

Onde:

I_h = valor *rms* da componente harmônica de corrente de ordem h;

h = ordem harmônica;

I_n = valor *rms* da corrente fundamental.

Tostes (2003) afirma que normalmente, os limites de distorção harmônica, tolerados nos sistemas de distribuição, são expressos por esses índices, de acordo com as normas que regulamentam a qualidade da energia elétrica em todo o mundo.

2.2.5 Cargas Lineares e Não Lineares

As cargas lineares mostram os sinais de tensão e corrente se seguindo muito próximo. Em um circuito de corrente alternada, isso significa que a aplicação de uma tensão senoidal resulta em uma corrente senoidal. À medida que a tensão instantânea muda ao longo do período da onda senoidal, a corrente instantânea aumenta e cai proporcionalmente à tensão, de modo que a forma de onda da corrente torne-se também uma onda senoidal. Este comportamento de tensão e corrente pode ser explicado com a lei de Ohm que afirma que a corrente através de uma resistência alimentada por uma fonte de tensão variável é igual à relação entre a tensão e a resistência (*IEEE* 519,1992; BAGGINI, 2008; *IEC* 61000-4-7, 2002 e Arrillaga e Watson, 2004).

A Tabela 2.6 mostra algumas cargas lineares, onde é possível observar que as cargas com as duas formas de onda em fase um com o outro (carga resistiva), mas também com a tensão avançada (carga indutiva) ou a corrente principal (carga capacitiva) são consideradas lineares, porque mesmo que as duas formas de onda estejam fora de fase de um ao outro, nenhuma distorção de forma de onda pode ser encontrada.

Tabela 2.6 – Cargas lineares.

Cargas Resistivas	Cargas Indutivas	Cargas Capacitivas
Lâmpadas Incandescentes	Motores de Indução	Correção de Fator de Potência
Aquecedores Elétricos	Geradores de Indução	

Fonte: Kamenka (2014).

Uma carga é considerada não linear se a corrente projetada pela carga não for senoidal mesmo quando estiver conectada a uma tensão senoidal. Essa corrente não linear contém componentes de frequência que são múltiplos da frequência do sistema de energia. Essas correntes harmônicas interagem com a impedância da rede de energia elétrica para criar distorção de tensão que pode afetar a própria rede elétrica e as cargas conectadas a ela. A Tabela 2.7 fornece uma lista com alguns dispositivos não lineares.

Tabela 2.7 – Cargas não lineares.

Eletrônica de Potência	Dispositivo de Arco
Variadores de velocidade	Máquinas de soldagem
Fontes chaveadas (*SMPS*)	Lâmpadas fluorescentes e compactas
Carregadores de bateria	Forno a arco
Equipamentos de Tecnologia da Informação	
Sistema de Potência Ininterrupto (*UPS*)	

Fonte: Kamenka (2014).

2.2.6 Fontes da Distorção Harmônica

Para Afonso e Martins (2005), grande parte dos problemas que surgem nos sistemas elétricos possuem origem na excessiva distorção das correntes ou tensões junto ao consumidor final. Toste (2003) afirma que o uso disseminado de equipamentos eletroeletrônicos tem mudado acentuadamente a natureza das cargas dos sistemas de distribuição e, por isso, provocado o aparecimento de componentes de corrente com frequências diferentes da fundamental em regime permanente. A principal causa deste fenômeno, afirmam Afonso e Martins (2005) se deve à crescente utilização de equipamentos eletrônicos alimentados pela rede elétrica, tais como computadores, impressoras, TVs, lâmpadas compactas e fluorescentes, controladores eletrônicos para uma enorme variedade de cargas industriais, etc.

Tostes (2003) apresenta a classificação das cargas não lineares em três categorias de acordo com a natureza da distorção harmônica por elas provocadas (*IEEE TASK FORCE ON HARMONICS MODELING AND SIMULATIONS*,1996; SILVA, 1997) conforme apresentada na Tabela 2.8:

Tabela 2.8 – Classificação das cargas não-lineares em categorias de acordo com a natureza da distorção.

Categoria	Tipos de Equipamentos	Natureza da Distorção
1	Equipamentos com característica operativa de arcos voltaicos, tais como os **fornos a arco**, **máquinas de solda**, **iluminação fluorescente** e outros.	A natureza da distorção da corrente é oriunda da não linearidade do arco voltaico.
2	Equipamentos de núcleo magnético saturado, tais como **motores, reatores e transformadores** de núcleo saturados.	A natureza da distorção da corrente é oriunda da não linearidade do circuito magnético.
3	Equipamentos eletrônicos, tais como **inversores**, **retificadores**, **UPS**, **televisores**, **fornos de micro-ondas**, **computadores, fontes chaveadas** e outros.	A natureza da distorção da corrente é oriunda da não linearidade dos componentes eletrônicos.

Fonte: Tostes (2003).

2.2.7 Normas e Padrões para Distorções Harmônicas

Com o objetivo de manter-se a convivência harmoniosa entre equipamentos perturbadores e equipamentos sensíveis, afirmam Tostes (2003) e Kamenka (2014) que é necessário o estabelecimento de limites e normas para controle de tais fenômenos. As emissões harmônicas de corrente e tensão estão sujeitas a vários padrões e regulamentos conforme afirma Kamenka (2014) através dos tópicos a seguir:

- Padrões de Emissões aplicáveis aos equipamentos que causam distorções harmônicas;
- Padrões de compatibilidade para redes de distribuição
- Recomendações emitidas por utilitários e aplicáveis às instalações

Existe atualmente um conjunto de normas relativas à compatibilidade eletromagnética elaborada pela *IEC* e a recomendação *STD* 519 de 1992 do *IEEE*, que estabelece limites de níveis de distorção harmônica no SEE. No Brasil, utilizam-se limites estabelecidos pelo ONS e pela Norma Brasileira (NBR) 5410, a qual apresenta medidas para o dimensionamento de condutores na presença de distorções harmônicas e estão vigentes o PRODIST – Módulo 8 (ANEEL-PRODIST, 2016).

2.2.7.1 Normas Internacionais

Dentre as normas internacionais que regulamentam a injeção de correntes harmônicas, Tostes (2003) e Kamenka (2014) apresentam como referência a *IEC* 1000, *IEC* 6000, EN50160 e a recomendação *IEEE* 519/1992.

2.2.7.1.1 Norma IEC 61000-3-2 para equipamentos de baixa tensão com corrente nominal igual ou superior a 16 A

A *IEC* 61000-3-2 trata da limitação de correntes harmônicas injetadas no sistema de suprimento público por qualquer aparelho com uma corrente nominal igual ou superior a 16 A por fase e destinado a ser conectado a sistemas públicos de distribuição de baixa tensão. Especifica limites de componentes harmônicos da corrente de entrada que podem ser produzidos por equipamentos testados em condições especificadas. O objetivo deste padrão é estabelecer limites para as emissões de correntes harmônicas de equipamentos dentro do seu escopo, o cumprimento dos limites garante que os níveis de perturbação harmônica não excedam os níveis de compatibilidade definidos na *IEC* 61000 -2-2. Para fins de limitação de corrente harmônica, o equipamento é classificado em 4 classes conforme Tabela 2.9.

Tabela 2.9 – Classes conforme norma *IEC* 61000-2-2.

Classe	Tipo de Equipamento
A	❖ Equipamentos trifásicos balanceados; ❖ Aparelhos domésticos, excluindo equipamentos identificados como classe D; ❖ Ferramentas, excluindo ferramentas portáteis; ❖ *Dimmers* para lâmpadas incandescentes; ❖ Equipamento de áudio; ❖ Equipamento não especificado em uma das outras três classes deve ser considerado como equipamento da classe A.
B	❖ Ferramentas portáteis; ❖ Equipamentos de soldagem a arco que não são equipamentos profissionais.
C	❖ Equipamento de iluminação.
D	❖ Equipamentos com uma potência especificada inferior ou igual a 600 W, dos seguintes tipos: Computadores pessoais e monitores de computadores pessoais; Receptores de televisão.

Fonte: IEC 61000-2-2.

Os limites para o equipamento de classe A são mostrados na Tabela 2.10. Os limites referem-se a valores fixos para correntes harmônicas 2ª a 40ª ordem. Para os equipamentos de classe B, esses limites podem ser multiplicados por um fator de 1,5. A Tabela 2.11 mostra os limites para equipamentos de classe C com potência ativa superior a 25W. Para os equipamentos de classe C com uma potência de entrada menor ou igual a 25W, os limites da tabela 2.11 se aplicam ou a terceira corrente harmônica não deve exceder 86% e a quinta harmônica de corrente não deve exceder 61% da corrente fundamental.

Para o equipamento da classe D, os limites são mostrados na tabela 2.12 como uma corrente de potência (mA / W).

Tabela 2.10 – Limites conforme norma *IEC* 61000-2-2 para Classe A.

Ordem Harmônica	Máxima corrente harmônica permitida (A)
Harmônicos Ímpares	
3ª	2,30
5ª	1,14
7ª	0,77
9ª	0,40
11ª	0,33
13ª	0,21
15ª ≤ n ≤ 39ª	$0{,}15 \times \frac{15}{n}$
Harmônicos Pares	
2ª	1,08
4ª	0,43
6ª	0,30
8ª ≤ n ≤ 40ª	$0{,}23 \times \frac{8}{n}$

Fonte: IEC 61000-2-2.

Tabela 2.11 – Limites conforme norma *IEC* 61000-2-2 para Classe C.

Ordem Harmônica	Máxima corrente harmônica permitida (A) expressa como um percentual da corrente de entrada da frequência fundamental
2ª	2
3ª	30 x λ (λ...fator de potência do circuito)
5ª	10
7ª	7
9ª	5
11ª ≤ n ≤ 39ª	3

Fonte: *IEC* 61000-2-2.

Tabela 2.12 – Limites conforme norma *IEC* 61000-2-2 para Classe D.

Ordem Harmônica	Máxima corrente harmônica permitida por Watt (mA/W)	Máxima corrente harmônica permitida (A)
3ª	3,40	2,30
5ª	1,90	1,14
7ª	1,00	0,77
9ª	0,50	0,40
11ª	0,35	0,33
13ª ≤ n ≤ 39ª	$\frac{3,85}{n}$	Ver Tabela 2.10

Fonte: *IEC* 61000-2-2.

Tostes (2003) apresenta um resumo evidenciado na tabela 2.13.

Tabela 2.13 – Resumo dos limites da norma *IEC* 61000-2-2 para as correntes harmônicas.

Ordem Harmônica	Classe A Máxima Corrente [A]	Classe B Máxima Corrente [A]	Classe C (>25W) % da fundamental	Classe D >10W, < 300W	Classe D&A Máxima Corrente [A]
Harmônicos Ímpares					
3ª	2,30	3,45	30 x o fator de potência	3,40	2,30
5ª	1,14	1,71	10,00	1,90	1,14
7ª	0,77	1,15	7,00	1,00	0,77
9ª	0,40	0,60	5,00	0,50	0,40
11ª	0,33	0,49	3,00	0,35	0,33
13ª	0,21	0,31	3,00	0,29	0,21
15ª ≤ *n* ≤ 39ª			3,00	3,85 / n	2,25 / n
Harmônicos Pares					
2ª	1,08	1,62	2,00		
4ª	0,43	0,64			
6ª	0,30	0,45			
8ª ≤ n ≤ 40ª	(0,23 x 8) / n				

Fonte: Tostes (2003).

2.2.7.1.2 Norma IEC 61000-3-12 para equipamentos de baixa tensão com corrente nominal superior a 16 A e inferior a 75 A

A norma *IEC* 61000-3-12 trata da limitação das correntes harmônicas injetadas no sistema de abastecimento público. Os limites indicados neste padrão internacional são aplicáveis a equipamentos elétricos e eletrônicos com uma

corrente nominal de entrada superior a 16 A e até 75 A por fase, destinados a ser conectados a sistemas públicos de distribuição de CA de baixa tensão dos seguintes tipos:

- Tensão nominal até 240 V, monofásico, dois ou três fios;
- Tensão nominal até 690 V, trifásica, três ou quatro fios;
- Frequência nominal 50 Hz ou 60 Hz.

Os limites de corrente harmônica estão especificados nas Tabelas de 2.14 a 2.17.

Tabela 2.14 – Limites de emissão de corrente da norma *IEC* 61000-3-12 para equipamentos que não sejam trifásicos balanceados.

Mínima R_{SCE}	**Corrente harmônica individual admissível** $\frac{I_h}{I_{ref}}$ **%**						**Parâmetros harmônicos admissíveis %**	
	I_3	I_5	I_7	I_9	I_{11}	I_{13}	THC/I_{ref}	$PWHC/I_{ref}$
33	21,6	10,7	7,2	3,8	3,1	2	23	23
66	24	13	8	5	4	3	26	26
120	27	15	10	6	5	4	30	30
250	35	20	13	9	8	6	40	40
≥ 350	41	24	15	12	10	8	47	47

Os valores relativos aos harmônicos pares até a 12ª ordem não devem exceder 16 / h%.
Mesmo os harmônicos acima da 12ª ordem são consideradas em *THC* e *PWHC* da mesma forma que harmônicos de ordem ímpares.
É permitida a interpolação linear entre valores R_{SCE} sucessivos.

I_{ref} = corrente de referência; I_h = componente de corrente harmônica

Fonte: *IEC* 61000-3-12.

Tabela 2.15 – Limites de emissão de corrente da norma *IEC* 61000-3-12 para equipamentos trifásicos balanceados.

Mínima R_{SCE}	**Corrente harmônica individual admissível $\frac{I_h}{I_{ref}}$ %**				**Parâmetros harmônicos admissíveis %**	
	I_5	I_7	I_{11}	I_{13}	THC/I_{ref}	$PWHC/I_{ref}$
33	10,7	7,2	3,1	2	13	22
66	14	9	5	3	16	25
120	19	12	7	4	22	28
250	31	20	12	7	37	38
≥ 350	40	25	15	10	48	46

Os valores relativos aos harmônicos pares até a 12ª ordem não devem exceder 16 / h%. Mesmo os harmônicos acima da 12ª ordem são consideradas em *THC* e *PWHC* da mesma forma que harmônicos de ordem ímpares.

É permitida a interpolação linear entre valores R_{SCE} sucessivos.

I_{ref} = corrente de referência; I_h = componente de corrente harmônica

Fonte: *IEC* 61000-3-12.

A Tabela 2.16 pode ser usada com equipamento trifásico equilibrado se qualquer uma destas condições for atendida:

a) As correntes harmônicas das 5ª e 7ª são cada uma menos de 5% da corrente de referência durante todo o período de observação do teste.

b) O design da peça de equipamento é tal que o ângulo de fase da 5ª harmônica de corrente não possui valor preferencial ao longo do tempo e pode levar qualquer valor em todo o intervalo [0°, 360°].

c) O ângulo de fase da 5ª harmônica de corrente relacionada à tensão fundamental de fase a neutro está na faixa de 90° a 150° durante todo o período de observação do teste.

Tabela 2.16 – Limites de emissão de corrente da norma *IEC* 61000-3-12 para equipamentos trifásicos balanceados sob condições específicas (a, b, c).

Mínima R_{SCE}	**Corrente harmônica individual admissível $\frac{I_h}{I_{ref}}$ %**				**Parâmetros harmônicos admissíveis %**	
	I_5	I_7	I_{11}	I_{13}	THC/I_{ref}	$PWHC/I_{ref}$
33	10,7	7,2	3,1	2	13	22
≥ 120	40	25	15	10	48	46

Os valores relativos aos harmônicos pares até a 12ª ordem não devem exceder 16 / h%. Mesmo os harmônicos acima da 12ª ordem são consideradas em *THC* e *PWHC* da mesma forma que harmônicos de ordem ímpares.

É permitida a interpolação linear entre valores R_{SCE} sucessivos.

I_{ref} = corrente de referência; I_h = componente de corrente harmônica

Fonte: *IEC* 61000-3-12.

A Tabela 2.17 pode ser usada com equipamento trifásico equilibrado se qualquer uma destas condições for atendida:

d) As correntes harmônicas das 5ª e 7ª são cada uma menos de 3% da corrente de referência durante todo o período de observação do teste.

e) O design da peça de equipamento é tal que o ângulo de fase da 5ª harmônica de corrente não possui valor preferencial ao longo do tempo e pode levar qualquer valor em todo o intervalo [0°, 360°].

f) O ângulo de fase da 5ª harmônica de corrente relacionada à tensão fundamental de fase a neutro está na faixa de 150° a 210° durante todo o período de observação do teste.

Tabela 2.17 – Limites de emissão de corrente da norma *IEC* 61000-3-12 para equipamentos trifásicos balanceados sob condições específicas (d, e, f).

Mínima R_{SCE}	Corrente harmônica individual admissível $\frac{I_h}{I_{ref}}$ %												Parâmetros harmônicos admissíveis %	
	I_5	I_7	I_{11}	I_{13}	I_{17}	I_{19}	I_{23}	I_{25}	I_{29}	I_{31}	I_{35}	I_{37}	THC/I_{ref}	$PWHC/I_{ref}$
33	10,7	7,2	3,1	2	2	1,5	1,5	1,5	1	1	1	1	13	22
≥ 250	25	17,3	12,1	10,7	8,4	7,8	6,8	6,5	5,4	5,2	4,9	4,7	35	70

Para R_{SCE} igual a 33, os valores relativos aos harmônicos pares até a 12ª ordem não devem exceder 16 / h%. Os valores relativos de todas os harmônicos de I_{14} a I_{40} não listados acima não devem exceder 1% de I_{ref}. Para $R_{SCE} \geq 250$, os valores relativos aos harmônicos pares até a 12ª ordem não devem exceder 16 / h%. Os valores relativos de todos os harmônicos de I_{14} a I_{40} não listados acima não devem exceder 3% de I_{ref}. É permitida a interpolação linear entre valores R_{SCE} sucessivos.

I_{ref} = corrente de referência; I_h = componente de corrente harmônica

Fonte: *IEC* 61000-3-12.

2.2.7.1.3 Norma EN 50160

O documento principal que trata da qualidade da distribuição na Europa e em outras partes do mundo é o padrão EN 50160. Caracteriza os parâmetros de tensão da energia elétrica nos sistemas públicos de distribuição e fornece os principais parâmetros de tensão e os seus intervalos de desvio permissíveis no PAC em sistemas públicos de baixa tensão (BT), média tensão (MT) e alta tensão (AT), em condições normais de operação. Define assim a chamada qualidade mínima de energia disponível para o equipamento do usuário no PAC. Esta qualidade mínima é necessária para ter uma boa chance de obter o equipamento instalado funcionando corretamente. Os limites indicados na EN 50160 devem ser garantidos pelo fornecedor. No entanto, como mencionado anteriormente, para muitos consumidores, mesmo cumprir os requisitos fornecidos na EN 50160 não significa automaticamente um nível satisfatório de QEE sem problemas. As características de tensão dentro deste padrão são definidas em termos de frequência, magnitude, forma de onda e simetria e incluem definições e, em alguns casos, métodos de medição e níveis de conformidade para 10 características da tensão de alimentação:

- Frequência de potência;
- Variações da tensão de alimentação;
- Mudanças rápidas de tensão e *Flicker*;

- Afundamentos de tensão de alimentação;
- Interrupções curtas;
- Longas interrupções;
- Sobretensões temporárias e transitórias;
- Desequilíbrio de tensão de alimentação;
- Tensão harmônica;
- Tensão de sinalização de rede.

Os requisitos de tensão harmônica em condições normais de operação são definidos para períodos de cada uma semana e 95% dos valores médios *rms* de 10 min de cada tensão harmônica individual devem ser menores ou iguais aos valores indicados na Tabela 2.18. As ressonâncias podem causar altas tensões para um harmônico individual. Além disso, o *THD* da tensão de alimentação (incluindo todos os harmônicos até a 40ª ordem) deve ser inferior ou igual a 8%.

Tabela 2.18 – Valores das tensões harmônicas individuais nos terminais de fornecimento dados em porcentagem da tensão fundamental u_1.

Harmônicos Ímpares				**Harmônicos Pares**	
Não múltiplas de 3		**Múltiplas de 3**			
Ordem h	Amplitude relativa U_h	Ordem h	Amplitude relativa U_h	Ordem h	Amplitude relativa U_h
5	6,0%	3	5,0%	2	2,0%
7	5,0%	9	1,5%	4	1,0%
11	3,5%	15	0,5%	6..24	0,5%
13	3,0%	21	0,5%		
17	2,0%				
19, 23 e 25	1,5%				

NOTA Nenhum valor é dado para harmônicos de ordens superiores a 25, pois geralmente são pequenos, mas em grande parte imprevisíveis devido a efeitos de ressonância.

Fonte: EN 50160.

2.2.7.1.4 Norma IEEE 519

A norma *IEEE* 519 apresenta uma abordagem conjunta entre fornecedores e clientes para limitar o impacto de cargas não lineares. Esta prática recomendada pretende estabelecer metas para o projeto de sistemas elétricos que incluem cargas

tanto lineares como não lineares. As formas de onda de tensão e corrente que podem existir em todo o sistema são descritas e as metas de distorção da forma de onda para o designer do sistema são estabelecidas. A interface entre fontes e cargas é descrita como o ponto de acoplamento comum; e a observância dos objetivos de projeto minimizará a interferência entre equipamentos elétricos. Esta prática recomendada aborda a limitação no estado estacionário. As condições transitórias que excedem essas limitações podem ser encontradas. Este documento define a qualidade de energia que deve ser fornecido no ponto de acoplamento comum.

A filosofia de desenvolvimento de limites harmônicos de corrente no *IEEE* 519 é:

a) Limite a injeção harmônica de clientes individuais para que eles não provoquem níveis de distorção de tensão inaceitáveis para características normais do sistema;
b) Limite a distorção harmônica geral do sistema fornecido pelo fornecedor.

A Tabela 2.19 lista as bases para limites de corrente harmônica, enquanto a Tabela 2.20 lista os limites da corrente harmônica com base no tamanho da carga em relação ao tamanho do sistema de energia ao qual a carga está conectada. A razão I_{SC} / I_L é a proporção do curto-circuito disponível no ponto de acoplamento comum (PAC), até a corrente de carga fundamental máxima. O padrão *IEEE* 519-1992 também introduz a distorção da demanda total (*TDD*), a distorção da corrente harmônica em % da carga máxima da demanda (demanda de 15 ou 30 min).

Tabela 2.19 – Base para limites de corrente harmônica.

Taxa de curto circuito no PAC	Harmônico de tensão individual máxima (%)	Hipótese relacionada
10	2,5 - 3,0	Sistema dedicado
20	2,0 – 2,5	1 a 2 grandes consumidores
50	1,0 – 1,5	Alguns consumidores relativamente grandes
100	0,5 – 1,0	5 a 20 consumidores de tamanho médio
1000	0,05 – 0,10	Muitos clientes pequenos

Fonte: *IEEE* 519.

Tabela 2.20 – Limites de distorção de corrente para sistemas de distribuição geral (120V a 69000V).

Máxima distorção harmônica de corrente em percentagem do I_L Ordem harmônica individual (harmônicos ímpares)						
I_{SC} / I_L	**< 11**	**11 ≤ h ≤ 17**	**17 ≤ h ≤ 23**	**23 ≤ h ≤ 35**	**35 ≤ h**	***THDi***
< 20*	4,0	2,0	1,5	0,6	0,3	5,0
20-50	7,0	3,5	2,5	1,0	0,5	8,0
50-100	10,0	4,5	4,0	1,5	0,7	12,0
100-1000	12,0	5,5	5,0	2,0	1,0	15,0
>1000	15,0	7,0	6,0	2,5	1,4	20,0

Onde:

I_{SC} = corrente máxima de curto-circuito no PAC.

I_L = corrente de carga de demanda máxima (componente de frequência fundamental) no PAC.

Fonte: IEEE 519.

A Tabela 2.21 mostra os limites de distorção de tensão que devem ser utilizados como valores de projeto do sistema. Como os limites atuais esses valores também são destinados ao "pior caso" para a operação normal do sistema.

Tabela 2.21 – Limites de distorção de tensão.

Tensão no PAC	**Distorção de tensão individual (%)**	***THDv (%)***
69 kV e abaixo	3.0	5.0
69,001 kV até 161 kV	1,5	2,5
Acima de 161 kV	1,0	1,5

Fonte: *IEEE* 519.

2.2.7.2 Normas Nacionais

Tostes (2003) afirma que a ONS elaborou limites com o objetivo de definir os padrões da rede básica e que isso foi resultado da compilação da experiência de planejamento e operação do Sistema Elétrico Brasileiro no âmbito do Grupo Coordenador de Planejamento dos Sistemas Elétricos (GCPS), do Grupo Coordenador para Operação Interligada (GCOI), do Comitê Coordenador de Operação Nacional (CCON), e também dos resultados preliminares das discussões técnicas promovidas no âmbito do Grupo de Trabalho Especial – Qualidade de Energia Elétrica, coordenado pelo ONS e constituído por representação dos diversos agentes, universidades, consumidores etc.

2.2.7.2.1 ANEEL - PRODIST - Módulo 8 – Qualidade da Energia Elétrica

O objetivo desta norma, afirma Leite (2015), é estabelecer os procedimentos relativos à qualidade da energia elétrica - QEE, abordando a qualidade do produto e a qualidade do serviço prestado.

A Tabela 2.22 indica os valores de referência para as distorções harmônicas totais que servem para referência do planejamento elétrico em termos de QEE e que, em breve será regulamentado, após período experimental de coleta de dados.

Tabela 2.22 – Valores de referência globais das distorções harmônicas totais.

Tensão nominal do barramento	Distorção harmônica total de tensão (%)
Vn ≤ 1 kV	10
1 kV < Vn ≤ 13,8 kV	8
13,8 kV < Vn ≤ 69 kV	6
69 kV < Vn ≤ 230 kV	3

Fonte: ANEEL-PRODIST-Módulo 8.

Os valores das distorções harmônicas individuais são mostrados na Tabela 2.23.

Tabela 2.23 – Níveis de referência para distorções harmônicas individuais de tensão.

Harmônicos	Ordem	Distorção Harmônica Individual de Tensão [%]			
		Vn ≤ 1 kV	Vn > 1 kV Vn ≤ 13,8 kV	Vn > 13,8 kV Vn ≤ 69 kV	Vn > 69 kV Vn ≤ 230 kV
Ímpares não múltiplos de 3	5ª	7,5	6,0	4,5	2,5
	7ª	6,5	5,0	4,0	2,0
	11ª	4,5	3,5	3,0	1,5
	13ª	4,0	3,0	3,0	1,5
	17ª	2,5	2,0	1,5	1,0
	19ª	2,0	1,5	1,5	1,0
	23ª	2,0	1,5	1,5	1,0
	25ª	2,0	1,5	1,5	1,0
	> 25ª	1,5	1,0	1,0	0,5
Ímpares múltiplos de 3	3ª	6,5	5,0	4,0	2,0
	9ª	2,0	1,5	1,5	1,0
	15ª	1,0	0,5	0,5	0,5
	21ª	1,0	0,5	0,5	0,5
	> 21ª	1,0	0,5	0,5	0,5

Harmônicos	Ordem	Distorção Harmônica Individual de Tensão [%]			
		Vn ≤ 1 kV	Vn > 1 kV Vn ≤ 13,8 kV	Vn > 13,8 kV Vn ≤ 69 kV	Vn > 69 kV Vn ≤ 230 kV
Pares	2ª	2,5	2,0	1,5	1,0
	4ª	1,5	1,0	1,0	0,5
	6ª	1,0	0,5	0,5	0,5
	8ª	1,0	0,5	0,5	0,5
	10ª	1,0	0,5	0,5	0,5
	12ª	1,0	0,5	0,5	0,5
	> 12ª	1,0	0,5	0,5	0,5

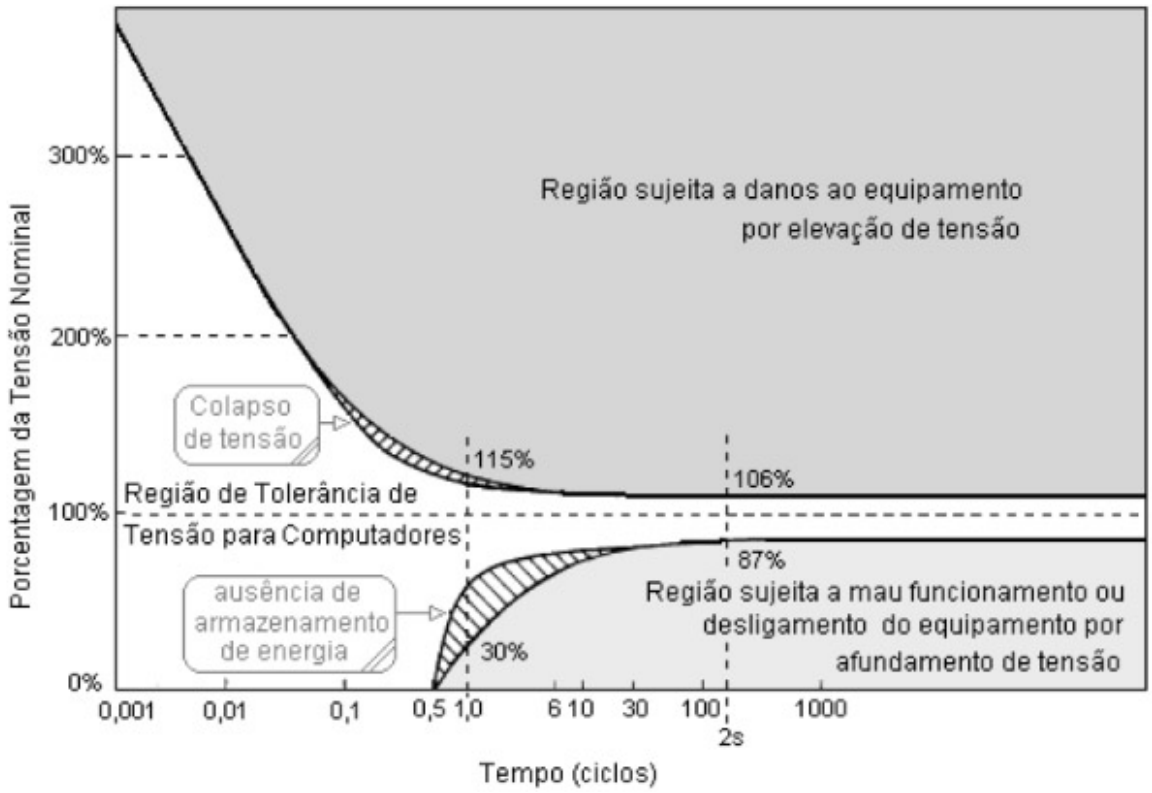
Fonte: ANEEL-PRODIST-Módulo 8.

2.2.7.3 Referência para QEE em computadores

2.2.7.3.1 Curva CBEMA/ITIC

Outro ponto importante para a motivação da presente pesquisa é que a mesma propõe análise dentro de uma empresa fabricante de computadores, pois desde de 1978 quando Thomas Key estudando a confiabilidade do suprimento de energia elétrica para instalações militares concluiu que afundamentos de tensão de curta duração poderiam prejudicar a operação normal dos grandes computadores daquela instalação. Como resultado desse estudo foi criada a curva *Computer Business Equipment Manufacturers Association-CBEMA*. Deckmann e Pomilio (2010) afirmam que essa curva é o guia para os fabricantes de equipamentos na área de informática, e que os EUA estabeleceram a norma *ANSI/IEEE* Std. 446, conhecida como curva *CBEMA* propondo metas a serem satisfeitas pelas fontes e dispositivos que alimentam computadores com relação às variações de tensão e respectivas durações suportáveis conforme evidenciado na Figura 2.3.

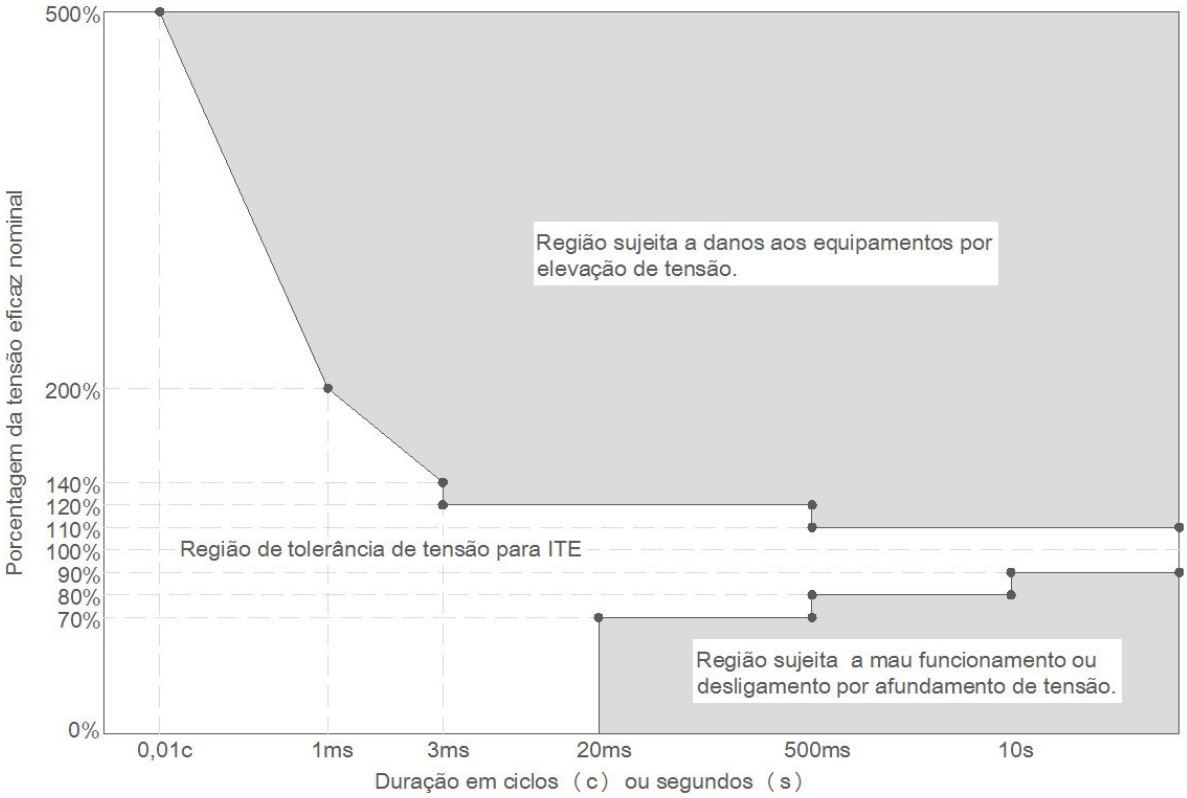
Figura 2.3 – Curva *CBEMA*.



Fonte: Deckmann e Pomilio (2010). Dugan *et al.*(2004).

Originariamente a Curva *CBEMA* foi criada para caracterizar a sensibilidade de computadores, no entanto acabou se tornando referência para outros equipamentos eletroeletrônicos. Desta forma foi revisada pela *Information Technology Industry Council (ITIC)* em 1994 em parceria com o Centro de Aplicações em Eletrônica de Potência da *EPRI* e é evidenciada na Figura 2.4.

Figura 2.4 – Curva *ITIC*.



Fonte: Deckmann e Pomilio (2010). Dungan *et al.*(2004).

Essa curva passou a ser uma referência para verificação do nível de vulnerabilidade de equipamentos afirmam Deckmann e Pomilio (2010).

2.2.7.3.2 Recomendações para montadoras de computadores

A *INTEL* (2016) afirma que a evolução do desempenho tem levado ao aumento da sensibilidade dos semicondutores e apresenta dados de problemas relacionados a falhas em campo em computadores tendo como principal causa problemas relacionados a descarga eletrostática e estresse elétrico acontecidas na empresa fabricante do computador e indica que entre as causas está distúrbios de QEE e recomendo o monitoramento e controle com base na norma *IEC 61000-4-30 Classe A* conforme Tabela 2.24.

Tabela 2.24 – Itens a serem monitorados para mitigar o estresse elétrico (*INTEL*).

Item auditado	Norma de Referência	Especificação	Nível	Onde	Periodicidade	O que	Como	Recomendação
QEE AC	***IEC* 61000-4-30 Classe A**	**Elevação e Afundamento de tensão e transientes**	**Placa e Sistema**	**Rep, TF e CQ**	**Mensal**	**Energia Elétrica**	**Analisador de QEE**	**Manter a QEE o tempo todo (7 dias / 24h)**
Rep= Reparo, TF= Teste Funcional, CQ= Controle de Qualidade								

Fonte: *INTEL* (2016).

Entre as diversas recomendações de fabricantes de processadores está o monitoramento e controle da QEE para evitar corrupção de dados durante a fabricação de computadores. Segundo a *INTEL* (2016), a corrupção de dados refere-se a eventos que causam uma modificação nos dados escritos sejam na diferença na armazenagem, erros durante o processo de escrita ou alteração do conteúdo já escrito e podem acontecer em diferentes etapas do processo produtivo como mostrada na Tabela 2.25.

Tabela 2.25. Onde pode ocorrer a corrupção de dados.

Ação	Nível de Componente	Nível de Placa			Nível de Sistema	
	Gravação do componente	*In-Circuit Tester ICT*	Teste Funcional	Controle de Qualidade Final	Montagem e Teste	Teste Final
Corrompido	x	x	x	x	x	x
Detectado		x	x	x	x	x

Fonte: Adaptado de *INTEL* (2016).

Como evidenciado na Tabela 2.25, a corrupção dos dados pode ocorrer em todos os níveis do processo produtivo, no entanto é mais predominante nas etapas teste funcional da placa e montagem e teste do sistema. Para monitoramento e ações de mitigação, a *INTEL* (2016) recomenda a medição da QEE em todas as etapas do processo tendo como base a norma *IEC* 61000-4-30 Classe A e recomenda auditoria mensal conforme é apresentado na Tabela 2.26.

Tabela 2.26. Medição da QEE na auditoria da *INTEL*.

	Forma de Onda	**Transientes**	***THDv***	**Tensão entre Neutro e Terra**	**Afundamento e Elevação**
Resultado Esperado	Senoidal sem distorções	< 1 em 60 segundos	< 5%	< 4 Volts	< 1 evento por hora

Fonte: Adaptado de *INTEL* (2016).

3 INTELIGÊNCIA COMPUTACIONAL E ASPECTOS GERAIS SOBRE O PROCESSO *KDD*

3.1 Considerações Iniciais

Nesse capítulo serão abordados os conceitos principais da IC, do processo *KDD*, bem como as técnicas de mineração de dados Árvore de Decisão e *Naïve Bayes*. Camilo e Silva (2009) afirmam que desde o surgimento dos sistemas computacionais, um dos principais objetivos das organizações tem sido o de armazenar dados e que nas últimas décadas essa tendência ficou ainda mais evidente com a queda nos custos para a aquisição de *hardware*, tornando possível armazenar quantidades cada vez maiores de dados.

Chegamos a era do ***big data*** conforme afirma Roiger (2017) e devido ao grande volume de dados, o grande desafio é encontrar novas técnicas para o seu processamento. Kazerooni *et al.* (2014), afirmam que popularidade do conceito "*big data*" tem favorecido a busca de conhecimento através de técnicas de IC como o processo de descoberta do conhecimento em base dados (*KDD*).

Segundo informações do *International Data Corporation-IDC* (2017), existem atualmente 13 bilhões de "coisas" conectadas e que a previsão é que até 2020, haverá 30 bilhões de "coisas" conectadas e oportunidades de receita projetadas de US $ 1,7 trilhões para o ecossistema, tendo como principais aplicações: operações de fabricação, com US$ 105 bilhões, tecnologias para redes inteligentes de energia elétrica, saneamento básico e gás (US$ 56 bilhões), transporte de cargas (US$ 50 bilhões), gerenciamento dos ativos de produção (US$ 45 bilhões) e tecnologias para cidades inteligentes (US$ 40 bilhões) e que nesse período o universo digital crescerá até 40.000 *exabytes*. É uma quantidade muito grande considerando que 5 *exabytes* representa o total de palavras que falamos afirma Roiger (2017).

É necessária a análise dos dados para se transformar esses dados em conhecimento útil em diversas áreas como a descoberta de cura para doenças

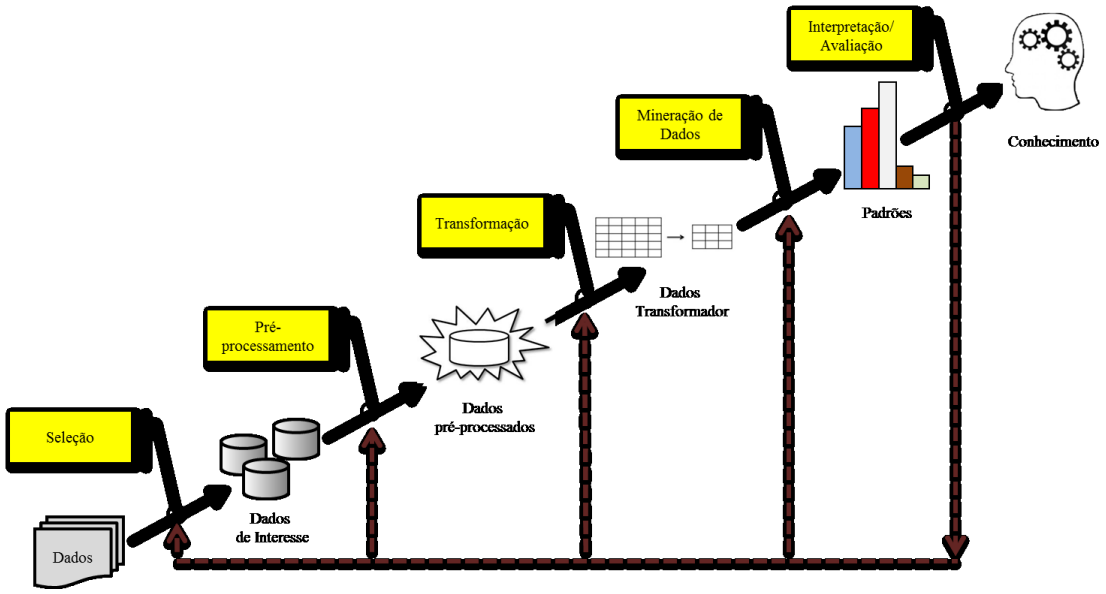
através de ferramentas de diagnóstico, análise política, monitorar e prever problemas ambientais, detectar comportamento fraudulento e analisar, monitorar e prever problemas de QEE. A análise de dados é definida como o processo de extração do conhecimento significativo a partir dos dados. Seus métodos vêm de várias disciplinas, incluindo ciência da computação, matemática, estatísticas e processamento distribuído entre outros.

3.2 Processo *KDD*

Nogueira (2015) afirma que o termo *Knowledge Discovery in Databases*, ou *KDD*, refere-se ao amplo processo de encontrar o conhecimento em dados. O *KDD* engloba, em sua natureza, as áreas de aprendizagem de máquina, reconhecimento de padrões, bancos de dados, estatísticas, inteligência artificial, e visualização de dados (FAYYAD et al, 1996). De acordo com Piatetsky-Shapiro e Frankley (1991), a frase "*Knowledge Discovery in Databases*" foi criada na primeira oficina *KDD* em 1989 por Piatetsky-Shapiro e representa um processo automatizado que se baseia em métodos de diversas áreas, como reconhecimento de padrões, estatísticas, aprendizado de máquina, redes neurais, etc. a encontrar padrões de dados na etapa de mineração de dados do processo de *KDD*.

Tecnicamente, o *KDD* é a aplicação do método científico onde o processo completo inclui (além da mineração de dados), extração, preparação de dados e tomada de decisão em relação aos dados minerados afirma Roiger (2017) e para Fayyad *et al.* (1996), o *KDD* pode ser divido em cinco fases: seleção, pré-processamento, transformação, mineração de dados e interpretação, de acordo com a Figura 3.1.

Figura 3.1 – Processo *KDD*.



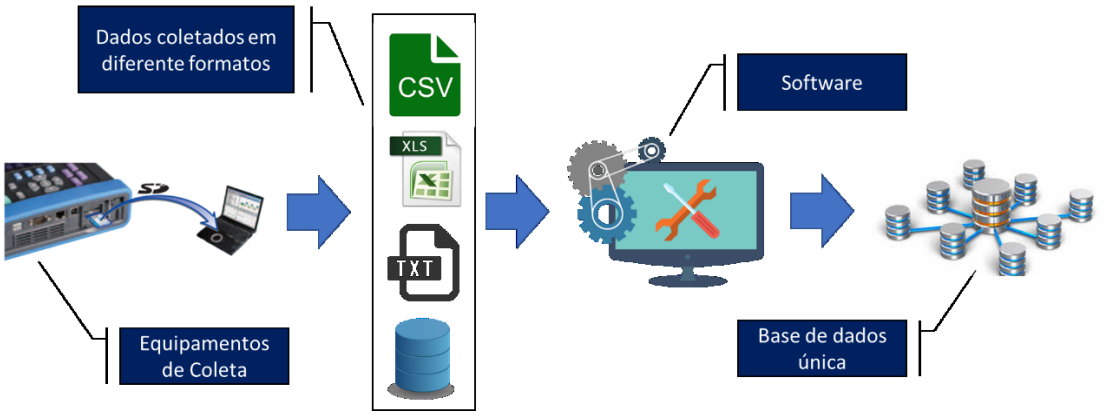
Fonte: Fayyad *et al.*(1996).

3.2.1 Seleção dos dados

É a primeira fase do processo de descoberta do conhecimento, e Nogueira (2015) afirma que é nessa fase que os principais dados, que pertencem a um domínio e contém todas as possíveis variáveis (características ou atributos) e os registros (casos ou observações) que serão selecionados e agrupados em uma única base de dados. Normalmente a escolha dos dados fica a critério de um especialista do domínio e esse processo é bastante complexo, pois pode possuir diversos formatos e como afirma Prass (2009) essa escolha impactará no resultado do processo.

Os dados, normalmente, são coletados em fontes diversas como Sistemas de Gerenciamento de Banco de Dados (SGBD), planilhas, arquivos de textos e ou outros sistemas, afirma Nogueira (2015) e algumas vezes, não apresentam um padrão que seja legível pelos softwares de mineração, então, é necessário o desenvolvimento de um *software* especifico para padronizar e adequar os dados coletado, como mostrado na Figura 3.2.

Figura 3.2 – Seleção dos dados do processo *KDD*.

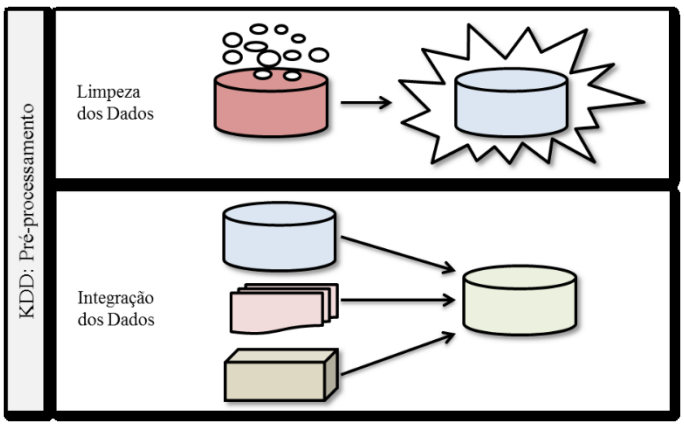


Fonte: Elaborada pelo autor (2017).

3.2.2 Pré-processamento e limpeza dos dados

É uma das partes mais determinantes no processo do *KDD* afirma Nogueira (2015), uma vez que a qualidade dos dados vai definir a eficácia dos algoritmos de mineração de dados. Para Prass (2009) nesta etapa deverão ser realizadas tarefas que eliminem dados redundantes e inconsistentes, recuperem dados incompletos e avaliem possíveis dados discrepantes ao conjunto (*outliers*). Essa etapa será decomposta em limpeza e integração dos dados conforme evidenciado na Figura 3.3.

Figura 3.3 – Limpeza e integração dos dados do processo *KDD*.



Fonte: Elaborada pelo autor (2017).

Nesta fase, afirma Prass (2009), também são utilizados métodos de redução ou transformação para diminuir o número de variáveis envolvidas no processo para

melhorar o desempenho do algoritmo de análise e que os dados podem ser classificados como:

- Dados ausentes;
- Dados discrepantes;
- Dados derivados.

3.2.2.1 Balanceamento de classes

Alberto e Almeida (2012) afirmam que a utilização de bases de dados reais apresentam diversos problemas, tais como grande quantidade de ruído e inconsistências, excesso de valores desconhecidos, conjunto de dados com classes desbalanceadas, entre outros. De acordo com Pyle (1999) e Hall *et al.* (2009) a tarefa de classificação seria mais bem sucedida se os problemas mencionados fossem identificados e tratados antes dos dados serem fornecidos a um algoritmo de classificação na fase de mineração de dados e para Alberto e Almeida (2012) afirmam que esse tratamento deve ser feito na fase de pré-processamento.

Os dados desequilibrados normalmente se referem a um problema com problemas de classificação onde as classes não são representadas de forma igual, afirma Brownlee (2015) e cita como exemplo de um problema com duas classes com 100 instâncias, onde 80 são rotuladas com *Classe 1* e 20 são rotuladas com *Classe 2*., observa-se que esse é um conjunto de dados desequilibrados e que a proporção entre as instâncias é de 80:20 ou 4:1.

A grande maioria dos conjuntos de dados de classificação não possui exatamente o mesmo número de instâncias. E esse desequilíbrio é natural e esperado, como por exemplo, no conjunto de dados como os que caracterizam transações fraudulentas são totalmente desequilibrados, pois a grande maioria das transações estará na classe “Sem Fraude” e uma minoria muito pequena estará classificada como “Fraude” afirma Brownlee (2015).

Brownlee (2015) apresenta oito táticas para combater os dados de treinamento desequilibrados:

1. Fazer a coleta de mais dados;
2. Mudar a métrica de avaliação do desempenho;
3. Refazer o conjunto de dados incluindo cópias dos dados e/ou excluindo dados;
4. Utilizar algoritmos para gerar amostras sintéticas;
5. Utilizar diferentes algoritmos para classificação;
6. Utilizar modelos de classificação penalizada;
7. Experimentar detecção de anomalias e de mudanças;
8. Ser criativo na solução do problema.

O método mais relevante é o que for testado empiricamente e o que permite melhores resultados afirma Brownlee (2015).

Alberto e Almeida (2012) apresentam abordagens em torno de técnicas de *sobreamostragem*, *subamostragem* e na *geração de dados sintéticos* para o balanceamento de dados, onde são citadas as seguintes técnicas:

- *CNN (Condensed Nearest Neighbor Rule).* Essa abordagem é geralmente conhecida como regra do vizinho mais próximo condensado. Para remover exemplos redundantes pode-se ainda criar um subconjunto consistente, *C*, do conjunto original de exemplos *S.* Por definição, um conjunto $C \subset S$ é consistente com *S* se, utilizando-se o algoritmo *1-vizinho-mais-próximo (1-NN)* ele corretamente classifica os exemplos de *S* conforme afirma Hart (1968).
- *ENN (Edited Nearest Neighbor Rule) proposto por* Wilson (1972). O método *ENN* faz *subamostragem* dos dados de forma contrária à maneira em que é feito a *subamostragem* no método *CNN* apresentado por Hart (1968) e consiste na eliminação de todos os objetos que são mal classificados por seus kvizinhos-mais-próximos (*KNN)*. O método *ENN* faz subamostragem do conjunto de entrada S retirando todo caso Ei $\in$ S cuja classe diverge da classe predita pelos kvizinhos-mais-próximos.

- *Tomek links* proposto por Tomek (1976). Para melhorar a precisão da classificação de dados, é uma boa ideia tentar remover o maior ruído do rótulo da classe possível, bem como exemplos limítrofes com maior probabilidade de serem incorretos. Um método para fazer isso é um processo conhecido como remoção do link Tomek. Os links do Tomek consistem em pontos que são os vizinhos mais próximos do outro, mas não compartilham o mesmo rótulo da classe.
- *OSS (One-Sided Selection)*, proposto em Kubat *et al.* (1997). também conhecido como Seleção Unilateral, é um método de *subamostragem* informativa;
- *Cluster-based Oversampling*. Nickerson *et al.* (2001) propõem uma abordagem para minimizar o problema de classes e exemplos raros. A abordagem proposta sugere a utilização de *sobreamostragem* não somente da classe minoritária, mas também dos exemplos raros. A ideia é agrupar os dados de treinamento em *clusters* e, então, balancear a distribuição de seus exemplos.
- *CNL (Cleaning Nearest Rule)*. O método *CNL* de Laurikkala (2001) utiliza a ideia do método *ENN* afirma Wilson (1972) para encontrar, através de *subamostragem*, os dados ruidosos que não pertencem à classe de interesse e, logo em seguida, fazer limpeza do limiar da vizinhança da classe de interesse.
- *SMOTE (Synthetic Minority Oversampling Technique)* proposto em Chawla *et al.* (2002). Originalmente proposto por Chawla *et al.* (2002), o método *SMOTE* é baseado em *sobreamostragem* informativa e cria novos exemplos da classe minoritária por meio da interpolação de exemplos da classe minoritária que se encontram próximos. Também foi citada por Brownlee (2015);
- *WWE (Weighted Wilson's Editing*). Intensifica o método *ENN* eliminando um número maior de exemplos da classe majoritária cujo rótulo difere da maioria de seus kvizinhos-mais-próximos afirmam Barandela *et al.* (2004).
- *Balance Cascade*. O algoritmo *Balance Cascade*, por sua vez, usa uma estratégia iterativa de geração de um *ensemble* de

classificadores para a escolha dos exemplos a serem removidos consideram Liu *et al.* (2009).

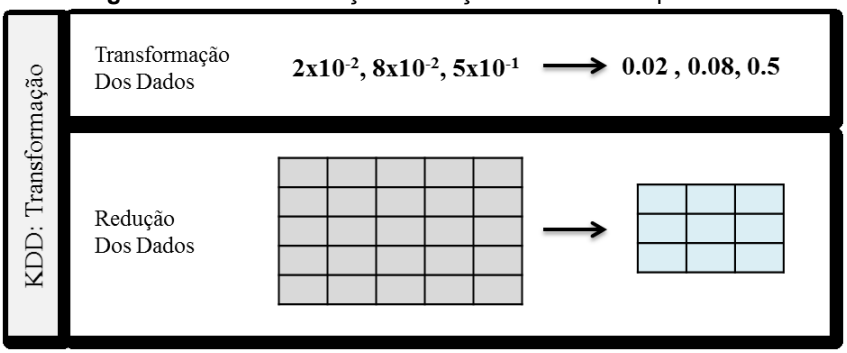
Elhassan *et al.* (2016) afirmam que em diversas aplicações já foram propostas as técnicas *Tomek-Link, CNN* e *OSS* para melhorar o desempenho e apresentam como métodos avançados de amostragem, as técnicas *Tomek Link (T-Link)* e *SMOTE*.

3.2.3 Transformação dos dados

Nessa fase os dados os dados já foram selecionados, pré-processados limpos e integrados e antecede a fase de mineração de dados. Segundo Nogueira (2015) os dados devem ser armazenados e formatados adequadamente para que os algoritmos possam atuar, extraindo conhecimento e que precisam ser organizados de acordo com as particularidades e necessidades das técnicas de extração de conhecimento que serão utilizadas na etapa seguinte.

Prass (2009) afirma que os dados que estão dispersos devem ser agrupados em um repositório único como é mostrado na Figura 3.4.

Figura 3.4 – Transformação e redução dos dados do processo *KDD*.



Fonte: Adaptado de Prass (2009).

3.2.4 Mineração dos dados

Segundo Camilo e Silva (2009), a mineração de dados é ser considerada multidisciplinar e variam com o campo de atuação dos autores, onde é destacada a

estatística, aprendizado de máquina e banco de dados. Em Zhu *et al.* (2007) é feita uma análise comparativa sobre as três perspectivas citadas abaixo.

1. Em Hand *et al.* (2001), a definição é dada de uma perspectiva estatística: "Mineração de Dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados".

2. Em Cabena *et al.* (1998), a definição é dada de uma perspectiva de banco de dados: "Mineração de Dados é um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados".

3. Em Fayyad *et al.* (1996), a definição é dada da perspectiva do aprendizado de máquina: "Mineração de Dados é um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados."

Prass (2009) afirma que é a etapa de Mineração de Dados (*data mining*) que recebe o maior destaque na Literatura e que conforme Berry e Linoff (1997), mineração de dados é a exploração e análise, de forma automática ou semiautomática, de grandes bases de dados com objetivo de descobrir padrões e regras e para Roiger (2017) é definida como o processo de encontrar, dentro de dados, estruturas interessantes que podem assumir muitas formas, como uma árvore, uma rede, um gráfico, um conjunto de regras, várias equações e muito mais, e usa um ou vários algoritmos para identificar tendências e padrões.

O objetivo principal do processo de data mining é fornecer as corporações informações que a possibilitem montar melhores estratégias de *marketing*, vendas e suporte, melhorando assim os seus negócios.

Segundo Nogueira (2015) nesta fase os dados são submetidos a um ou mais algoritmos de aprendizagem para extrair o conhecimento de aprendizagem e de regras, a partir dos dados que foram selecionados, pré-processados e transformados por um especialista na área.

As tarefas de mineração de dados, segundo Tan *et al.* (2009), geralmente são divididas em duas categorias principais de acordo com a Tabela 3.1.

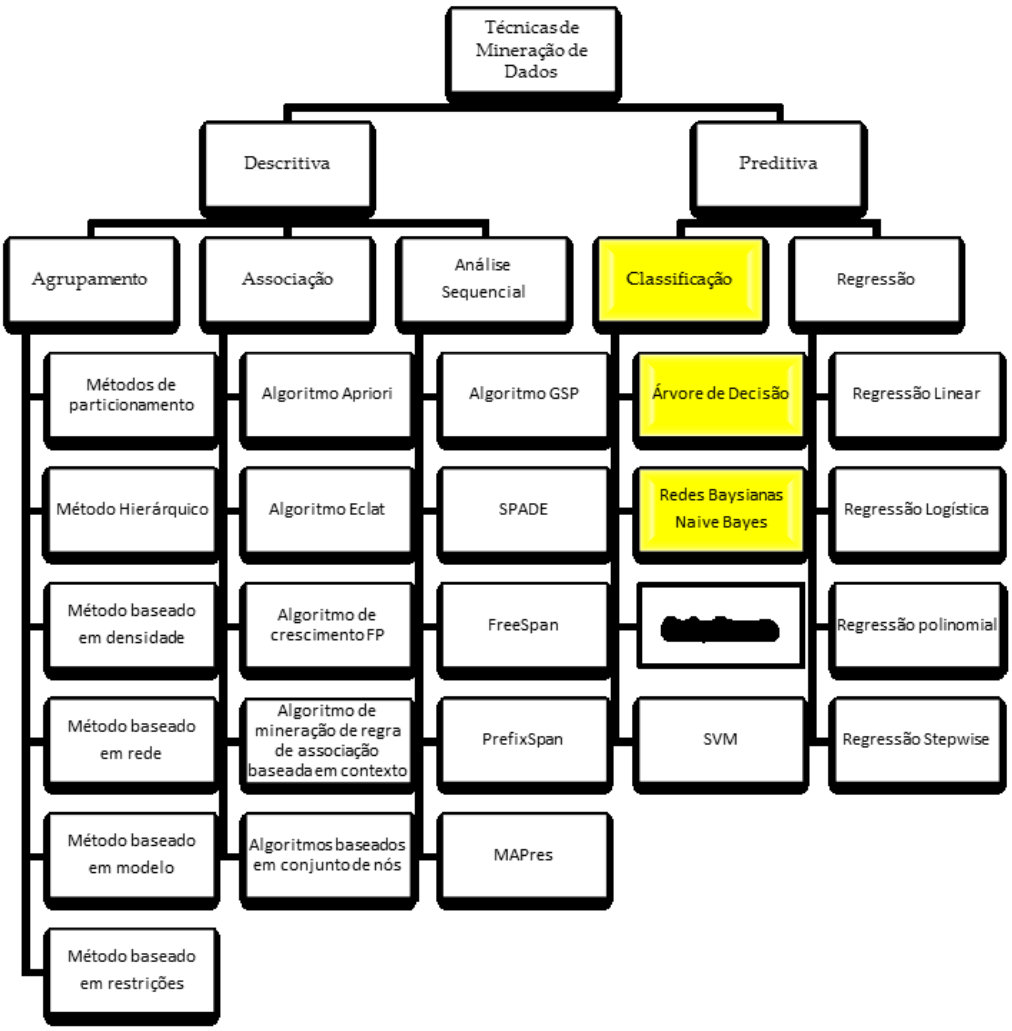
Tabela 3.1 – Categoria da mineração de dados.

Categoria	Objetivo
Tarefa preditiva	O objetivo dessas tarefas é prever o valor de um determinado atributo com base nos valores de outros atributos. O atributo a ser previsto é comumente conhecido com a variável ou o alvo dependente, ao passo que os atributos utilizados para fazer a previsão são conhecidos como variáveis independentes ou explicativas.
Tarefa descritiva	Aqui, o objetivo é derivar padrões (correlações, tendências, grupos, trajetórias e anomalias) que resumem as relações subjacentes nos dados. As tarefas descritivas da mineração de dados são muitas vezes de natureza exploratória e muitas vezes requerem técnicas de pós-processamento para validar e explicar os resultados.

Fonte: Tan *et al.* (2009).

Considerando as duas categorias, existem diversas técnicas de mineração de dados conforme mostrado na Figura 3.5.

Figura 3.5 – Técnicas de mineração de dados.



Fonte: Elaborada pelo autor (2017).

Para Nogueira (2015), a mineração de dados pode utilizar uma grande quantidade de algoritmos de aprendizagem que compõem os paradigmas de aprendizagem disponíveis. Podem ser destacados alguns como: simbólico, conexionista, baseados em regras, evolutivos ou genéticos e o estatístico.

As técnicas de mineração podem ser usadas em qualquer área do conhecimento conforme exemplos abaixo:

- ***Marketing*:** redução dos custos com o envio de correspondências através de sistemas de mala direta a partir da identificação de grupos de clientes potenciais;

- **Detecção de fraude:** reclamações indevidas de seguro, chamadas clonadas de telefones celulares, compras fraudulentas com cartão de crédito;
- **Investimento:** modelos de redes neurais têm sido aplicados no mercado de ações e na previsão da cotação do ouro e do dólar;
- **Produção:** empresas desenvolvem sistemas para detectar e diagnosticar erros na fabricação de produtos. Estas falhas são normalmente agrupadas por técnicas de Análise de Agrupamentos.

Segundo FAYYAD *et al.*, (1996), as técnicas de mineração de dados podem ser aplicadas a tarefas como:

- **Associação:** determina quais fatos ou objetos tendem a ocorrerem juntos num mesmo evento;
- **Classificação:** construção um modelo que possa ser aplicado a dados não classificados visando categorizar os objetos em classes;
- **Predição/Previsão:** usada para definir um provável valor para uma ou mais variáveis;
- **Segmentação:** visa dividir uma população em subgrupos o mais heterogêneos possível entre si;
- **Sumarização:** métodos para encontrar uma descrição compacta para um subconjunto de dados.

Apesar das definições sobre a Mineração de Dados levarem a crer que o processo de extração de conhecimento se dá de uma forma totalmente automática, sabe-se hoje que de fato isso não é verdade (Larose, 2005). Apesar de encontrarmos diversas ferramentas que nos auxiliam na execução dos algoritmos de mineração, os resultados ainda precisam de uma análise humana. Porém, ainda assim, a mineração contribui de forma significativa no processo de descoberta de conhecimento, permitindo aos especialistas concentrarem esforços apenas em partes mais significativa dos dados.

3.2.4.1 Treinamento e Teste

A avaliação dos modelos na etapa mineração de dados é significativa e necessita que os dados sejam separados em conjuntos de treinamento e tese. Em condições normais, quando o conjunto dos dados são separados em um conjunto de treinamentos e em um conjunto de testes, a maior quantidade dos dados é utilizada para os treinamentos e a menor é utilizada para os testes.

A MICROSOFT (2018) utiliza o Assistente de Mineração de Dados e por padrão é divido em 70% dos dados para o conjunto de treinamento e 30% para o conjunto de teste e afirma que esse padrão foi escolhido devido a proporção 70-30 ser a mais utilizada em Mineração de Dados.

Para a utilização de conjuntos de dados para treinamentos e testes no software WEKA existem quatro opções:

- ***Use training set*** é a opção onde todo o conjunto de dados é utilizado como conjunto de treinamento;
- ***Supplied test set*** é possível fornecer um conjunto diferente de dados para construir o modelo*;*
- ***Cross-validation*** é a opção que permite o WEKA construir um modelo baseado em subconjuntos dos dados fornecidos e então calcular sua média para criar um modelo final;
- ***Percentage Split*** é a opção onde o WEKA toma um subconjunto percentual dos dados fornecidos para construir um modelo final. Essa opção permite que o percentual do conjunto de dados seja utilizado como conjunto de treinamento e o restante dos dados será utilizado como conjunto de teste.

Considerando um conjunto com 1000 dados se tem os seguintes cenários:

Usando o *Use training set*:

- Conforme informações da WEKA, o software pegará os 1000 dados;
- Aplicará um algoritmo para construir um classificador a partir desses 1000 dados;
- Aplicará esse classificador novamente sobre os 1000 dados;
- Fornecerá as métricas de desempenho do classificador com base nos 1000 dados.

Usando o *Cross-Validation* com 10 *folds*:

- Conforme informações da WEKA, o software pegará os 1000 dados;
- Produzirá 10 conjuntos de tamanhos iguais. Cada conjunto será dividido em dois grupos: 900 dados serão usados para treinamento e 100 dados serão usados para teste;
- Produzirá um classificador com um algoritmo sobre os 900 dados do conjunto de treinamento e aplicará isso nos 100 dados do conjunto de teste do conjunto 1;
- Repetirá para os conjuntos de 2 a 10 e produzirá 9 outros algoritmos;
- Calculará a média das métricas de desempenho dos 10 classificadores produzidos a partir dos 10 conjuntos de tamanhos iguais (900 dados de treinamento e 100 dados de teste).

Usando o *Percentage Split* com 70%:

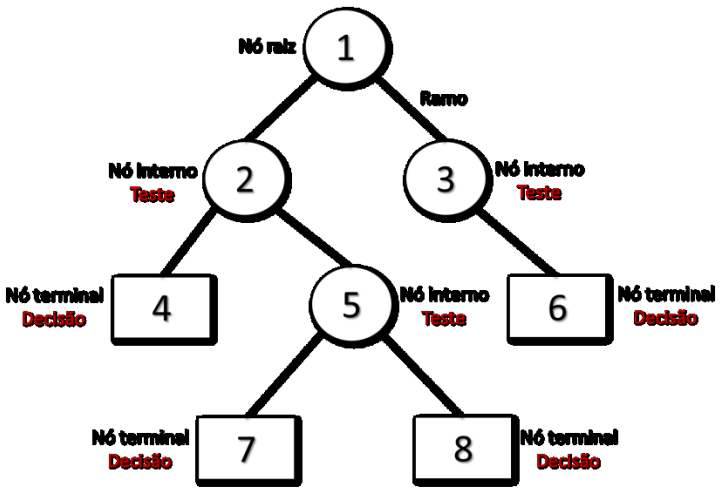
- Conforme informações da WEKA, o software pegará os 1000 dados;
- Os 1000 dados serão distribuídos aleatoriamente com base no número informado em “Mais opções”;
- Produzirá um classificador com um algoritmo sobre os 70% das instâncias nos dados embaralhados como conjunto de treinamento;
- Repetirá esse classificador para os 30% dos dados embaralhados restantes como conjunto de teste;
- Fornecerá as métricas de desempenho do classificador com base nos dados de teste (30% dos dados embaralhados).

3.2.4.2 Árvore de Decisão

Para Roiger (2017), as Árvores de Decisão são estruturas muito usadas para a aprendizagem supervisionada e que inúmeros artigos apresentaram aplicações bem-sucedidas de modelos de Árvore de Decisão para problemas do mundo real.

De acordo com Prass (2009), Árvore de Decisão é um modelo preditivo, conforme classificação mostrada na Figura 3.5, que pode ser visualizado na forma de uma árvore, daí seu nome. Cada ramo da árvore é uma questão de classificação e cada folha é uma partição do conjunto de dados com sua classificação. Para Nogueira (2015), as Árvores de Decisões constroem modelos de classificação, também mostrada na Figura 3.5, ou de regressão na forma de uma estrutura de árvore, transforma um conjunto de dados em subconjuntos cada vez menores na forma de uma árvore, com o objetivo de transformar um grande problema em problemas menores, facilitando assim a tomada de decisões. Já para Roiger (2017), uma Árvore de Decisão é uma estrutura simples em que todos os nós não terminais representam testes em um ou mais atributos e os nós terminais refletem resultados da decisão conforme mostrada na Figura 3.6, além disso, possuem inúmeras vantagens, pois são fáceis de serem entendidas, podem ser transformadas regras e experimentalmente demonstram funcionalidade.

Figura 3.6: Estrutura genérica de uma Árvore de Decisão.



Fonte: Elaborada pelo autor (2017).

As árvores de decisões, segundo Mitchell (1997), classificam instâncias, classificando-os da raiz da árvore para baixo até algum nó terminal, que prevê a classificação da instância. Cada nó da árvore especifica um ensaio de algum atributo do exemplo, e cada ramo descendente, a partir desse nó corresponde a um dos valores possíveis para este atributo. Para Prass (2009), a forma de execução é simples: dado um conjunto de dados cabe ao usuário escolher uma das variáveis como objeto de saída. A partir daí o algoritmo encontra o fator mais importante correlacionado com a variável de saída e defini-lo como o primeiro ramo (chamado de raiz), os demais fatores são subsequentemente são classificados como nós até que se chegue ao último nível, a folha ou nó terminal.

Desta forma, a árvore de decisão utiliza a estratégia de dividir para conquistar, um problema complexo é decomposto em subproblemas mais simples e recursivamente a mesma estratégia é aplicada a cada subproblema.

3.2.4.2.1 Exemplo simples de Árvore de Decisão

Com o objetivo de evidenciar o funcionamento de uma árvore de decisão, Roiger (2017) apresenta um exemplo apresentado na Tabela 3.2, onde a mesma é configurada no formato do valor do atributo e na primeira linha são apresentados os nomes de cada atributo e os respectivos valores são apresentados nas demais linhas. Os atributos de dor de garganta, febre, glândulas inchadas, congestionamento e dor de cabeça são possíveis sintomas experimentados por indivíduos que sofrem de uma aflição particular (faringite, resfriado ou alergia). Esses atributos são conhecidos como atributos de entrada e são usados para criar um modelo para representar os dados. O diagnóstico é o atributo cujo valor se deseja prever e é conhecido como classe ou atributo de saída.

Tabela 3.2 - Dados de treinamento hipotético para diagnóstico de doenças.

Identificação do Paciente	Dor de Garganta	Febre	Glândulas Inchadas	Congestio-namento	Dor de Cabeça	Diagnóstico
1	Sim	Sim	Sim	Sim	Sim	Faringite
2	Não	Não	Não	Sim	Sim	Alergia
3	Sim	Sim	Não	Sim	Não	Resfriado
4	Sim	Não	Sim	Não	Não	Faringite
5	Não	Sim	Não	Sim	Não	Resfriado
6	Não	Não	Não	Sim	Não	Alergia
7	Não	Não	Sim	Não	Não	Faringite
8	Sim	Não	Não	Sim	Sim	Alergia
9	Não	Sim	Não	Sim	Sim	Resfriado
10	Sim	Sim	Não	Sim	Sim	Resfriado

Fonte: Roiger (2017).

Cada instância de dados está representada a partir da segunda linha.

Cada linha individual mostra os sintomas e a aflição de um único paciente. Por exemplo, o paciente com identificação igual a 2 teve o diagnóstico de alergia, pois tem congestionamento e dor de cabeça.

Embora o conjunto de dados da Tabela 3.2 ser pequeno, é difícil desenvolver uma representação geral, no entanto um algoritmo de aprendizagem supervisionado apropriado pode fazer esse trabalho conforme afirma Roiger (2017).

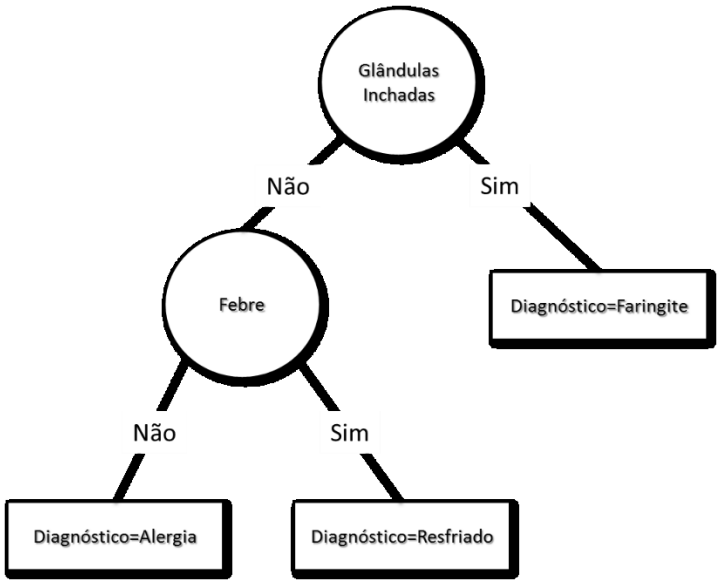
A Árvore de Decisão criada a partir dos dados da Tabela 3.2 é mostrada na Figura 3.7 onde pode ser observado:

• Se um paciente tiver glândulas inchadas, o diagnóstico da doença é faringite;

• Se um paciente não tem glândulas inchadas e tem febre, o diagnóstico é resfriado e

• Se um paciente não tem glândulas inchadas e não tem febre, o diagnóstico é alergia.

Figura 3.7 - Árvore de Decisão para os dados da Tabela 3.2.



Fonte: Adaptado de Roiger (2017).

A Árvore de Decisão mostrada na Figura 3.7 evidencia que os atributos de dor de garganta, congestionamento e dor de cabeça não contribuem com o diagnóstico e que um diagnóstico preciso com o conjunto de dados apresentados basta levar em consideração as glândulas inchadas e febre do paciente.

Para a realização da validação dos dados é apresentada a Tabela 3.3 que não possui dados para diagnóstico.

Tabela 3.3 - Instância de dados com classificação desconhecida.

Identificação do Paciente	Dor de Garganta	Febre	Glândulas Inchadas	Congestio-namento	Dor de Cabeça	Diagnóstico
11	Não	Não	Sim	Sim	Sim	?
12	Sim	Sim	Não	Não	Sim	?
13	Não	Não	Não	Não	Sim	?

Fonte: Roiger (2017).

Utilizando-se a Árvore de Decisão da Figura 3.7 para classificar as duas primeiras instâncias da Tabela 3.3 encontra-se:

- Considerando que o paciente 11 tem o valor *Sim* para *Glândula Inchadas*, segue-se a linha do ramo direito do nó raiz da Árvore de Decisão. Esse ramo chega até o nó terminal indicando que o diagnóstico é faringite;

- Considerando que o paciente 12 tem o valor *Não* para *Glândula Inchadas*, segue-se a linha do ramo esquerdo do nó raiz da Árvore de Decisão até chegar ao atributo *febre*. Considerando que *febre* possui valor *Sim*, segue-se esse ramo até chegar ao nó terminal indicando que o diagnóstico é resfriado.

A Árvore de Decisão apresentada na Figura 3.3 pode ser representada por um conjunto de regras conforme abaixo:

1. **SE** *Glândulas Inchadas* = *Sim*

   ENTÂO *Diagnóstico* = *Faringite*

2. **SE** *Glândulas Inchadas* = *Não* **E** *Febre* = *Sim*

   ENTÂO *Diagnóstico* = *Resfriado*

3. **SE** *Glândulas Inchadas* = *Não* **E** *Febre* = *Não*

   ENTÂO *Diagnóstico* = *Alergia*

Utilizando-se as regras de produção para classificar o paciente 13 temos que a regra 1 deve ser passada, pois o valor de *Glândulas Inchadas* é *Não*. Como *Febre* é *Não*, então a regra 2 não deve ser aplicada. A regra 3 deve ser aplicada, pois o paciente possui *Glândulas Inchadas* igual a *Não* e *Febre* igual a *Não*, desta forma o *Diagnóstico* é *Alergia.*

Roiger (2017) afirma que as instâncias usadas para criar o modelo da Árvore de Decisão são conhecidas como *dados de treinamento* e que para validar a acurácia são utilizados os dados com classificação conhecida chamados *conjunto de teste* e isso indicará o desempenho futuro do modelo.

No final da década de 70, início da década de 80, J. Ross Quinlan desenvolve o ID3 (*Iterative Dichotomiser*), um algoritmo para geração de Árvores de Decisão. Depois Quinlan desenvolveu o C4.5 (uma versão otimizada do ID3), e que

até hoje serve como benchmark para novos métodos supervisionados (Quinlan, 1992). Foi na mesma época (1984) que um grupo de estatísticos (L. Breiman, J. Friedman, R. Olshen e C. Stone), sem conhecer o trabalho de Quinlan, desenvolveram um algoritmo e publicaram um livro chamado *Classification and Regression Trees (CART)* (Breiman *et al*.,1984). Ambos os algoritmos são considerados precursores e diversas variações que surgiram deles. Eles utilizam a estratégia de dividir-e-conquistar recursiva aplicada de cima para baixo (*top-down*).

Com o argumento de que os algoritmos tradicionais de árvore de decisão precisam carregar todo o conjunto de dados na memória, novos algoritmos capazes de acessar repositórios persistentes foram desenvolvidos: *SLIQ* (Mehta *et al.*, 1996) e *SPRINT* (Shafer et al, 1996). Gehrke (2000) apresenta um *framework* para auxiliar na execução de algoritmos de classificação e separá-los de questões relativas a escalabilidade. O *BOAT (Bootstrapped Optimistic Algorithm for Tree Construction)* utiliza-se de uma estratégia chamada de "*bootstrapping*" conforme apresenta Gehrke (1999). Chandra e Varghese (2007) apresenta uma otimização do BOAT e Chandra e Varghese (2008) uma variação usando lógica nebulosa para o *SLIQ*.

3.2.4.2.2 Construindo uma Árvore de Decisão por meio de um algoritmo.

Para construção de uma Árvore de Decisão inicialmente é selecionado um subconjunto de instâncias de um conjunto de treinamento, onde este subconjunto é utilizado para construir uma Árvore de Decisão. O restante das instâncias do conjunto de treinamento é utilizado para testar a precisão da árvore construída. Desta forma o procedimento termina quando a Árvore de Decisão classificar as instâncias corretamente. Caso uma instância for classificada de forma incorreta, a instância é inserida ao subconjunto de instâncias de treinamento e é construída uma nova árvore e esse processo se repetirá até criação de uma árvore que classifique corretamente todas as instâncias não selecionadas ou até que a árvore de decisão seja construída a partir do conjunto de treinamento completo.

Roiger (2017) sugere as etapas seguintes para o entendimento sobre a construção de um algoritmo e considerando o conjunto de treinamento total.

1. Considerar CIT o Conjunto de Instâncias de Treinamento;

2. Escolher um atributo que possa melhor diferenciar as instâncias contidas em CIT.

3. Criar um nó na árvore cujo valor seja o atributo escolhido. Criar ramos para os nós filhos deste nó onde cada ramo representa um valor exclusivo para o atributo escolhido. Usar os valores do ramo filho para subdividir as instâncias em subclasses;

4. Considerar os seguintes passos para as subclasses criadas na etapa 3:

a) Especificar a classificação para novas instâncias seguindo este caminho de decisão, se as instâncias na subclasse satisfizerem critérios pré-definidos ou se o conjunto de opções de atributo restantes para esse caminho da árvore for nulo;

b) Considerar CIT o conjunto atual de instâncias de subclasse e retornar ao passo 2 se a subclasse não satisfizer os critérios predefinidos e existe pelo menos um atributo para subdividir ainda mais o caminho da árvore.

Para melhor entendimento, considera-se desenvolver um modelo preditivo onde os atributos de entrada são limitados a Faixa de Renda, Seguro de Cartão de Crédito e Gênero e Idade.

Considerando o passo 1, a Tabela 3.4 apresenta os dados de treinamento selecionados.

Tabela 3.4 - Base de dados da promoção do cartão de crédito.

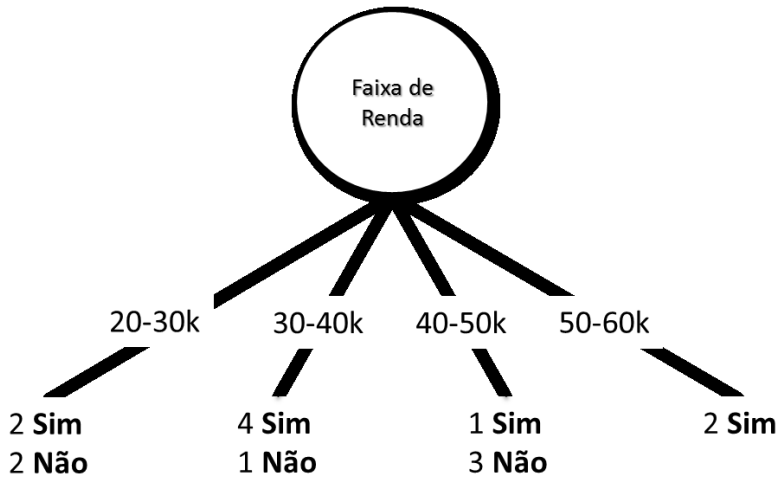
Identificação	Faixa de Renda	Promoção de Seguro de Vida	Seguro de Cartão de Crédito	Gênero	Idade
1	40-50k	Não	Não	Masculino	45
2	30-40k	Sim	Não	Feminino	40
3	40-50k	Não	Não	Masculino	42
4	30-40k	Sim	Sim	Masculino	43
5	50-60k	Sim	Não	Feminino	38
6	20-30k	Não	Não	Feminino	55
7	30-40k	Sim	Sim	Masculino	35
8	20-30k	Não	Não	Masculino	27
9	30-40k	Não	Não	Masculino	43
10	30-40k	Sim	Não	Feminino	41
11	40-50k	Sim	Não	Feminino	43
12	20-30k	Sim	Não	Masculino	29
13	50-60k	Sim	Não	Feminino	39
14	40-50k	Não	Não	Masculino	55
15	20-30k	Sim	Sim	Feminino	19

Fonte: Roiger (2017).

Após a seleção dos dados, é necessário executar o passo 2. Para isso entre os atributos Faixa de Renda, Seguro de Cartão de Crédito, Gênero e Idade, foi selecionado inicialmente o atributo Faixa de Renda.

Na Figura 3.8 é mostrada a Árvore de Decisão parcial criada no passo 3 com o intervalo da Faixa de Renda escolhido como nó principal.

Figura 3.8 - Árvore de Decisão parcial com nó raiz Faixa de Renda



Fonte: Adaptado de Roiger (2017).

Na parte inferior de cada ramo são mostrados os totais de respostas sim e não para o atributo de saída Promoção do Seguro de Vida.

A análise da árvore segue abaixo:

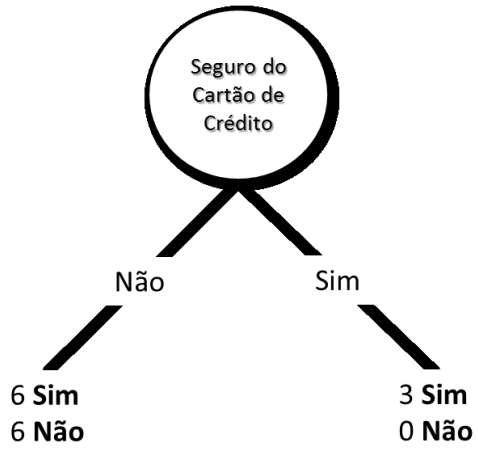
- Faixa de Renda 20-30k: Empate, pois se tem 2 Sim e 2 Não para a Promoção do Seguro de Vida. Para desempatar opta-se pelo mais frequente na árvore, como existem 9 respostas Sim e 6 respostas Não, tem-se que para esse ramo a resposta é Sim. Como a resposta foi Sim, então classificou corretamente 2, pois teve 2 Sim;
- Faixa de Renda 30-40k: A resposta é Sim, pois se tem 4 Sim e 1 Não. Como a resposta foi Sim, então classificou corretamente 4, pois teve 4 Sim;
- Faixa de Renda 40-50k: A resposta é Não, pois se tem 1 Sim e 3 Não. Como a resposta foi Não, então classificou corretamente 3, pois teve 3 Não;
- Faixa de Renda 50-60k: A resposta é Sim, pois o Não é zero. Como a resposta foi Sim, então classificou corretamente 2, pois teve 2 Sim;

Para esse resultado parcial se observa que a árvore classificou corretamente 11 (2 + 4 + 3 + 2) das 15 instâncias do conjunto de treinamento, então teve percentual de acerto acima de 73%.

O índice de qualidade da raiz é considerado pela divisão do percentual de acerto (73%) pela quantidade de ramos (4), onde se encontra que para esse caso é de 0,18.

Para o Seguro do Cartão de Crédito como nó raiz encontra-se a Árvore de Decisão apresentada na Figura 3.9 na execução do passo 3 do algoritmo apresentado.

Figura 3.9 - Árvore de Decisão parcial com nó raiz Seguro do Cartão de Crédito



Fonte: Adaptado de Roiger (2017).

Aplicando o mesmo conceito do anterior tem-se:

- Seguro do Cartão de Crédito Não: Encontra-se empate, pois se tem 6 Sim e 6 Não para a Promoção do Seguro de Vida. Para desempatar opta-se pelo mais frequente na árvore, como existem 9 respostas Sim e 6 respostas Não, tem-se que para esse ramo a resposta é Sim. Como a resposta foi Sim, então classificou corretamente 6, pois teve 6 Sim;
- Seguro do Cartão de Crédito Sim: A resposta é Sim, pois se tem 3 Sim e 0 Não. Como a resposta foi Sim, então classificou corretamente 3, pois teve 3 Sim;

Para esse resultado parcial se observa que a árvore classificou corretamente 9 (6 + 3) das 15 instâncias do conjunto de treinamento, então teve percentual de acerto de 60%.

Como o índice de qualidade da raiz é considerado pela divisão do percentual de acerto (60%) pela quantidade de ramos (2), onde se encontra que para esse caso é de 0,30, onde se conclui que o Seguro de Cartão de Crédito pode ser melhor para o nós raiz do que a Faixa de Renda.

O atributo Idade é um dado numérico e desta forma precisa ser agrupado conforme o atributo Seguro do Cartão de Crédito e desta forma os valores são classificados conforme Tabela 3.5.

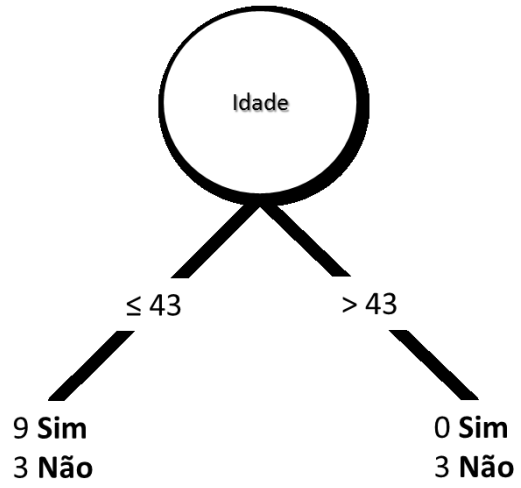
Tabela 3.5 - Atributo numérico Idade ordenado.

19	**27**	**29**	**35**	**38**	**39**	**40**	**41**	**42**	**43**	**43**	**43**	**45**	**55**	**55**
S	N	S	S	S	S	S	S	N	S	S	N	N	N	N

Fonte: Adaptado de Roiger (2017).

Observando os dados ordenados se tem que a melhor divisão acontece em 43, pois acima de 43 o resultado é Não. Desta forma organiza-se o atributo Promoção do Seguro de Vida = Sim com Idade ≤ 43 e Promoção do Seguro de Vida = Não com Idade > 43, pois a maioria dos dados abaixo de 43 é Sim (9 Sim e 3 Não) e acima de 43 é Não (3 Não) conforme mostrado na Figura 3.10. Assim sendo, a precisão do conjunto de treinamento é de 80% (12 de 15) e o índice de qualidade é de 0,40.

Figura 3.10 - Árvore de Decisão parcial com nó raiz Idade



Fonte: Adaptado de Roiger (2017).

Para finalizar, foi executado o passo 3 com o atributo Gênero e o índice de qualidade encontrado foi 0,367. Desta forma entre as 4 Árvores de Decisão parciais o melhor resultado foi para o atributo idade que atingiu um índice de qualidade de 0,40.

Para o passo 4 o processo é repetido até encontrar um critério de encerramento ou até que todas as possibilidades de seleção de atributos tenham sido esgotadas.

3.2.4.2.3 Algoritmos C4.5 e J48

Nogueira (2015) afirma que o algoritmo C4.5 é sucessor do algoritmo ID3 que foi idealizado por Ross Quinlan em 1979 e que entre as principais características está o fato de poder auxiliar em tarefas de classificação, que em alguns momentos necessitam somente de uma resposta de sim ou não, ou seja, C4.5 gera um classificador que é capaz de atuar como um especialista, classificando inclusive os casos desconhecidos. Segundo Roiger (2017), o C4.5 seleciona o atributo que divide os dados de modo a mostrar a maior quantidade de ganho em informações para qualquer ponto de escolha na árvore.

As fórmulas usadas pelo algoritmo C.45 para seleção dos atributos, onde o atributo de maior índice de ganho é o atributo selecionado para subdividir a estrutura da árvore. Em (3.1) pode ser visto a fórmula para calcular a relação de ganho para o atributo A.

$$\text{Taxa de Ganho (A)} = \text{Ganho (A)} / \text{Info Dividida (A)} \tag{3.1}$$

Para o conjunto de I Instâncias, a fórmula para calcular o Ganho(A) é dados como:

$$\text{Ganho (A)} = \text{Info (I)} - \text{Info (I,A)} \tag{3.2}$$

Onde:

Info (I) é a informação contida no conjunto de instâncias atualmente examinadas e Info (I, A) é a informação depois de particionar as instâncias em I de acordo com os possíveis resultados para o atributo A.

As fórmulas para o cálculo de Info (I), Info (I,A) e Info Dividida (A) são apresentadas em (3.3), (3.4) e (3.5).

$$\text{Info (I)} = -\sum_{i=1}^{n} \frac{\#\text{ na classe i}}{\#\text{ em I}} \log_2\left(\frac{\#\text{ na classe i}}{\#\text{ em I}}\right) \qquad (3.3)$$

Depois de I ser dividido em em k resultados, Info (I,A) é calculado como:

$$\text{Info (I,A)} = \sum_{j=1}^{k} \frac{\#\text{ na classe j}}{\#\text{ em I}} \text{ Info (classe j)} \qquad (3.4)$$

Info Dividida (A) normaliza o cálculo do ganho para eliminar uma tendência para a escolha de atributos com muitos resultados conforme fórmula abaixo:

$$\text{Info Dividida (A)} = -\sum_{j=1}^{k} \frac{\#\text{ na classe j}}{\#\text{ em I}} \log_2\left(\frac{\#\text{ na classe j}}{\#\text{ em I}}\right) \qquad (3.5)$$

Para aplicação das fórmulas vamos considerar os dados relativos a Tabela 3.4 correspondendo a base de dados da promoção do cartão de crédito.

De acordo com a Tabela 3.4, existem 15 instâncias, logo:

em I = 15.

Para a saída *Promoção de Seguro de Vida*, tem-se 9 valores *Sim* e 6 valores *Não*. Logo n=2, pois existem 2 classes (*Sim* e *Não*), então:

em classe 1 = 9 (*Classe Sim)*,

em classe 2 = 6 (Classe *Não*).

O cálculo de Info (I) é

$$\text{Info (I)} = -[9/15 \log_2 9/15 + 6/15 \log_2 6/15] = 0{,}97095$$

A saída *Faixa de Renda* tem quatro possíveis resultados, então k=4 e o cálculo de Info (I, *Faixa de Renda*) é:

$$\text{Info (I,}\textit{Faixa de Renda}\text{)} = 4/15 \text{ Info (20-30k)} + 5/15 \text{ Info (30-40k)} + 4/15 \text{ Info (40-50k)} + 2/15 \text{ Info (50-60k)}$$

$$\text{Info (I,}\textit{Faixa de Renda}\text{)} = 0{,}72365$$

Onde:

$$\text{Info (20-30k)} = -[2/4 \log_2 2/4 + 2/4 \log_2 2/4]$$
$$\text{Info (30-40k)} = -[4/5 \log_2 4/5 + 1/5 \log_2 1/5]$$
$$\text{Info (40-50k)} = -[3/4 \log_2 3/4 + 1/4 \log_2 1/4]$$
$$\text{Info (50-60k)} = -[2/2 \log_2 2/2]$$

e

$$\text{Info Dividida (}\textit{Faixa de Renda}\text{)} = -[4/15 \log_2 4/15 + 5/15 \log_2 5/15 + 4/15 \log_2 4/15 + 2/15 \log_2 2/15] = 1{,}93291$$

Desta forma o cálculo do Ganho é:

$$\text{Ganho (}\textit{Faixa de Renda}\text{)} = \text{Info (I)} - \text{Info (I,}\textit{Faixa de Renda}\text{)}$$
$$\approx 0{,}97905 - 0{,}72365 = 0{,}25540$$

Finalmente,

$$\text{Taxa de Ganho (}\textit{Faixa de Renda}\text{)} =$$
$$= \text{Ganho (}\textit{Faixa de Renda}\text{) / Info Dividida (}\textit{Faixa de Renda}\text{)}$$
$$= 0{,}25540 / 1{,}93291 = 0{,}13302.$$

Para os atributos categóricos *Seguro do Cartão de Crédito* e *Gênero,* os cálculos são realizados de forma similar. Para o atributo numérico *Idade*, é necessário discretizá-lo. Finalmente, ao fazer cálculos para a *Faixa de Renda*, *Seguro do Cartão de Crédito*, *Gênero* e *Idade,* descobrimos que o *Seguro do Cartão de Crédito* tem o melhor Taxa de Ganho de 3.610.

Nogueira (2015) afirma que o algoritmo C4.5 é escrito em originalmente em linguagem C e que o algoritmo J48 é uma recodificação do C4.5 escrito em linguagem Java e que tem boa aceitação no meio acadêmico e por especialistas pela característica de ser adequado nos procedimentos que relacionam variáveis quantitativas discretas e contínuas.

3.2.4.2.4 Vantagens e Desvantagens das Árvores de Decisão

Segundo Roiger (2017), as Árvores de Decisão possuem inúmeras vantagens e desvantagens conforme apresentas na Tabela 3.6:

Tabela 3.6 - Vantagens e desvantagens das Árvores de Decisão

Vantagens	Desvantagens
• São fáceis de entender e delineiam satisfatoriamente o conjunto de regras de produção; • Já foram aplicadas com sucesso em casos reais; • Não fazem pressupostos anteriores sobre a natureza dos dados; • São capazes de construir modelos com conjuntos de dados contendo números, bem como dados categóricos.	• Os atributos de saída devem ser categóricos e vários atributos de saída não são permitidos; • Os algoritmos de árvore de decisão são instáveis em que pequenas variações nos dados de treinamento podem resultar em diferentes seleções de atributos em cada ponto de escolha dentro da árvore. O efeito pode ser significante, pois as escolhas de atributo afetam todas as sub-árvores descendentes; • As árvores criadas a partir de conjuntos de dados numéricos podem ser bastante complexas, pois as divisões de atributos para dados numéricos normalmente são binárias.

Fonte: Adaptado de Roiger (2017).

3.2.4.3 Rede Bayesiana – Naïve Bayes

Veiga e Da Silva (2002) afirmam que as redes bayesianas são conhecidas como um modelo gráfico que representa uma distribuição de probabilidade conjunta e que se utiliza os gráficos da computação com fundamentos matemáticos da probabilidade e estatística.

Redes Bayesianas, tem origem no teorema de Bayes do reverendo Thomas Bayes, porém desenvolvido por Simon de LaPlace em 1812 e são ferramentas poderosas para decisão e raciocínio sob incertez e que mostra de forma clara as dependências de causas entre as variáveis em função de sua forma gráfica. Uma forma muito simples de redes Bayesianas é chamada de classificadores *Naïve Bayes* dedicada à tarefa de classificação conforme afirmam Langley et al (1992, 1994), Grossman e Domingos (2004) e Duda, Hart e Stork (1995).

Comparativos mostram que os algoritmos Bayesianos, chamados de *Naïve Bayes*, obtiveram resultados compatíveis com os métodos de Árvore de Decisão e redes neurais. Devido a sua simplicidade e o alto poder preditivo, é um dos algoritmos mais utilizados segundo afirma Zhang (2004). O algoritmo de *Naïve Bayes* parte do princípio que não exista relação de dependência entre os atributos. No entanto, nem sempre isto é possível. Nestes casos, uma variação conhecida como *Bayesian Belief Networks*, ou *Bayesian Networks*, deve ser utilizada conforme afirma Niedermayer (2008). Em Hall & Frank (2008), é proposta uma combinação dos algoritmos de *Naïve Bayes* e Árvore de Decisão para realizar a classificação.

3.2.4.3.1 Teorema de Bayes

O teorema de Bayes é expresso na forma conforme a equação 3.6.

$$P(A|B)=\frac{P(B|A)\,P(A)}{P(B)} \tag{3.6}$$

Onde *A* e *B* são eventos e $P(B) \neq 0$.

O teorema de Bayes pode ser reescrito da seguinte forma:

$$P(A|B)\,P(B) = P(A \cap B) = P(B \cap A) = P(B|A)\,P(A) = P(A|B)\,P(B)$$

Onde:

P(A) e *P(B)* são as probabilidades a *priori* de *A* e *B*;

P(A|B) é a probabilidade a *posteriori* (probabilidade condicionada) de *A* condicional a *B;*

P(B|A) é a probabilidade a *posteriori* (probabilidade condicionada) de *B* condicional a *A.*

3.2.4.3.2 Classificador Naïve Bayes

A rede Bayesiana é a forma mais simples existente e utiliza conceitos da probabilidade Bayesiana ingênua, onde uma variável meta é escolhida com valores desconhecidos D o qual é colocada em evidência. Conforme afirma Roiger (2017), o classificador Naïve Bayes oferece uma técnica de classificação supervisionada simples e poderosa.

Considera que todas as outras variáveis do domínio são independentes condicionais dado a ocorrência de D e são conhecidas como características Fi. Desta forma se chega a uma distribuição conjunta de todas as características que satisfaz:

$$P(D,F_1,...,F_K) = (\prod_{i=1}^{K} P(F_1|D)) P(D) \quad (3.7)$$

A equação 3.7 considera que para a distribuição da variável meta D é necessário uma entrada numérica e a distribuição condicional de cada uma das características para cada uma das classes de metas, onde esta pode ser facilmente estimada em um conjunto de treinamento se tiver um exemplo aleatório Desta forma, calcular a probabilidade posterior de cada meta sobre a base valores (ocorridos) observados é extremamente direto. Este modelo simples é conhecido como Bayes Ingênuo (“Naïve Bayes”) citado em (Titterington et al., 1981).

Warner et al. (1961) utilizou pela primeira vez o modelo Naïve Bayes para diagnósticos de doenças congênitas do coração e até hoje é utilizado em várias outras aplicações. Este modelo permite uma simples aplicação do teorema de Bayes, e a partir das suposições de independência condicional indicamos que cada item de evidência pode ser considerado em ordem, com a distribuição de probabilidade posterior para a meta D depois de observar cada item F_i , tornando-se a distribuição de probabilidade à priori para a próxima. Cowell (1999) afirma que a esparsidade do modelo gráfico leva diretamente para uma forma modular de inferência.

Roiger (2017) afirma que o classificador Naïve Bayes é baseado no teorema de Bayes conforme apresentado na equação 3.6 e reescrito na equação 3.8.

$$P(H/E)=\frac{P(E/H)\,P(H)}{P(E)} \quad (3.8)$$

Onde *H* é a hipótese a ser testada e *E* é a evidência associada com a hipótese e do ponto de vista da classificação, a hipótese é a variável dependente e representa a classe prevista e a evidência é determinada por valores dos atributos de entrada.

P(H|E) é a *probabilidade condicional* de que *H* é verdadeiro, dada a evidência *E*. *P(H)* é uma *probabilidade a priori*, que denota a probabilidade da hipótese antes da apresentação de qualquer evidência. As probabilidades condicionais e a priori são facilmente calculadas a partir dos dados de treinamento.

3.2.4.3.3 Exemplo com o classificador Naïve Bayes

Para melhor entendimento do funcionamento do classificador Naïve Bayes através de uma aplicação, consideram-se os dados apresentados na Tabela 3.7, que é um subconjunto da base de dados apresentada na Tabela 3.4.

Para a aplicação do Naïve Bayes utilizou-se *gênero* como atributo de saída cujo valor deve ser previsto e a Tabela 3.8 apresenta a listagem dos valores dos atributos de saída para cada atributo de entrada e evidencia que quatro homens aproveitaram a promoção da revista e que esses quatro homens representam dois terços da população masculina total e indica também que três das quatro instâncias do conjunto de dados femininas adquiriram a promoção da revista.

Tabela 3.7 - Dados para o classificador Naïve Bayes.

Promoção de Revista	Promoção de Relógio	Promoção de Seguro de Vida	Seguro de Cartão de Crédito	Gênero
Sim	Não	Não	Não	Masculino
Sim	Sim	Sim	Sim	Feminino
Não	Não	Não	Não	Masculino
Sim	Sim	Sim	Sim	Masculino
Sim	Não	Sim	Não	Feminino
Não	Não	Não	Não	Feminino
Sim	Sim	Sim	Sim	Masculino
Não	Não	Não	Não	Masculino
Sim	Não	Não	Não	Masculino
Sim	Sim	Sim	Não	Feminino

Fonte: Adaptado de Roiger (2017).

Tabela 3.8 - Contagem e probabilidade para o atributo *Gênero*

Gênero	Promoção de Revista		Promoção de Relógio		Promoção de Seguro de Vida		Seguro de Cartão de Crédito	
	Masculino	Feminino	Masculino	Feminino	Masculino	Feminino	Masculino	Feminino
Sim	4	3	2	2	2	3	2	1
Não	2	1	4	2	4	1	4	3
Taxa: Sim/total	4/6 (2/3)	3/4	2/6 (1/3)	2/4 (1/2)	2/6 (1/3)	3/4	2/6 (1/3)	1/4
Taxa: Não/total	2/6 (1/3)	1/4	4/6 (2/3)	2/4 (1/2)	4/6 (2/3)	1/4	4/6 (2/3)	3/4

Fonte: Adaptado de Roiger (2017).

Para exemplificação será utilizado o classificador Bayes para realizar uma nova classificação nos dados da Tabela 3.8. Considera-se a novas instâncias abaixo:

- *Promoção de Revista = Sim;*
- *Promoção de Relógio = Sim;*
- *Promoção de Seguro de Vida = Não;*
- *Seguro de Cartão de Crédito = Não;*
- *Gênero = ?*

Para esse exemplo, têm-se duas hipóteses a serem testadas:

- A primeira hipótese afirma que o titular do cartão de crédito é masculino;
- A segunda hipótese vê a instância como um titular de cartão feminino;

Considerando as hipóteses apresentadas e para determinar qual hipótese é correta, será aplicado Naïve Bayes para calcular uma probabilidade para cada hipótese. A equação 3.9 apresenta o cálculo da probabilidade de que o cliente seja masculino.

$$P(\text{gênero=masculino}/E) = \frac{P(E/\text{gênero=masculino})\, P(\text{gênero=masculino})}{P(E)} \qquad (3.9)$$

Considerando-se a equação 3.9 é possível observar que a probabilidade condicional *P(E|gênero=masculino)* é calculada multiplicando-se os valores da

probabilidade condicional por cada evidência. A probabilidade condicional global é o produto das quatro probabilidades condicionais seguintes:

1. *P(promoção de revista=sim | gênero=masculino)*=4/6;
2. *P(promoção de relógio=sim | gênero=masculino)*=2/6;
3. *P(promoção de seguro de vida=não | gênero=masculino)*=4/6;
4. *P(seguro de cartão de crédito=não | gênero=masculino)*=4/6;

Os valores foram retirados da Tabela 3.8, portanto a probabilidade condicional *P(gênero=masculino)* é calculada como:

$$P(E|gênero=masculino) = (4/6)(2/6)(4/6)(4/6) = 8/81$$

Para o cálculo da probabilidade a priori *P(gênero=masculino)* da equação 3.9, considera-se que é a probabilidade de um cliente masculino sem o conhecimento do histórico de oferta promocional das instâncias. Como existem dez cliente, onde seis clientes são do gênero masculino e quatro clientes são do gênero feminino, a probabilidade a priori para *gênero=masculino* é 6/10 ou 3/5. Desta forma o numerador da equação 3.9 fica:

$$P(E|gênero=masculino)\ P(gênero=masculino) = (8/81)\ (3/5) \approx 0{,}0593$$

Desta forma, tem-se:

$$P(gênero= masculino|E) \approx 0{,}0593 / P(E)$$

Para o cálculo da probabilidade que o cliente seja feminino tem-se a equação 3.10.

$$P(gênero=feminino/E) = \frac{P(E/gênero=feminino)\ P(gênero=feminino)}{P(E)} \qquad (3.10)$$

Utilizando-se os mesmos passos do anterior com base nos dados da Tabela 3.8, tem-se o seguinte para o gênero feminino:

1. *P(promoção de revista=sim | gênero=feminino)*=3/4;
2. *P(promoção de relógio=sim | gênero=feminino)*=2/4;

3. *P(promoção de seguro de vida=não | gênero=feminino)*=1/4;
4. *P(seguro de cartão de crédito=não | gênero=feminino)*=3/4;

A probabilidade condicional global é:

$$P(E|gênero=feminino) = (3/4)(2/4)(1/4)(3/4) = 9/128$$

Como existem 4 clientes femininas, a probabilidade a priori para *P(gênero=feminino)* é 4/10 ou 2/5, portanto o numerador da equação 3.10 fica:

$$P(E|gênero=feminino)\ P(gênero=feminino) = (9/128)\ (2/5) \approx 0{,}0281$$

Desta forma, tem-se:

$$P(gênero=feminino|E) \approx 0{,}0281 / P(E)$$

Nas duas hipóteses é apresentada *P(E)* como denominado e representa tanto a probabilidade da evidência quando o *gênero = masculino* ou *gênero = feminino,* logo não é necessário considerar *P(E).* Como 0,0593 > 0.0281, o classificador Naïve Bayes apresenta como a probabilidade maior é de que um cliente de cartão de crédito ser do gênero masculino.

3.2.4.3.4 Atributo com valor zero

A técnica Naïve Bayes apresenta um problema significante quando o número de um atributo tem valor 0. Considerando-se que o número de clientes femininos com valor *Não* para *Seguro do Cartão de Crédito* fosse 0, o numerador da equação 3.10 seria 0. Isto significa que todos os atributos são irrelevantes, pois a multiplicação terá como resultado 0.

Para solucionar esse problema, pode-se acrescenta uma constante *k* para o numerado (*n*) e para o denominador (*d*) para cada cálculo. Portanto cada taxa da forma *n/d* torna-se:

$$\text{Taxa n/d} = \frac{n+(k)(p)}{d+k} \qquad (3.11)$$

Onde:

k é um valor entre 0 e 1 (normalmente 1)

p é escolhido como uma parte fracionada igual do número total de valores possíveis para o atributo (por exemplo, se um atributo tiver dois valores possíveis, *p* será 0,5)

Utilizando-se essa técnica no cálculo da probabilidade condicional *P(E|gênero=feminino)* para o exemplo anterior, considerando *k*=1 e *p*=0,5, logo:

$$\frac{(3+0,5)(2+0,5)(1+0,5)(3+0,5)}{(4+1)(4+1)(4+1)(4+1)} \approx 0,0176$$

3.2.4.3.5 Dados faltantes

O classificar Naïve Bayes não apresenta problema com dados faltantes conforme pode ser evidenciado no exemplo abaixo com base na Tabela 3.8:

- *Promoção de Revista = Sim;*
- *Promoção de Relógio = Desconhecido;*
- *Promoção de Seguro de Vida = Não;*
- *Seguro de Cartão de Crédito = Não;*
- *Gênero = ?*

Como o valor do atributo *Promoção de Relógio* é desconhecido, o mesmo pode ser ignorado no cálculo da probabilidade condicional. Assim tem-se:

$$P(E|gênero=masculino) = (4/6)(4/6)(4/6) = 8/27$$

$$P(E|gênero=feminino) = (3/4)(1/4)(3/4) = 9/64$$

$$P(gênero=masculino|E) \approx 0,1778/P(E)$$

$$P(gênero=feminino|E) \approx 0,05625/P(E)$$

Se for atribuído qualquer valor para o atributo *Promoção de Relógio* o resultado não será alterado, pois afetará da mesma forma as duas hipóteses.

3.2.4.3.6 Dado numérico

Uma vez conhecida a função de densidade de probabilidade que representa a distribuição dos dados é possível tratar os dados numéricos de forma semelhante. A equação 3.12, onde uma variável X tem distribuição normal com parâmetros μ e σ^2 se sua função densidade de probabilidade é dada por:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (3.12)$$

Onde:

e = a função exponencial
μ = a média da classe para o atributo numérico dado
σ = o desvio padrão da classe para o atributo
x = o valor do atributo

Aparentemente a equação 3.12 parece muito complicada, no entanto é de fácil aplicação como demonstrado a seguir. Considere os dados da Tabela 3.9 que tem como base os dados da Tabela 3.7 com a adição do atributo numérico *idade*.

Tabela 3.9 - Adição do atributo *idade* ao conjunto de dados do classificador Naïve Bayes.

Promoção de Revista	Promoção de Relógio	Promoção de Seguro de Vida	Seguro de Cartão de Crédito	Idade	Gênero
Sim	Não	Não	Não	45	Masculino
Sim	Sim	Sim	Sim	40	Feminino
Não	Não	Não	Não	42	Masculino
Sim	Sim	Sim	Sim	30	Masculino
Sim	Não	Sim	Não	38	Feminino
Não	Não	Não	Não	55	Feminino
Sim	Sim	Sim	Sim	35	Masculino
Não	Não	Não	Não	27	Masculino
Sim	Não	Não	Não	43	Masculino
Sim	Sim	Sim	Não	41	Feminino

Fonte: Adaptado de Roiger (2017).

Utilizando-se esse novo atributo para calcular as probabilidades condicionais para as classes masculino e feminino para as seguintes instâncias.

- *Promoção de Revista = Sim;*
- *Promoção de Relógio = Sim;*
- *Promoção de Seguro de Vida = Não;*
- *Seguro de Cartão de Crédito = Não;*
- *Idade = 45*
- *Gênero = ?*

Para a probabilidade condicional global, tem-se

$$P(E|gênero=masculino) = (4/6)(2/6)(4/6)(4/6)[P(idade=45|gênero=masculino)]$$

$$P(E|gênero=feminino) = (3/4)(2/4)(1/4)(3/4)[P(idade=45|gênero=feminino)]$$

Para determinar a probabilidade condicional do atributo *idade* considerando que *gênero=masculino*, é assumido que a idade seja normalmente distribuída e aplique a função de densidade de probabilidade. Com base nos dados da Tabela 3.9 para encontrar os escores de desvio padrão e da média. Para a classe *gênero = masculino* temos σ = 7,69, μ = 37,00 e x = 45. Portanto, a probabilidade de que a *idade = 45* dado *gênero = masculino* seja calculada como:

$$P(idade=45|gênero=masculino)=1/(\sqrt{2\pi}\ 7,69)e^{-(45-37,00)^2/2(7,69)^2}$$

Processando o cálculo, tem-se:

$$P(idade=45|gênero=masculino) \approx 0,030$$

Para determinar a probabilidade condicional do atributo *idade* para o *gênero=feminino*, e substituindo σ = 7,77, μ = 43,50 e x = 45, tem-se

$$P(idade=45|gênero=feminino)=1/(\sqrt{2\pi}\ 7,77)e^{-(45-43,50)^2/2(7,77)^2}$$

Processando o cálculo, tem-se:

$$P(idade=45|gênero=feminino) \approx 0,050$$

Considerando os valores da probabilidade do atributo *idade*, tem-se o valor da probabilidade condicional global

$$P(E|gênero=masculino) = (4/6)(2/6)(4/6)(4/6)(0,030) \approx 0,003$$

$$P(E|gênero=feminino) = (3/4)(2/4)(1/4)(3/4)(0,050) \approx 0,004$$

Aplicando o resultado nas equações 3.9 e 3.10 tem-se:

$$P(gênero = masculino|E) \cong (0,003)\ (0,60)/P(E) \cong 0,0018/P(E)$$

$$P(gênero = feminino|E) \cong (0,004)\ (0,40)/P(E) \cong 0,0016/P(E)$$

Mais uma vez *P(E)* é desprezado e a conclusão é que a instância pertence ao gênero masculino.

3.2.4.3.7 Vantagens e desvantagens do Naïve Bayes

Entre as principais vantagens e desvantagens do classificador Naïve Bayes encontram-se:

Vantagens:

- Treinamento rápido (varredura única);
- Rápido para classificar;
- Não sensível a características irrelevantes;
- Trabalha com dados reais e discretos;
- Trabalha bem com dados contínuos.

Desvantagens:

- Assume independência das características.

3.2.4.4 Análise de desempenho

A Tabela 3.10 apresenta a matriz confusão para problemas de duas classes. Considerando os dados da Tabela 3.10 é possível calcular a *acurácia* (precisão da

classificação) por meio da equação 3.13 e conforme afirmam Prati *et al.*(2003) é uma das medidas de desempenho mais utilizadas para avaliar a qualidade dos modelos.

Tabela 3.10 - Matriz de confusão para problemas de duas classes.

	Predição positiva	**Predição negativa**
Classe positiva	Verdadeiro positivo (V*P*)	Falso negativo (*FN*)
Classe negativa	Falso positivo (*FP*)	Verdadeiro negativo (V*N*)

Fonte: Prati *et al.* (2003).

$$Acc = \frac{VP+VN}{VP+FP+FN+VN} \tag{3.13}$$

Na Tabela 3.10 pode ser observado que o relacionamento entre a primeira e segunda linha representa a distribuição entre as classes (a proporção entre exemplos positivos e negativos). Desta forma, qualquer medida de desempenho que utilize valores de ambas as colunas será, necessariamente, sensível a desproporção entre as classes. Métricas como precisão, taxa de erro, entre outras, utilizam valores de ambas as linhas da matriz de confusão. Desta forma a taxa de erro na classificação é definida como $TE_R = 1 - A_{cc}$. Havendo uma mudança na distribuição de classes, os valores dessas métricas também mudarão, mesmo que o desempenho global do modelo não melhore. Os elementos ao longo da diagonal principal da matriz de confusão representam as decisões corretas: número de verdadeiros positivos (*VP*) e verdadeiros negativos (*VN*). Os elementos fora dessa diagonal representam os erros cometidos: número de falsos positivos (*FP*) e falsos negativos (*FN*).

Alberto e Almeida (2012) afirma que para conjuntos de dados com classes desbalanceadas é necessária uma medida de desempenho mais apropriada que deve desassociar os erros, ou acertos, ocorridos para cada classe. É possível derivar quatro medidas de desempenho que medem o desempenho de classificação nas classes negativa e positiva conforme equações de 3.14 a 3.17.

Taxa de Verdadeiro Positivo (Sensibilidade ou Recall): É a porcentagem de casos positivos classificados corretamente como pertencentes à classe positiva. Mede quão provável é que o teste dê positivo em alguém com a característica. Dentre todas as pessoas que têm a característica, qual proporção dará positivo?

$$TV_P = \frac{VP}{VP+FN} = 1 - TF_N \tag{3.14}$$

Taxa de Falso Positivo: É a porcentagem de casos negativos classificados incorretamente como pertencentes à classe positiva.

$$TF_P = \frac{FP}{FP+VN} \tag{3.15}$$

Taxa de Verdadeiro Negativo (Especificidade): É a porcentagem de casos negativos classificados corretamente como pertencentes à classe negativa. Mede quão provável é que o teste dê negativo em alguém que não tem a característica. Dentre todas as pessoas sem a característica, qual proporção dará negativo?

$$TV_N = \frac{VN}{FP+VN} = 1-TF_P \tag{3.16}$$

Taxa de Falso Negativo: É a porcentagem de casos positivos classificados incorretamente como pertencentes à classe negativa.

$$TF_N = \frac{FN}{VP+FN} \tag{3.17}$$

Na literatura são encontradas outras métricas conforme 3.18 a 3.20.

Precision (P) corresponde à razão entre o número de exemplos positivos corretamente classificados e o número total de exemplos identificados como positivos pelo classificador:

$$P = \frac{VP}{VP+FP} \tag{3.18}$$

O **valor preditivo positivo** (**VP_P** ou **precisão**) mede quão provável é que a pessoa tenha a característica se o teste der positivo. Dentre todas as pessoas com teste positivo, qual proporção realmente tem a característica?

Fmeasure considera somente o desempenho para a classe positiva. Ela é calculada a partir das métricas adotadas em *Recuperação de Informação*: *Recall* e *Precision* (Tan *et al.*, 2009):

$$Fmeasure = \frac{(1+\beta).R.P}{\beta^2.R+P} \tag{3.19}$$

Onde *β* é usado para ajustar a importância relativa entre *RecallI* e *Precision*. Tipicamente, *β=1.*

Gmean foi proposta por Kubat *et al.* (1998) e corresponde à média geométrica entre a sensibilidade (S_E) ou taxa de verdadeiros positivos (TV_P) e especificidade (E_S) ou taxa de verdadeiros negativos (TV_N):

$$Gmean = \sqrt{S_E.E_S} \quad \text{ou} \quad Gmean = \sqrt{TV_P.TV_N} \qquad (3.20)$$

Gmean mede o desempenho equilibrado de um classificador em relação às taxas de acertos de ambas as classes (Sun, Y. et al., 2007). Valores elevados de ***Gmean*** refletem taxas de acerto elevadas e equilibradas para ambas as classes.

O valor preditivo negativo (**VP_N**) mede quão provável é que alguém não tenha a característica se o teste der negativo. Dentre todas as pessoas que tiveram resultado negativo, qual proporção realmente não tem a característica?

O principal objetivo de qualquer classificador é minimizar as taxas de falso positivo e de falso negativo, ou, de forma similar, maximizar as taxas de verdadeiro positivo e verdadeiro negativo. Entretanto, para a maioria das aplicações do mundo real, existe uma relação de perda e ganho entre FN e FP, ou, de forma similar, entre VN e VP.

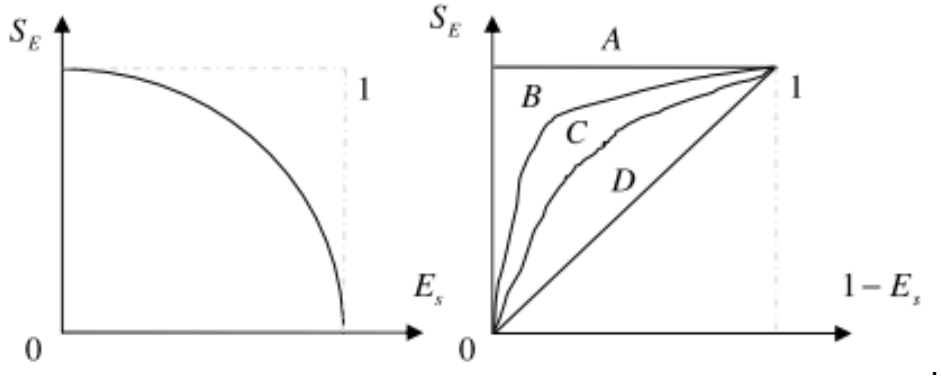
Um método alternativo para a avaliação do desempenho desses algoritmos é a análise de curvas ROC, a qual representa o compromisso entre a sensitividade no eixo *y* e falso alarme no eixo *x*. Curvas ROC são obtidas pela variação do limiar de um modelo probabilístico de classificação.

A Curva ROC descreve o desempenho de um classificador sem levar em conta a distribuição de classe ou de custos de erros. Ela apresenta a TV_P (Sensibilidade) no eixo vertical contra a TV_N (Especificidade) no eixo horizontal.

A partir de um gráfico ROC, é possível calcular uma medida da qualidade global do modelo: a área abaixo da curva ROC (AUC^2). A AUC (*Area Under ROC Curve)* é a fração da área total que está abaixo da curva ROC. Essa média é

equivalente a várias outras medidas estatísticas para a avaliação e ranqueamento de modelos de classificação [Hand, 1997] conforme apresentado na Figura 3.11.

Figura 3.11 - Curva ROC (S_E vs. 1-E_S).



Fonte: Alberto e Almeida (2012).

3.2.5 Interpretação e Avaliação dos dados

Na fase de interpretação e avaliação, segundo afirma Nogueira (2015) e Prass (2009), um ou mais especialistas na área interpretam e avaliam o conhecimento adquirido na fase de mineração dos dados. Em caso onde o resultado não seja satisfatório, o que não é raro, o processo pode retornar a qualquer um dos estágios anteriores ou até mesmo ser recomeçado, conforme pode ser observado na Figura 3.1 através das linhas pontilhadas.

4 METODOLOGIA APLICADA NA ANÁLISE DA DISTORÇÃO HARMÔNICA COM O PROCESSO *KDD*

4.1 Considerações Iniciais

As melhorias e modernizações dos processos industriais com a utilização de produtos e equipamentos microprocessados e microcontrolados com cargas predominantemente não lineares têm proporcionado à informatização das áreas de escritório e automatização no chão de fábrica, no entanto com tudo isso surgem novos desafios com relação à qualidade de energia elétrica através dos impactos oriundos desses equipamentos e isso é um consenso entre todos os trabalhos que possuem a QEE como propósito. Para isso é necessário analisar, quantificar e qualificar o impacto na qualidade de energia elétrica e a ANEEL com o objetivo de estabelece base para o monitoramento e controle faz recomendações para os limites mínimos e máximos dos indicadores de qualidade através do PRODIST.

Sousa (2017) afirma que os problemas associados aos harmônicos do sistema elétrico de potência estão relacionados diretamente ao uso de cargas não lineares oriundas das novas tecnologias e que se estima que mais de 50% das cargas americanas e europeias conterão eletrônica de potência no futuro.

As fontes de energia não lineares tornaram-se muito populares hoje em dia, sendo amplamente presentes em quase todos os dispositivos eletrônicos conectados à rede de alimentação, afirmam Ciufu *et al.* (2017), e onde a natureza distorcida destes tipos de cargas tem influenciado fortemente a qualidade da energia elétrica gerando altos níveis de harmônicos nas formas de onda de corrente e tensão dos sistemas de energia.

Para Tarasiuk (2011) a medição de conteúdo harmônico tem sido usada para caracterizar o comportamento de cargas não lineares, para localizar fontes harmônicas e quantificar a distorção harmônica em sistemas de energia. Considerando os novos conceitos de distribuição de energia elétrica, segundo

Vlahinic et. al.(2009) há uma preocupação com outros indicadores, principalmente com a distorção harmônica total (*THD*), sendo esse último o índice harmônico mais comum usado para avaliar e medir as condições de variações da Qualidade da Energia em condições não senoidais.

Tendo como premissa uma indústria de produtos de informática onde aproximadamente 1500 produtos estão ligados simultaneamente na rede interna, a presente pesquisa tomou como base essas cargas não lineares para avaliar o impacto das mesmas através dos seus componentes harmônicos no *THD* de entrada.

O estudo teve apoio de ferramentas computacionais para analisar os dados que foram monitorados e coletados em cada um dos processos, estes por sua vez, foram submetidos ao processo de *data mining* e duas técnicas de IC que deu apoio ao processo de tomada de decisão.

4.2 A Metodologia Aplicada

Na referente tese foram aplicados os conceitos anteriormente tratados e seguiram os seguintes passos:

1. Escolher uma indústria para aplicação;
2. Definir o ambiente da pesquisa;
3. Definir os pontos de coleta;
4. Selecionar os analisadores de QEE;
5. Definir o período de coleta;
6. Coletar os dados;
7. Executar o Processo KDD;
    a. Selecionar os dados;
    b. Realizar a limpeza dos dados;
    c. Integrar os dados;
    d. Transformar os dados;
    e. Reduzir os dados;
    f. Minerar os dados;
        i. Aplicar a técnica Árvore de Decisão;

ii. Aplicar a técnica Naïve Bayes.

g. Interpretar e avaliar os dados;

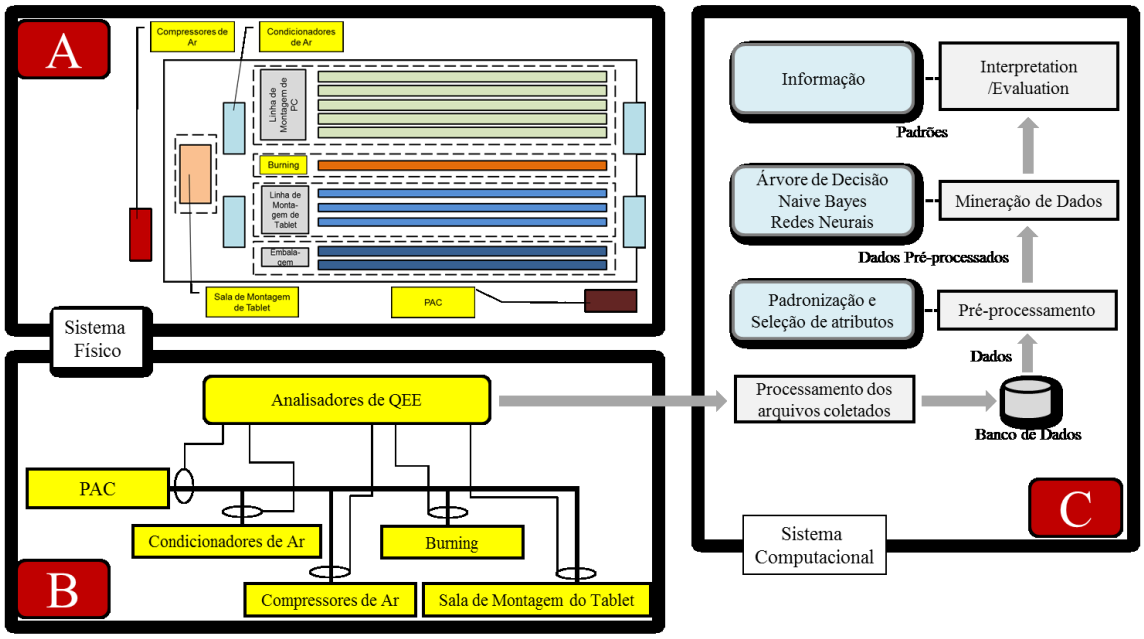
8. Analisar e discutir os resultados;
   a. Refazer a etapa 7 nos 3 turnos (T0, T1 e T2);
   b. Refazer a etapa 7 com diferentes percentuais de treinamento e teste (30/70, 50/50 e 70/30);
   c. Refazer a etapa 7 com diferentes percentuais de balanceamento de classes (100%, 200%, 300%, 1000% e 2000%).
9. Concluir o processo.

4.3 A Indústria Analisada

Para a aplicação da pesquisa proposta e com objetivo de instalação dos equipamentos, selecionou uma fabricante de produtos de tecnologia como *desktops*, *notebooks*, *netbooks*, celulares e TVs dentre outros. A empresa está entre os líderes em vendas do segmento no varejo e conta com parceiros renomados como INTEL, Microsoft e Qualcomm. Nos últimos 5 anos, foram aproximadamente 2,3 milhões de computadores fabricados e 7,4 milhões de televisores produzidos. Com tradição de quase 50 anos, é um dos maiores conglomerados do Polo Industrial em Manaus.

Para o desenvolvimento deste trabalho, foram estabelecidos dois sistemas: sistemas físico e computacional, conforme mostrado na Figura 4.1. No sistema físico visualizado na Figura 4.1 (A), o layout da indústria analisada é mostrado e a Figura 4.1 (B) mostra a instalação proposta dos cinco analisadores de QEE. Na Figura 4.1 (C), o sistema computacional que cobre todas as etapas do processo KDD é mostrado.

Figura 4.1 – Sistema físico que contém o layout (A), os pontos de instalação dos analisadores de QEE (B) e o processo completo do *KDD* no sistema computacional (C).



Fonte: Oliveira *et al.* (2017).

Na parte produtiva do sistema físico, apresentado na Figura 4.1 (A), é composta por:

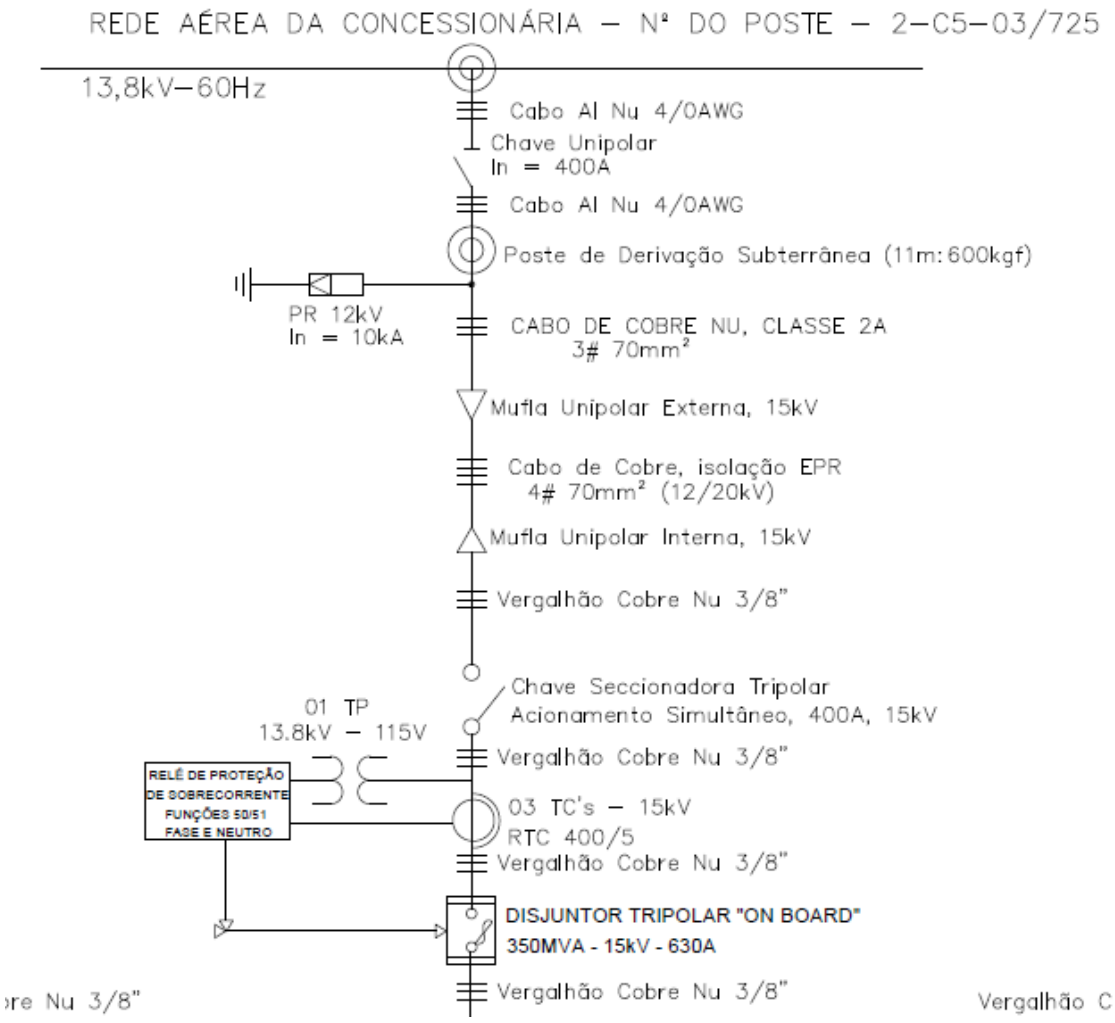
- **Linhas de Montagem de PC** compostas por cinco linhas de produção onde as duas primeiras montam *desktops* com potência total de 15 kW cada e as três últimas montam *notebooks e netbooks* com potência 3 kW cada. **Totalizando 39 kW**.
- **Burning** é uma linha destinada à realização dos testes de estresses dos produtos *desktops, notebooks, netbooks* e *tablets*. Onde os produtos são ligados nessa linha e ficam em média duas horas ligados executando software de estresses para a detecção de possíveis problemas de funcionamento. Ficando em média 480 *desktops* e 720 *notebooks e netbooks*, **totalizando em média 167 kW**;
- **Linhas de Montagem de Tablets** compostas por três linhas de montagem de tablets com **potência total de 2,5 kW** ;

- Linhas de Embalagem são duas linhas onde acontece o processo de embalagem de todos os produtos que passaram de forma satisfatória nos testes e inspeções;
- **Sala de Montagem de *Tablets*** são duas linhas em células destinadas a montagem de *tablets* corporativos. A sala de montagem possui infraestrutura com 10 impressoras laser de 445 W cada, 4 balanças eletrônicas de 6 W cada, 17 computadores de 300 W cada, 25 lâmpadas tubulares de LED de 19 W cada e 140 tablets em produção de 8 W cada. **Totalizando 11 kW** na sala de produção de *Tablet*;
- **Condicionadores de Ar** são quatro equipamentos de refrigeração posicionados entre o início e final das linhas de montagem. Cada central de ar com potência de 52 kW, **totalizando 208 kW** de potência para refrigeração da unidade produtiva;

A parte externa do sistema físico, apresentado na Figura 4.1 (A), é composta por:

- **Sala dos Compressores de Ar** onde existe um compressor de ar com potência média de 55 kW e duas bombas de 18 kW, **totalizando 91 kW**. As bombas e compressores funcionam durante o expediente da produção com a função de produzir ar comprimido para as linhas e equipamentos da unidade produtiva e bombear água do poço artesiano para a torre de água;
- **Subestação abaixadora** que recebe a energia da concessionária em 13.800 V conforme Figura 4.2. A mesma possui três transformadores abaixadores de 13,8 kV para 220V/127V, sendo um transformador de 300 kVA para a alimentação do prédio de escritório, um transformador de 1000 kVA para a alimentação da unidade produtiva e outro de 1000 kVA como reserva. Possui, ainda, um grupo gerador de 105 kVA para o prédio e outro de 450 kVA para a unidade produtiva.

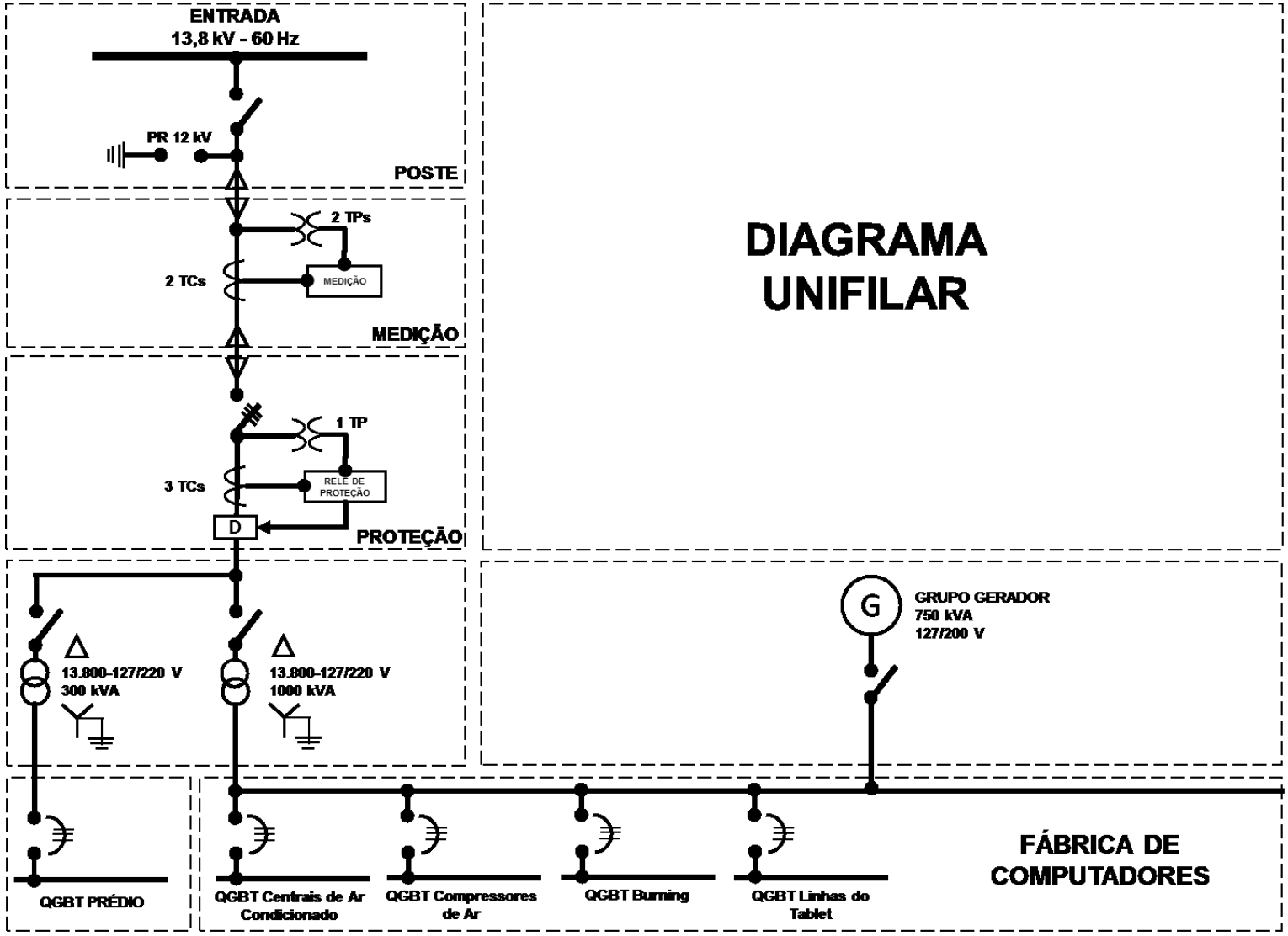
Figura 4.2 - Diagrama unifilar da entrada da empresa selecionada



Fonte: Empresa selecionada (2017).

As cargas estão distribuídas conforme diagrama unifilar apresentado na Figura 4.3.

Figura 4.3 - Diagrama unifilar da empresa selecionada



Fonte: Empresa selecionada (2017).

4.3.1 Definição do ambiente da pesquisa

A ausência de dados reais relacionados a distorções harmônicas em fábricas de computadores, além da necessidade de atendimento aos requisitos de controle de processo sugerido pela *INTEL* e das recomendações à curva *CBEMA-ITIC* e com o objetivo de contribuir para a ciência, uma indústria de computadores localizada no Polo Industrial de Manaus foi selecionada contendo linhas de produção de *notebooks*, *desktops*, monitores de computador e *tablets*, como mostrado na Figura 4.1 (A), onde a grande maioria são cargas não lineares distribuídas como mostrado na Tabela 4.1.

Tabela 4.1. Pontos de coleta.

Ponto de Coleta	Característica do Ponto	Período de Funcionamento
PAC-Ponto de Acoplamento Comum	Possui m transformador de potência com 1000 kVA 220V/127 V e outro de 300 kVA.	24h por dia.
Condicionadores de Ar	Possui quatro centrais de ar condicionado, cada um com dois compressores de 15 toneladas de refrigeração (TR) aproximadamente 52 kW.	Das 5:30 da manhã às 1:30 da manhã de segunda a sexta-feira. Ficam desligadas nos finais de semana.
Compressores de Ar	Existe um compressor de ar de 75 HP, uma bomba de repressão de 25 HP e uma bomba de torre de 25 HP.	Das 5:30 da manhã às 1:30 da manhã de segunda a sexta-feira. Ficam desligadas nos finais de semana.
Burning	Existem aproximadamente 1200 computadores conectados simultaneamente no teste de queima, onde cada computador permanece por 1 hora e é substituído por outro	Computadores permanecem conectados no teste à noite e nos fins de semana.
Sala de montagem de tablets	Área produtiva de *tablets*, contendo entre as principais cargas 10 impressoras, 4 balanças eletrônicas, 17 computadores, 25 lâmpadas *LED* e aproximadamente 140 tablets no teste.	Os *Tablets* permanecem conectados no teste à noite e nos fins de semana.

Fonte: Elaborada pelo autor (2017).

É importante observar que algumas cargas como os condicionadores de ar e os compressores de ar ficam ligadas somente no horário de produção e que as demais cargas ficam ligadas 24h diariamente. A distribuição dos horários nos turnos está mostrada na Tabela 4.2.

Tabela 4.2. Turnos de trabalho.

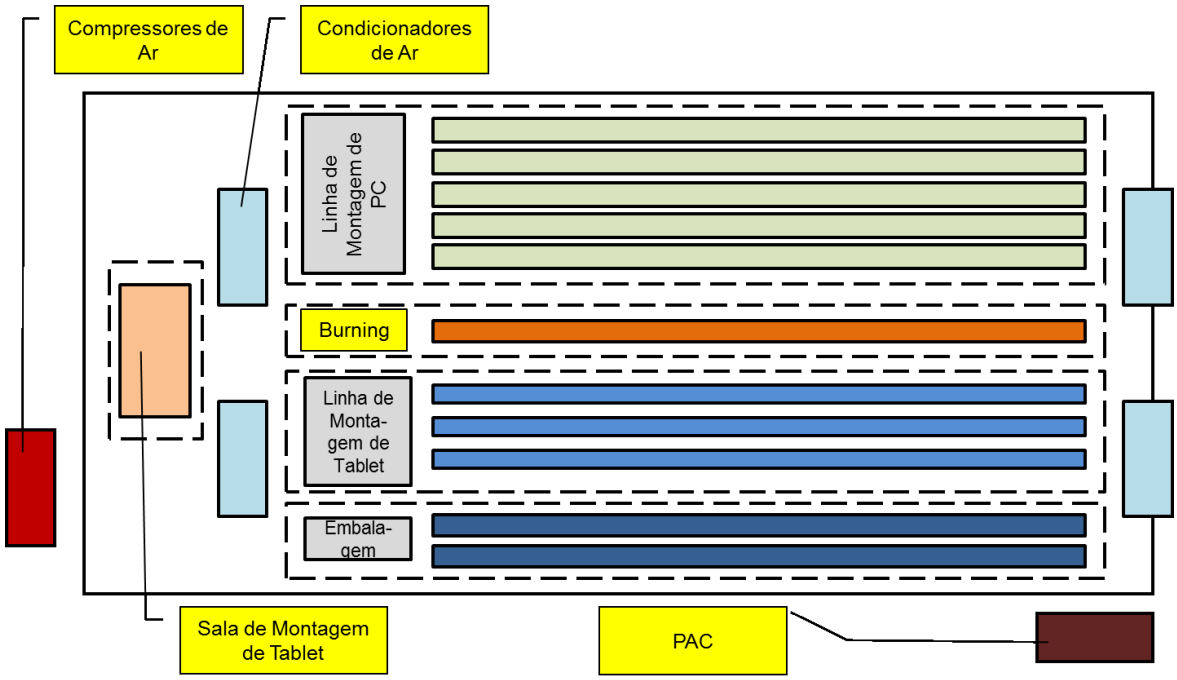
Turno	Início	Fim
T1 (Primeiro)	6h	15h48
T2 (Segundo)	15h48	1h10
T0 (Vazio)	1h10	6h

Fonte: Elaborada pelo autor (2017).

4.3.2 Definição dos pontos de coleta

A presente tese apresenta uma análise completa usando cinco analisadores de QEE instalados em pontos estratégicos como observado no *layout* da indústria apresentado na Figura 4.1 (A) e ampliada na Figura 4.4 conforme detalhado na Seção 4.3.1, para que a coleta dos dados ocorra simultaneamente.

Figura 4.4 – *Layout* da indústria de computadores com os pontos de coleta (amarelo).



Fonte: Oliveira *et al.* (2017).

Os pontos de coleta foram instalados nos Quadros Gerais relacionados as cargas e conforme são apresentados na Tabela 4.3.

Tabela 4.3. Quadros para os pontos de coleta.

Localização	Descrição
QGBT	Quadro Geral de Baixa Tensão (220V/127V) localizado na sub-estação.
Centrais de AR	Quadro que alimenta 4 Centrais de Ar Condicionado
Compressores	Quadro que alimenta a sala de compressores e bomba d'água
Burning	Quadro que alimenta a linha de teste de *notebooks*
Tablet	Quadro que alimenta a sala de montagem de *tablets*

Fonte: Elaborada pelo autor (2017).

4.4 Seleção dos Analisadores de QEE

Para que a coleta tivesse qualidade nos dados foram escolhidos os equipamentos Analisadores de Qualidade de Energia PW3198 da fabricante HIOKI conforme Figura 4.5. Este é um analisador de qualidade de energia para monitoramento e gravação de anomalias de fornecimento de energia, permitindo que suas causas sejam rapidamente investigadas, e também para avaliar os problemas de fornecimento de energia, tais como quedas de tensão, *flicker*, harmônicos e outros problemas elétricos.

Figura 4.5 – Analisador de QEE PW3198 (HIOKI).

Fonte: HIOKI.

Como características principais o PW3198 possui:

- **Tipos de medições:**
  - Para os tipos de linhas de medição monofásica de 2 fios, monofásico 3 fios, trifásica 3 fios ou trifásica 4 fios mais um canal de entrada extra (deve ser sincronizado com referência canal durante a medição AC / DC);

- **Faixa de Tensão:**
  - Medição de tensão: 600,00 V *rms* e
  - Medida de transiente: 6,0000 kV pico;
- **Faixa de Corrente:**
  - De 500,00 mA a 5,0000 kA AC (dependente do sensor de corrente em uso);
- **Faixa de Potência:**
  - 300,00 W a 3,000 MW (determinado automaticamente baseado na faixa de tensão e corrente em uso);
- **Precisão Básica:**
  - Tensão: ± 0,1% da tensão nominal;
  - Corrente: ± 0,2% da leitura. ± 0,1% fs + precisão do sensor de corrente e
  - Potência Ativa: ± 0,2% da leitura. ± 0,1% fs + precisão do sensor de corrente;
- **Possui os seguintes itens de medição:**
  - Transitória sobre a tensão: 2 MHz de amostragem;
  - Ciclo Frequência: Calculado como um ciclo, de 40 a 70 Hz.
  - Tensão (1/2) *RMS*: um cálculo ciclo atualizado a cada ciclo de metade Corrente (1/2) *RMS*: cálculo de semi-ciclo;
  - Elevação de tensão, afundamento de tensão, interrupção de tensão;
  - Corrente de pico;
  - Comparação de forma de onda de tensão;
  - Valor cintilação instantânea: Conforme *IEC* 61000-4-15;
  - Frequência: Calculado como 10 ou 12 ciclos, de 40 a 70 Hz;
  - Frequência de 10 segundos: Calculado como o tempo de ciclo durante todo o período especificado 10 s, 40 a 70 Hz;
  - Tensão de pico da forma de onda, corrente de pico da forma de onda;
  - Tensão, corrente, potência ativa, potência aparente, potência reativa, energia ativa, energia reativa, fator de potência, fator de potência de deslocamento, o fator de desbalanceamento de tensão, Fator de desbalanceamento de corrente (fase negativa, fase zero);
  - Componentes harmônicos de alta ordem (tensão / corrente): 2 kHz a 80 kHz;
  - Harmônicos, ângulo de fase dos harmônicos (tensão / corrente), potência harmônica: 0 a 50 ª ordens ;
  - Ângulo de fase harmônica de tensão-corrente: 1th 50 ª ordens;
  - Fator de distorção harmônica total (tensão / corrente);
  - Inter Harmônicos (tensão / corrente): 0,5 Hz a 49,5 Hz ;
  - Fator K (fator de multiplicação);
  - *IEC Flicker*, Δ V10 *Flicker*.
- **Registros:**
  - Registros dos dados até 55 semanas (com repetição do conjunto de gravação para [1] Semana, 55 iterações);
  - 35 dias (com repetição do conjunto de gravação para [*OFF*]);
- **Interfaces:**

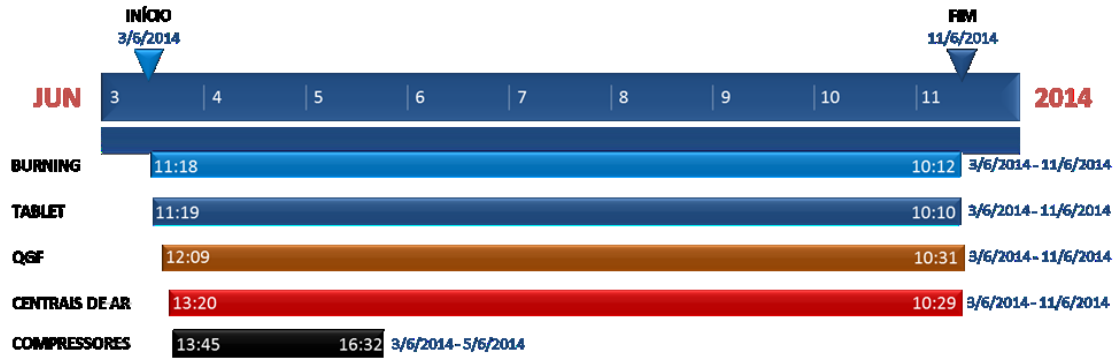
- Cartão *SD* / *SDHC*, RS-232C, *LAN (function servidor HTTP)* e USB2.0.

- **Tela:**
  - Tela de LCD colorida de 6,5 polegadas com resolução 640x480;
- **Fonte de Alimentação:**
  - Adaptador AC Z1002 (12 VDC, Tensão de entrada entre 100 VAC a 240 VAC, 60/50 Hz);
  - Bateria Z1003 (Ni-MH 7,2 VDC 4500 mAh).
- **Dimensões e Peso:**
  - Largura 300 mm x Altura 211 mm x Profundidade 68 mm;
  - Peso: 2,6 kg (incluindo a bateria).
- **Acessórios:**
  - 1 Manual de instruções, 1 conjunto de cabo de força L1000, conjunto de cabos (vermelho, amarelo, azul e cinza com 1 cabo de cada, preto com 4 cabos e 8 garras jacaré), 20 tubos espirais, 1 etiqueta de entrada dos cabos, 1 adaptador AC Z1102, 1 alça, 1 cabo USB de 1m de comprimento, 1 embalagem de bateria, 1 cartão de memória SD 2GB Z4001.

4.5 Período da Coleta dos Dados

O período da coleta dos dados aconteceu entre os dias 03/06/14 a 11/06/2014 conforme Figura 4.6.

Figura 4.6 – Período da coleta de dados.



Fonte: Adaptado de Oliveira *et al.* (2017).

Como podem ser observados, os dados coletados mostram a contribuição de ocorrência simultânea do processo de medição no período entre 13h45 do dia 3 de junho a 16h32 do dia 5 de junho para todos os pontos coletados. Caso o ponto compressor seja excluído, a análise vai de 13h20 do dia 3 de junho até às 10h10 do dia 11 de junho. Todos os analisadores de QEE foram removidos em 11 de junho de 2014.

4.6 Dados Coletados

Para a análise de dados como descrito anteriormente, o processo completo de *KDD* foi usado como mostrado na Figura 3.1 que usa o conceito de Fayyad (1996), pelo qual o *KDD* pode ser dividido em cinco etapas: seleção, pré-processamento, transformação, mineração de dados e avaliação. Os dados coletados pelos analisadores PW3198 foram armazenados em um cartão digital seguro (*SD*) e transferidos para um computador, antes do processo *KDD* conforme Figura 4.7.

Figura 4.7 – Transferência dos dados por cartão *SD* do analisador de QEE para o computador.

Fonte: Hioki.

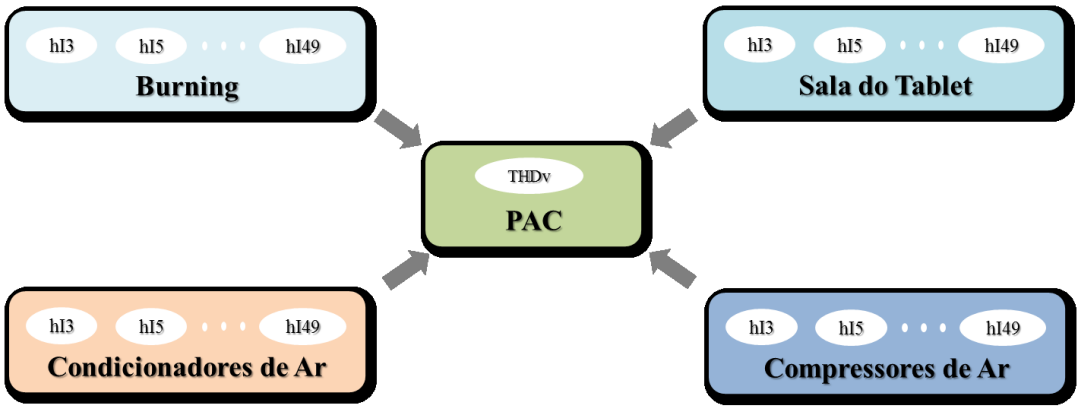
4.7 Execução do Processo KDD

O presente trabalho aplicou todas as fases do processo *KDD* conforme apresentado na seção 3.2 e está descritos nos tópicos a seguir.

4.7.1 Processo KDD: Seleção dos Dados

Essa fase é a aplicação da seção 3.2.1, sendo que o presente artigo avaliou o impacto da contribuição de cada harmônico de corrente de cada carga instalada na indústria de computadores na *THD* de tensão do PAC para cada fase, conforme é mostrado na Figura 4.8.

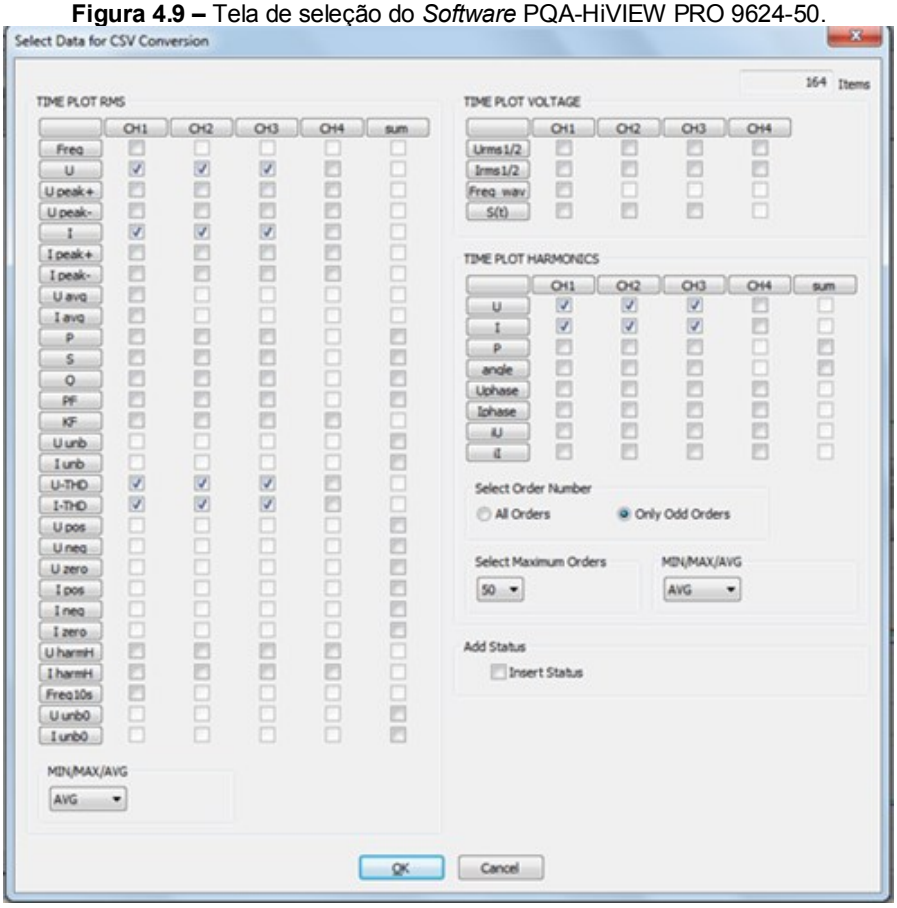
Figura 4.8 – Impacto das correntes harmônicas atuais de cada carga no *THDv* no PAC.



Fonte: Oliveira *et al.* (2017).

O *software* PQA-HiVIEW PRO 9624-50 foi utilizado para a fase seleção dos dados do processo *KDD* e os itens selecionados para cada fase (CH1, CH2 e CH3) foram os seguintes: Tensão *RMS* (U), Corrente *RMS* (I), *THDv* (*U-THD*), *THDi* (*I-THD)*, Harmônicos Impares de Tensão e Corrente até a 49ª (*Only Odd Orders e Select Maximum Orders*=50). A Figura 4.9 evidencia os itens selecionados e a conversão para o formato *CSV (Select Data for CSV Conversion).*

Figura 4.9 – Tela de seleção do *Software* PQA-HiVIEW PRO 9624-50.



Fonte – Tela capturada pelo autor (2017).

Além dos itens selecionados, os arquivos também possuem a data e hora de cada leitura realizada que foi utilizada nas etapas seguintes.

4.7.2 Processo *KDD*: Pré-processamento: limpeza e integração dos dados coletados

Essa fase é a aplicação da seção 3.2.2 e para realizar estas etapas, os dados foram analisados para cada ponto coletado e os ajustes necessários foram feitos de acordo com a estrutura mostrada na Figura 4.10. Os dados iniciais entre os blocos ID e RMS I T da são apresentados na Tabela 4.4 em virtude do tamanho do arquivo.

Figura 4.10 – Estrutura dos dados coletados.

Fonte: Oliveira *et al.* (2017).

A estrutura dos dados para cada ponto de coleta está representada conforme abaixo:

- *ID*: Identificador do registro;
- *Date:* Data da coleta do dado;
- *Time:* Hora da coleta do dado;
- *RMS U RST:* Dados da tensão *RMS* para as fases R,S e T;
- *RMS I RST:* Dados da corrente *RMS* para as fases R,S e T;
- *THDU RST*: Distorção Harmônica Total de tensão para as fases R, S e T;
- H I 1-49 RST: Harmônico da corrente da 1ª a 49ª ordem para as fases R, S e T;
- H U 1-49 RST: Harmônico da tensão da 1ª a 49ª ordem para as fases R, S e T.

Tabela 4.4. Dados coletados de um analisador de QEE em formato CSV.

id	Date	time	RMS U R	RMS U S	RMS U T	RMS I R	RMS I S	RMS I T
1	3-jun-14	1:20:00 PM	128.03E+0	127.20E+0	126.72E+0	0.267E+3	0.231E+3	0.219E+3
2	3-jun-14	1:21:00 PM	128.37E+0	127.53E+0	127.03E+0	0.175E+3	0.150E+3	0.141E+3

id	Date	time	RMS U R	RMS U S	RMS U T	RMS I R	RMS I S	RMS I T
3	3-jun-14	1:22:00 PM	128.55E+0	127.73E+0	127.19E+0	0.174E+3	0.150E+3	0.140E+3
4	3-jun-14	1:23:00 PM	128.43E+0	127.57E+0	127.03E+0	0.175E+3	0.150E+3	0.140E+3
5	3-jun-14	1:24:00 PM	128.49E+0	127.64E+0	127.17E+0	0.174E+3	0.151E+3	0.141E+3
6	3-jun-14	1:25:00 PM	128.75E+0	127.93E+0	127.54E+0	0.175E+3	0.151E+3	0.142E+3
7	3-jun-14	1:26:00 PM	128.36E+0	127.50E+0	127.13E+0	0.175E+3	0.151E+3	0.142E+3
8	3-jun-14	1:27:00 PM	128.45E+0	127.55E+0	27.16E+0	0.175E+3	0.151E+3	0.142E+3
9	3-jun-14	1:28:00 PM	128.58E+0	127.67E+0	127.27E+0	0.176E+3	0.151E+3	0.143E+3
10	3-jun-14	1:29:00 PM	128.53E+0	127.61E+0	127.17E+0	0.175E+3	0.150E+3	0.142E+3
11	3-jun-14	1:30:00 PM	128.73E+0	127.85E+0	127.41E+0	0.176E+3	0.151E+3	0.143E+3
12	3-jun-14	1:31:00 PM	128.66E+0	127.79E+0	127.37E+0	0.175E+3	0.152E+3	0.142E+3
13	3-jun-14	1:32:00 PM	128.75E+0	27.87E+0	127.42E+0	0.176E+3	0.152E+3	0.142E+3

Fonte: Software PQA-HiVIEW PRO 9624-50 (HIOKI).

4.7.2.1 Limpeza dos dados

A Tabela 4.5 resume a análise de cada ponto de coleta e as ações a serem tomadas para limpar os dados. No total, haviam 17.070.240 itens de dados coletados.

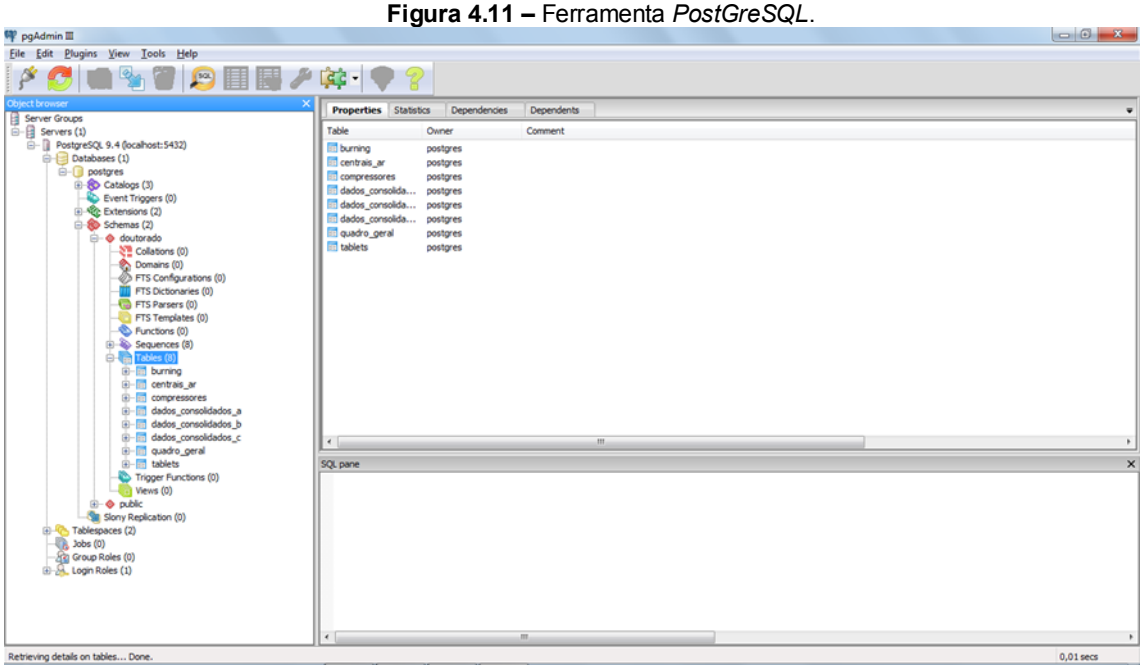
Tabela 4.5. Análise dos dados coletados de um analisador de QEE em formato *CSV.*

Ponto	Taxa de Amostragem	Linhas	Colunas	Total de Dados	Análise
Burning	1 minuto	9,871.00	165.00	1,628,715.00	Os valores dos harmônicos de corrente devem ser convertidos em níveis, uma vez que o equipamento foi configurado para porcentagens. Toda tensão, corrente, distorção harmônica total (*THD*) e valores harmônicos devem ser convertidos de notação científica ao normal.
Sala do *Tablet*	1 minuto	9,872.00	165.00	1,628,880.00	Os valores das correntes devem ser divididos por 1000 para ajustar a ponta utilizada. Todos os valores de tensão, corrente, *THD* e harmônicos devem ser convertidos de notação científica para normal.
PAC	1 minuto	11,423.00	165.00	1,884,795.00	Os valores das correntes devem ser divididos por 1000 para ajustar a sonda utilizada. Todos os valores de tensão, corrente, *THD* e harmônicos devem ser convertidos de notação científica para normal.
Condicionadores de Ar	1 minuto	11,350.00	165.00	1,872,750.00	Todos os valores de tensão, corrente, *THD* e harmônicos devem ser convertidos de notação científica para normal.
Compressores de Ar	3 segundos	60,940.00	165.00	10,055,100.00	Todos os valores de tensão, corrente, *THD* e harmônicos devem ser convertidos de notação científica para normal.

Fonte: Elaborada pelo autor (2017).

Para facilitar o processo de limpeza os dados foram convertidos para o banco de dados da ferramenta *Postgresql Tool* conforme Figura 4.11 abaixo e aplicadas as sugestões anteriormente mencionadas através de vários scripts da linguagem *SQL*.

Figura 4.11 – Ferramenta *PostGreSQL*.



Fonte: Tela capturada pelo autor (2017).

Após a aplicação dos scripts, os dados limpos entre os blocos ID e RMS I estão mostrados na Tabela 4.6.

Tabela 4.6. Dados limpos após a aplicação dos scripts.

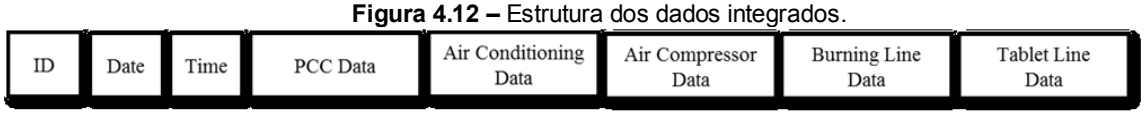
id	date	time	RMS U R	RMS U S	RMS U T	RMS I R	RMS I S	RMS I T
1	3-jun-14	1:20:00 PM	128.03	127.20	126.72	267.00	231.00	219.00
2	3-jun-14	1:21:00 PM	128.37	127.53	127.03	175.00	150.00	141.00
3	3-jun-14	1:22:00 PM	128.55	127.73	127.19	174.00	150.00	140.00
4	3-jun-14	1:23:00 PM	128.43	127.57	127.03	175.00	150.00	140.00
5	3-jun-14	1:24:00 PM	128.49	127.64	127.17	174.00	151.00	141.00
6	3-jun-14	1:25:00 PM	128.75	127.93	127.54	175.00	151.00	142.00
7	3-jun-14	1:26:00 PM	128.36	127.50	127.13	175.00	151.00	142.00
8	3-jun-14	1:27:00 PM	128.45	127.55	127.16	175.00	151.00	142.00
9	3-jun-14	1:28:00 PM	128.58	127.67	127.27	176.00	151.00	143.00
10	3-jun-14	1:29:00 PM	128.53	127.61	127.17	175.00	150.00	142.00
11	3-jun-14	1:29:00 PM	128.73	127.85	127.41	176.00	151.00	143.00

Fonte: Elaborada pelo autor (2017).

4.7.2.2 Integração dos dados

Para permitir que os dados fossem usados através da mineração de dados, era necessário integrar os dados em uma única tabela contendo os dados de entrada (*pcc_THDv*) e os dados de carga, de acordo com a estrutura como mostrado na Figura 4.12. A Tabela 4.7 mostra os dados integrados entre os blocos *ID* e *Air Condiotining Data* até a 9ª componente harmônica de corrente. Essa integração foi sincronizada de tal forma que a data e a hora de cada ponto coletado eram as mesmas, além disso para a realização das análise foram separadas por fase.

Figura 4.12 – Estrutura dos dados integrados.



Fonte: Oliveira *et al.* (2017).

Tabela 4.7. Dados Integrados.

id	date	time	PAC Data THDv	Air Conditioning Data 3rd harmonic current air_hi_a_3	Air Conditioning Data 5th harmonic current air_hi_a_5	Air Conditioning Data 7th harmonic current air_hi_a_7	Air Conditioning Data 9th harmonic current air_hi_a_9
43	3-jun-14	2:27:00 PM	2.08	5.20	2.30	1.80	0.90
44	3-jun-14	2:28:00 PM	2.15	5.20	2.30	1.90	0.90
45	3-jun-14	2:29:00 PM	2.18	5.20	2.40	1.90	0.90
46	3-jun-14	2:30:00 PM	2.15	5.10	2.40	1.80	0.80
47	3-jun-14	2:31:00 PM	2.16	5.20	2.40	1.90	0.70
48	3-jun-14	2:32:00 PM	2.13	5.20	2.40	1.80	0.70
49	3-jun-14	2:33:00 PM	2.10	5.20	2.40	1.80	0.70
50	3-jun-14	2:34:00 PM	2.16	5.20	2.40	1.90	0.80
51	3-jun-14	2:35:00 PM	2.18	5.20	2.50	1.90	0.80

Fonte: Oliveira *et al.* (2017).

4.7.3 Processo *KDD*: Transformação dos Dados

Nesta fase, os dados já foram selecionados, pré-processados, limpos e integrados; Tudo isso precede a fase de mineração de dados. Após a integração por fase dos dados em uma única tabela, havia uma necessidade de transformar alguns dados de valores numéricos para valores categóricos.

Para a transformação de dados, nos referimos a Leite (2013), e os padrões *IEEE* 519-1992 que estabelecem limites para distorção harmônica de corrente e tensão em redes de distribuição e transmissão para sistemas de baixa tensão. Os dados encontrados em Schneider (2005) estabelecem intervalos de valores normais (NORM), risco (RISCO) e críticos (CRIT) para o *THDv*, de acordo com a Tabela 4.8.

Tabela 4.8. Faixa de valores para o *THDv*.

STATUS	*THDv*
NORM	< 5 %
RISK	> 5 % and < 8 %
CRIT	> 8 %

Fonte: Leite (2013), IEEE 519-1992 e Schneider (2005).

O processo *KDD* apresentado na seção 3.2, mostra a necessidade de retornar às etapas anteriores. Considerando a interpretação dos resultados preliminares, foi necessário dividir os dados em turnos para aprofundar e validar as análises, conforme evidenciado na Tabela 4.9.

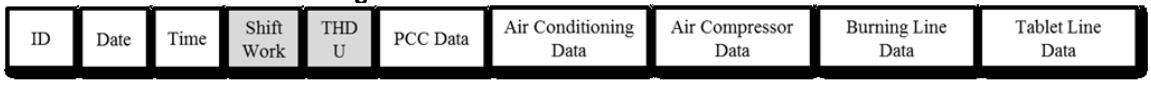
Tabela 4.9. Horário de cada turno.

STATUS	Horário
T1	hora >= '06:00' & hora <= '15:48'
T2	hora > '15:48' & hora <= '23:59' OU hora >= '00:00' & hora <= '01:10'
T0	hora > '01:10' & hora < '06:00'

Fonte: Elaborada pelo autor (2017).

Os dados transformados são mostrados na Figura 4.13, onde as etapas de mudança e *THDv f*oram adicionadas como mostrado nas Tabelas 4.8 e 4.9.

Figura 4.13 – Estrutura dos dados transformados.



Fonte: Oliveira *et al.* (2017).

4.7.3.1 Processo KDD: Redução dos Dados

Considerando que as componentes harmônicas de corrente mais significativas das cargas estão entre as componentes harmônicas 3º e 15º (Harm I),

outras componentes harmônicas de corrente e os dados que não foram necessários pela metodologia analisada foram excluídos, como mostrado na Tabela 4.10.

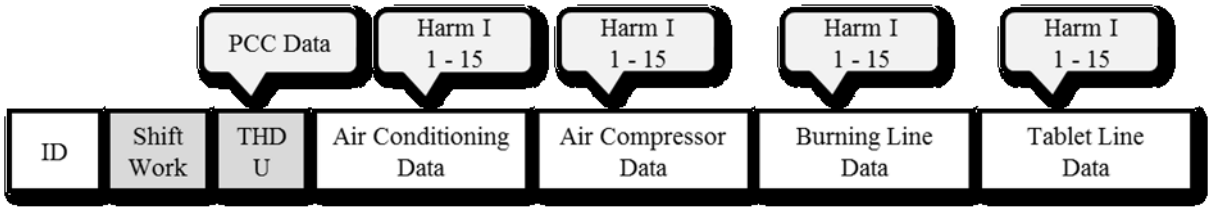
Tabela 4.10. Dados descartados.

PAC	Distorção Harmônica Total de corrente (*THDi*) Tensão *RMS* Corrente *RMS* Tensão fundamental Corrente fundamental Harmônicos de tensão de ordem 17 a 49 Harmônicos de corrente de ordem 17 a 49
Cargas: Burning Sala do Tablet Condicionadores de Ar Compressores de Ar	Tensão *RMS* Corrente *RMS* Distorção Harmônica Total de tensão (*THDv*) Distorção Harmônica Total de corrente (*THDi*) Tensão fundamental Corrente fundamental Harmônicos de tensão de ordem 17 a 49

Fonte: Elaborada pelo autor (2017).

Após a redução de dados, a estrutura de dados foi organizada como mostrado na Figura 4.14.

Figura 4.14 – Estrutura dos dados reduzidos.

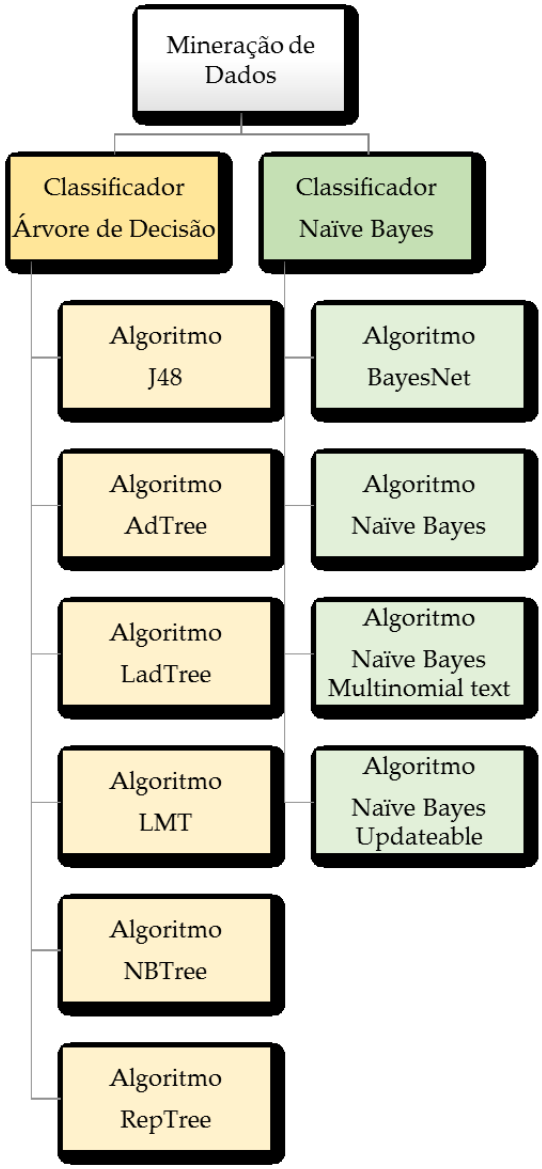


Fonte: Oliveira *et al.* (2017).

4.7.4 Processo *KDD*: Mineração dos Dados

A presente tese utilizou os conceitos da seção 3.2.4 e para essa finalidade foram utilizados os *softwares WEKA* versão 3.7.13 (2015) e *Rapid Miner Studio Basic* versão 6.5.02 (2015) e os classificadores Árvore de Decisão e *Naïve Bayes* com diferentes algoritmos conforme evidenciado na Figura 4.15.

Figura 4.15 – Algoritmos de Mineração de Dados.



Fonte: Elaborada pelo autor (2017).

4.7.5 Processo *KDD*: Interpretação e avaliação dos Dados

Na presente tese, a etapa de interpretação e avaliação dos dados segue os conceitos da seção 3.2.5. Com objetivo de validação os dados foram avaliados com diferentes tipos de algoritmos e apresentado de forma comparativa, conforme apresentado na Figura 4.15, para os classificadores Árvore de Decisão e *Naïve* Bayes por meio de dois *softwares* de mineração de dados: *WEKA* e *Rapid Miner*.

Tendo-se como premissas que o processo *KDD* é dinâmico e que pode retornar as etapas anteriores para aprofundamento ou ajustes dos dados, foram

realizadas análises com diferentes cenários para garantir o balanceamento das classes conforme é apresentado na seção 3.2.2.1, nessa ação a técnica utilizada foi o *SMOTE* como técnica sintética de *sobre amostragem* com diferentes valores de ajustes. Como o mesmo propósito, também foi avaliado utilizando-se diferentes percentuais para a separação dos dados em conjunto de treinamento e de teste.

Com o propósito de identificação, validação e comparação dos resultados, foram realizadas análises com os dados separados em grupos de turnos, desta forma foi possível comparar o resultado com o cenário completo, cenário com os dados dos tuno T0, T1 e T2. Finalmente foram excluídas as medidas do ponto do compressor de ar para aumentar o número de dias avaliados. Os dados originais foram avaliados de acordo com a sanidade dos resultados encontrados nas técnicas de mineração de dados.

4.8 Considerações Finais

A indústria escolhida para realização do trabalho apresenta características muito interessantes, pois concentra uma quantidade grande de produtos de informática como *notebooks, desktops* e *tablets* ligados simultaneamente durante o período de 24 horas diárias em 7 dias da semana, ou seja, não desliga. Outra característica importante são as demais cargas analisadas que funcionam somente no período produtivo que vai de 05h30 até 01h30, ficando desligados durante o período de 01h30 até 05h30 e o dia inteiro no final de semana. Esses dados são importantes para direcionar a análise dos dados e entender o comportamento das cargas que mais impactam no *THD* da entrada.

5 RESULTADOS ENCONTRADOS

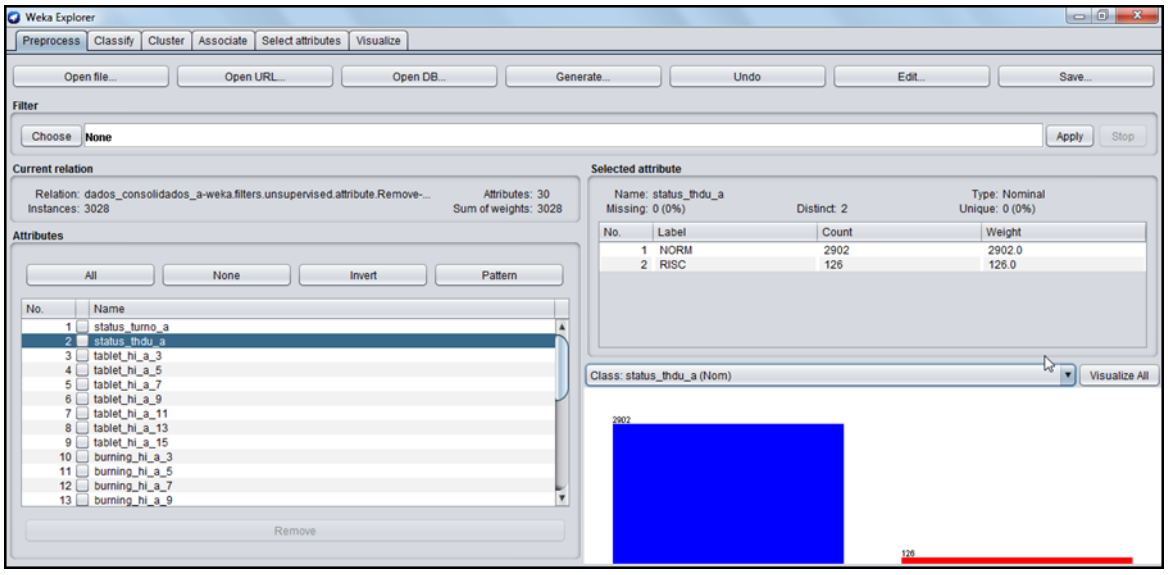
5.1 Considerações Iniciais

Na tese presente foram coletados 17.070.240 dados por meio dos cinco analisadores de QEE instalados estrategicamente em pontos de uma indústria de computadores conforme apresentado no capítulo 4.

Inicialmente os dados totais coletados foram submetidos ao processo de mineração de dados utilizando-se o *Software WEKA* com o classificador Árvore de Decisão e utilizando-se diferentes cenários sem aplicar filtro para balanceamento de classes e com a opção de teste *cross-validation* com *folds* ajustado para 10.

Os dados consolidados apresentam 3028 instâncias sendo 2902 classificadas como NORM e 126 classificadas como RISC conforme apresentados na Figura 5.1.

Figura 5.1 - Dados gerais consolidados visualizados no *software WEKA*.



Fonte: Tela do *WEKA* capturada pelo autor (2017).

5.2 Resultado com classificador Árvore de Decisão no *Software WEKA*

5.2.1 Análise da Árvore de Decisão com dados gerais, sem balanceamento de classes, com *cross-validation* em 10 *folds* e com os algoritmos J48, NBTREE, LADTREE, ADTREE, LMT e REPTREE.

Para a mineração de dados inicialmente utilizou-se o *software WEKA* com classificador Árvore de Decisão e utilizou-se diversos algoritmos e a opção de teste foi ajustada com 10 *folds* na técnica *cross-validation* para identificar o resultado mais relevante baseado nas métricas Instâncias Classificadas Corretamente, Instância Classificadas Incorretamente, Área ROC, Falsos Positivos e Falsos Negativos conforme Tabela 5.1

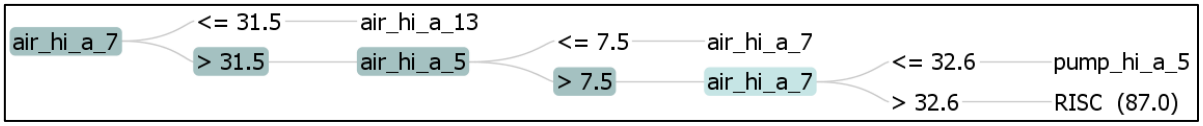
Tabela 5.1 - Resultado dos algoritmos do classificador Árvore de Decisão do software WEKA.

ALGORÍTMO	Instâncias Classificadas Corretamente	Instâncias Classificadas Incorretamente	ROC Área	Positivos Falsos	Negativos Falsos
J48	99,9009 %	0,0991 %	1,000	3	0
NBTREE	99,8349 %	0,1651 %	1,000	2	3
LADTREE	99,8018 %	0,1982 %	1,000	4	2
ADTREE	99,7358 %	0,2642 %	1,000	5	3
LMT	99,7358 %	0,2642 %	1,000	4	4
REPTREE	99,1744 %	0,8256 %	0,961	14	11

Fonte: Elaborada pelo autor (2017).

De acordo com a Tabela 5.1, o algoritmo J48 apresentou o resultado mais significativo devido aos itens **Instâncias Classificado Corretamente**, **Instâncias Classificadas Incorretamente**, **ROC Área**, **Positivos Falsos** e **Negativos Falsos.** O resultado indicou como a maior impactante a 7ª componente harmônica de corrente do ponto **Condicionadores de Ar** conforme evidenciado na Figura 5.2 para o conjunto de dados coletados.

Figura 5.2 - Resultado com algoritmo J48 no *THDv* com os dados gerais.



Fonte: Tela capturada pelo autor (2017).

A Figura 5.2 mostra que se a sétima harmônica das centrais de ar condicionado for maior que 31,5% e a quinta harmônica de corrente das centrais de ar for maior que 7,5% então o *THDv* entrará na condição de risco em 87 casos do total de 126 considerando o conjunto de dados coletados durante o período da campanha de medição.

Utilizou-se, então, o algoritmo J48 com para fazer as análises com o *THDv* e *THDi* conforme Tabela 5.2.

Tabela 5.2. Resultado com algoritmo J48 no *THDv* e *THDi*.

ALGORÍTMO	Instâncias Classificadas Corretamente	Instâncias Classificadas Incorretamente	ROC Área	Falsos Positivos	Falsos Negativos
J48 – *THDv*	99,9009 %	0,0991 %	1,000	3	0
J48 – *THDi*	99,9670 %	0,0330 %	0,999	0	1

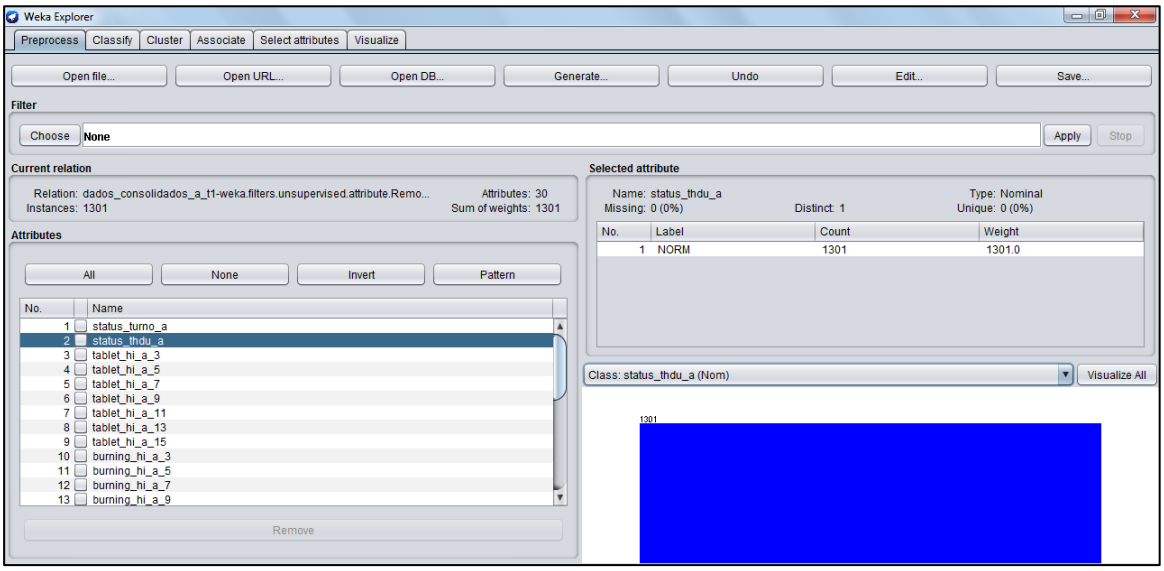
Fonte: Elaborada pelo autor (2017).

5.2.2 Análise da Árvore de Decisão com dados de cada turno, sem balanceamento de classes, com *cross-validation* em 10 *folds* e com o algoritmo J48.

O resultado realizado nas seções anteriores continha os dados coletados em todas as cargas e considerando todos os turnos no período de 03/06 a 05/06 e todos indicavam como a maior impactante a 7ª componente harmônica de corrente do ponto **Condicionadores de Ar** considerando o conjunto de dados coletados. Para confirmação do resultado dividiu-se a base de dados em 3 bases organizadas por turnos: T1, T2 e T0.

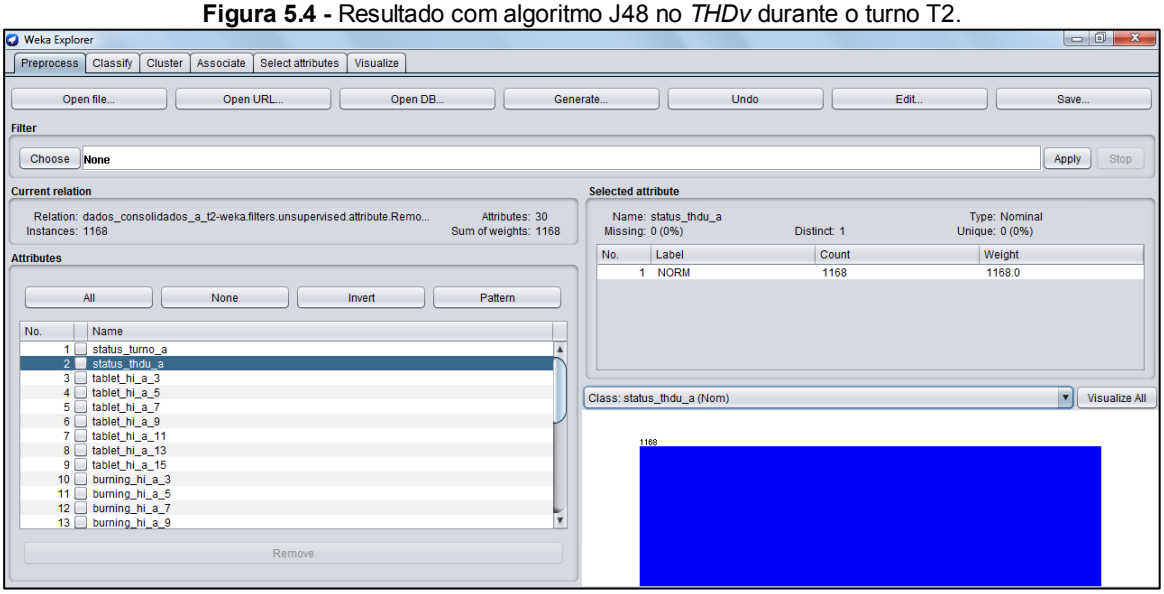
Para os dados coletados somente nos turnos T1 e T2 não foi possível a execução da Árvore de Decisão, pois não houve *THDv* além dos valores normais conforme evidenciado nas Figuras 5.3 e 5.4.

Figura 5.3 - Resultado com algoritmo J48 no *THDv* durante o turno T1.



Fonte: Tela do *WEKA* capturada pelo autor (2017).

Figura 5.4 - Resultado com algoritmo J48 no *THDv* durante o turno T2.



Fonte: Tela do *WEKA* capturada pelo autor (2017).

Para os dados coletados somente no Turno 0, o resultado é apresentado na Tabela 5.3 para o *THDv*.

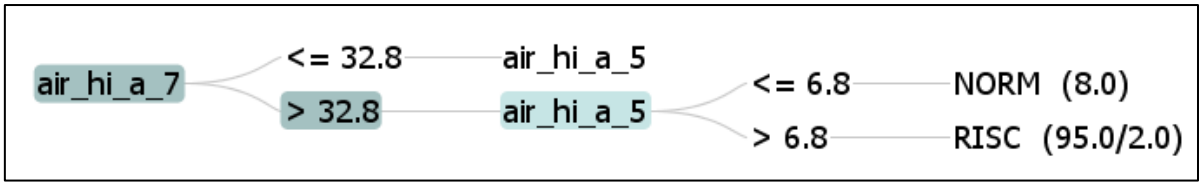
Tabela 5.3. Resultado com algoritmo J48 no *THDv* durante o turno T0.

ALGORÍTMO	Instâncias Classificadas Corretamente	Instâncias Classificadas Incorretamente	ROC Área	Positivos Falsos	Negativos Falsos
J48 – *THDv*-Total	99,9009 %	0,0991 %	1,000	3	0
J44 – *THDv*-T1	*THDv* Normal - sem resultado				
J44 – *THDv*-T2	*THDv* Normal - sem resultado				
J48 – *THDv*-T0	99,6422 %	0,3578 %	0,998	2	0

Fonte: Elaborada pelo autor (2017).

E a Árvore ficou conforme a Figura 5.5 e o resultado da 7ª componente harmônica de corrente do ponto **Condicionadores de Ar** para o conjunto de dados coletados é mantido e confirmando a análise anterior e conclui-se com as análises dos turnos T1, T2 e T0 que o impacto do *THD* acontece no turno T0.

Figura 5.5 - Árvore de Decisão do turno T0.



Fonte: Tela capturada pelo autor (2017).

5.2.3 Análise da Árvore de Decisão com dados gerais sem o ponto Compressor de Ar, sem balanceamento de classes, com *cross-validation* em 10 *folds* e com o algoritmo J48.

Para aumentar a quantidade de dias analisados e com base nas análises anteriores que demonstravam que o ponto **Compressores de Ar** não estava impactando o resultado, foi criada uma nova base de dados excluindo-se os dados do ponto **Compressores de Ar** com todos os turnos conforme pode ser visto na Figura 5.6.

Figura 5.6 - Dados coletados com COMPRESSORES e Sem COMPRESSORES.

Fonte: Elaborada pelo autor (2017).

Com essa estratégia a quantidade de linhas da tabela consolidada passou de 3028 para 9741, aumentando o período de 51h para 179h de dados a serem analisados. Essa estratégia não afetou a qualidade da análise conforme pode ser evidenciado na Tabela 5.4.

Tabela 5.4. Resultado com algoritmo J48 no *THDv* com e sem **Compressores de Ar**.

ALGORÍTMO	Instâncias Classificadas Corretamente	Instâncias Classificadas Incorretamente	ROC Área	Falsos Positivos	Falsos Negativos
J48 – *THDv* com COMPRESSORES	99,9009 %	0,0991 %	1,000	3	0
J48 – *THDv* sem COMPRESSORES	99,9281 %	0,0719 %	1,000	2	5

Fonte: Elaborada pelo autor (2017).

5.2.4 Resumo das análises da Árvore de Decisão sem balanceamento de classes, com *cross-validation* em 10 *folds* e com o algoritmo J48.

Após as diversas análises tem-se a tabela com o resumo dos resultados das Árvores de Decisão com o algoritmo J48 para o impacto no *THD* de Tensão do Quadro Geral de Entrada conforme mostrado na Tabela 5.5.

Tabela 5.5 - Resumo dos resultados com algoritmo J48 no THDv.

CENÁRIO	CARGA	TURNO	IMPACTANTE
1	Todas as cargas: **03/06 – 05/06**	Todos os turnos	**7ª componente harmônica de corrente Condicionadores de Ar**
2	Todas as cargas: **03/06 – 05/06**	Somente t1	Não possível executar, pois o *THDv* estava normal
3	Todas as cargas: **03/06 – 05/06**	Somente t2	Não possível executar, pois o *THDv* estava normal
4	Todas as cargas: **03/06 – 05/06**	Somente t0	**7ª componente harmônica de corrente Condicionadores de Ar**
5	Sem compressores: **03/06 – 11/06**	Todos os turnos	**7ª componente harmônica de corrente Condicionadores de Ar**
6	Sem compressores: **03/06 – 11/06**	Somente t1	Não possível executar, pois o *THDv* estava normal
7	Sem compressores: **03/06 – 11/06**	Somente t2	**7ª componente harmônica de corrente Condicionadores de Ar**
8	Sem compressores: **03/06 – 11/06**	Somente t0	**7ª componente harmônica de corrente Condicionadores de Ar**

Fonte: Elaborada pelo autor (2017).

Os cenários de 1 a 4 representam as análises com todos os dados coletados e aplicados as etapas do processo *KDD*, sendo o cenário 1 a análise com todos dos dados e os cenários 1, 2 e 3 as análises com os turnos T0, T1 e T3.

Os cenários de 5 a 8 representam as análises realizada com a retirada dos dados coletados no ponto compressores de ar e aplicados as etapas do processo *KDD*, sendo cenário 5 a análise com todos dos dados e os cenários 6, 7 e 8 as análises com os turnos T0, T1 e T3.

Com base nos resultados da árvore de decisão pôde-se concluir que a 7ª componente harmônica de corrente no ponto **Condicionadores de Ar** é a maior impactante para levar o *THD* de Tensão do sinal de entrada ao nível de risco durante o turno T0 em qualquer as hipóteses analisadas considerando o conjunto de dados coletados.

5.2.5 Análise da Árvore de Decisão com dados gerais, sem balanceamento de classes, com *cross-validation* em 10, 50, 100 e 1000 *folds* e com o algoritmo J48.

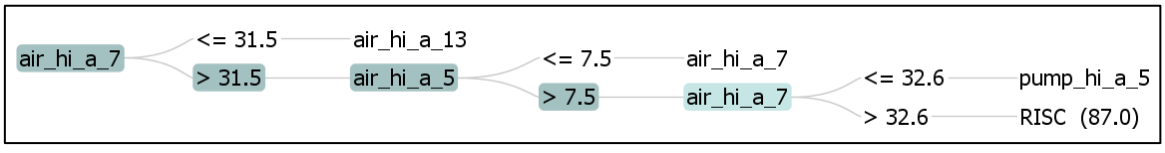
Com o propósito de análise de consistência os dados gerais com 3028 instâncias sendo 2902 classificadas com NORM e 126 classificadas com RISC, as quais foram submetidos a diferentes números de *folds* da técnica de validação cruzada para comparação do resultado final e da acurácia de cada execução conforme apresentadas na Tabela 5.6 e o resultado apresentado com a Árvore na Figura 5.7.

Tabela 5.6 – Comparativo entre análise do *cross-validation* com *folds* 10, 50, 100 e 1000.

Métricas	***Folds* 10**	***Folds* 50**	***Folds* 100**	***Folds* 1000**
VP	2892	2890	2889	2891
VN	114	111	112	113
FP	12	15	14	13
FN	10	12	13	11
Acurácia	99,3%	99,1%	99,1%	99,2%
TV_P Sensibilidade	99,7%	99,6%	99,6%	99,6%
TF_P Recall	9,5%	11,9%	11,1%	10,3%
TV_N Especificidade	90,5%	88,1%	88,9%	89,7%
TF_N	0,3%	0,4%	0,4%	0,4%
VP_P Precisão	99,6%	99,5%	99,5%	99,6%
VP_N	91,9%	90,2%	89,6%	91,1%
Gmean	95,0%	93,7%	94,1%	94,5%
ROC *Area*	0,980	0,958	0,962	0,967
Resultado	**7ª componente harmônica de corrente dos Condicionadores de Ar**			

Fonte: Elaborada pelo autor (2018).

Figura 5.7 – Árvore de Decisão para a condição sem balanceamento e *cross-validation*.



Fonte: Tela capturada pelo autor (2018).

O resultado da Tabela 5.6 e da Figura 5.7 evidenciaram que a componente harmônica de corrente mais impactante no *THDv* da entrada é a sétima e que as métricas em todos os cenários são muito semelhantes. A Figura 5.7 apresenta a

Árvore de Decisão para todos os cenários de utilização da validação cruzada, ou seja, o resultado foi exatamente o mesmo para os *folds* 10, 50, 100 e 100.

Similarmente o aconteceu na Figura 5.2, na Figura 5.7 também evidencia que se a sétima harmônica das centrais de ar condicionado for maior que 31,5% e a quinta harmônica de corrente das centrais de ar for maior que 7,5%, então o THDv entrará na condição de risco em 87 casos do total de 126 considerando o conjunto de dados coletados durante o período da campanha de medição.

5.2.6 Análise da Árvore de Decisão com dados gerais, sem balanceamento de classes, com taxa de treinamento e teste de 30/70%, 50/50% e 70/30% e com o algoritmo J48.

Utilizar diferentes cenários de taxa de treinamento e teste teve o propósito de analisar a coerência do resultado encontrado até então conforme apresentado na Tabela 5.7. Desta forma os dados Verdadeiro Positivo-VP vão sendo reduzido, pois correspondem ao percentual do conjunto de teste, assim em taxa 30/70 o valor de VP corresponde a 70% do conjunto de dados, em taxa 50/50 corresponde a 50% do conjunto de dados e em taxa 70/30 somente 30% do conjunto de dados.

Tabela 5.7 – Comparativo entre análise com taxas de treinamento e teste de 30/70, 50/50 e 70/30.

Métricas	***Cross-Validation Folds* 10**	***Taxa 30/70***	***Taxa 50/50***	***Taxa 70/30***
VP	2892	2020	1436	855
VN	114	65	58	47
FP	12	26	18	4
FN	10	9	2	2
Acurácia	99,3%	98,3%	98,7%	99,3%
TV_P Sensibilidade	99,7%	99,6%	99,9%	99,8%
TF_P Recall	9,5%	28,6%	23,7%	7,8%
TV_N Especificidade	90,5%	71,4%	76,3%	92,2%
TF_N	0,3%	0,4%	0,1%	0,2%
VP_P Precisão	99,6%	98,7%	98,8%	99,5%

Métricas	*Cross-Validation* *Folds* 10	*Taxa* *30/70*	*Taxa* *50/50*	*Taxa* *70/30*
VP_N	91,9%	87,8%	96,7%	95,9%
Gmean	95,0%	84,3%	87,3%	95,9%
ROC *Area*	0,980	0,903	0,856	0,961
Resultado	**7ª componente harmônica de corrente dos Condicionadores de Ar**			

Fonte: Elaborada pelo autor (2018).

Os dados da Tabela 5.7 demonstram que as métricas melhoram com o aumento do percentual de treinamento e que na condição 70/30 os resultados são melhores do que a validação cruzada. Apresenta ainda que o resultado é exatamente igual ao da validação cruzada, além disso a Árvore de Decisão apresentada na Figura 5.7 é a mesma para as três condições de taxa de treinamento e teste.

5.2.7 Análise da Árvore de Decisão com dados gerais com balanceamento de classes e com diferentes taxas de treinamento e teste.

A utilização de técnicas de balanceamento de classes é fundamental para a melhoria do desempenho das técnicas de mineração de dados. Como nos dados originais a distribuição entre as instâncias NORM com 2902 e RISC com 126 demonstram que há um desbalanceamento muito grande, pois NORM representa 95,8% dos dados e RISC apenas 4,2%.

Para o equilíbrio dos dados buscando a melhoria do desempenho foi aplicado o filtro SMOTE que está enquadrado como uma técnica sintética de sobre amostragem. Para as análises foram aplicados diferentes percentuais de sobre amostragem nos dados originais correspondendo a 100%, 200%, 300%, 1000% e 2000% em diferentes análise de taxa de treinamento e teste conforme é apresentado na Tabela 5.8.

Tabela 5.8 – Comparativo dos dados gerais entre análise de balanceamento com SMOTE.

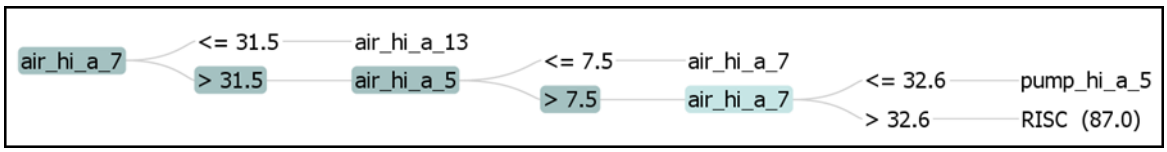
Instâncias	Original	SMOTE 100%	SMOTE 200%	SMOTE 300%	SMOTE 1000%	SMOTE 2000%
Total	3028	3154	3280	3406	4288	5548
NORM	2902	2902	2902	2902	2902	2902
RISC	126	252	378	504	1386	2646
%	95,8/4,2	92,0/8,0	88,4/11,6	85,2/14,8	67,7/32,3	52,3/47,7
% Treinamento/Teste = 30/70						
VP	2020	2029	2034	2021	2016	1998
VN	65	155	240	334	948	1854
FP	26	18	14	19	15	7

Instâncias	Original	SMOTE 100%	SMOTE 200%	SMOTE 300%	SMOTE 1000%	SMOTE 2000%
FN	9	6	8	10	23	25
Acurácia	98,3%	98,9%	99,0%	98,8%	98,7%	99,2%
TV_P Sensibilidade	99,6%	99,7%	99,6%	99,5%	98,9%	98,8%
TF_P Recall	28,6%	10,4%	5,5%	5,4%	1,6%	**0,4%**
TV_N Especificidade	71,4%	89,6%	94,5%	94,6%	98,4%	**99,6%**
TF_N	0,4%	0,3%	0,4%	0,5%	1,1%	1,2%
VP_P Precisão	98,7%	99,1%	99,3%	99,1%	99,3%	**99,7%**
VP_N	87,8%	96,3%	96,8%	97,1%	97,6%	98,7%
Gmean	84,3%	94,5%	97,0%	97,0%	98,7%	99,2%
ROC Area	0,903	0,931	0,961	0,974	**0,988**	0,995
% Treinamento/Teste = 50/50						
VP	1436	1448	1446	1440	1444	1401
VN	58	115	173	251	679	1349
FP	18	7	9	9	8	9
FN	2	7	12	3	13	15
Acurácia	98,7 %	**99,1%**	98,7%	**99,3%**	**99,0%**	99,1%
TV_P Sensibilidade	99,9%	99,5%	99,2%	99,8%	99,1%	98,9%
TF_P Recall	23,7%	**5,7%**	4,9%	3,5%	**1,2%**	0,7%
TV_N Especificidade	76,3%	**94,3%**	95,1%	96,5%	**98,8%**	99,3%
TF_N	0,1%	0,5%	0,8%	0,2%	0,9%	1,1%
VP_P Precisão	98,8%	**99,5%**	99,4%	99,4%	**99,4%**	99,4%
VP_N	96,7%	94,3%	93,5%	98,8%	98,1%	98,9%
Gmean	87,3%	**96,9%**	97,1%	98,2%	**99,0%**	99,1%
ROC Area	0,856	**0,970**	0,996	0,986	0,987	0,995
% Treinamento/Teste = 70/30						
VP	855	862	878	869	858	851
VN	47	65	101	145	413	800
FP	4	15	0	3	7	5
FN	2	4	5	5	8	8
Acurácia	**99,3%**	98,0%	**99,5%**	99,2%	98,8%	**99,2%**
TV_P Sensibilidade	99,8%	99,5%	99,4%	99,4%	99,1%	**99,1%**
TF_P Recall	**7,8%**	18,8%	**0,0%**	**2,0%**	1,7%	0,6%
TV_N Especificidade	**92,2%**	81,3%	**100,0%**	**98,0%**	98,3%	99,4%
TF_N	0,2%	0,5%	0,6%	0,6%	0,9%	0,9%
VP_P Precisão	**99,5%**	98,3%	**100,0%**	**99,7%**	99,2%	99,4%
VP_N	95,9%	94,2%	95,3%	96,7%	98,1%	99,0%
Gmean	**95,9%**	89,9%	**99,7%**	98,7%	98,7%	**99,2%**
ROC Area	**0,961**	0,869	**0,997**	**0,992**	0,987	0,994

Fonte: Elaborada pelo autor (2018).

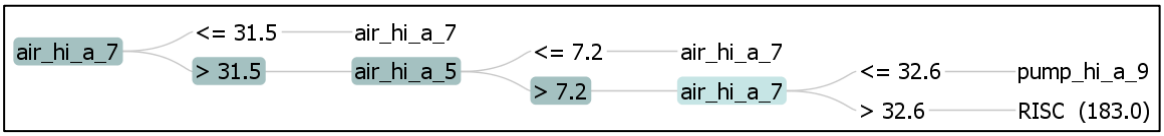
Observa-se que conforme os dados vão sendo equilibrados as métricas tendem a melhorar. Considerando os dados coletados na semana da campanha de qualidade de energia, os resultados das Árvores de Decisão permanecem muito parecidos e também indicam a sétima componente harmônica de corrente dos Condicionadores de Ar como a principal impactante conforme visto nas Figuras de 5.8 a 5.13.

Figura 5.8 – Árvore de Decisão com treinamento 70/30 e sem balanceamento.



Fonte: Tela capturada pelo autor (2018).

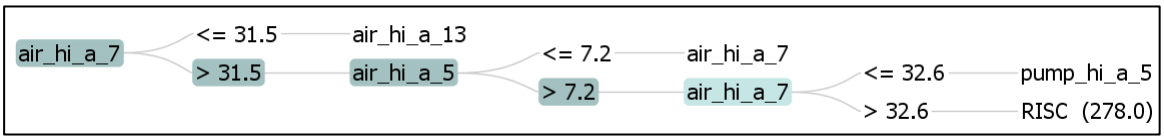
Figura 5.9 – Árvore de Decisão com treinamento 70/30 e com balanceamento SMOTE 100%.



Fonte: Tela capturada pelo autor (2018).

As diferenças entre as Figuras 5.8 e 5.9 são o índice da quinta componente harmônica de corrente das Centrais de Ar que muda de 7,5% para 7,2% e o valor do RISC que muda de 87,0 para 183,0 casos, considerando os dados coletados.

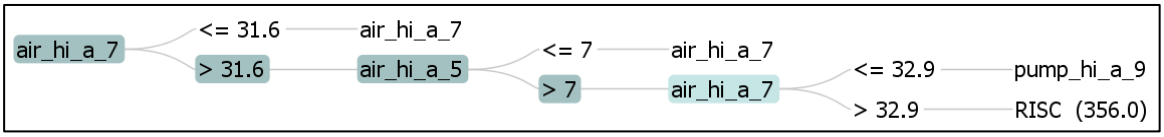
Figura 5.10 – Árvore de Decisão com treinamento 70/30 e com balanceamento SMOTE 200%.



Fonte: Tela capturada pelo autor (2018).

A única diferença entre as Figuras 5.9 e 5.10 é o valor do RISC que muda de 183,0 para 278,0 casos, considerando os dados coletados.

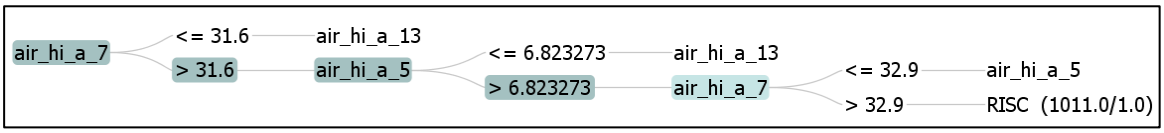
Figura 5.11 – Árvore de Decisão com treinamento 70/30 e com balanceamento SMOTE 300%.



Fonte: Tela capturada pelo autor (2018).

As diferenças entre as Figuras 5.10 e 5.11 são o índice da sétima componente harmônica de corrente das Centrais de Ar que muda de 31,5% para 31,6%, da quinta componente harmônica de corrente das Centrais de Ar que muda de 7,2% para 7,0%, novamente o índice da sétima componente harmônica de corrente das Centrais de Ar que muda de 32,6% para 32,9% e o valor do RISC que muda de 278,0 para 356,0 casos, considerando os dados coletados.

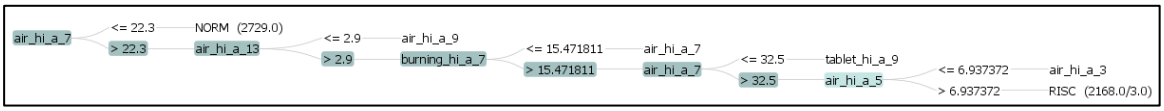
Figura 5.12 – Árvore de Decisão com treinamento 70/30 e com balanceamento SMOTE 1000%.



Fonte: Tela capturada pelo autor (2018).

As diferenças entre as Figuras 5.11 e 5.12 são o índice da quinta componente harmônica de corrente das Centrais de Ar que muda de 7,0% para 6,823273% e o valor do RISC que muda de 356,0 para 1011,0/1,0 casos, considerando os dados coletados.

Figura 5.13 – Árvore de Decisão com treinamento 70/30 e com balanceamento SMOTE 2000%.



Fonte: Tela capturada pelo autor (2018).

Com a configuração do SMOTE com 2000% a Árvore de Decisão evidenciada na Figura 5.13, já apresenta uma quantidade maior de ramos e aparecem a décima terceira componente harmônica de corrente das Centrais de Ar e a sétima componente harmônica de corrente do *Burning*, no entanto continua mantendo a sétima componente harmônica de corrente das Centrais de Ar como a maior impactante.

5.2.8 Análise da Árvore de Decisão com dados do turno T0 com balanceamento de classes e com diferentes taxas de treinamento e teste.

Utilizando-se a mesma metodologia anterior foram testados os mesmos cenários da seção 5.2.5 com os dados do turno T0 conforme apresentado na Tabela 5.9.

Tabela 5.9 – Comparativo dos dados do turno T0 com análise de balanceamento com SMOTE.

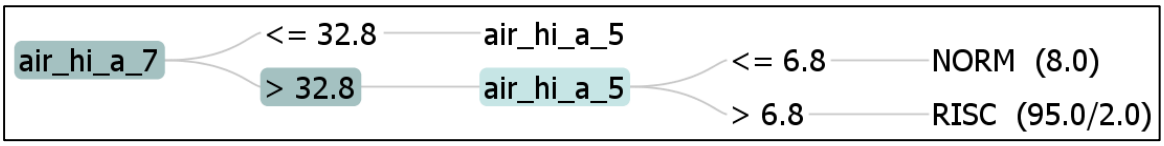
Instâncias	Original	SMOTE 100%	SMOTE 200%	SMOTE 300%	SMOTE 1000%	SMOTE 2000%
Total	559	685	811	937	1819	3079
NORM	433	433	433	433	433	433
RISC	126	252	378	504	1386	2646
%	77,5/22,5	63,2/36,8	53,4/46,6	46,2/53,8	23,8/76,2	14,0/86,0
% Treinamento/Teste = 30/70						
VP	291	281	285	282	262	263
VN	79	167	247	343	956	1859
FP	10	8	23	9	24	16

Instâncias	Original	SMOTE 100%	SMOTE 200%	SMOTE 300%	SMOTE 1000%	SMOTE 2000%
FN	11	23	13	22	31	17
Acurácia	94,6%	93,5%	93,7%	95,3%	95,7%	98,5%
TV_P Sensibilidade	96,4%	92,4%	95,6%	92,8%	89,4%	93,9%
TF_P Recall	11,2%	**4,6%**	8,5%	2,6%	2,4%	0,9%
TV_N Especificidade	88,8%	95,4%	91,5%	97,4%	97,6%	99,1%
TF_N	3,6%	7,6%	4,4%	7,2%	10,6%	6,1%
VP_P Precisão	96,7%	97,2%	92,5%	96,9%	91,6%	94,3%
VP_N	87,8%	87,9%	95,0%	94,0%	96,9%	**99,1%**
Gmean	92,5%	93,9%	93,5%	95,1%	93,4%	96,5%
ROC Area	0,908	0,968	0,936	0,949	0,914	0,969
% Treinamento/Teste = 50/50						
VP	211	197	205	206	205	176
VN	55	123	181	251	679	1343
FP	8	8	10	4	17	6
FN	5	14	9	7	8	14
Acurácia	95,3%	93,6%	95,3%	**97,6%**	97,2%	**98,7%**
TV_P Sensibilidade	97,7%	93,4%	95,8%	96,7%	96,2%	92,6%
TF_P Recall	12,7%	6,1%	5,2%	1,6%	2,4%	**0,4%**
TV_N Especificidade	87,3%	**93,9%**	94,8%	98,4%	97,6%	**99,6%**
TF_N	2,3%	6,6%	4,2%	3,3%	3,8%	7,4%
VP_P Precisão	96,3%	**96,1%**	95,3%	98,1%	92,3%	**96,7%**
VP_N	91,7%	89,8%	95,3%	97,3%	98,8%	99,0%
Gmean	92,3%	93,6%	95,3%	**97,6%**	96,9%	96,0%
ROC Area	0,928	**0,935**	0,961	**0,991**	0,969	0,964
% Treinamento/Teste = 70/30						
VP	128	124	130	122	127	111
VN	33	72	106	151	414	800
FP	4	6	2	2	3	6
FN	3	3	5	6	2	7
Acurácia	**95,8%**	**95,6%**	**97,1%**	97,2%	**99,1%**	98,6%
TV_P Sensibilidade	97,7%	97,6%	96,3%	95,3%	98,4%	94,1%
TF_P Recall	**10,8%**	7,7%	**1,9%**	**1,3%**	**0,7%**	0,7%
TV_N Especificidade	**89,2%**	92,3%	**98,1%**	**98,7%**	**99,3%**	99,3%
TF_N	2,3%	2,4%	3,7%	4,7%	1,6%	5,9%
VP_P Precisão	**97,0%**	95,4%	**98,5%**	**98,4%**	**97,7%**	94,9%
VP_N	91,7%	96,0%	95,5%	96,2%	99,5%	99,1%
Gmean	**93,4%**	**94,9%**	**97,2%**	97,0%	**98,9%**	**96,6%**
ROC Area	**0,944**	0,930	**0,979**	0,973	**0,991**	**0,980**

Fonte: Elaborada pelo autor (2018).

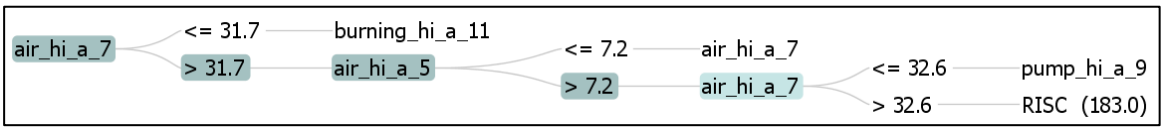
O mesmo comportamento da seção 5.2.5 é observado com os dados do turno T0 e os resultados das Árvores de Decisão permanecem muito parecidos e também indicam a 7ª componente harmônica de corrente dos Condicionadores de Ar como a principal impactante, considerando os dados coletados, com exceção da última análise conforme visto nas Figuras de 5.14 a 5.19.

Figura 5.14 – Árvore de Decisão com treinamento 70/30 e sem balanceamento no turno T0.



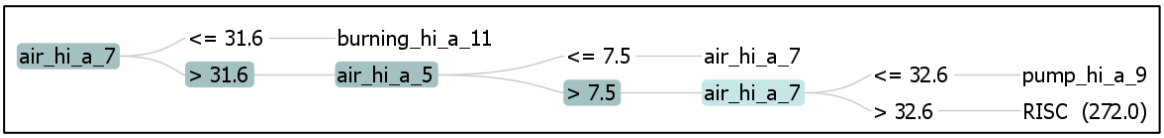
Fonte: Tela capturada pelo autor (2018).

Figura 5.15 – Árvore de Decisão com treinamento 70/30 e com SMOTE 100% no turno T0.



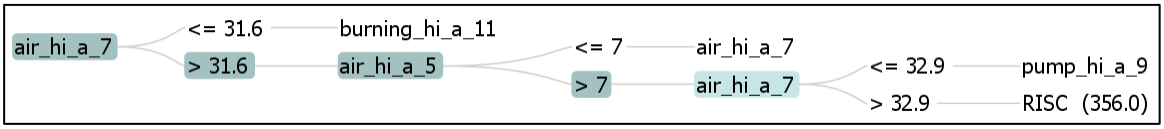
Fonte: Tela capturada pelo autor (2018).

Figura 5.16 – Árvore de Decisão com treinamento 70/30 e com SMOTE 200% no turno T0.



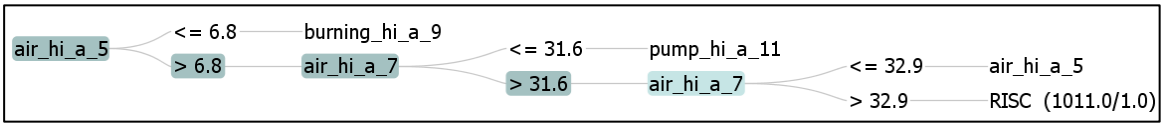
Fonte: Tela capturada pelo autor (2018).

Figura 5.17 – Árvore de Decisão com treinamento 70/30 e com SMOTE 300% no turno T0.



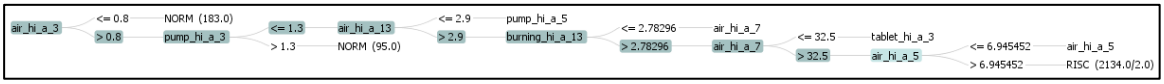
Fonte: Tela capturada pelo autor (2018).

Figura 5.18 – Árvore de Decisão com treinamento 70/30 e com SMOTE 1000% no turno T0.



Fonte: Tela capturada pelo autor (2018).

Figura 5.19 – Árvore de Decisão com treinamento 70/30 e com SMOTE 2000% no turno T0.



Fonte: Tela capturada pelo autor (2018).

É importante observar que diferentemente dos dados gerais, nas análises realizadas no turno T0 o equilíbrio dos dados aconteceu com SMOTE 200%, enquanto que nos dados gerais aconteceu com SMOTE 2000% devido a diferença grande entre as instâncias NORM e RISC.

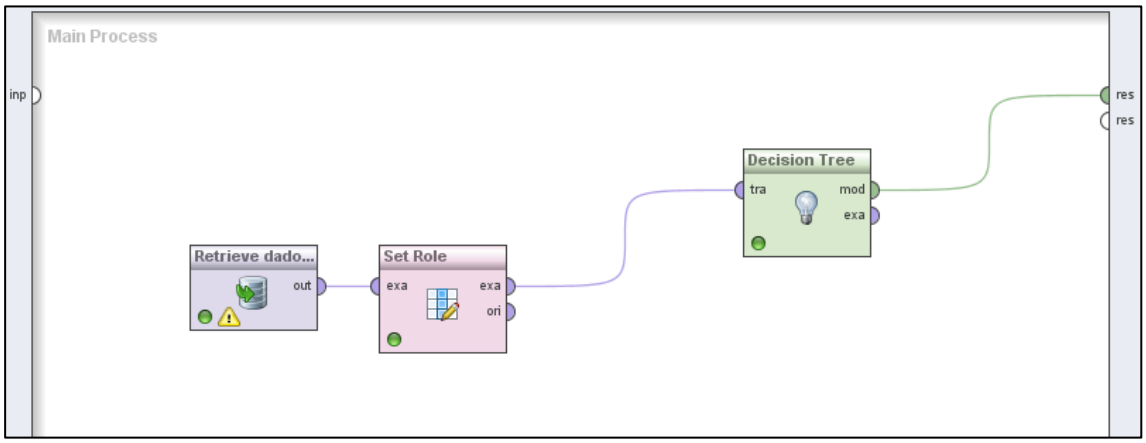
5.3 Resultado com classificador Árvore de Decisão no *Software Rapid Miner Studio*.

Para validar os dados utilizou-se, ainda, como ferramenta de Mineração de Dados o *software Rapid Miner Studio Basic* versão 6.5.02 (2015).

Os dados consolidados foram convertidos do Banco de Dados *PostgreSQL* para o *CSV* para facilitar a leitura do mesmo pelo *Rapid Miner*.

Após o carregamento dos dados, a ferramenta foi configurada para o processo com o classificador Árvore de Decisão conforme Figura 5.20.

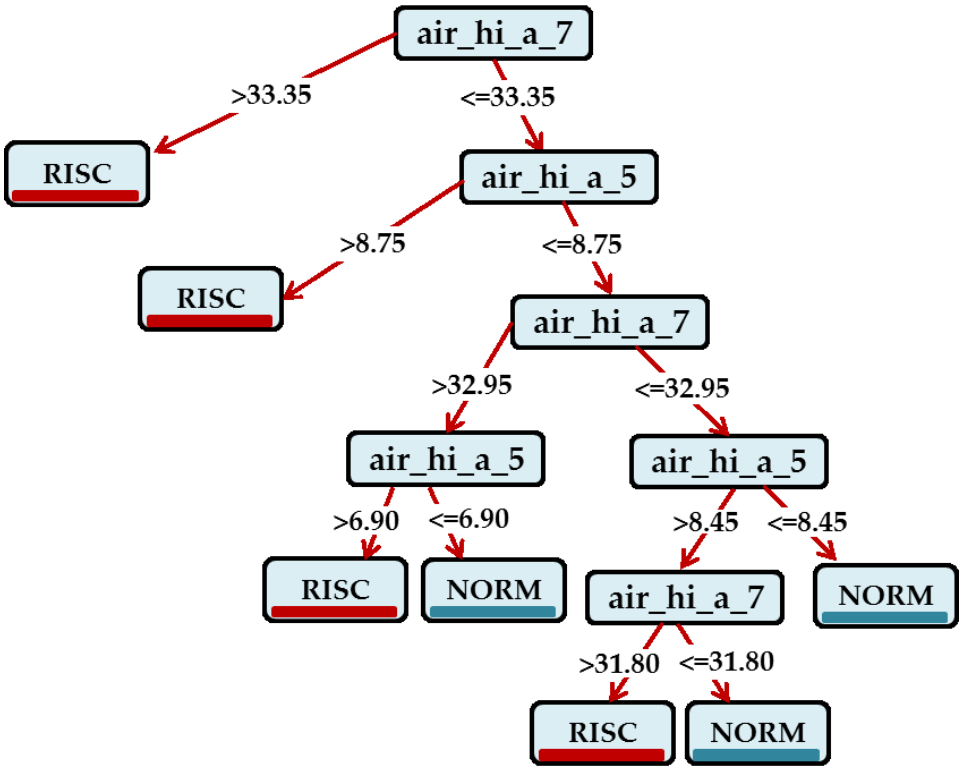
Figura 5.20 - Configuração do *software Rapid Miner* para Árvore de Decisão.



Fonte: Tela capturada pelo autor (2017).

A Árvore de Decisão gerada no *Rapid Miner* evidenciou que a sétima componente harmônica de corrente do ponto **Condicionadores de Ar** é a maior impactante conforme evidenciado pela Figura 5.21 e considerando os dados coletados na semana de realização da campanha de medição.

Figura 5.21 - Resultado do classificador Árvore de Decisão no *software Rapid Miner*.



Fonte. Adaptada do *software Rapid Miner* (2017).

5.4 Resultado com classificador Naïve Bayes no *Software WEKA*.

Com o propósito de completar a análise dos dados na etapa da mineração dos dados com dois classificadores, inicialmente as análises foram feitas com o classificador Árvore de Decisão em diferentes cenários. Nessa seção será abordada a utilização do classificador Naïve Bayes com os principais cenários utilizados nas seções anteriores desse capítulo.

5.4.1 Análise dos dados gerais com Naïve Bayes, sem balanceamento de classes, com *cross-validation* em 10 *folds* e com os algoritmos BAYESNET, NAÏVE BAYES, NAÏVE BAYES MULTINOMIAL TEXT e NAÏVE BAYES UPDATEABLE.

Na análise com o classificador Naïve Bayes no *software WEKA* utilizou-se diversos algoritmos e a opção de teste foi ajustada com 10 *folds* na técnica *cross-validation* para identificar o resultado mais significativo baseado nas métricas

Instâncias Classificadas Corretamente, Instância Classificadas Incorretamente, Área ROC, Falsos Positivos e Falsos Negativos conforme Tabela 5.10.

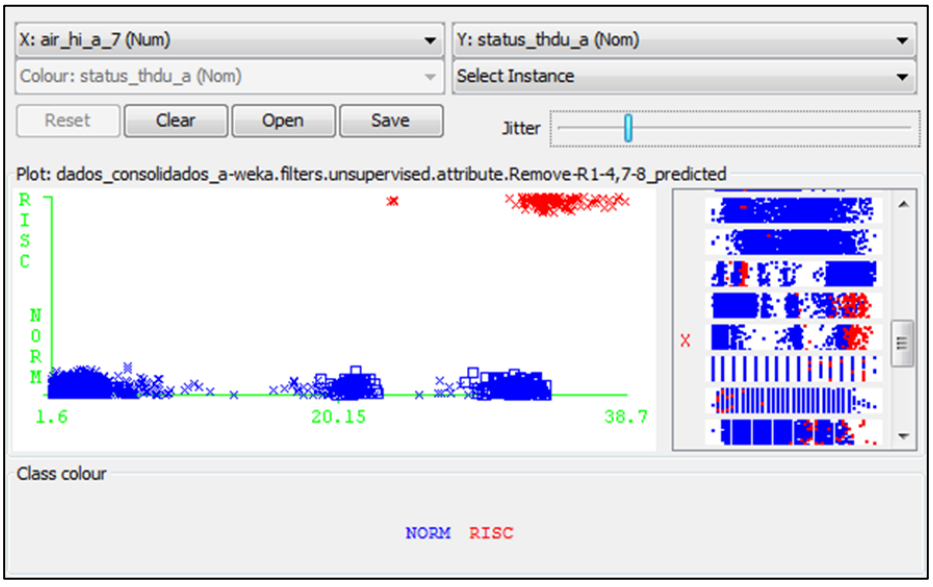
Tabela 5.10 - Resultado dos algoritmos do classificador Naïve Bayes do software WEKA.

ALGORÍTMO	Instâncias Classificadas Corretamente	Instâncias Classificadas Incorretamente	ROC Área	Positivos Falsos	Negativos Falsos
BAYESNET	89,6962 %	10,3038 %	0,996	312	0
NAÏVE BAYES	92,4373%	7,5627 %	0,984	229	0
NAÏVE BAYES MULTINOMIAL TEXT	95,8388 %	4,1612 %	0,500	0	126
NAÏVE BAYES UPDATEABLE	92,4373 %	7,5627 %	0,984	229	0

Fonte: Elaborada pelo autor (2018).

De acordo com a Tabela 5.10, os algoritmos NAÏVE BAYES e NAÏVE BAYES UPDATEABLE apresentaram os resultados mais significativos devido aos itens **Instâncias Classificado Corretamente**, **Instâncias Classificadas Incorretamente**, **ROC Área**, **Positivos Falsos** e **Negativos Falsos.** O resultado indicou como a maior impactante a sétima componente harmônica de corrente do ponto **Condicionadores de Ar** conforme evidenciado na Figura 5.22, considerando os dados coletados durante a campanha de medição.

Figura 5.22 – Resultado do Naïve Bayes sem balanceamento e com 10 *folds* no *cross-validation*.



Fonte: Tela capturada pelo autor (2018).

A figura 5.22 pode ser observada o quadrado que contém as distribuições para cada instância, onde os pontos em azul são os classificados como normal (NORM) e os vermelhos são os classificados como em risco (RISC). Desta forma observa-se que os que possuem as maiores concentrações são os das quinta e sétima componentes harmônicas de corrente no pontos Ar Condicionado, considerando os dados coletados durante uma semana de campanha de medição. Na Figura 5.22 é mostrada na tela maior o valor numérico da sétima harmônica de corrente conforme o eixo **X:air_hi_a_7 (Num)** contra o valor paramétrico conforme o eixo **Y:status_thdu_a (Nom)**, assim sendo observa-se que a maior concentração está em torno do valor máximo de 38.7%.

5.4.2 Análise dos dados gerais com Naïve Bayes com balanceamento de classes e com diferentes taxas de treinamento e teste.

Nessa seção foi utilizado o filtro SMOTE com os mesmos valores aplicados nas análises com o classificador Árvore de Decisão e com as mesmas taxas de treinamento e teste conforme é apresentado na Tabela 5.11.

Tabela 5.11 – Comparativo dos dados gerais entre análise de balanceamento com SMOTE no classificador Naïve Bayes.

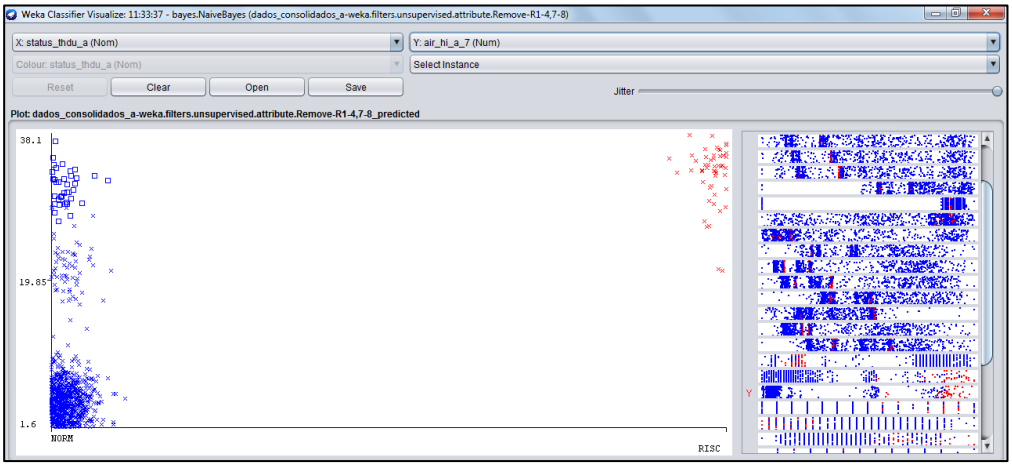
Instâncias	Original	SMOTE 100%	SMOTE 200%	SMOTE 300%	SMOTE 1000%	SMOTE 2000%
Total	3028	3154	3280	3406	4288	5548
NORM	2902	2902	2902	2902	2902	2902
RISC	126	252	378	504	1386	2646
%	95,8/4,2	92,0/8,0	88,4/11,6	85,2/14,8	67,7/32,3	52,3/47,7
% Treinamento/Teste = 30/70						
VP	1886	1930	1883	1834	1858	1826
VN	91	173	254	351	963	1859
FP	0	0	0	2	0	2
FN	143	105	159	197	181	197
Acurácia	93,3%	**95,2%**	**93,1%**	**91,7%**	**94,0%**	94,9%
TV_P Sensibilidade	93,0%	94,8%	92,2%	90,3%	91,1%	90,3%
TF_P Recall	0,0%	**0,0%**	**0,0%**	**0,6%**	**0,0%**	**0,1%**
TV_N Especificidade	100,0%	**100,0%**	**100,0%**	**99,4%**	**100,0%**	**99,9%**
TF_N	7,0%	5,2%	7,8%	9,7%	8,9%	9,7%
VP_P Precisão	100,0%	**100,0%**	**100,0%**	**99,9%**	**100,0%**	**99,9%**
VP_N	38,9%	62,2%	61,5%	64,1%	84,2%	90,4%
Gmean	96,4%	**97,4%**	**96,0%**	**94,8%**	**95,5%**	95,0%
ROC Area	0,987	**0,987**	0,983	**0,984**	**0,983**	**0,980**
% Treinamento/Teste = 50/50						
VP	1358	1347	1336	1296	1321	1289
VN	76	121	181	258	685	1354
FP	0	1	1	2	2	4
FN	80	108	122	147	136	127

Instâncias	Original	SMOTE 100%	SMOTE 200%	SMOTE 300%	SMOTE 1000%	SMOTE 2000%
Acurácia	94,7%	93,1%	92,5%	91,3%	93,6%	95,3%
TV_P Sensibilidade	94,4%	92,6%	91,6%	89,8%	90,7%	91,0%
TF_P Recall	**0,0%**	0,8%	0,5%	0,8%	0,3%	0,3%
TV_N Especificidade	**100,0%**	99,2%	99,5%	99,2%	99,7%	99,7%
TF_N	5,6%	7,4%	8,4%	10,2%	9,3%	9,0%
VP_P Precisão	**100,0%**	99,9%	99,9%	99,8%	99,8%	99,7%
VP_N	48,7%	52,8%	59,7%	63,7%	83,4%	91,4%
Gmean	97,2%	95,8%	95,5%	94,4%	95,1%	95,3%
ROC Area	0,987	0,980	0,981	0,982	0,981	0,978
% Treinamento/Teste = 70/30						
VP	820	810	811	785	790	785
VN	51	80	101	146	419	804
FP	0	0	0	2	1	1
FN	37	56	72	89	76	74
Acurácia	**95,9%**	94,1%	92,7%	91,1%	**94,0%**	**95,5%**
TV_P Sensibilidade	95,7%	93,5%	91,8%	89,8%	91,2%	**91,4%**
TF_P Recall	**0,0%**	0,0%	**0,0%**	1,4%	0,2%	**0,1%**
TV_N Especificidade	**100,0%**	100,0%	**100,0%**	98,6%	99,8%	**99,9%**
TF_N	4,3%	6,5%	8,2%	10,2%	8,8%	8,6%
VP_P Precisão	**100,0%**	100,0%	**100,0%**	99,7%	99,9%	**99,9%**
VP_N	58,0%	58,8%	58,4%	62,1%	84,6%	91,6%
Gmean	**97,8%**	96,7%	95,8%	94,1%	95,4%	**95,5%**
ROC Area	**0,992**	0,984	**0,985**	0,979	0,981	**0,980**

Fonte: Elaborada pelo autor (2018).

Diferentemente das análises com a Árvore de Decisão, as métricas mais significativas foram na condição SMOTE 100% com taxa de treinamento 30/70 e SMOTE 2000% com taxa de treinamento 70/30. Os resultados com o classificador Naïve Bayes permanecem muito parecidos e também indicam a 7ª componente harmônica de corrente dos Condicionadores de Ar como a principal impactante conforme visto nas Figuras de 5.23 a 5.28.

Figura 5.23 – Naïve Bayes com treinamento 70/30 e sem balanceamento.



Fonte: Tela capturada pelo autor (2018).

Figura 5.24 – Naïve Bayes com treinamento 70/30 e com balanceamento SMOTE 100%.

Fonte: Tela capturada pelo autor (2018).

Figura 5.25 – Naïve Bayes com treinamento 70/30 e com balanceamento SMOTE 200%.

Fonte: Tela capturada pelo autor (2018).

Figura 5.26 – Naïve Bayes com treinamento 70/30 e com balanceamento SMOTE 300%.

Fonte: Tela capturada pelo autor (2018).

Figura 5.27 – Naïve Bayes com treinamento 70/30 e com balanceamento SMOTE 1000%.

Fonte: Tela capturada pelo autor (2018).

Figura 5.28 – Naïve Bayes com treinamento 70/30 e com balanceamento SMOTE 2000%.

Fonte: Tela capturada pelo autor (2018).

5.4.3 Análise dos dados do turno T0 com Naïve Bayes com balanceamento de classes e com diferentes taxas de treinamento e teste.

A Tabela 5.12 apresenta o resumo dos resultados das análise com os dados do turno T0 utilizando o balanceamento de classes por meio do SMOTE e diferentes taxas de treinamento.

Tabela 5.12 – Comparativo dos dados do turno T0 com análise de balanceamento com SMOTE.

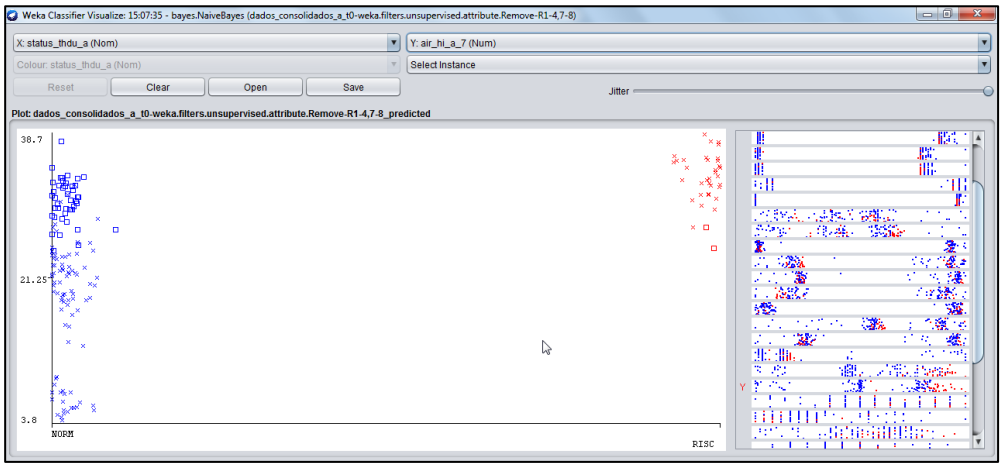
Instâncias	Original	SMOTE 100%	SMOTE 200%	SMOTE 300%	SMOTE 1000%	SMOTE 2000%
Total	559	685	811	937	1819	3079
NORM	433	433	433	433	433	433
RISC	126	252	378	504	1386	2646
%	77,5/22,5	63,2/36,8	53,4/46,6	46,2/53,8	23,8/76,2	14,0/86,0

Instâncias	Original	SMOTE 100%	SMOTE 200%	SMOTE 300%	SMOTE 1000%	SMOTE 2000%
% Treinamento/Teste = 30/70						
VP	199	194	196	198	191	177
VN	85	172	264	342	952	1835
FP	4	3	6	10	28	40
FN	103	110	102	106	102	103
Acurácia	72,6%	**76,4%**	81,0%	82,3%	89,8%	93,4%
TV_P Sensibilidade	65,9%	**63,8%**	65,8%	65,1%	65,2%	63,2%
TF_P Recall	4,5%	1,7%	2,2%	2,8%	2,9%	2,1%
TV_N Especificidade	95,5%	98,3%	**97,8%**	97,2%	97,1%	**97,9%**
TF_N	34,1%	36,2%	34,2%	34,9%	34,8%	36,8%
VP_P Precisão	98,0%	98,5%	**97,0%**	95,2%	87,2%	81,6%
VP_N	45,2%	61,0%	72,1%	76,3%	90,3%	**94,7%**
Gmean	79,3%	**79,2%**	80,2%	79,5%	79,6%	78,7%
ROC Area	0,893	0,904	O,916	0,906	**0,935**	0,902
% Treinamento/Teste = 50/50						
VP	141	127	143	140	139	125
VN	61	129	186	250	681	1319
FP	2	2	5	5	15	30
FN	75	84	71	73	74	65
Acurácia	72,4%	74,9%	81,2%	**83,3%**	**90,2%**	**93,8%**
TV_P Sensibilidade	65,3%	60,2%	66,8%	65,7%	65,3%	65,8%
TF_P Recall	**3,2%**	1,5%	2,6%	2,0%	2,2%	2,2%
TV_N Especificidade	**96,8%**	98,5%	97,4%	98,0%	**97,8%**	97,8%
TF_N	34,7%	39,8%	33,2%	34,3%	34,7%	34,2%
VP_P Precisão	**98,6%**	98,4%	96,6%	**96,6%**	**90,3%**	80,6%
VP_N	44,9%	60,6%	72,4%	77,4%	90,2%	95,3%
Gmean	**79,5%**	77,0%	80,7%	**80,3%**	**79,9%**	**80,2%**
ROC Area	0,928	**0,916**	0,921	0,911	0,920	**0,922**
% Treinamento/Teste = 70/30						
VP	87	74	94	82	84	77
VN	35	77	105	150	408	789
FP	2	1	3	3	9	17
FN	44	53	41	46	45	41
Acurácia	**72,6%**	73,7%	**81,9%**	82,6%	90,1%	93,7%
TV_P Sensibilidade	**66,4%**	58,3%	**69,6%**	64,1%	65,1%	65,3%
TF_P Recall	5,4%	1,3%	2,8%	**2,0%**	**2,2%**	2,1%
TV_N Especificidade	94,6%	**98,7%**	97,2%	**98,0%**	**97,8%**	**97,9%**
TF_N	33,6%	41,7%	30,4%	35,9%	34,9%	34,7%
VP_P Precisão	97,8%	**98,7%**	96,9%	96,5%	**90,3%**	**81,9%**
VP_N	44,3%	59,2%	71,9%	76,5%	90,1%	95,1%
Gmean	79,3%	75,8%	**82,3%**	79,3%	79,8%	**79,9%**
ROC Area	0,916	0,878	**0,926**	**0,928**	0,924	0,920

Fonte: Elaborada pelo autor (2018).

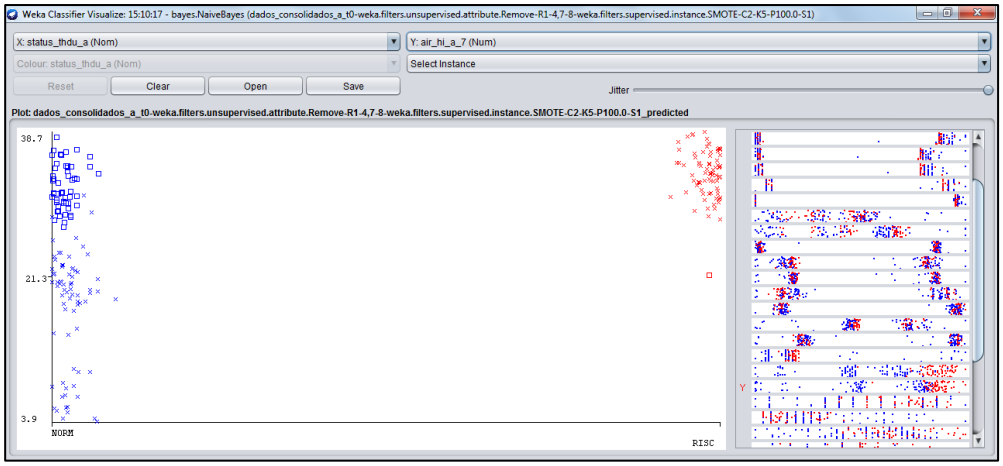
Diferentemente das análises com a Árvore de Decisão, as métricas mais significativas foram na condição SMOTE 200% com taxa de treinamento 30/70 e SMOTE 2000% com taxa de treinamento 70/30. Os resultados com o classificador Naïve Bayes permanecem muito parecidos e também indicam a sétima componente harmônica de corrente dos Condicionadores de Ar como a principal impactante conforme visto nas Figuras de 5.29 a 5.34, considerando os dados coletados.

Figura 5.29 – Naïve Bayes com treinamento 70/30 e sem balanceamento no turno T0.



Fonte: Tela capturada pelo autor (2018).

Figura 5.30 – Naïve Bayes com treinamento 70/30 e com balanceamento SMOTE 100% no turno T0.

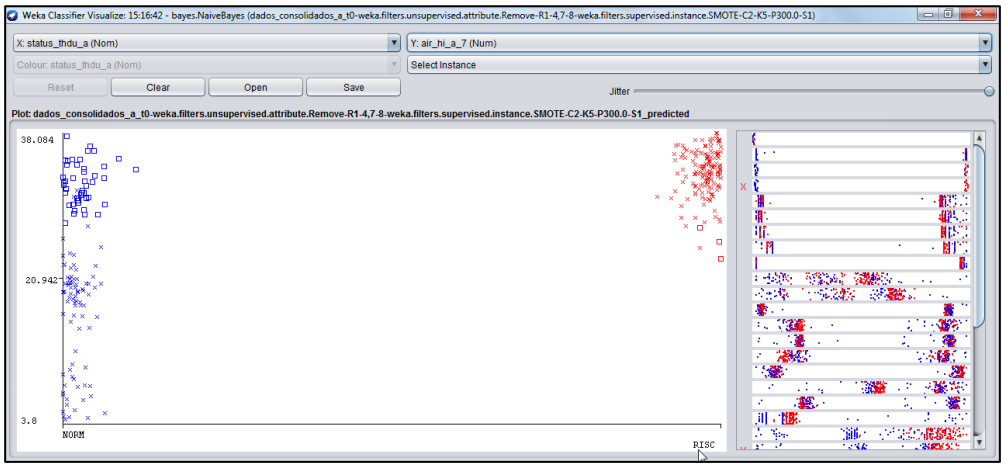


Fonte: Tela capturada pelo autor (2018).

Figura 5.31 – Naïve Bayes com treinamento 70/30 e com balanceamento SMOTE 200% no turno T0.

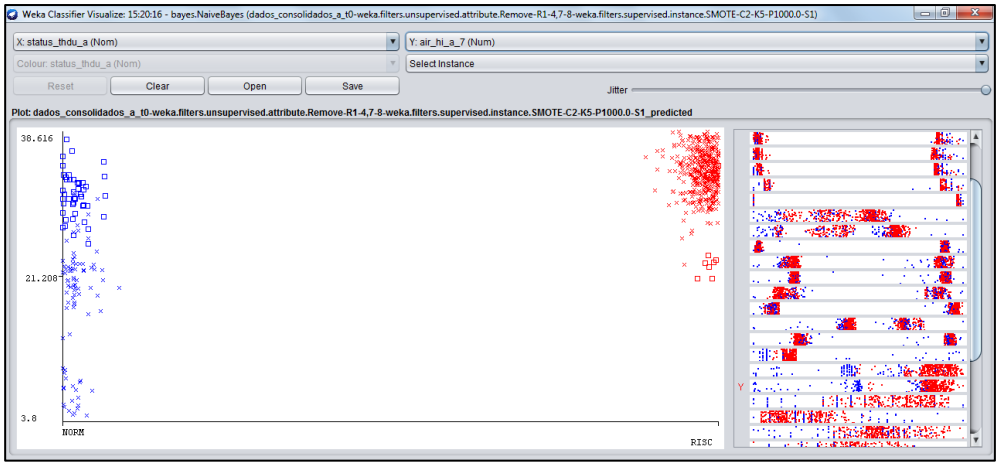
Fonte: Tela capturada pelo autor (2018).

Figura 5.32 – Naïve Bayes com treinamento 70/30 e com balanceamento SMOTE 300% no turno T0.



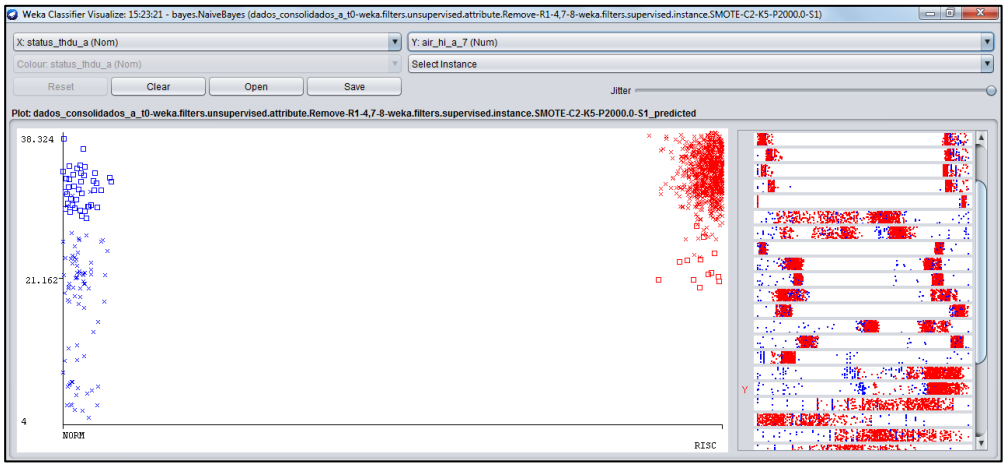
Fonte: Tela capturada pelo autor (2018).

Figura 5.33 – Naïve Bayes com treinamento 70/30 e com balanceamento SMOTE 1000% no turno T0.



Fonte: Tela capturada pelo autor (2018).

Figura 5.34 – Naïve Bayes com treinamento 70/30 e com balanceamento SMOTE 2000% no turno T0.

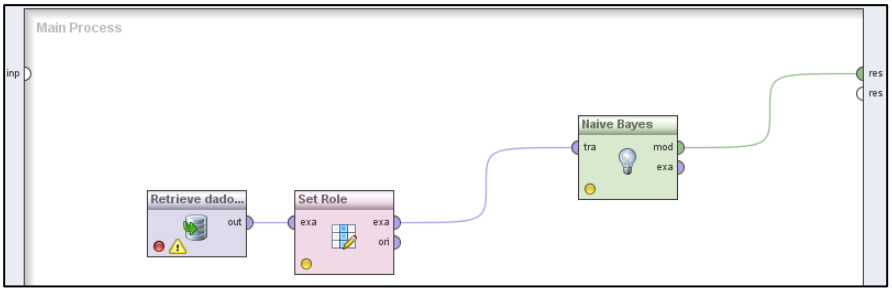


Fonte: Tela capturada pelo autor (2018).

5.5 Resultado com classificador Naïve Bayes no *Software Rapid Miner*.

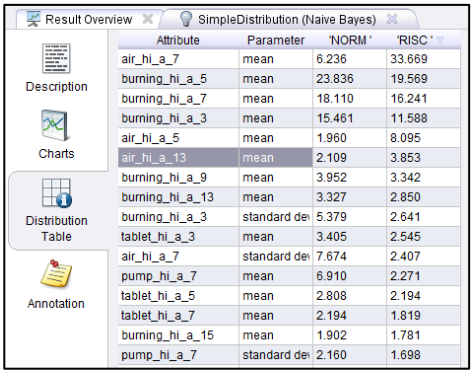
Objetivo de utilizar o *Rapid Miner* com o classificador Naïve Bayes teve o mesmo propósito da Árvore de Decisão, ou seja, verificar a sanidade das análises e dos resultados conforme apresentando nas Figuras 5.35 e 5.36.

Figura 5.35 – Naïve Bayes configurado no *Rapid Miner.*



Fonte: Tela capturada pelo autor (2018).

Figura 5.36 – Resultado do Naïve Bayes no *Rapid Miner.*



Attribute	Parameter	'NORM '	'RISC '
air_hi_a_7	mean	6.236	33.669
burning_hi_a_5	mean	23.836	19.569
burning_hi_a_7	mean	18.110	16.241
burning_hi_a_3	mean	15.461	11.588
air_hi_a_5	mean	1.960	8.095
air_hi_a_13	mean	2.109	3.853
burning_hi_a_9	mean	3.952	3.342
burning_hi_a_13	mean	3.327	2.850
burning_hi_a_3	standard de	5.379	2.641
tablet_hi_a_3	mean	3.405	2.545
air_hi_a_7	standard de	7.674	2.407
pump_hi_a_7	mean	6.910	2.271
tablet_hi_a_5	mean	2.808	2.194
tablet_hi_a_7	mean	2.194	1.819
burning_hi_a_15	mean	1.902	1.781
pump_hi_a_7	standard de	2.160	1.698

Fonte: Tela capturada pelo autor (2018).

Conforme evidenciado na Figura 5.36, o resultado aponta como a maior impactante a sétima componente harmônica das Centrais de Ar com 33,669%. Nessa análise é possível perceber que a segunda mais impactante é a quinta componente harmônica de corrente do ponto *Burning*, considerando os dados coletados durante a campanha de medição.

5.6 Validação dos Resultados

Com base nos dados consolidados de todos os pontos coletados na fase que antecedeu a mineração de dados, buscaram-se os valores mínimos e máximos

para as componentes harmônicas de corrente de terceira a décima quinta ordens e para cada carga conforme Tabela 5.13.

Tabela 5.13 - Valores mínimos e máximos considerando as correntes harmônicas ímpares (3ª a 15ª).

Pontos de Coleta	Qtde de registros	3ª componente harmônica de corrente		5ª componente harmônica de corrente		7ª componente harmônica de corrente		9ª componente harmônica de corrente		11ª componente harmônica de corrente		13ª componente harmônica de corrente		15ª componente harmônica de corrente	
		Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Burning	9966	2.69	31.74	16.50	36.18	13.90	24.03	0.90	7.09	0.09	4.10	2.09	5.78	0.57	3.40
Condicionadores de Ar	11350	0.30	6.60	0.20	9.40	1.60	38.70	0.20	1.70	0.20	3.40	0.60	5.30	0.50	1.80
Compressores de Ar	60940	0.00	6.40	0.00	12.10	0.10	14.80	0.00	0.80	0.00	1.00	0.00	0.40	0.00	0.30
Sala de Tablet	11350	1.62	5.98	1.31	4.77	1.13	3.48	0.00	2.23	0.00	1.18	0.00	0.00	0.00	0.00

Fonte: Elaborada pelo autor (2017).

Da Tabela 5.13, o valor mais alto (38,70) é o da sétima componente harmônica de corrente do ponto **Condicionadores de Ar**, o segundo maior valor (36,18) está na quinta componente harmônica de corrente do ponto ***Burning*** e o terceiro maior valor (31.74) está na terceira componente harmônica de corrente do ponto ***Burning***.

5.7 Discussão dos Resultados

Os dados encontrados com o processo *KDD* completo e especificamente na fase de mineração de dados demonstraram que o ponto **Condicionadores de Ar** afeta o *THD* de tensão no PAC, levando a uma situação de risco. A corrente de sétima componente harmônica é a principal contribuidora desse impacto e ocorre principalmente no turno T0, cobrindo o período de 1h10 a 6h da manhã, considerando os dados coletados durante o período da campanha de medição. Os resultados mostram a importância de usar o processo completo de *KDD* com técnicas de mineração de dados com classificadores para auxiliar na análise de impacto da QEE em fabricantes de computadores ou indústrias similares ou mesmo outros tipos de empresas. Este tipo de análise destaca a possibilidade para futuras penalidades em decorrência da nota técnica 0083/2012-SRD/ANEEL (2012) para consumidores que excedam os valores da distorção harmônica total de tensão no ponto de acoplamento comum.

Com base nos resultados encontrado foram feitas análise adicionais e constatou-se que a utilização de bancos de capacitores nos circuitos que alimentavam as centrais de ar eram provavelmente a causa dos resultados, desta forma foram sugeridos a aplicação de filtros para as centrais de ar e a utilização de equipamentos de sistema ininterrupto de energia com controle *online* para as linhas de produção de computadores.

6 CONCLUSÕES GERAIS E SUGESTÕES PARA TRABALHOS FUTUROS

O aumento do número de cargas não lineares devido à evolução industrial nos últimos anos impactou a QEE em consumidores de diferentes países. Devido aos impactos na *THDv*, particularmente nas indústrias, especialistas tem procurado diagnosticar e mitigar esses efeitos com o uso crescente de técnicas de Inteligência Computacional. Esta tese apresenta resultados coerentes que evidenciam que a análise do impacto das cargas no *THDv* no PAC de uma indústria usando o processo *KDD* completo é possível.

Nesta tese foi utilizada a metodologia baseada no processo *KDD* para análise e identificação do impacto na distorção harmônica (*THDv)* no PAC proporcionada pelas cargas principais instaladas na rede elétrica de um consumidor industrial. Esse processo foi descrito com detalhes compreendendo as etapas de escolha do consumidor industrial que possibilitasse a análise devido a utilização de cargas não lineares e desta forma o ambiente de pesquisa foi selecionado e na sequência a definição de quais pontos de cargas deveriam ser avaliados. Para operacionalização e coleta de dados selecionaram-se os medidores de qualidade de energia na quantidade necessária para a coleta simultânea nos cinco pontos de coleta. Com os dados coletados foram aplicadas e descritas todas as fases do processo *KDD* com especial atenção as técnicas de mineração de dados com o propósito de atingir os objetivos propostos e para contribuir com documentação no desenvolvimento de trabalhos futuros.

De acordo com o que foi apresentado na seção 4.2, a metodologia seguiu o processo *KDD* e foi necessário em diversas situações retroceder a etapa anterior para análise de cenários específicos. Inicialmente utilizou-se o classificador Árvore de Decisão como técnica de mineração de dados e com todos os dados coletados e foi obtido o primeiro resultado parcial apontando a sétima componente harmônica de corrente do ponto Ar Condicionado como item impactante no THDv.

Para aprofundamento, os dados foram separados por turnos de trabalho e novamente aplicado processo *KDD* e obteve novos resultados parciais onde os turnos T1 e T2 não apresentaram riscos e indicou que o impacto estava acontecendo no turno T0 com as mesmas caraterísticas dos dados gerais.

Com o propósito de aumentar a quantidade de instâncias forma removidos os dados o ponto Compressor de Ar e novamente foi aplicado o processo *KDD* com a Árvore de Decisão considerando os dados gerais e os dados por turnos e o resultado parcial foi o mesmo das análises anteriores.

Para validação dos resultados aplicou-se também a Árvore de Decisão com o *software Rapid Miner* e o resultado foi o mesmo.

Após os primeiros resultados parciais foi possível observar que seria necessário ter mais uma técnica de mineração de dados para análise, fazer o balanceamento das classes, também utilizar os dados separados em diferentes percentuais de treinamento e teste e outras métricas de desempenho para comparar com os resultados anteriores. Assim sendo foram utilizadas inicialmente diferentes análises com taxas de validação cruzada de 10, 50, 100 e 100 *folds* e com distribuição entre treinamento e teste com percentuais de 30/70, 50/50 e 70/30. Os resultados foram comparados e os resultados foram idênticos às análises anteriores. Aplicou-se também a técnica SMOTE com taxas de balanceamento de classes correspondente a 100%, 200%, 300%, 1000% e 2000% para equilibrar os dados nos diferentes cenários anteriores e os resultados foram também similares. Os resultados foram comparados com diversas métricas de desempenho validando os mesmos. Finalizando foi utilizado o classificador Naïve Bayes e foram aplicados nos mesmos cenários anterior e os resultados foram similares com os *Software WEKA* e *Rapid Miner*.

Em todas as análises o processo retrocedia para o início e eram novamente executadas as etapas do processo KDD e os resultados foram evidenciados por meio de tabelas com os dados nos diferentes cenários, com tabelas com as métricas de cada cenário e através de figuras coletadas de cada técnica e configurações diferentes.

Em todas as análises procurou-se comparar os resultados nos diferentes cenários e com dois *softwares* diferentes, quais sejam WEKA e *Rapid Miner,* para consolidar os resultados que apontaram como a corrente de maior impacto para levar a condição de risco como a sétima componente harmônica de corrente do ponto Centrais de Ar, considerando os dados coletados no período da campanha de medição.

A solução inovadora com a aplicação da mineração de dados através dos classificadores Árvore de Decisão e Naïve Bayes para análise e comparação dos resultados no processo KDD demonstra a eficácia das técnicas utilizadas. Os resultados obtidos pelas técnicas computacionais foram validados com diferentes resultados parciais e por meio da análise dos valores mínimo e máximo das componentes harmônicas de corrente de cada carga.

A tese conclui com relação ao objetivo principal que de acordo com os resultados apresentados, é evidente que o processo KDD apresenta aplicabilidade na análise do impacto na distorção harmônica total de tensão (THDv) no ponto de acoplamento comum (PAC) proporcionada pelas principais cargas instaladas na rede elétrica de uma indústria de computadores.

Com relação aos objetivos específicos, a tese deixa como contribuição a descrição de cada etapa do processo KDD na aplicação em um consumidor industrial com diferentes cenários e condições, o que proporciona a replicação em outros consumidores. Além disso, vem contribuir com o detalhamento na utilização dos classificadores Árvore de Decisão e Naïve Bayes aplicados em diferentes cenários pesquisados e considerando métricas, conjunto de treinamento e teste e técnica de balanceamento de classes. Os resultados foram satisfatórios para indicar que a carga mais impactante são as Centrais de Ar Condicionado com a sétima harmônica de corrente e em seguida a quinta harmônica de corrente, considerando os dados coletados no período da campanha de medição.

6.1 Propostas para trabalhos futuros

Como sugestões para trabalhos futuros têm-se:

- Aplicação da metodologia em outros tipos de consumidores, como por exemplo, na UFPA;
- Utilização de outros tipos de filtro para balanceamento das classes;
- Utilização de outros classificadores de mineração de dados como Redes Neurais e SVM, Árvore de Regressão, Amostragem por Importância ou técnicas de regressão;
- Desenvolvimento de sistema de monitoramento da QEE em tempo real.

REFERÊNCIAS

ABDELGAYED, T.; MORSI, W.; SIDHU, T. **A New Approach for Fault Classification in Microgrids Using Optimal Wavelet Functions Matching Pursuit**. IEEE Transactions on Smart Grid, 2017.

AFONSO, J.; MARTINS, J. **Qualidade da energia eléctrica**. 2005.

ALBERTO, B.; ALMEIDA, P. **Abordagens de pré-processamento de dados em problemas de classificaçao com classes desbalanceadas**. 2012. Tese de Doutorado. Master's Thesis, Centro Federal de Educação Tecnológica de Minas Gerais (Mestrado em Modelagem Matemática e Computacional).

ALBU, M.; SĂNDULEAC, M.; STĂNESCU, C. **Syncretic use of smart meters for Power Quality monitoring in emerging networks**. IEEE Transactions on Smart Grid, v. 8, n. 1, p. 485-492, 2017.

ANAR, M. **Partial Discharge In Electronic Equipments: Dissertation Thesis**. 2009. Tese de Doutorado. Brno University of Technology.

ANEEL. **Nota Técnica, n° 0083 /2012-SRD/ANEEL**. Processo: 48500.002798/2012-61.12 Junho 2012.

ANEEL. **Módulo 8 – Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional – PRODIST**. Revisão 8. 2016.

ARRILLAGA, J.; WATSON, N. **Power system harmonics**. John Wiley & Sons, 2004.

BAGGINI, A. **Handbook of power quality**. John Wiley & Sons, 2008.

BARANDELA, R. et al. **The imbalanced training sample problem: Under or over sampling?** In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer, Berlin, Heidelberg, 2004. p. 806-814.

BARROS, J. **Estudo de viabilidade econômica e das proteções da subestação de 69-13,8 kV do Campus do Pici da Universidade Federal do Ceará**. 2010. 174f. Monografia (conclusão do curso) - Universidade Federal do Ceará, Curso de Engenharia Elétrica, Fortaleza-CE.

BERRY, M.; LINOFF, G. **Data mining techniques: for marketing, sales, and customer support**. John Wiley & Sons, Inc., 1997.

BOLLEN, M.; GU, I. **Signal processing of power quality disturbances**. John Wiley & Sons, 2006.

BOLLEN, M. **Understanding Power Quality Problems: Voltage Sags andInterruptions***,* IEEE Press, ISBN: 0-7803-4713-7, 2000.

BOLLEN, M. et al*.* **"Trends, challenges and opportunities in power quality research"**, EUROPEAN, *Transactions on Electrical Power*, v. 20, pp. 3–18, 2009.

BORGES, F. et al. **Feature extraction and power quality disturbances classification using smart meters signals**. IEEE Transactions on Industrial Informatics, v. 12, n. 2, p. 824-833, 2016.

BORGES, F. **Extração de características combinadas com árvore de decisão para detecção e classificação dos distúrbios de qualidade da energia elétrica**. Tese de Doutorado. Universidade de São Paulo. 2013.

BREIMAN, L. et al. **Classification and Regression Trees**. Chapman and Hall/CRC, 1984.

BROWNLEE, J. **Tactics to combat imbalanced classes in your machine learning dataset.** Machine Learning Process, 2015.

CABENA, P. et al. **Discovering Data Mining: From Concept to Implementation**. Prentice Hall, 1998.

CAMILO, C.; SILVA, J. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas**. Universidade Federal de Goiás (UFC), p. 1-29, 2009.

CHAKRAVORTY, D. et al. **Impact of Modern Electronic Equipment on the Assessment of Network Harmonic Impedance**. IEEE Transactions on Smart Grid, v. 8, n. 1, p. 382-390, 2017.

CHANDRA, B.; VARGHESE, P. **On improving efficiency of sliq decision tree algorithm**. International Joint Conference on Neural Networks - IJCNN, p. 66–71, 2007.

CHANDRA, B.; VARGHESE, P. **Fuzzy sliq decision tree algorithm**. IEEE Transactions on Cybernetics, 38:1294–1301, 2008.

CHAWLA, N. et al. ***SMOTE: synthetic minority over-sampling technique***. Journal of artificial intelligence research, v. 16, p. 321-357, 2002.

CIUFU, L.; POPESCU, C.L.; POPESCU, M.O. **Experimental mitigation techniques to reduce the Total Harmonic Distortion of low voltage non-linear power sources**. In: Advanced Topics in Electrical Engineering (ATEE), 2017 10th International Symposium on. IEEE, 2017. p. 138-141.

COWELL, R. **Advanced inference in Bayesian networks, Learning in graphical models.** 1999.

DA SILVA, I. et al. **Recognition of Disturbances Related to Electric Power Quality Using MLP Networks**. In: Artificial Neural Networks. Springer International Publishing, 2017. p. 241-246.

DA SILVA PESSOA, A.; OLESKOVICZ, M. **Fault location in radial distribution systems based on decision trees and optimized allocation of power quality meters**. In: PowerTech, 2017 IEEE Manchester. IEEE, 2017. p. 1-6.

DE ALMEIDA, A.; MOREIRA, L.; DELGADO, J. **Power quality problems and new solutions. In: International conference on renewable energies and power quality**. 2003.

DE ARAÚJO, V. et al. **Dedicated hardware implementation of a high precision power quality meter**. In: Instrumentation and Measurement Technology Conference (I2MTC), 2015 IEEE International. IEEE, 2015. p. 393-398.

DECKMANN, S.; POMILIO, J. **Avaliação da qualidade da energia elétrica**. Campinas: UNICAMP, 2010.

DUGAN, R.; GRANAGHAN, M. **Electrical Power Systems Quality**, McGraw-Hill. ISBN 0-07-018031-8 1996, 2004.

ELHASSAN, T. et al. **Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method**. Journal of Informatics and Data Mining, v. 1, n. 2, 2016.

EN 50160, **Voltage characteristics of electricity supplied by public distribution systems**, 2011

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery in databases**. AI magazine, 1996. 17(3): p. 37.

FERNANDES, R. et al. **Identificação de cargas lineares e não-lineares em sistemas elétricos residenciais usando técnicas para seleção de atributos e redes neurais artificiais**. Sba: Controle & Automação Sociedade Brasileira de Automática, v. 21, n. 4, p. 389-405, 2010.

FERREIRA, D. **Análise de distúrbios elétricos em Sistemas de Potência**. Tese de Doutorado, COPPE, Universidade Federal do Riode Janeiro, Dezembro de 2010.

FERREIRA, V. et al. **A survey on intelligent system application to fault diagnosis in electric power system transmission lines**. Electric Power Systems Research, v. 136, p. 135-153, 2016.

F II, I. **IEEE recommended practices and requirements for harmonic control in electrical power systems**. 1993. Standard IEEE Std 519-1992.

GEHRKE, J.; RAMAKRISHNAN, R.; GANTI, V. **RainForest—a framework for fast decision tree construction of large datasets**. Data Mining and Knowledge Discovery, v. 4, n. 2, p. 127-162, 2000.

GEHRKE, J. et al. **BOAT—optimistic decision tree construction**. In: ACM SIGMOD Record. ACM, 1999. p. 169-180.

GOMES, A. **Detecção e Classificação de Faltas em Linhas de Transmissão utilizando Análise Funcional E Inteligência Computacional.** 2011. Dissertação de Mestrado. Universidade Federal de Minas Gerais.

GRANADOS-LIEBERMAN, D. et al. **Techniques and methodologies for power quality analysis and disturbances classification in power systems: a review**. IET Generation, Transmission & Distribution, v. 5, n. 4, p. 519-529, 2011.

GROSSMAN, D.; DOMINGOS, P. **Learning Bayesian maximizing conditional likelihood**. Proceedings on machine learning, p. 46-57, 2004.

GUERRERO, J. **Guest Editorial Special Issue on Power Quality in Smart Grids**. IEEE Transactions on Smart Grid, v. 8, n. 1, p. 379-381, 2017.

HALL, M.; FRANK, E. **Combining Naïve Bayes and Decision Tables**. In 2008 FLAIRS Conference - AAAI, 2008.

HALL, M. et al. **The WEKA data mining software: an update**. ACM SIGKDD explorations newsletter, v. 11, n. 1, p. 10-18, 2009.

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of Data Mining**. MIT Press, 2001.

HAND, D. **Construction and assessment of classification rules**. Wiley, 1997.

HART, P. **The condensed nearest neighbor rule (Corresp.)**. IEEE transactions on information theory, v. 14, n. 3, p. 515-516, 1968.

HASAN, A.; EBOULE, P.; TWALA, B. **The use of machine learning techniques to classify power transmission line fault types and locations**. In: Optimization of Electrical and Electronic Equipment (OPTIM) & 2017 Intl Aegean Conference on Electrical Machines and Power Electronics (ACEMP), 2017 International Conference on. IEEE, 2017. p. 221-226.

HEARN, S. **Basic Understanding Of Harmonics In Electrical Systems**". Hearn Engineering Tech. Report, 2010.

IAGAR, A.; POPA, G.; DINIS, C. **The influence of home nonlinear electric equipment operating modes on power quality**. WSEAS Transactions on Systems 13 357-367, 2014

IEC. **IEC-91- Draft Classification of Electromagnetic Environments**. Relatório técnico, IEC: TC77WG6 (Secretary) 110-R5, Janeiro 1991.

IEC 61000-2-2: **Electromagnetic compatibility — part 2-2, environment compatibility levels for low-frequency conducted disturbances and signaling in public and low voltage power supply systems**.

IEC 61000-2-4 **Electromagnetic compatibility (EMC) - Part 2-4: Environment - Compatibility levels in industrial plants for low-frequency conducted disturbances**.

IEC 61000-3-2. **Electromagnetic compatibility (EMC) - Part 3-2: Limits - Limits for harmonic current emissions (equipment input current <= 16 A per phase) (IEC 61000-3-2:2005 + A1:2008 + A2:2009)**; German version EN 61000-3-2:2006 + A1:2009 + A2:2009

IEC 61000-3-12 ed. 2 **Electromagnetic compatibility (EMC) - Part 3-12: Limits - Limits for harmonic currents produced by equipment connected to public low-voltage systems with input current >16 A and ≤ 75 A per phase.**

IEC 61000-4-7:2002 ed. 2.0 **Electromagnetic compatibility (EMC) – Part 4-7: Testing and measurement techniques –General guide on harmonics and interharmonics measurements and instrumentation, for power supply systems and equipment connected thereto**.

IEC 61000-4-30, 2003: **Power quality measurement methods**.

IEEE Std. 1159-1995. **IEEE Recommended Practice for Monitoring Electric Power Quality**. IEEE Standards Board, 1995.

IGNATOVA, V.; VILLARD, D.; HYPOLITE, J.M. **Simple indicators for an effective Power Quality monitoring and analysis**. In: Environment and Electrical Engineering (EEEIC), 2015 IEEE 15th International Conference on. IEEE, 2015. p. 1104-1108.

INTEL. **Electrostatic Discharge and Electrical Overstress Guide.** 2016

INTERNATIONAL DATA CORPORATION-IDC. **Connecting the IOT: The Road to Success.**
Disponível em: < https://www.idc.com/infographics/IoT/ATTACHMENTS/IoT.pdf>. Acessado em 03 de novembro de 2017.

KAHLE, K. **Power Converters and Power Quality**. arXiv preprint arXiv:1607.01556, 2016.

KAMENKA, A. **Six tough topics about harmonic distortion and Power Quality indices in electric power systems**. A White Paper of Schaffner Group, 2014.

KANDEV, N.; CHENARD, S. **Method for determining customer contribution to harmonic variations in a large power network**. In: Harmonics and Quality of Power (ICHQP), 2010 14th International Conference on. IEEE, 2010. p. 1-7.

KAZEROONI, M.; ZHU, H.; OVERBYE, T. **Literature review on the applications of data mining in power systems**. In: Power and Energy Conference at Illinois (PECI), 2014. IEEE, 2014. p. 1-8.

KHOKHAR, S. et al. **A new optimal feature selection algorithm for classification of power quality disturbances using discrete wavelet transform and probabilistic neural network**. Measurement, v. 95, p. 246-259, 2017.

KUBAT, M. et al. **Addressing the curse of imbalanced training sets: one-sided selection**. In: ICML. 1997. p. 179-186.

KUBAT, M.; HOLTE, R.; MATWIN, S. **Machine learning for the detection of oil spills in satellite radar images.** Machine learning, v. 30, n. 2-3, p. 195-215, 1998.

KUMAR, R. et al. **Recognition of power-quality disturbances using S-transform-based ANN classifier and rule-based decision tree**. IEEE Transactions on Industry Applications, v. 51, n. 2, p. 1249-1258, 2015.

LANGLEY, P. et al. **An analysis of Bayesian classifiers**. In: Aaai. 1992. p. 223-228.

LANGLEY, P.; SAGE, S. **Induction of selective Bayesian classifiers**. In: Proceedings of the Tenth international conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1994. p. 399-406.

LAROSE, D. **Discovering Knowledge in Data: An Introduction to Data Mining**. John Wiley and Sons, Inc, 2005.

LAURIKKALA, J. **Improving identification of difficult small classes by balancing class distribution**. In: Conference on Artificial Intelligence in Medicine in Europe. Springer, Berlin, Heidelberg, 2001. p. 63-66.

LEÃO, R.; SAMPAIO, R.; ANTUNES, F. **Harmônicos em Sistemas Elétricos**. Elsevier, 376 pg, 2014.

LEITE, J. et al., **Multicriteria design of passive harmonic filters for industrial installations using evolutionary computation techniques**. Journal of Engineering and Technology for Industrial Application - JETIA, 2015. 01(03): p. 55-60.

LEITE, J. **Projeto multicritério de filtros harmônicos passivos para instalações industriais utilizando técnicas de inteligência computacional**. 2013. Tese de Doutorado. Universidade Federal do Pará.

LIU, X.Y.; WU, J.; ZHOU, Z.H. **Exploratory undersampling for class-imbalance learning**. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), v. 39, n. 2, p. 539-550, 2009.

LOPES, R. **Estudos elétricos para expansão e melhoria da rede elétrica do Campus do Pici da UFC**. 2011. 181f. Monografia (conclusão do curso) - Universidade Federal do Ceará, Curso de Engenharia Elétrica, Fortaleza - CE.

LOURENÇO, T. **Avaliação da qualidade de energia elétrica no Centro de Tecnologia da Universidade Federal do Ceará**. 2012. Tese de Doutorado. Universidade Federal do Ceará.

MAHELA, O.; SHAIK, A. **Recognition of Power Quality Disturbances Using S-Transform Based Ruled Decision Tree and Fuzzy C-Means Clustering Classifiers**. Applied Soft Computing, 2017.

MARQUES, C. **Técnicas de processamento de sinais para a estimação da frequência e harmônicos de Sistemas Elétricos De Potência**. 2011. Tese de Doutorado. Universidade Federal do Rio de Janeiro.

MATOS, E. et al. **Using linear and non-parametric regression models to describe the contribution of non-linear loads on the voltage harmonic distortions in the electrical grid**. IET Generation, Transmission & Distribution, v. 10, n. 8, p. 1825-1832, 2016.

MATOS, E. et al. **Análise não paramétrica para identificação de fontes de distorções harmônicas em sistemas de energia elétrica: um estudo aplicado no campus universitário do Guamá da Universidade Federal do Pará**. 2016. Tese de Doutorado. Universidade Federal do Pará.

MEHTA, M.; AGRAWAL, R.; RISSANEN, J. **Sliq: A fast scalable classifier for data mining**. Procs. of the 5th EDBT, p. 18U˝ 32, 1996.

MELO, T.; FILHO, J.; CARDOSO, P. **Sistema de Gerenciamento Pessoal do Consumo de Energia Elétrica**. In: IEEE/IAS Induscon-IX International Conference on Industry Applications. 2010.

MICROSOFT. **Conjunto de dados de teste e treinamento.** Disponível em: <https://docs.microsoft.com/pt-br/sql/analysis-services/data-mining/training-and-testing-data-sets#creating-test-and-training-sets-for-data-mining-structures> Acessado em: 12 de abril de 2018.

MITCHELL, T. **Does machine learning really work?** AI magazine, v. 18, n. 3, p. 11, 1997.

MONTOYA, F. et al. **Power quality techniques research worldwide: A review**. Renewable and Sustainable Energy Reviews, v. 54, p. 846-856, 2016.

MOURA, C. **Estudo para implantação de um sistema de recomposição automática para a rede de distribuição do Campus do Pici**. 2010. 102f. Monografia (conclusão do curso) - Universidade Federal do Ceará, Curso de Engenharia Elétrica, Fortaleza - CE.

NAIK, C. et al. **Classification of power quality events using wavelet packet transform and extreme learning machine**. In: Power Electronics Conference (SPEC), IEEE Annual Southern. IEEE, 2016. p. 1-6.

NICKERSON, A.; JAPKOWICZ, N.; MILIOS, E. **Using Unsupervised Learning to Guide Resampling in Imbalanced Data Sets**. In: AISTATS. 2001.

NIEDERMAYER, D. **An introduction to Bayesian networks and their contemporary applications**. Innovations in Bayesian Networks, p. 117-130, 2008.

NOGUEIRA, R. et al. **Harmonic Impact analysis coming from the manufacturing processes of a Eletroeletronic Industry Using KDD and Decision Trees**. Journal of Engineering and Technology for Industrial Applications - JETIA, Manaus-Amazonas, p. 80 - 87, 10 mar. 2015.

NOGUEIRA, R. et al. **Análisis de los Impactos Armónicos en la Industria de la Electrónica utilizando Árboles De Decisión**. CUBA: CIIC - CUJAE, 2014 (Trabalhos completos publicados em anais de congressos).

NOGUEIRA, R. et al. **Análise dos impactos harmônicos em uma indústria de manufatura de eletroeletrônicos utilizando árvores de decisão**. 2015. Dissertação de Mestrado. Universidade Federal do Pará.

OLIVEIRA, C. et al. **Estudo de caso de eficiência energética e qualidade de energia elétrica**. 6p. Anais do VIII Induscon Conferência Internacional de Aplicações Industriais, Poços de Caldas/MG, 2008.

OLIVEIRA, E. et al. **Voltage THD Analysis Using Knowledge Discovery in Databases with a Decision Tree Classifier**. IEEE Access, v. 6, p. 1177-1188, 2017.

OUBRAHIM, Z. et al. **Disturbances Classification based on a Model Order Selection Method for Power Quality Monitoring**. IEEE Transactions on Industrial Electronics, 2017.

OZGONENEL, O.; THOMAS, D.; YALCIN, T. **Superiority of decision tree classifier on complicated cases for power system protection**. 2012.

PEEK, F. ***Voltage and current harmonics caused by corona***. Journal of the American Institute of Electrical Engineers, v. 40, n. 6, p. 455-461, 1921.

PIATETSKY-SHAPIRO, G.; FRAWLEY, W. **Knowledge Discovery in Databases**. Cambridge, Mass.: AAAI. 1991.

PRASS, F. **KDD–Uma visão geral do processo**. 2009.

PRATI, R.; BATISTA, G.; MONARD, M. **Uma experiência no balanceamento artificial de conjuntos de dados para aprendizado com classes desbalanceadas utilizando análise ROC.** In: Proc. of the Workshop on Advances & Trends in AI for Problem Solving. 2003. p. 28-33.

PYLE, D. **Data preparation for data mining**. Morgan Kaufmann Publishers Inc., 1999.

QUINLAN, J. **C4.5: Programs for Machine Learning.** Morgan Kaufmann Publishers Inc., 1992.

RAHMATIAN, M. et al. **Transient stability assessment via decision trees and multivariate adaptive regression splines**. Electric Power Systems Research, v. 142, p. 320-328, 2017.

RAY, P. et al. **Optimal feature and decision tree-based classification of power quality disturbances in distributed generation systems**. IEEE Transactions on Sustainable Energy, v. 5, n. 1, p. 200-208, 2014.

RIBEIRO, T.; ROCHA, J. **Aplicação de Filtro de Harmônicos em Indústria: um Estudo de Caso**. IV Simpósio Brasileiro de Sistemas Elétricos, Goiânia-GO. 2012

DUDA, R.; HART, P.; STORK, D. **Pattern classification and scene analysis**. 2nd. Edition. 1995.

RODRÍGUEZ, A. et al. **A Decision Tree and S-transform based approach for power quality disturbances classification**. In: Power Engineering, Energy and Electrical Drives (POWERENG), 2013 Fourth International Conference on. IEEE, 2013. p. 1093-1097.

RODRIGUEZ-GUERRERO, M. et al. **A novel methodology for modeling waveforms for power quality disturbance analysis**. Electric Power Systems Research, v. 143, p. 14-24, 2017.

ROIGER, R. **Data mining: a tutorial-based primer**. CRC Press, 2017.

SAMANTARAY, S. **Decision tree-initialised fuzzy rule-based approach for power quality events classification**. IET generation, transmission & distribution, v. 4, n. 4, p. 538-551, 2010.

SANTOS, I.; DE OLIVEIRA, J. **Critical analysis of the current and voltage superposition approaches at sharing harmonic distortion responsibility**. IEEE Latin America Transactions, v. 9, n. 4, p. 516-521, 2011.

SARMANHO, U. **Influência dos distúrbios elétricos em média tensão na qualidade de energia – Estudo em um ambiente universitário**. Dissertação. PUC RS, Porto Alegre, Agosto de 2005.

SCHNEIDER-ELECTRIC. **Qualidade de Energia-Harmônicas**. in Workshop de Instalações Eléctricas de Baixa Tensão. 2005.

SEERA, M. et al. **Power Quality Analysis Using a Hybrid Model of the Fuzzy Min–Max Neural Network and Clustering Tree**. IEEE transactions on neural networks and learning systems, v. 27, n. 12, p. 2760-2767, 2016.

SHAFER, J.; AGRAWAL, R.; MEHTA, M. ***Sprint: A scalable parallel classifier for data mining***. Procs. of the 22nd VLDB, p. 544U˝ 555, 1996.

SILVA, M. **Modelação e análise da vida útil (metrológica) de medidores tipo indução de energia elétrica ativa**. 2010. Dissertação de Mestrado. Universidade Estadual Paulista “Júlio de Mesquita Filho” – SP.

SILVA, W. et al. ***Naïve Bayes Aplicados na Análise de Impactos Harmônicos em Sistemas Elétricos Industriais***. 2014 (Trabalhos completos publicados em anais de congressos).

SOUSA, V. et al. **Analysis of harmonic distortion generated by PWM motor drives.** In: Power Electronics and Power Quality Applications (PEPQA), 2017 IEEE Workshop on. IEEE, 2017. p. 1-6.

SOUZA, A. **Inteligência computacional aplicada na identificação e classificação de problemas de medição de energia elétrica**. in Brazilian Conference on Dynamics Control and Their Applicatioons. 2010.

SUN, Y. et al. **Cost-sensitive boosting for classification of imbalanced data**. Pattern Recognition, v. 40, n. 12, p. 3358-3378, 2007.

TAN, P.N.; STEINBACH, M.; KUMAR, Vi. **Introdução ao datamining: mineração de dados**. Ciência Moderna, 2009.

TARASIUK, T. **Estimator-analyzer of power quality. Part I. Methods and algorithms**. Measurement 44(1) 238–247, 2011

TITTERINGTON, M. et al. **Comparison of discrimination techniques applied to a complex data set of head injured patients.** Journal of the Royal Statistical Society. Series A (General), p. 145-175, 1981.

TOMEK, I. **Two modifications of CNN**. IEEE Trans. Systems, Man and Cybernetics, v. 6, p. 769-772, 1976.

TOSTES, M. **Avaliação dos impactos causados pela geração de harmônicos na rede de distribuição em Consumidores em baixa tensão**. Tese de Doutorado do Programa de Pós-Graduação em Engenharia Elétrica do Instituto de Tecnologia da Universidade Federal do Pará (ITEC-UFPA). 2003.

UNSAR, O. et al. **Identification of harmonic current contributions of iron and steel plants based on time-synchronized field measurements—Part I: At PCC**. IEEE Transactions on Industry Applications, v. 50, n. 6, p. 4336-4347, 2014.

UNSAR, O. et al. **Identification of Harmonic Current Contributions of Iron and Steel Plants Based on Time-Synchronized Field Measurements—Part II: Inside Plants.** IEEE Transactions on Industry Applications, v. 6, n. 50, p. 4348-4355, 2014.

VAID, K.; SRIKANTH, P.; SOOD, Y. **Critical impedance based automatic identification of harmonic sources in deregulated power industry**. In: Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), 2011 International Conference on. IEEE, 2011. p. 653-658.

VALE, J. **Projeto da subestação 69/13,8 kV da UFC-Campus do Pici**. 2011. 207f. Monografia (conclusão do curso) - Universidade Federal do Ceará, Curso de Engenharia Elétrica, Fortaleza - CE.

VEIGA, S.; DA SILVA, W. **Redes Bayesianas: Uma Visão Geral**. 2002.

VLAHINIC, S.; BRNOBIC, D.; STOJKOVIC, N. **Indices for harmonic distortion monitoring of power distribution systems**. IEEE Transactions on Instrumentation and Measurement 58 (5) 1771–1777, 2009.

WARNER, H. et al. **A mathematical approach to medical diagnosis: application to congenital heart disease**. Jama, v. 177, n. 3, p. 177-183, 1961.

WILSON, D. **Asymptotic properties of nearest neighbor rules using edited data**. IEEE Transactions on Systems, Man, and Cybernetics, n. 3, p. 408-421, 1972.

XU, L. et al. **Information security in big data: privacy and data mining**. IEEE Access, v. 2, p. 1149-1176, 2014.

YIN, Z.; SUN, Y.; YU, T. **New methods exploration for harmonic source identification technologies**. In: Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), 2011 4th International Conference on. IEEE, 2011. p. 399-402.

ZHANG, H. **The optimality of Naïve Bayes**. In 2004 FLAIRS Conference - AAAI, 2004.

ZHU, F. et al. **Mining colossal frequent patterns by core pattern fusion**. IEEE 23rd International Conference on Data Engineering, 2007.