**ANDERSON ROGES TEIXEIRA GÓES**

# UMA METODOLOGIA PARA A CRIAÇÃO DE ETIQUETA DE QUALIDADE NO CONTEXTO DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS: APLICAÇÃO NAS ÁREAS ELÉTRICA E EDUCACIONAL

Tese apresentada ao Programa de Pós-Graduação em Métodos Numéricos em Engenharia – área de concentração: Programação Matemática - dos setores de Tecnologia e de Ciências Exatas da Universidade Federal do Paraná, como requisito parcial à obtenção do grau de Doutor.

**Orientadora:** Profª. Dra. Maria Teresinha Arns Steiner.

CURITIBA

2012

G598u Góes, Anderson Roges Teixeira.
Uma metodologia para a criação de etiqueta de qualidade no contexto de descoberta de conhecimento em bases de dados: aplicação nas áreas elétrica e educacional / Anderson Roges Teixeira Góes – Curitiba, 2012.
145 p.: Il.

Orientadora: Profª. Dra. Maria Teresinha Arns Steiner.
Tese (Doutorado – Programa de Pós-Graduação em Métodos Numéricos em Engenharia) - Setor de Tecnologia e Setor de Ciências Exatas, Universidade Federal do Paraná.
Inclui Bibliografia.

1. Descoberta de Conhecimento em Bases de Dados. 2. Etiqueta de Qualidade. 3. Qualidade da Energia Elétrica. 4. Qualidade Educacional.
I. Título. II. Steiner, Maria Teresinha Arns. III. Universidade Federal do Paraná.

CDD 519.1

## TERMO DE APROVAÇÃO

ANDERSON ROGES TEIXEIRA GÓES

**UMA METODOLOGIA PARA A CRIAÇÃO DE ETIQUETA DE QUALIDADE NO CONTEXTO DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS: APLICAÇÃO NAS ÁREAS ELÉTRICA E EDUCACIONAL**

Tese aprovada como requisito parcial para obtenção do grau de Doutor, no Programa de Pós-Graduação em Métodos Numéricos em Engenharia – área de concentração: Programação Matemática - da Universidade Federal do Paraná, pela seguinte banca examinadora:

**Prof.ª Dr.ª Maria Teresinha Arns Steiner**
***Orientadora***
*Programa de Pós-Graduação em Métodos Numéricos em Engenharia, UFPR*
*Programa de Pós-Graduação em Engenharia de Produção e Sistemas, PUC/PR*

**Prof.ª Dr.ª Deise Maria Bertholdi Costa**
*Programa de Pós-Graduação em Métodos Numéricos em Engenharia, UFPR.*

**Prof.ª Dr.ª Maria Tereza Carneiro Soares**
*Programa de Pós-Graduação em Educação, UFPR.*

**Prof. Dr. Júlio Cesar Nievola**
*Programa de Pós-Graduação em Informática Aplicada, PUC/PR*

**Prof. Dr. Rodrigo Jardim Riella**
*Curso de Pós-Graduação em Desenvolvimento de Tecnologia e LACTEC*

Curitiba, 19 de junho de 2012.

*Dedico a Deus este trabalho,*
*pois tudo que acontece é vontade Dele;*

*à Maria, mãe de Deus, a quem tantas vezes*
*recorri pedindo sua intercessão;*

*aos meus pais,*
*Jorge Teixeira Góes e Ides Prado Góes,*
*que sempre acreditam e me apóiam*
*a conseguir tudo o que almejo;*

*à minha esposa,*
*Heliza Colaço Góes,*
*minha eterna namorada.*

# AGRADECIMENTOS

Tantos são os agradecimentos a fazer, que escrevê-los em algumas linhas é impossível, pois nestes anos de vida acadêmica e profissional conheci muitas pessoas e algumas deixaram lições de perspicácia, de humildade, de garra e de vontade de fazer sempre o melhor. Lições e exemplos para seguir e lembrar.

Ouso citar algumas destas pessoas que nestes últimos anos estiveram sempre ao meu lado.

Agradeço toda orientação que a professora Dra. Maria Teresinha Arns Steiner me proporcionou, sempre me direcionando para a busca do conhecimento. Uma pessoa extraordinária que não deixou que seus títulos acadêmicos e reconhecimentos no meio científico ofuscassem sua humildade. Serei eternamente grato pela "aula" de orientação.

Agradeço aos professores Dr. Júlio Cesar Nievola e Dr. Rodrigo Jardim Riella pelas valiosas contribuições para a aprimoração deste trabalho.

Agradeço à professora Dra. Deise Maria Bertholdi Costa, também membro desta banca. Ressalto que suas orientações no mestrado me fizeram chegar à conquista deste título de "Doutor".

Também agradeço a professora Dra. Maria Tereza Carneiro Soares, que desde a graduação admiro. Sua presença na banca de defesa desta tese mostrou que o trabalho que desenvolvi pode ser apreciado por pesquisadores de outras áreas distintas.

Agradeço a todos os professores com quem tive contato, seja em sala de aula ou numa simples conversa sobre a pesquisa. Tenho certeza que aproveitei cada momento para a consolidação deste trabalho.

Não posso deixar de agradecer a Maristela Bandil e Jucilei do Carmo Estradioto Reinhardt, que compõem o quadro de servidores públicos desta Universidade, pois o carisma, a competência e a dedicação à profissão que vocês possuem é o que se espera de todos nesta instituição.

Dentre meus agradecimentos não posso esquecer os familiares que sempre estiveram ao meu lado, em especial, meus pais, que sempre sonharam em ter um "doutor" na família, hoje realizo nosso sonho, pois já sou "Doutor"; meus tios, Eloi e Dulce, que me acolheram quando sai do interior do estado e vim à capital cursar licenciatura em matemática, o resultado deste gesto é este trabalho que concluo; e a minha esposa Heliza, pois quando eu tinha desistido de terminar esta pesquisa, ela soube dizer as palavras certas que me fizeram repensar tal atitude.

Agradeço aos meus alunos, seja do Ensino Fundamental ou do Ensino Superior, pois com o objetivo de sempre me aprimorar para proporcionar a vocês uma qualidade de ensino cada vez melhor é que almejei este título.

Agradeço aos meus ex-colegas de trabalho, que não arrisco citar nomes, pois foram muitos e me orgulho de ter trabalhado com eles no Colégio Social Madre Clélia (Curitiba, Pr), Escola João Paulo (Araucária, Pr), Faculdade Pilares (atual FAE São José dos Pinhais, Pr), Colégio Bagozzi (Curitiba, Pr) e Centro Integrado de Educação de Jovens e Adultos (Curitiba, Pr). E aos que hoje me incentivam tanto no Município de Araucária, onde sou professor do Ensino Fundamental, quanto na Universidade Federal do Paraná – Departamento de Expressão Gráfica.

Por fim, todos que passaram por mim sintam-se agradecidos e tenham certeza que saberei retribuir à sociedade tudo que o Ensino Público me proporcionou desde a Pré-escola à obtenção deste título.

A todos meus sinceros agradecimentos.

*"Quando te disserem que não pode perguntar mais nada e*
*que faz bem ficar sempre de cabeça baixa,*
*diante desta hipocrisia, acredite que exista,*
*não um muro, mas um futuro para você.*
*Confie em mim, eu também sofri quando*
*com coragem eu vi o mundo à minha maneira."*
*Laura Pausini*

## RESUMO

A busca pela qualidade é muito discutida, existem muitos trabalhos sobre o assunto aplicado às mais diversas áreas de conhecimento (ciências da terra, saúde, ciências da informação, educação, engenharia elétrica, dentre outras), mas parece não haver técnicas que explorem registros contidos em bases de dados com a finalidade de obter a classificação da qualidade. Diante deste contexto, nesta tese é proposta uma metodologia para a criação de etiqueta de qualidade, utilizando o processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - *KDD*), com a finalidade de classificar, com base em registros históricos, a qualidade de elementos de uma mesma região e/ou grupo, preenchendo assim, a referida lacuna. Para mostrar que tal metodologia pode ser aplicada a problemas de diversas áreas, foram aqui realizados dois estudos de caso, um na área elétrica e outro na área educacional. No primeiro, os dados para a criação de etiqueta de Qualidade de Energia Elétrica são provenientes de duas bases de dados de uma concessionária, onde verificamos a qualidade dos alimentadores de uma determinada subestação, levando em consideração a magnitude do afundamento, sua duração e sua freqüência (quantidade de ocorrências no decorrer de certo período). Já no segundo, os dados para a criação de etiqueta de Qualidade Educacional foram obtidos junto às escolas municipais da cidade de Araucária/Pr. Tais dados são originados da avaliação Prova Brasil que compõe o Índice de Desenvolvimento da Educação Básica (IDEB) - notas das provas de língua portuguesa e matemática dos anos iniciais e finais do Ensino Fundamental. Na busca pela classificação da qualidade, principal etapa da criação da etiqueta, foram aqui utilizadas as metaheurísticas Redes Neurais, *Support Vector Machine* e Algoritmos Genéticos, além de heurísticas baseadas em distâncias euclidiana e de Malahanobis. A versatilidade para a obtenção da etiqueta foi constatada pela aplicação da mesma a dois estudos em áreas totalmente distintas, já que em tal metodologia é possível construir um quadro para todos os elementos (independente da área de aplicação) onde se queira classificar a qualidade através das classes. Já a simplicidade da etiqueta pode ser verificada visualmente, já que a mesma ficou definida através de uma escala com seis níveis (“A” a “F”), com diferentes cores. Por fim conclui-se que o processo *KDD* é fundamental nesta metodologia que ainda tem muito a ser explorada, através da utilização de outras técnicas e de outras aplicações.



**Palavras-chave:** Descoberta de Conhecimento em Bases de Dados, Etiqueta de Qualidade, Qualidade da Energia Elétrica, Qualidade Educacional.

## ABSTRACT

The search for quality is much discussed today, there are many papers on this topic, applied to various areas of knowledge (earth science, health, information sciences, education, electrical engineering, among others), but it seems does not have techniques which explore records contained in databases in order to get the classification of the quality. In this context, this research proposes a methodology for creating quality label, using the process of Knowledge Discovery in Databases (KDD) in order to classify, based on historical records contained in data bases, the quality elements of the same region and/or group, filling in thereby, such lacuna. To show that this methodology can be applied to problems of various areas, we realized two case studies, one in the electrical area and the other in the education area. In the first one, the data to create the Electric Power Quality label, come from two databases from a electrical company, where we verify the quality of the feeders of a particular substation, taking into account the magnitude sag, its duration and its frequency (amount of events during a certain period). In the second one, the data for the creation of Educational Quality label were obtained from the schools of the city of Araucaria/Pr. These data were originated from the Prova Brasil evaluation which forms the Índice de Desenvolvimento da Educação Básica (IDEB or Index of Development of Basic Education) - test scores of Portuguese language and mathematics of the initial and final year of elementary school. In the search for classification in the quality, the main step of label creation, it was used the metaheuristics Neural Networks, Support Vector Machine and Genetic Algorithms, and the heuristics based on Euclidean and Malahanobis distances. The versatility to obtain the label was verified by applying it in two studies in areas totally different, since in this methodology it is possible to construct a framework for all elements (independent of the application area) where want to classify the quality through the classes. The simplicity of the label can be checked visually, since it was defined using a scale with six levels ("A" through "F") with different colors. Finally we can conclude that the KDD process is fundamental in this methodology that still has much to be explored through the use of other techniques and other applications.



**Keywords:** Knowledge Discovery in Databases, Quality Label, Power Quality, Education Quality.

# LISTA DE FIGURAS

# LISTA DE QUADROS

# LISTA DE TABELAS

# LISTA DE SIGLAS E ABREVIAÇÕES

| | | |
|---|---|---|
| AG | - | Algoritmos Genéticos ou *Genetic Algorithms.* |
| ANEEL | - | Agência Nacional de Energia Elétrica. |
| CAQI | - | Custo-Aluno Qualidade Inicial. |
| CEB | - | Câmara de Educação Básica. |
| CFB | - | Constituição Federal Brasileira. |
| *DM* | - | *Data Mining* ou Mineração de dados. |
| EQ | - | Etiqueta de Qualidade. |
| IDEB | - | Índice de Desenvolvimento da Educação Básica. |
| INEP | - | Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. |
| *KDD* | - | *Knowledge Discovery in Databases* ou Descoberta de Conhecimento em Bases de dados. |
| MEC | - | Ministério da Educaççção. |
| PPGMNE | - | Programa de Pós-Graduação em Métodos Numéricos. |
| QEd | - | Qualidade Educacional. |
| QEE | - | Qualidade de Energia Elétrica. |
| RNA | - | Redes Neurais ou *Neural Network.* |
| *SVM* | - | *Support Vector Machine* ou Máquinas de Vetores Suporte. |
| UFPR | - | Universidade Federal do Paraná. |

## SUMÁRIO

# 1 INTRODUÇÃO

Esta pesquisa surge de projeto registrado na Agência Nacional de Energia Elétrica (ANEEL), firmado entre a Universidade Federal do Paraná (UFPR), através do Programa de Pós-Graduação em Métodos Numéricos (PPGMNE), e uma concessionária de energia elétrica brasileira, com o objetivo principal de classificar a Qualidade de Energia Elétrica (QEE) em relação aos afundamentos de tensão ocorridos na rede de distribuição de tal empresa.

Na revisão de literatura realizada para atingir o objetivo do projeto, muitos trabalhos sobre QEE foram encontrados, em sua maioria estes tratavam de detectar ou classificar os eventos/distúrbios relacionados à QEE e não a QEE propriamente dita. No entanto, foram encontrados dois trabalhos (Casteren *et al.*, 2005; Cobben e Casteren, 2006) que tratavam da classificação da QEE através de etiqueta de qualidade, mas estes se limitavam a mostrar aspectos iniciais, não apresentando técnicas para sua efetivação e não realizaram estudos de casos, ou seja, os dados apresentados são fictícios.

Com isso, existem algumas questões no trabalho desses autores que parecem não ter respostas na literatura pesquisada: Como utilizar dados reais para a criação da etiqueta de qualidade? Como definir uma qualidade “normal” com base em dados reais? Como classificar um elemento na etiqueta de qualidade, que não se enquadra em nenhuma das faixas de classificação?

Destas perguntas surge esta tese com o objetivo de apresentar uma metodologia para a criação de etiqueta de qualidade, de forma comparativa, tendo em seu contexto o processo Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* ou *KDD*).

Apesar dos trabalhos mencionados serem da área de energia elétrica, a metodologia aqui proposta é versátil, podendo ser aplicada às mais diversas áreas, sendo isso comprovado por meio dos dois estudos de casos realizados: o primeiro na área elétrica e o segundo na área de educação.

A etiqueta aqui proposta é denominada de “forma comparativa”, pois indica em um grupo/região a qualidade dos elementos em uma escala de seis faixas (A, B, C, D, E e F), através da comparação destes.

Para determinar a classificação de cada elemento na etiqueta de qualidade, são realizados testes com diversas técnicas da Pesquisa Operacional, desde as

mais sofisticadas, como Redes Neurais Artificiais (RNA), Máquina de Vetores de Suporte (*Support Vector Machine* ou *SVM*) e Algoritmos Genéticos (AG), às mais simples, que envolvem distâncias, como a euclidiana e a de Malahanobis. Ao final desta pesquisa é indicada a técnica com resultado mais próximo à *moda estatística* das soluções encontradas nos estudos de casos, sendo esta a técnica para compor a metodologia em aplicações futuras.

Assim, esta tese se destaca pelo fato de apresentar uma metodologia para a criação de etiqueta de qualidade, de forma comparativa, explorando bases de dados, que pode ser aplicada às diversas áreas, o que parece não haver na literatura, caracterizando, desta forma, uma contribuição inédita.

## 1.1 OBJETIVOS DO TRABALHO

Para o desenvolvimento desta pesquisa foram traçados alguns objetivos.

### 1.1.1 Objetivo Geral

Desenvolver uma metodologia para criação de etiqueta de qualidade, de forma comparativa, que possa ser aplicada a diversas áreas, tendo em seu contexto o processo *KDD*.

### 1.1.2 Objetivos Específicos

Como objetivos específicos, podem-se listar os seguintes:

- Utilizar o processo *KDD* como suporte à criação da etiqueta de qualidade, de forma comparativa, para explorar as informações provenientes de bases de dados;
- Aplicar diversas técnicas da Pesquisa Operacional, compará-las com a finalidade de indicar a que apresenta a solução mais próxima da obtida pela “classificação por voto” (moda estatística) entre todas as técnicas aplicadas aos estudos de caso para a criação de tal etiqueta de qualidade; e

- Aplicar a metodologia a duas bases de dados de áreas diferentes, com a finalidade de mostrar a sua versatilidade.

## 1.2 JUSTIFICATIVA

As empresas dos mais diversos ramos possuem informações armazenadas em bases de dados que, muitas vezes, vão simplesmente sendo guardadas sem nenhuma pretensão de uso, ou quando há tal pretensão, seus especialistas não sabem como explorá-los de forma a obter conhecimento.

Por exemplo, na área elétrica existem equipamentos para medir e, consequentemente, gerar informações a respeito da QEE, mas não exploram as informações captadas por equipamentos. Já na área educacional há instrumentos de avaliação do Governo Federal que são utilizados, geralmente, para a definição de algum índice, mas que podem ser administrados para gerar outras informações aos Estados, Municípios e toda comunidade escolar.

Assim, uma metodologia para criação de etiqueta de qualidade, de forma comparativa, de fácil visualização e que tenha em seu contexto o processo *KDD,* processo este que busca descobrir conhecimento em bases de dados, tornará estes dados úteis às mais diversas áreas. Além disso, com a etiqueta de qualidade é possível fornecer padrões de comparação, verificar se o serviço/produto foi ofertado de forma a contento tanto para o consumidor quanto ao fornecedor.

## 1.3 LIMITAÇÃO DO TRABALHO

Esta tese se limita a propor uma metodologia para criação de etiqueta de qualidade, de forma comparativa, no contexto *KDD* e a indicar uma das técnicas para aplicações futuras da metodologia, por meio dos estudos de casos.

Dessa forma, o trabalho não pretende definir índices de qualidade absolutos, mas sim relativos, já que se estará trabalhando de forma comparativa considerando uma região ou grupo.

# 2 REVISÃO DE LITERATURA

Neste capítulo são apresentadas definições de qualidade e trabalhos presentes na literatura a fim de verificar métodos que são aplicados na busca pela indicação de qualidade em serviços e produtos, bem como estes se apresentam.

## 2.1 O QUE É QUALIDADE

Segundo Paladini (1995), na pré-história o homem já buscava a qualidade embora não fosse claro seu significado, desde então ela é percebida nas diversas áreas do conhecimento.

Sua definição pode ter muitos significados e depende de onde seu uso é empregado, pois para cada conceito existem vários níveis de abstração, sendo assim, não tem um único sentido. No entanto, partindo da etimologia da palavra "qualidade", sua origem vem do latim *qualitas* e significa "de que natureza". Já seu significado na língua portuguesa é "algo que o distingue de outras coisas similares".

> "**1.** Propriedade, atributo ou condição das coisas ou das pessoas que as distingue das outras e lhes determina a natureza. **2.** Superioridade, excelência de alguém ou algo. **3.** Dote, virtude. **4.** Condição social, civil, jurídica, etc.; casta, laia." (FERREIRA, 2001)

Por possuir vários sentidos e significados, cinco abordagens são propostas por Garvin (1992), e estas englobam todos os sentidos de qualidade: transcendental; baseada em produto; baseada na produção; baseada no usuário; e baseada no valor.

Na *abordagem transcendental*, a qualidade é considerada como inata, ou seja, não se pode medir ou definir com precisão, é algo que existe ou não existe e é reconhecida pela experiência. Um caso deste tipo de qualidade é a atribuída a relógios da marca Rolex, por exemplo, em que ao apenas ouvir o nome desta marca todos sabem que são produtos de alta qualidade.

Na *abordagem baseada em produto*, a qualidade é mensurada pela quantidade de características que este possui, ou seja, quanto mais atributos, maior será sua qualidade. Um exemplo aparece na escolha de um carro novo, pois ao

comparar dois carros com mesmas características, diferenciando apenas pelo fato de um ter ar-condicionado e outro não, o que possui tal item terá maior qualidade.

Na *abordagem baseada na produção*, a qualidade é atribuída às características do produto que estão em conformidade com as especificações, ou seja, livre de erros. Pode-se citar como exemplo a produção de camisetas bordadas com logo de empresa. Todos os logos estão na posição correta? Quanto mais camisetas com o logo na posição correta, maior será a qualidade da produção.

Na *abordagem baseada no usuário*, a qualidade é verificada se o produto ou serviço fornecido está adequado ao que se propõe. Nesta abordagem ela é subjetiva, pois a avaliação dos usuários em relação às especificações são os padrões próprios à qualidade. Exemplos: O ensino nas escolas está atendendo as necessidades dos alunos e da sociedade? A energia elétrica distribuída à residência do consumidor não causa desconforto, por exemplo, com interrupções?

Por fim, na *abordagem baseada no valor*, a qualidade é entendida como a relação entre o preço e seu uso/custo, ou seja, o preço que o usuário/consumidor está disposto a pagar pelo serviço/produto. Um exemplo pode ser vivenciado na programação de uma viagem: considerando uma mesma localização, o usuário pode se hospedar em um hotel com mais ou menos “estrelas” em sua classificação.

E como já descrito anteriormente, as definições de Garvin (1992) mostram que não existe uma única “verdade” sobre qualidade, até pelo fato que uma ou mais abordagens deste autor podem coexistir no mesmo cenário. No entanto, percebe-se que este autor consegue abranger todas as definições.

## 2.2 TRABALHOS PRESENTES NA LITERATURA SOBRE QUALIDADE

Na literatura há muitos trabalhos relacionados à qualidade, sendo que estes estão presentes nas mais diversas áreas e alguns são expostos a seguir.

Em sua grande maioria as pesquisas estão relacionadas à qualidade do solo, as quais analisam indicadores de qualidade física, química e microbiológicas, quase sempre em relação ao manejo de diversos cultivos como, por exemplo, laranjeiras (Fidalski, Tormena e Scapim, 2007), bananeiras (Fialho *et al.*, 2006) e eucalipto (Chaer e Tótola, 2007), utilizando de técnicas estatísticas, mais especificamente da análise multivariada.

Distinto dos trabalhos acima mencionados, ainda na área de ciências da terra, Deponti, Eckert e Azambuja (2002) propõem metodologia para criação de indicadores para avaliar a sustentabilidade de diferentes sistemas: técnico, econômico, ambiental e social. Estes autores destacam que a metodologia não aponta uma fórmula (ou técnica) para definir um número. Afirmam, sim, que "a construção de indicadores para avaliação da sustentabilidade é um trabalho que exige uma equipe interdisciplinar, pois não há uma fórmula pronta, é necessário análise, interpretação e compreensão por parte dos envolvidos".

Trabalho semelhante é desenvolvido na área de saúde por Bittar (2001), enfatizando no planejamento, organização, coordenação e controle das atividades desenvolvidas. A metodologia para criação de parâmetros surge de comparações entre metas, fatos, dados e informações que, segundo o autor, são elementos fundamentais para se conhecer as mudanças ocorridas em uma instituição de saúde.

Já Hartz *et al.* (1996) desenvolvem metodologia para calcular "índices de mortes evitáveis", através de técnica que compara a mortalidade infantil observada com a esperada na base populacional analisada, gerando assim um percentual.

Kurcgant, Tronchin e Melleiro (2006) apresentam indicadores de qualidade para a avaliação de serviços de enfermagem, construídos mediante "uma expressão matemática, na qual o numerador representa o total de eventos predefinidos e o denominador a população de risco selecionada, observando-se a confiabilidade, a validade, a objetividade, a sensibilidade, a especificidade e o valor preditivo dos dados". No entanto, não apresentam resultados numéricos ou estudos de caso.

A qualidade na área de ciências da informação pode ser evidenciada pelos trabalhos de Naumann e Rolker (2000) e Vergueiro e Carvalho (2001). O primeiro explora a qualidade da informação na *internet* propondo três critérios de avaliação visto que, segundo os autores, os critérios são de natureza subjetiva e não podem ser avaliados automaticamente, pois as fontes de informação são autônomas e nem sempre publicam informações úteis e de qualidade. Assim, identificam três critérios que envolvem: o usuário, a fonte de informação e o processo de consulta. Não apresentam metodologia de classificação da qualidade, muito menos formas de pontuar cada critério e subcritérios. Por fim, sugerem que esta classificação seja automatizada.

Vergueiro e Carvalho (2001) investigam a qualidade de bibliotecas públicas brasileiras, para isto elaboram uma lista com 16 indicadores (comunicação, acesso, confiança, cortesia, efetividade, eficiência, qualidade, resposta, tangíveis, credibilidade, segurança, extensividade, garantia, satisfação do usuário, custo benefício e tempo de resposta) e com estes analisam e discutem sobre diferentes pontos de vista a qualidade no serviço de informação dos usuários e administradores. Com isso, são sugeridas atitudes a serem desenvolvidas pelas bibliotecas universitárias, com a finalidade de aprimorar a qualidade dos serviços/atendimentos prestados.

Os trabalhos citados até o momento não fazem uso de técnicas que mensurem a qualidade, e quando o fazem, estas ocorrem de forma bastante simplificada. O que se pode afirmar é que a análise realizada por estes autores é subjetiva.

Na sequência são apresentados trabalhos que utilizam de técnicas da área da Pesquisa Operacional para determinar a qualidade.

Na área de educação, poucos trabalhos apresentam métodos que expressam a qualidade. E ainda os que são encontrados discursam sobre o posicionamento da sociedade para se ter qualidade na educação (Oliveira e Araujo, 2005), ou sobre procedimentos para definir indicadores nesta área (Ribeiro, Ribeiro e Gusmão, 2005; Carreira e Pinto, 2007), ou ainda, sobre a análise da qualidade de sites educacionais (Carvalho, 2006; Graells, 1999). Todavia, há, ao menos, dois trabalhos que abordam o problema da qualidade utilizando o método estatístico ServQUAL (Mahapatra, 2007; Figueiredo Neto *et al.*, 2006)

O ServQUAL é um método que indica a qualidade através de vários itens em serviços, em que, por meio das informações quantitativas, procura expressar a análise qualitativa. Para isto, utiliza duas declarações afirmativas, sendo uma referente à expectativa e a outra à percepção da qualidade do serviço. Os entrevistados avaliam cada um dos itens do instrumento com opções do tipo “discordo totalmente” a “concordo totalmente”, e este resultado compõe uma escala de 5 ou 7 pontos. São utilizados elementos da estatística, como médias e desvio padrão, para analisar as respostas e verificar se os serviços satisfazem as expectativas e percepções do cliente.

Em seu trabalho, Mahapatra (2007) desenvolve um instrumento de medição da qualidade na área educacional (instituições de ensino técnico) baseado no ServQUAL. Para isto, utiliza quatro topologias de RNA, tendo como algoritmo de aprendizagem o *backpropagation*, com a finalidade de prever a qualidade na educação para as diferentes partes interessadas (alunos, ex-alunos, pais, recrutadores, faculdades, pessoal de apoio, governo, sociedade e administradores). O instrumento é validado pela análise fatorial, seguido pelo método varimax. Entretanto, assim como nos demais trabalhos já descritos, o autor não apresenta a qualidade em uma escala de classificação.

Na área elétrica, os trabalhos concentram-se na busca da identificação, localização, classificação e/ou previsão de distúrbios relacionados à QEE: afundamentos de tensão, sobretensão, *Total Harmonic Distortion* (*THD*), frequência, desequilíbrio do circuito, entre outros. Assim, os trabalhos não tratam diretamente de classificar a qualidade da energia elétrica, mas sim dos distúrbios que afetam tal qualidade.

Oleskovicz *et al.* (2006) realizaram um estudo comparativo, para analisar duas grandezas elétricas (frequência e tensão), utilizando as ferramentas: Transformada de Fourier com Janela, Transformada de Wavelet e RNA. As duas primeiras ferramentas detectam, localizam e classificam os distúrbios agregados às formas de ondas de tensão de um sistema de distribuição. A última é utilizada como forma alternativa e paralela para classificar, segundo sua natureza, os eventos de qualidade de energia.

Um método para detecção e classificação de eventos da QEE é proposto por Silva *et al.* (2007). Este método é dividido em duas etapas: a primeira consiste em detectar os eventos, em que aplica-se a Transformada de Wavelet Discreta; e na segunda, as amostras de tensão e correntes são normalizadas. Em seguida é realizada a reamostragem desses sinais, convertendo-os da frequência de amostragem original do equipamento empregado no monitoramento para a frequência padrão e também são apresentados a RNA que classifica a falta.

No trabalho desenvolvido por Adepoju, Ogunjuyigbe e Alawode (2007) são utilizadas as RNA supervisionadas para prever a carga de energia elétrica na hora seguinte. A escolha desta técnica é justificada pelo fato deste tipo de rede aprender as relações complexas entre padrões de entrada e de saída, o que é difícil para modelos em métodos convencionais. Os resultados mostram o alto grau de precisão

na capacidade das RNA para previsão de carga elétrica, pois, em geral, o valor de erro médio absoluto foi de 2,54% para a rede. Concluem, ainda, que esta técnica foi capaz de determinar a relação não-linear existente entre os dados.

Propondo um sistema de previsão da carga diária, Caciotta, Giarnetti e Leccese (2009) aplicaram RNA devido, principalmente, à capacidade de identificar automaticamente a correlação não-linear em séries de dados. Para os autores os resultados obtidos confirmam a aplicabilidade desta técnica para previsão de carga horária de energia elétrica, uma vez que o erro médio absoluto é de 2,75% em um ano. No entanto, ressaltam que no verão este erro chega a 3,5% para o ano de teste.

Trindade (2005) desenvolveu um sistema capaz de detectar, armazenar em mídia digital e classificar eventos da QEE. A classificação do evento é realizada com base no sinal de erro obtido pelo algoritmo de detecção proposto. O algoritmo realiza um processo de alinhamento dos distúrbios e, em seguida, obtém uma sub-amostragem com o objetivo de reduzir seu número de amostras. Essa nova amostra é aplicada como entrada de uma RNA que se mostra capaz de diferenciar e classificar alguns tipos de eventos.

De todos os trabalhos analisados na revisão de literatura, foi possível encontrar dois artigos (não descritos até o presente momento) que tratam da classificação da QEE e que propõem a criação de uma etiqueta de qualidade. Um destes trabalhos é de autoria de Casteren *et al.* (2005) e o outro é de Cobben e Casteren (2006). E foi com a leitura destes, mais especificamente no trabalho de Casteren *et al.* (2005), que surgiu a inspiração para a proposta da metodologia desta tese.

## 2.3 A INSPIRAÇÃO PARA A CRIAÇÃO DA ETIQUETA DE QUALIDADE, DE FORMA COMPARATIVA, NO CONTEXTO *KDD*.

O trabalho desenvolvido por Cobben e Casteren (2006) apresenta métodos para a classificação da QEE, no qual descrevem os níveis de qualidade para afundamentos de tensão de energia, principais responsáveis pelas queixas de clientes e custos associados, com base em: pequenas variações de tensão, oscilação de tensão e quedas de tensão.

Estes três métodos de classificação combinam transparência com simplicidade, utilizam um sistema de classificação por meio de etiqueta de qualidade proveniente de Casteren *et al.* (2005), conforme ilustrado na figura 2.1.

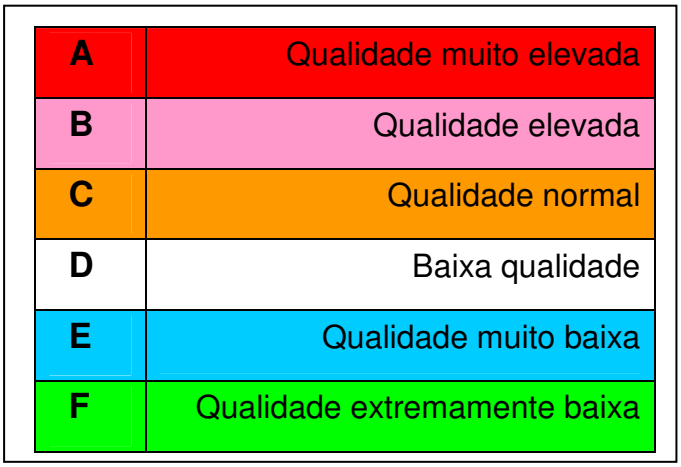


**FIGURA 2.1 –** ETIQUETA DE QEE
**FONTE**: COBBEN E CASTEREN (2006)

Já o trabalho Casteren *et al.* (2005) busca classificar os afundamentos de tensão de tal forma a indicar a responsabilidade pela causa do evento e mitigações (consumidor, fabricante do equipamento ou concessionária), examinando a duração e o valor remanescente de tais afundamentos.

De posse destes dados, os autores criam uma etiqueta de qualidade, de acordo com a frequência (número de ocorrências) ocorrida em cada uma das responsabilidades. Para tal, classificam os afundamentos em uma tabela dividida em nove tipos de faixas (figura 2.2), agrupados em três regiões, onde cada região representa uma área de responsabilidade.

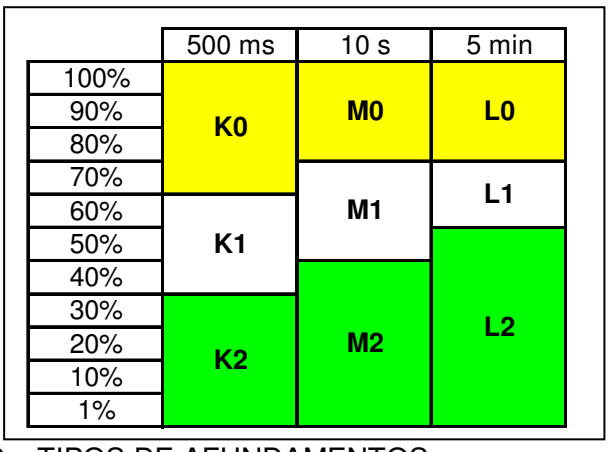


**FIGURA 2.2 –** TIPOS DE AFUNDAMENTOS
**FONTE:** CASTEREN *ET AL.* (2005)

Na figura 2.2, anterior, a região superior (K0, M0, L0) representa a área em que é impossível reduzir ou aliviar os afundamentos e é de responsabilidade do fabricante, ou seja, cabe ao mesmo desenvolver produtos que sejam capazes de resistir a estes afundamentos. Já a região intermediária (K1, M1, L1), consiste na área de responsabilidade do consumidor, onde, segundo os autores, deve se encontrar um equilíbrio quanto à disposição do consumidor pagar pela qualidade da energia ou arcar com equipamentos que reduzam os danos provocados pelos afundamentos. Por fim, na região inferior (K2, M2, L2), de responsabilidade das concessionárias, não se espera que os equipamentos resistam a estes afundamentos e economicamente é inviável a instalação de equipamentos para suportá-los. Assim, os autores afirmam que cabe às Agências Reguladoras impor normas de ocorrências às concessionárias.

Os autores não dispõem de dados detalhados de afundamento, medidos ou simulados; assim os números apresentados na elaboração da norma são fictícios (figura 2.3).

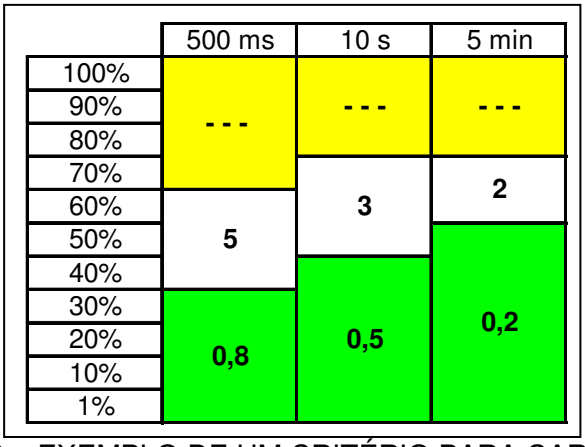


**FIGURA 2.3 –** EXEMPLO DE UM CRITÉRIO PARA CARACTERIZAR OS AFUNDAMENTOS
**FONTE:** CASTEREN *ET AL.* (2005)

Observando a figura 2.3, tem-se que um consumidor pode experimentar por ano cinco afundamentos do tipo K1, três afundamentos do tipo M1 e dois afundamentos do tipo L1. Afundamentos do tipo M2 são permitidos apenas um a cada dois anos.

Para facilitar a comunicação entre consumidores e concessionária, os autores elaboraram uma etiqueta de qualidade da QEE com base nos critérios para

caracterizar os afundamentos (figura 2.4). Nesta classificação, “A” denota alta qualidade de energia e “E” baixa qualidade.

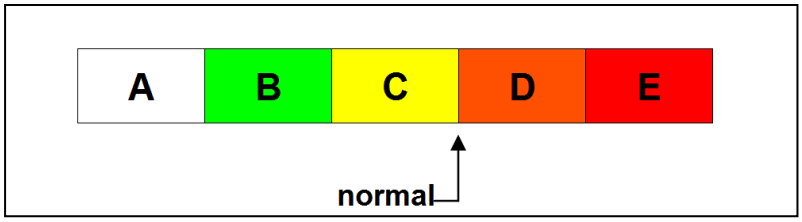


**FIGURA 2.4 –** ETIQUETA DE QUALIDADE DE ENERGIA
**FONTE:** CASTEREN *ET AL.* (2005)

A classificação da QEE (figura 2.4) deve estar vinculada aos critérios que caracterizam os afundamentos (figura 2.3) e, para isto, os autores utilizaram o critério do limite superior da etiqueta “C”, como mostrado na figura 2.5. De forma análoga podem ser criadas tabelas de critérios adicionais definindo os limites superiores de A, B e D.

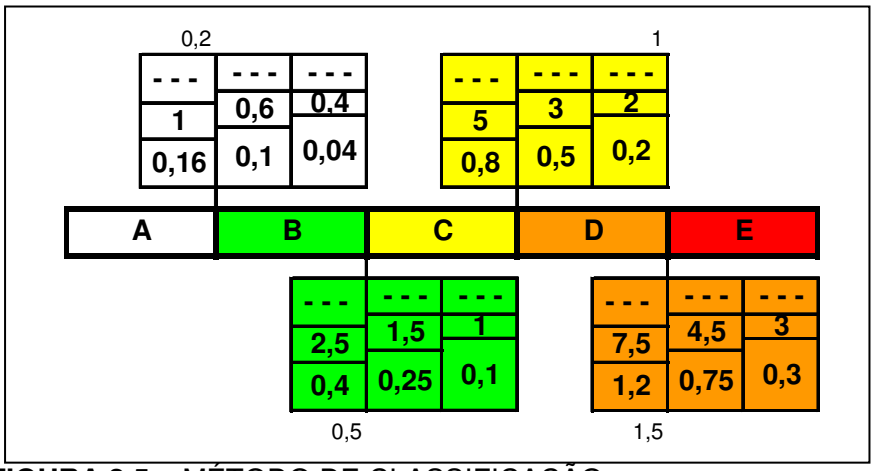


**FIGURA 2.5 –** MÉTODO DE CLASSIFICAÇÃO
**FONTE:** CASTEREN *ET AL.* (2005)

Os autores concluem que este método de classificação é simples e consistente, pois requer apenas alguns fatores de multiplicação adicionais. No entanto, não indicam como obter índices e muito menos como classificar ocorrências de afundamentos de tensão em que, por exemplo, K1, K2, M1 e L1 possuem valores na faixa de classificação “B”, mas M2 e L2 possuem valores na faixa “D” (figura 2.6).

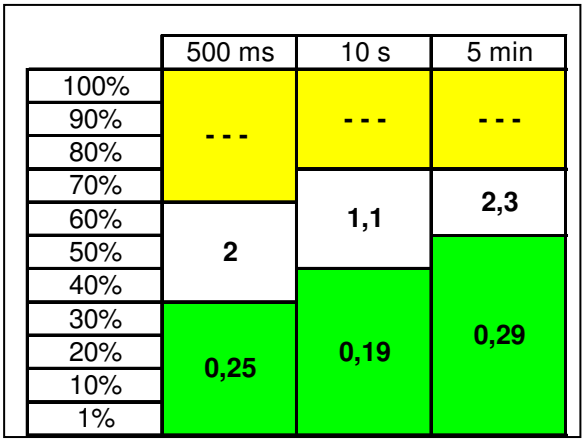


**FIGURA 2.6 –** EXEMPLO DE OCORRÊNCIAS DE AFUNDAMENTO
**FONTE:** O AUTOR (2012)

Assim, esta tese indica em suas conclusões uma técnica para classificação de elementos como o descrito no parágrafo anterior. Para isto, realiza testes com diversas técnicas nos estudos de casos.

Com a finalidade de situar a metodologia apresentada no capítulo 4, o capítulo 3, a seguir, apresenta o processo *KDD*, suporte para a criação da etiqueta de qualidade, de forma comparativa.

# 3 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS (*KNOWLEDGE DISCOVERY IN DATABASES - KDD*)

A geração de bases de dados ocorre de forma natural, pois os meios computacionais são bastante capacitados e práticos para seu armazenamento. Os dados são originados de indústrias dos mais diversos ramos de produção, empresas de telecomunicações, instituições educacionais, hospitais, instituições financeiras, dentre tantas outras. No entanto, a tarefa de simplesmente armazenar tais dados não é suficiente; é necessário, também, verificar se os dados coletados possuem informações relevantes e se há algum conhecimento a ser descoberto.

A "Descoberta de Conhecimento em Bases de Dados" (*Knowledge Discovery in Databases*, ou simplesmente, *KDD*) é um processo que visa encontrar informações em bases de dados de uma maneira automatizada, criando relações de interesse que, muitas vezes, não são observadas por especialistas no assunto.

O termo "*KDD*" surgiu no final da década de 80 e se mantém fortemente nos dias atuais, o que se pode verificar pelos trabalhos de Han e Kamber (2006), Steiner *et al.* (2006), Atami e Beldjilali (2007), Li e Kuo (2008), Bang *et al.* (2009), Alves e Falsarella (2009), Liu, Tian e Zhang (2010), Tronchoni *et al.* (2010), Bradwaj e Pal (2011), Mao *et al.* (2011), Arora e Bhatia (2012), Liu, Qi e Li (2012), dentre outros.

Na busca pela definição de *KDD*, sua primeira menção está no trabalho de Frawley, Piatetsky-Shapiro e Matheus (1991) como sendo "o processo, não trivial, de extração de informação, implícitas, previamente desconhecidas e úteis, a partir dos dados armazenados em uma base de dados".

Três anos depois foram Braschman e Anand (1994) que definiram *KDD* como "uma tarefa de descoberta de conhecimento intensivo, consistindo de interações complexas, feitas ao longo do tempo entre o homem e uma grande base de dados, possivelmente suportada por um conjunto heterogêneo de ferramentas".

No entanto, a definição mais comum na literatura é de Fayyad *et al.*, (1996) na qual se tem que *KDD* é "o processo não-trivial de identificação válida, em dados, novos, potencialmente úteis e finalmente com padrões compreensíveis".

Através desta última definição, podem-se definir alguns termos utilizados neste contexto, conforme apresentado no quadro 3.1.

**QUADRO 3.1 –** TERMOS PRESENTES NA DEFINIÇÃO DE *KDD* POR FAYYAD *ET AL.*, 1996

| **TERMO** | **DESCRIÇÃO** |
|---|---|
| Dados | Um conjunto de registros. |
| Padrão | Uma expressão $E$ em uma linguagem $L$ que descreve registros em um subconjunto dos dados. |
| Processo | Envolve várias etapas como a preparação de dados, busca de padrão, avaliação do conhecimento e refinamento com repetição após a modificação. |
| Válido | Padrões descobertos devem ser verdadeiros em novos dados com algum grau de certeza e generalizáveis no futuro para outros dados. |
| Novo | Os padrões devem ser novos, ou seja, não conhecidos previamente. |
| Útil | Os padrões levam a ações úteis. |
| Compreensível | O processo deve levar à percepção humana.<br>Padrões devem ser transformados em conhecimento compreensível, a fim de facilitar uma melhor interpretação dos dados subjacentes. |

Anterior a sua denominação, o processo *KDD* era denominado por muitos de *Data Mining* (Mineração de Dados ou, simplesmente, *DM*). *DM* é a principal das cinco etapas do *KDD*. Estas etapas são as seguintes, por ordem sequencial: seleção dos dados; limpeza dos dados ou pré-processamento; transformação dos dados; aplicação do *DM* e, finalmente, interpretação do conhecimento gerado. (FAYYAD *et al.*, 1996).

> "[...] *KDD* refere-se ao processo global de descoberta de conhecimento útil a partir de mineração de dados que é uma etapa particular neste processo. A mineração de dados é a aplicação de algoritmos específicos para extrair padrões dos dados. [...] Os passos adicionais no processo de *KDD*, como preparação de dados, seleção de dados, limpeza de dados, a incorporação de conhecimento prévio adequado, e a interpretação adequada dos resultados da mineração, são essenciais para garantir que o conhecimento útil seja derivado dos dados. A aplicação direta de métodos da mineração de dados pode ser uma atividade perigosa, que leva facilmente à descoberta de padrões sem nexo e inválidos." (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

A figura 3.1 apresenta a sequência da metodologia *KDD*, segundo Fayyad *et al.* (1996).

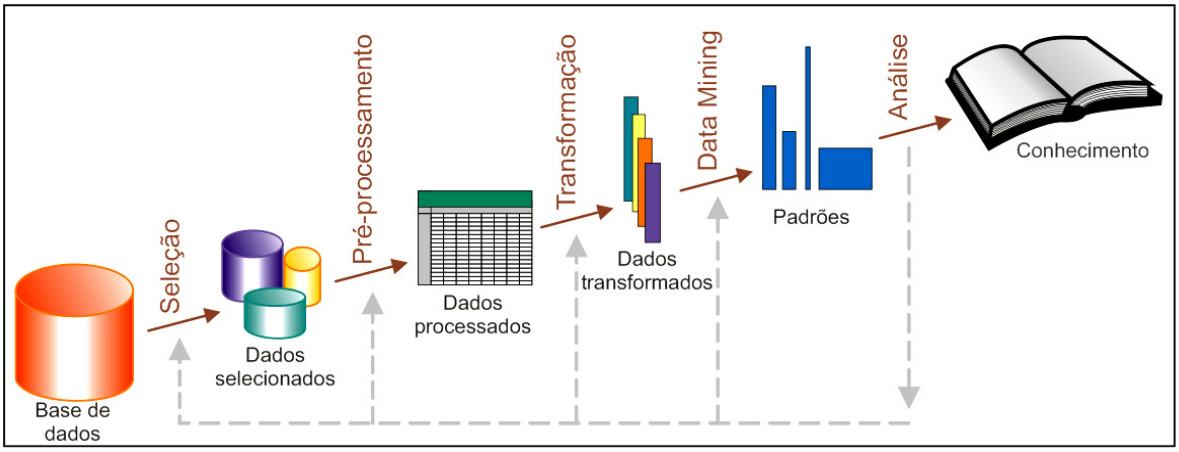


**FIGURA 3.1 –** ETAPAS DO PROCESSO *KDD* POR FAYYAD *ET AL.*, 1996
**FONTE:** MODIFICADO PELO AUTOR (2012)

No entanto, para outros autores, como Feldens e Castilho (1997) e Bonchi *et al.* (2009), a metodologia *KDD* possui três etapas: pré-processamento, mineração de dados e pós-processamento, mas ao analisar as etapas definidas por estes autores, verifica-se que as três primeiras de Fayyad *et al.* (1996) estão inseridas na primeira de Feldens e Castilho (1997); o mesmo ocorre com última etapa definida por Fayyad *et al.* (1996) que está contemplada pela última de Feldens e Castilho (1997) e Bonchi *et al.* (2009).

No decorrer deste processo pode-se encontrar conhecimentos explícitos ou não, isto é, as informações podem ser as que já se tinha conhecimento ou informações inesperadas que ao analisar a base de dados não se verificava nenhuma relação óbvia. Podem ainda ocorrer informações sem nenhuma relação significativa, pela falta de atributos ou por não haver conhecimento novo a ser descoberto.

Geralmente, têm-se como respostas informações que não se podem detectar quando se aplicam métodos tradicionais na análise de dados para posterior tomada de decisão, pois, em sua grande maioria, os métodos tradicionais são capazes de verificar apenas as relações explícitas nos bases de dados.

## 3.1 ETAPAS DO PROCESSO *KDD*

No desenvolvimento deste trabalho é utilizado o processo definido por Fayyad *et al.* (1996) e, assim sendo, as etapas são explicitadas a seguir.

A fase da **seleção dos dados** consiste na escolha do conjunto de dados que se pretende analisar, definindo assim os atributos e os registros. Em sua grande maioria, esta seleção é realizada por um especialista da área proveniente dos dados, pois possui papel fundamental no resultado final.

A **limpeza dos dados ou pré-processamento** é a fase em que se determinam quais dados serão eliminados por serem redundantes, ou por possuírem ruídos detectáveis e dados discrepantes dos demais. Além disto, é verificada a possibilidade de diminuir o número de variáveis. Para isto, podem ser aplicados métodos estatísticos, a fim de melhorar a eficácia dos algoritmos de classificação, como apresentado por Steiner *et al.* (2006).

Na **transformação dos dados**, estes precisam ser armazenados e formatados de forma adequada à aplicação do algoritmo na próxima fase. Também é nesta fase que são determinados atributos faltantes que podem ser obtidos de outros como, por exemplo, a duração de certo evento por meio do horário inicial e horário final da ocorrência do mesmo.

No ***Data Mining***, etapa mais importante do processo *KDD*, são aplicadas as técnicas para análise dos dados por meio de heurísticas ou metaheurísticas, para a descoberta de padrões. O tempo de execução desta fase deve ser compatível com o tempo disponível na espera da solução. Muitos são os métodos, sendo que os mais conhecidos são Redes Neurais e Algoritmos Genéticos.

Na **interpretação do conhecimento** gerado deve-se interpretar o conhecimento apresentado, verificando a relevância ou não dos padrões e, com isso, verificar também a eficácia do método aplicado na etapa do *DM*. Caso o analista julgue que o conhecimento não é válido, o processo deverá ser reiniciado, analisando-se todas as etapas em busca de melhorar e/ou refazer o que for necessário até que o conhecimento obtido seja julgado como verdadeiro por quem o analisa.

Sendo o *DM* a etapa mais importante do processo *KDD*, é evidente que métodos eficientes são exigidos à medida que as bases de dados se mostram cada vez mais ilimitados em relação número de informações que podem armazenar, sem muitas vezes explicitar as relações existentes entre os vários atributos.

## 3.2 *DATA MINING* (*DM* OU MINERAÇÃO DE DADOS)

O *DM* é a etapa mais importante do processo *KDD*, conforme já comentado, pois é nesta etapa que se aplicam técnicas para análise dos dados, seja através de procedimentos heurísticos ou metaheurísticos, para a descoberta de padrões.

> "este processo deve ser automático ou (mais geralmente) semi-automático. Os padrões descobertos devem ser significativos na medida em que leva a alguma vantagem, geralmente uma vantagem econômica. Os dados estão invariavelmente presentes em quantidades substanciais." (WITTEN e FRANK, 2005)

Ainda sobre o reconhecimento de padrões por meio do *DM*, tem-se que Côrtes, Procaro e Lifschitz (2002) afirmam:

> "é um processo altamente cooperativo entre homens e máquinas, que visa à exploração de grandes bases de dados, com o objetivo de extrair conhecimentos através do reconhecimento de padrões e relacionamento entre variáveis, conhecimentos esses que possam ser obtidos por técnicas comprovadamente confiáveis e validados pela sua expressividade estatística."

Sobre o que se espera obter ao aplicar o *DM*, Thuraisingham *et al.* (2005) afirmam que se realizam "várias consultas e extração de informações úteis e que, muitas vezes, são desconhecidas e inesperadas, como padrões e tendências".

No entanto, a busca em grandes bases de dados só é possível devido aos métodos computacionais e a análise do ser humano. Com base nisso, Weiss e Indurkhya (1998) definem *DM* como sendo:

> "... a busca de informações valiosas de grandes volumes de dados. É um esforço cooperativo entre os seres humanos e computadores. Os humanos projetam as bases de dados, descrevem os problemas e definem metas. Computadores analisam os dados, procurando por padrões relacionados às metas definidas pelos humanos."

De forma mais simples, Han e Kamber (2006) definem *DM* como "o processo de extração ou mineração de conhecimento em grandes quantidades de dados."

Esta fase não elimina o conhecimento específico da área de origem dos dados, mas sim, auxilia os analistas a encontrar os padrões nas bases de dados. Além disso, estes analistas devem verificar e assegurar se os resultados são significativos para a solução do que se pretende. São eles que podem determinar os dados relevantes para o processo.

A aplicação deste processo tem como principais resultados, a obtenção de: associação; agrupamento; previsão; classificação de padrões; dentre outros.

A *classificação de padrões* é a tarefa que generaliza uma estrutura conhecida a ser aplicada aos novos dados, ou seja, procura descobrir uma função a ser aplicada aos novos dados que consiga classificar registros em um conjunto de dados pré-definidos.

O *agrupamento ou clusterização de padrões* consiste em agrupar registros em subconjuntos que de alguma forma compartilhem propriedades, onde os grupos variáveis não são pré-definidos.

A *associação de padrões* busca relações entre as variáveis do problema, sendo que estas podem ser independentes ou exploratórias.

Por ser baseado em Inteligência Artificial, com apoio da Matemática e/ou Estatística, têm-se diversas técnicas apresentadas na literatura para a realização de *DM*. Dentre estas técnicas se destacam, dentre outras, os Algoritmos Genéticos (AG), as Redes Neurais Artificiais (RNA) e *Support Vector Machines* (*SVM*).

# 4 METODOLOGIA PARA CRIAÇÃO DE ETIQUETA DE QUALIDADE, DE FORMA COMPARATIVA, NO CONTEXTO KDD

Realizado o estudo sobre qualidade e sobre o processo de *KDD*, neste capítulo é apresentada a metodologia para criação da etiqueta de qualidade, de forma comparativa, no contexto *KDD*, em que são descritos os passos deste processo em caracteres em negrito, sendo detalhados os tratamentos realizados nos dados, técnicas e demais considerações pertinentes à obtenção de tal etiqueta.

A primeira etapa da metodologia consiste na **seleção dos dados**. Para isso, é necessário entender o problema, definir objetivos e só então selecionar a base de dados, sendo importante o trabalho junto a um profissional da área em que o problema está inserido.

Obtida a base de dados, a próxima etapa da metodologia consiste em realizar a **limpeza dos dados ou pré-processamento**, ou seja, os atributos da base de dados devem ser analisados, observando se há possibilidade de descartar alguns deles que contenham informações idênticas, verificar se há a necessidade de acrescentar novos atributos utilizando outros já existentes ou se há dados incompletos e que possam ser descartados. Este passo é fundamental, pois são estas as informações a serem consideradas na metodologia de criação de etiqueta de qualidade, de forma comparativa.

Realizada esta etapa, deve-se quantificar os elementos que serão utilizados para verificar a qualidade. Exemplificando, tem-se que se o objetivo é verificar a qualidade da energia elétrica classificando alimentadores de uma subestação, pode-se, para isso, considerar a quantidade de ocorrências de eventos de afundamentos de tensão em cada deles, comparando-os; se o objetivo é verificar a qualidade de escolas de certa região, em relação ao desempenho escolar, pode-se considerar a média das notas dos alunos destas escolas, por disciplina, comparando-as.

Nessa metodologia é sugerido que esta quantificação seja traduzida em quadros como os ilustrados a seguir. No quadro 4.1, está um exemplo da quantificação de eventos ocorridos para um alimentador de subestação de energia elétrica onde é considerada a magnitude do evento e a duração, para isto foram geradas dez classes de classificação ($C_1$, $C_2$, ..., $C_{10}$).

**QUADRO 4.1 –** CLASSIFICAÇÃO CONSIDERANDO A DURAÇÃO E A TENSÃO REMANESCENTE

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | $C_1$ | $C_2$ |
| 60 a 79% | $C_3$ | $C_4$ |
| 40 a 59% | $C_5$ | $C_6$ |
| 20 a 39% | $C_7$ | $C_8$ |
| 10 a 19% | $C_9$ | $C_{10}$ |

Já no quadro 4.2, há a quantificação de notas obtidas em avaliações em uma escola, onde são geradas quatro classes de classificação ($C_1$, $C_2$, $C_3$ e $C_4$).

**QUADRO 4.2 –** CLASSIFICAÇÃO CONSIDERANDO NOTAS EM AVALIAÇÕES

| **Nível de** | **Área do conhecimento** | |
|---|---|---|
| **Ensino** | X | Y |
| Z | $C_1$ | $C_2$ |
| W | $C_3$ | $C_4$ |

Os quadros 4.1 e 4.2 anteriores podem ser adaptados aos mais diversos problemas, sendo que para cada problema haverá uma quantidade de classes $C_i$.

Construídos os quadros para cada elemento que se quer conhecer a qualidade, a próxima etapa da metodologia consiste na mudança de escala dos dados (uma forma de **transformação dos dados**) com a finalidade de obter um melhor desempenho das técnicas de *DM*. Para isso procedemos da seguinte maneira:

*Para cada conjunto $C_i$.*

*Verificar o menor valor e subtrair este de todos os elementos, formando o conjunto $C_i$*.*

*Dividir todos os elementos de $C_i$* pelo maior valor deste conjunto, obtendo assim $C_i$ **.*

*Se a melhor qualidade consiste em ter os menores valores em $C_i$ (problema de minimização) então os dados já estão com a mudança de escala realizada, faça $C_i = C_i$**. Caso contrário, obtenha $C_i$***, onde cada elemento é o resultado da subtração: 1 menos o elemento correspondente de $C_i$**. Desta forma, faça $C_i = C_i$***.*

Com os dados no intervalo [0, 1] e definidas as faixas de classificação da etiqueta (figura 4.1, a seguir), que neste trabalho são apresentadas seis faixas (A, B, C, D, E e F), é possível criar os limites superior e inferior para cada uma destas faixas.


**FIGURA 4.1 –** REPRESENTAÇÃO DA ETIQUETA DE CLASSIFICAÇÃO DA QUALIDADE
**FONTE:** O AUTOR (2012)

Supondo um problema que possua $n$ classes $C_i$ de classificação, $i = 1,..,n$, é preciso determinar o intervalo de cada $C_i$ para cada faixa da etiqueta (figura 4.2). No entanto, é evidente que o limite superior de uma faixa de classificação da etiqueta é o limite inferior da faixa subsequente.

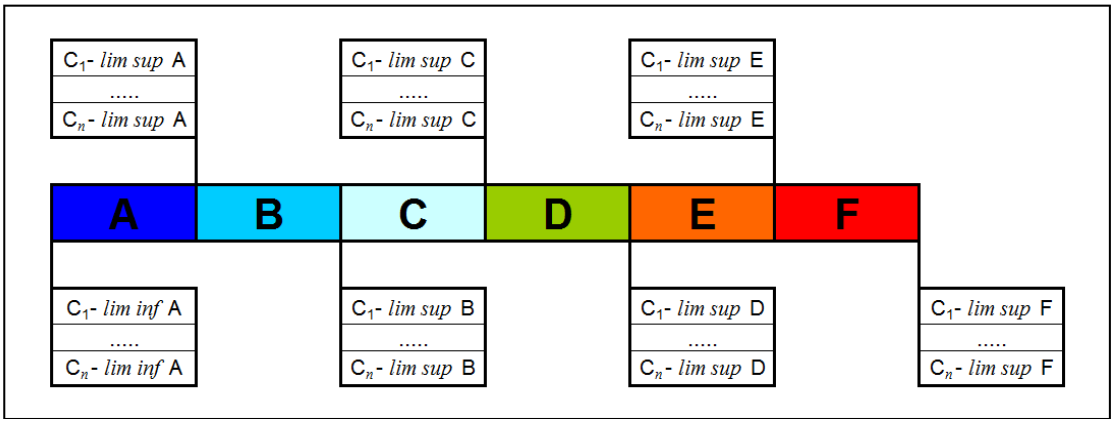


**FIGURA 4.2 –** REPRESENTAÇÃO DA ETIQUETA DE CLASSIFICAÇÃO DA QUALIDADE COM DEFINIÇÕES DE LIMITES
**FONTE:** O AUTOR (2012)

Nesta metodologia é considerado que o limite superior da faixa de classificação “C” é a média dos valores de cada classe $C_i$, o limite inferior da faixa de classificação “A”, é $C_i=0$, para todo $i$, e o limite superior da faixa de classificação “F” é $C_i = 1$, para todo $i$.

As demais faixas de classificação podem ser obtidas, por exemplo, multiplicando $lim\ sup$ C por fatores elaborados junto a um profissional da área em que problema está inserido. Dessa forma, pode-se restringir algumas faixas e deixar

outras com maior comprimento, por exemplo, restringir as faixas A e B como se segue: *lim sup* A = *lim sup* C*0,25, *lim sup* B = *lim sup* C*0,50, *lim sup* D = *lim sup* C*1,50 e *lim sup* E = *lim sup* C*2,00. Neste caso, [*lim inf* A, *lim sup* A] e [*lim sup* A, *lim sup* B] possuem mesmo comprimento, os intervalos [*lim sup* B, *lim sup* C], [*lim sup* C, *lim sup* D] e [*lim sup* D, *lim sup* E] também possuem mesmo comprimento, mas em relação aos anteriores possuem o dobro do tamanho. Por fim, o comprimento do intervalo [*lim sup* E, *lim sup* F] não é possível determinar, nesta análise, pois depende de *lim sup* C.

Outra forma de determinar *lim sup* A e *lim sup* B é fazer com que os intervalos [*lim inf* A, *lim sup* A], [*lim sup* A, *lim sup* B] e [*lim sup* B, *lim sup* C] tenham o mesmo comprimento, o mesmo ocorrendo na determinação de *lim sup* D e *lim sup* E, onde os intervalos [*lim sup* C, *lim sup* D], [*lim sup* D, *lim sup* E] e [*lim sup* E, *lim sup* F] têm o mesmo comprimento. Desta forma, nenhuma das faixas acima ou abaixo da média é privilegiada em seu comprimento.

Essas duas abordagens de obtenção dos valores que definem cada intervalo são melhores visualizadas nos estudos de casos presentes nos próximos capítulos. No primeiro estudo de caso é utilizada a primeira abordagem descrita e no segundo estudo de caso, a segunda abordagem.

Posto isto, basta verificar em qual faixa de classificação da etiqueta de qualidade os elementos, representados pelos quadros (quadro 4.1 e quadro 4.2), se enquadram. Mas esta tarefa não é tão simples, uma vez que há vários $C_i$ e ocorre, quase sempre, que nem todos os $C_i$ estejam em intervalos de mesma faixa. Por exemplo, se todos os $C_i$, com $i=1,...,\ n$ e $i \neq k$, com $k \leq n$, pertencem ao intervalo [*lim sup* A, *lim sup* B] e $C_k$ pertence ao intervalo [*lim sup* E, *lim sup* F], qual a classificação deste elemento?

Assim, com a finalidade de obter a classificação destes elementos que não se enquadram em nenhuma faixa da etiqueta diretamente, são aplicadas algumas técnicas de *DM,* mais precisamente, duas metaheurísticas (Redes Neurais e Algoritmos Genéticos) e seis métodos heurísticos (*Support Vector Machine*, somatório das distâncias euclidiana do novo elemento aos elementos de cada faixa da etiqueta de qualidade, distância euclidiana do novo elemento a ponto central de cada faixa da etiqueta de qualidade, $k$-vizinhos mais próximos, distância euclidiana do novo elemento a média dos elementos de cada faixa da etiqueta de qualidade e a distância de Mahalanobis), todas relacionadas à classificação de padrões. Estas

técnicas são descritas brevemente nas próximas seções, com algumas particularidades das aplicações aqui realizadas.

Na metodologia proposta, após a aplicação das técnicas de *DM* são realizadas as **análises dos resultados**. Nesta análise, são comparadas as técnicas aplicadas e verifica-se a classificação que apresentou maior ocorrência para cada elemento, determinando desta forma a sua classificação. Este procedimento é conhecido como “classificação por voto” ou, da estatística, “moda”. No entanto, como conclusão desta tese, por meio dos estudos de caso, é indicada a técnica com resultados mais adequados.

Antes de realizar breve comentário sobre cada técnica aplicada na etapa do *DM*, são expostas algumas considerações com relação aos dados e a metodologia de aplicação de tais técnicas.

## 4.1 CONSIDERAÇÕES ACERCA DAS FAIXAS DE CLASSIFICAÇÃO DA ETIQUETA DE QUALIDADE

Antes de indicar como as técnicas de *Data Mining* são utilizadas, faz-se necessário tecer alguns comentários sobre os dados na etiqueta de qualidade, de forma comparativa.

Se os quadros que representam os elementos (quadros 4.1 e 4.2) dos quais se quer saber a classificação da qualidade, possuem apenas uma classe $C_i$, tem-se que os conjuntos de cada faixa de classificação da etiqueta são representados por intervalos (figura 4.3). Nesse caso, vale salientar que não é necessária a aplicação de técnicas de *DM*, uma vez que a etiqueta de qualidade é unidimensional e os elementos vão se enquadrar diretamente em alguma faixa de classificação da qualidade.

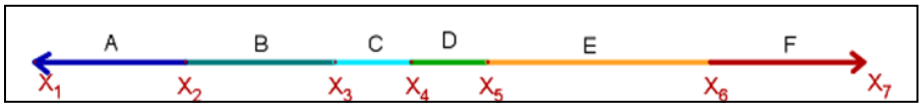


**FIGURA 4.3 –** REPRESENTAÇÃO GRÁFICA DA ETIQUETA COM APENAS UMA CLASSE $C_i$
**FONTE:** O AUTOR (2012)

Se os elementos possuem duas classes $C_i$, tem-se que a etiqueta de qualidade pode ser representada no plano cartesiano e os conjuntos que representam cada faixa de classificação são linearmente separáveis, como pode ser visualizados na figura 4.4, onde os $X_i$ representam os limites para as classes $C_1$ e $Y_i$ os limites para a classe $C_2$.

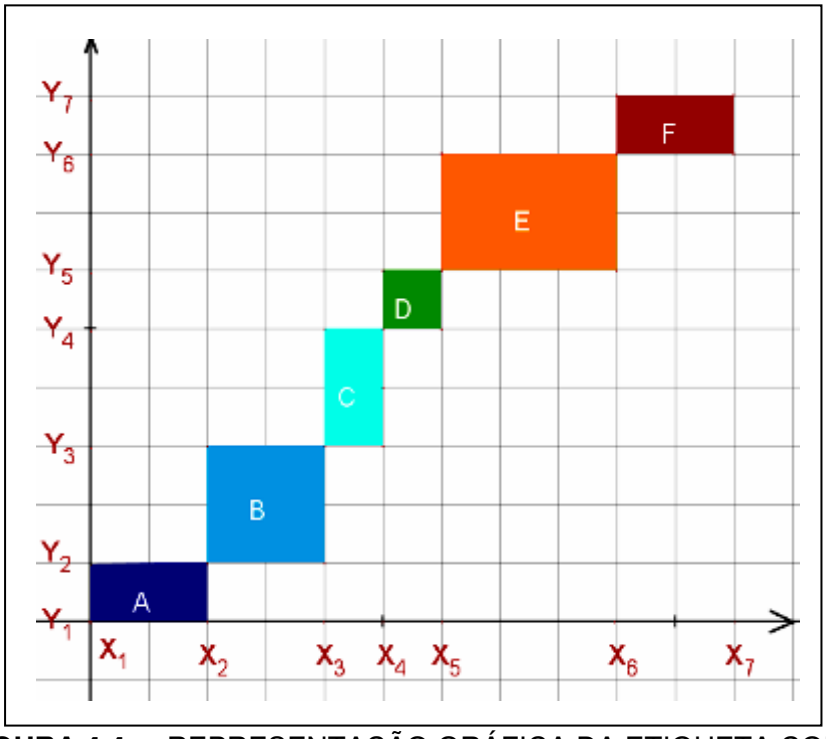


**FIGURA 4.4 –** REPRESENTAÇÃO GRÁFICA DA ETIQUETA COM DUAS CLASSES $C_{is}$
**FONTE:** O AUTOR (2012)

Se os elementos que se pretende saber a classificação possuem três classes, a etiqueta de qualidade pode ser representada no espaço, conforme figura 4.5, e é evidente que estas faixas são separáveis por planos.

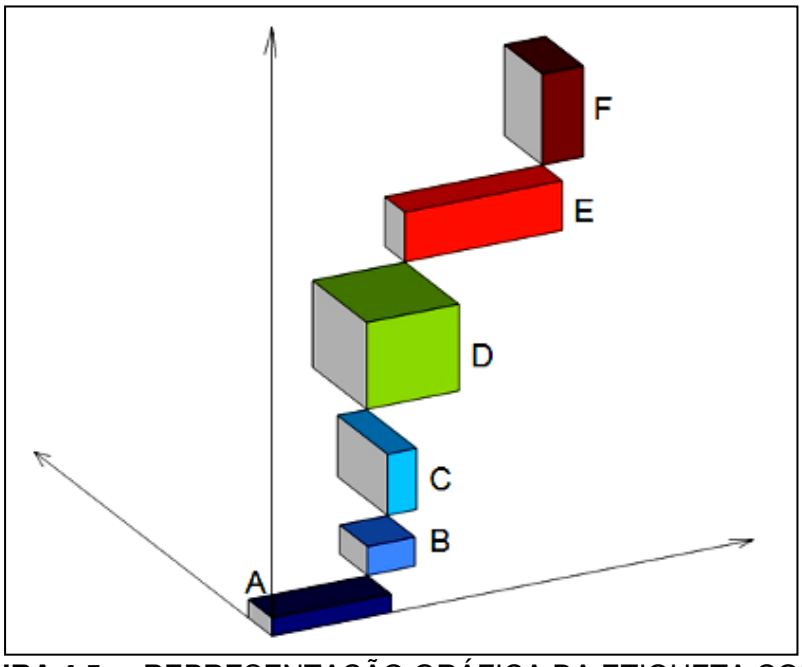


**FIGURA 4.5 –** REPRESENTAÇÃO GRÁFICA DA ETIQUETA COM TRÊS CLASSES $C_{is}$
**FONTE:** O AUTOR (2012)

Se os elementos possuem quatro classes de classificação ou mais, estes não podem ser representados graficamente. No entanto, as faixas de classificação da etiqueta são separáveis por hiperplanos, no caso de $n$ classes $C_i$, por hiperplano de $\Re^n$.

Assim, na etapa *DM* procurou-se aplicar técnicas (cujos desempenhos são comparados) que separassem as faixas da etiqueta de classificação. Quando o elemento se enquadra diretamente na faixa, a sua classificação é obtida diretamente; caso contrário, a sua classificação é dada pela técnica de *DM*.

## 4.2 METODOLOGIA UTILIZADA NA APLICAÇÃO DAS TÉCNICAS DE *DATA MINING*

Nesta seção são apresentados alguns comentários sobre a metodologia utilizada na aplicação das técnicas de *DM*, sendo a primeira referente à obtenção de dados para a aplicação.

Nesta metodologia é utilizado o aprendizado supervisionado, em que são conhecidas para o conjunto de dados as entradas e as saídas, ou seja, para cada elemento do conjunto já se sabe, de antemão, a sua classificação. Assim, as técnicas extraem deste conjunto características que os classifiquem como tal e, a

partir disso, é possível classificar um novo elemento que compunha o conjunto de treinamento. As melhores técnicas são as que possuem melhor capacidade de generalização, ou seja, são as que têm melhor capacidade de prever corretamente a classificação de um novo padrão apresentado ao classificador.

Para obter os conjuntos de dados que são utilizados no treinamento das técnicas (aprendizado supervisionado), já que não se dispunha de dados para tal, estes foram gerados de forma aleatória, da seguinte maneira: para cada faixa de classificação da etiqueta – respeitando o limite inferior e o superior – foram geradas certa quantidade $X$ de padrões, tendo assim um total de $Y=6*X$ padrões para o treinamento e teste em cada técnica. Este número $X$, nos estudos de caso apresentados neste trabalho, é igual a $30*(n - 2)$, onde $n$ é a quantidade de classes de classificação $C_i$, valor obtido através de diversos testes realizados.

Como ilustração, tem-se que os diferentes símbolos contidos no conjunto $Y$, apresentado na Figura 4.6, representam as diferentes faixas de classificação da etiqueta de qualidade.

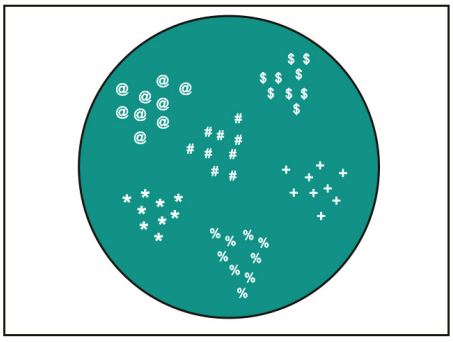

**FIGURA 4.6 –** REPRESENTAÇÃO DE DADOS GERADOS PARA CADA FAIXA DE CLASSIFICAÇÃO DA ETIQUETA DE QUALIDADE
**FONTE:** O AUTOR (2012)

A forma de avaliação das técnicas é a *validação cruzada*, com o *método* $k$*-fold* estratificado, com $k=3$, ou seja, o conjunto de dados $Y$ é dividido, aleatoriamente, em dois subconjuntos: $2/3$ para o conjunto de treinamento e $1/3$ para o conjunto de teste. Para isso, o conjunto de dados é dividido em três subconjuntos, aqui denominados de SCa, SCb e SCc (Subconjunto A, Subconjunto B e Subconjunto C, respectivamente), cada um contendo a mesma quantidade $X/3$ de padrões de cada faixa de classificação da etiqueta de qualidade. (Figura 4.7)

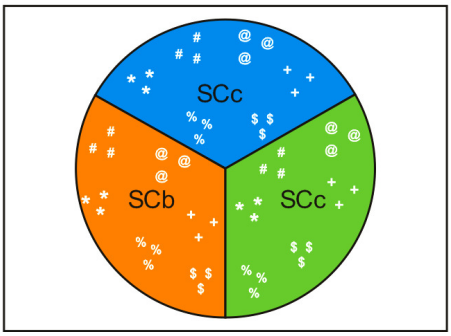


**FIGURA 4.7 –** REPRESENTAÇÃO DOS SUBCONJUNTOS DE DADOS GERADOS PARA CADA FAIXA DE CLASSIFICAÇÃO DA ETIQUETA DE QUALIDADE
**FONTE:** O AUTOR (2012)

Com os subconjuntos são realizados os testes da seguinte forma: na 1ª etapa tem-se os subconjuntos SCa e SCb formando o conjunto de treinamento (2/3 dos dados gerados) e SCc o conjunto de testes (1/3 dos dados gerados) (Figura 4.8a); na 2ª etapa tem-se os subconjuntos SCa e SCc formando o conjunto de treinamento e SCb o conjunto de testes (Figura 4.8b); e na 3ª etapa tem-se os subconjuntos SCb e SCc formando o conjunto de treinamento e SCa o conjunto de testes (Figura 4.8c).

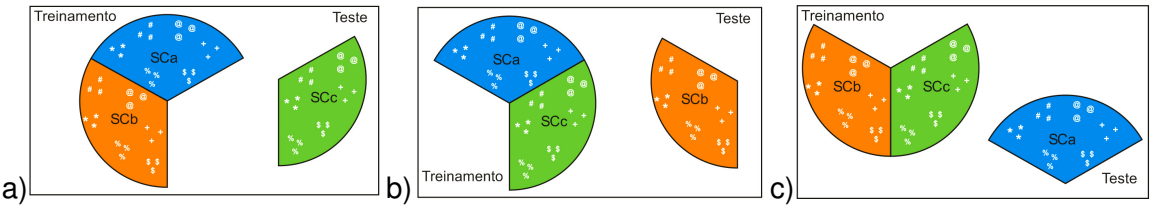


**FIGURA 4.8 –** CONJUNTOS DE TREINAMENTO E TESTE - VALIDAÇÃO CRUZADA
**FONTE:** O AUTOR (2012)

Além disto, para todas as técnicas foi adotada a metodologia de Steiner (1995), que afirma que ao se aplicar a RNA (técnica utilizada pela autora em problemas em que são necessários vários neurônios na camada de saída) - lembrando que nesta tese são seis, cada um dos quais representando uma faixa de classificação da etiqueta de qualidade - se obtém melhores resultados quando o problema é dividido em subproblemas da seguinte maneira: ao invés de se ter a saída, por exemplo, $A=100000$, $B=010000$,..., e $F=000001$, realiza-se um treinamento com $A=1$ e um conjunto $W=0$ composto dos demais registros do treinamento ($B$, $C$, $D$, $E$ e $F$) (Figura 4.9a). Em seguida realiza-se outro treinamento com os registros de $B=1$ e um conjunto $K^*=0$ composto dos registros de treinamento de $A$, $C$, $D$, $E$ e $F$

(Figura 4.9b). Da mesma forma, tem-se para os registros de $C$, $D$, $E$ e $F$, sendo esta técnica conhecida como “um contra todos”.

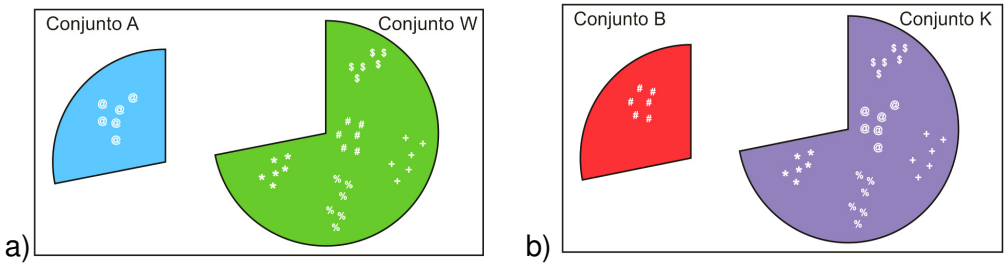


**FIGURA 4.9 –** REPRESENTAÇÃO DA METODOLOGIA DE STEINER, 1995.
**FONTE:** O AUTOR (2012)

Assim, na procura pela classificação de um novo exemplo, após o treinamento, este deve ser “apresentado” a todas as redes treinadas e terá sua classificação conforme o valor mais próximo a uma saída da RNA.

Assim, foram iniciadas as aplicações das técnicas, sendo a primeira a RNA. Ocorre que no treinamento com o segundo conjunto, conforme metodologia de Steiner (1995), $B=1$ e $K^*=0$, onde $K^*=\{A, C, D, E \ e \ F\}$, os resultados não foram satisfatórios, uma vez que a rede não conseguiu aprender o conjunto $B$, nas diversas topologias. Com isso, foi realizada modificação nesta metodologia: uma vez que uma técnica realizou o treinamento com o primeiro conjunto $A=1$ e $W=0$, onde $W=\{B, C, D, E \ e \ F\}$, o próximo conjunto não mais terá os elementos de A, ou seja, o novo conjunto para treinamento contém $B=1$ e $K=0$, *onde* $K=\{C, D, E \ e \ F\}$. (Figura 4.10)

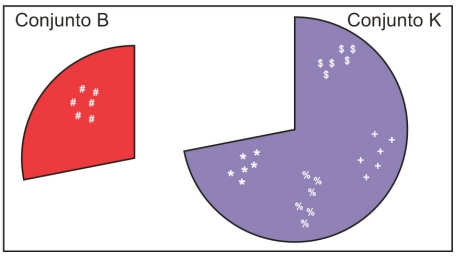


**FIGURA 4.10 –** REPRESENTAÇÃO DA MODIFICAÇÃO REALIZADA NA METODOLOGIA DE STEINER, 1995, A PARTIR DO SEGUNDO CONJUNTO DE TREINAMENTO
**FONTE:** O AUTOR (2012)

Com esta modificação os resultados foram satisfatórios e serão expostos nos capítulos 5 e 6, que tratam dos estudos de casos (áreas elétrica e educacional).

Na sequência são apresentadas as técnicas de *DM* aplicadas aos estudos de casos.

## 4.3 REDES NEURAIS

As Redes Neurais (RNA ou *Neural Network*), também conhecidas como Redes Neurais Artificiais, é uma metaheurística inspirada na biologia do sistema nervoso que simula o cérebro humano e mantém comportamentos como o aprender, errar e descobrir (HAYKIN, 1999; MITCHELL, 1997).

A RNA teve sua origem na década de 40 no trabalho de McCulloch e Pitts (1943), o primeiro, neurofisiologista do *Massachusetts Institute of Technology* (MIT) e o segundo, matemático da Universidade de Illinois. Eles realizaram uma analogia entre as células nervosas do corpo humano com o processo eletrônico.

Cada neurônio possui certo número de entradas (que são os dendritos do corpo celular biológico) e saídas. Em todos os nós, também chamados de sinapses (referência ao biológico), é utilizada uma combinação linear que produz um único valor de entrada. Esse único valor é a soma ponderada (entrada x peso) que reproduz os estímulos captados no processo biológico. A função transferência (limiar de disparo do neurônio biológico) geralmente é não-linear como, por exemplo, sigmóide, função degrau, função de limiar, função linear por partes, função logística e função tangente hiperbólica.

A figura 4.11 representa o neurônio proposto por McCulloch e Pitts (1943), onde $x_1$, $x_2$ e $x_3$ são as entradas, $wj_1$, $wj_2$ e $wj_3$ são os pesos referentes a cada entrada, $\Sigma$ é a representação da combinação linear, $T$ é a função de transferência e *y* representa a saída.

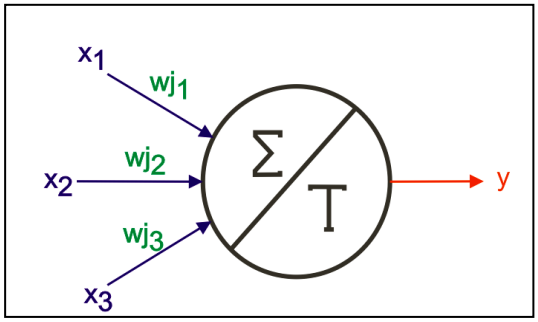


**FIGURA 4.11 –** REPRESENTAÇÃO DO NEURÔNIO PROPOSTO POR MCCULLOCH E PITTS (1943)
**FONTE:** O AUTOR (2012)

Seis anos depois Hebb (1949) introduz o conceito de treinamento:

> "Se as entradas de um sistema produzem o mesmo padrão de atividade repetidamente, o conjunto de elementos ativos que compõem o modelo se tornará cada vez mais fortemente interligados. Ou seja, cada elemento tende a se unir com todos os outros e (com pesos negativos) desconecta-se dos elementos que não fazem parte do padrão. Dito de outra forma, o padrão como um todo vai se tornar 'auto-associado'". (HEBB, 1949, p.44)

A partir disto, as RNA ganham espaço em aplicações nas mais diversas áreas, como a medicina, energia elétrica, indústrias, engenharia, química, biologia, física, geralmente com a finalidade de classificar e reconhecer padrões, otimização, previsão e controle automático.

A forma mais simples de uma Rede Neural é o *perceptron,* que é composto de uma única camada de neurônios, na qual cada vetor (elemento de um conjunto de padrões) apresentado à rede possui pesos que são ajustados no decorrer da aprendizagem. Se a saída deste vetor estiver correta, nenhuma mudança é realizada, caso contrário, os pesos são atualizados e utilizados nas regras de aprendizado do *perceptron*. Ao final de uma época (iteração) de treinamento (passagem de todos os vetores) pode-se verificar se a rede "aprendeu" o conceito que lhe foi apresentado. Em caso negativo, o aprendizado continua, apresentando os vetores novamente a rede, com os pesos atualizados.

Por ser uma forma simples de RNA, a aplicação do *perceptron* está limitada à aprendizagem de padrões dicotômicos, ou seja, do tipo "0" ou "1", padrão A ou B. Além disso, os padrões devem ser separáveis linearmente.

Como o *perceptron* é uma estrutura muito simples, aplicado apenas a problemas dicotômicos, pode-se utilizar vários deles em forma encadeada na busca da solução de problemas mais complexos. Esse encadeamento de *perceptron*, geralmente é realizado em forma de camadas, tendo o nome de Redes Neurais de Múltiplas Camadas.

Esta arquitetura de RNA pode ser representada graficamente como um grafo, no qual cada nó é um *perceptron* com conexões entre eles (figura 4.12). Para o treinamento desse tipo de rede, é utilizado, em sua grande maioria, o algoritmo *backpropagation* que tem sua base na aprendizagem por correção de erro. Este tem duas fases: a propagação e retropropagação (HAYKIN, 1999).

A primeira fase consiste na propagação, quando é apresentado um padrão à rede, este sinal se propaga (daí o nome desta fase) por todas as camadas até que passe pela camada de saída indicando um valor. Com esse valor obtido na camada de saída, é realizada a retropropagação que emite um sinal de erro da camada de saída para a camada de entrada, com isso ocorre o ajuste dos pesos para que a rede consiga aprender o conhecimento que se quer.

As passagens na RNA destes padrões ocorrem até que algum critério de parada seja satisfeito como: número máximo de épocas, quantidade de padrões classificados corretamente, erro médio quadrático atingido, entre outros.

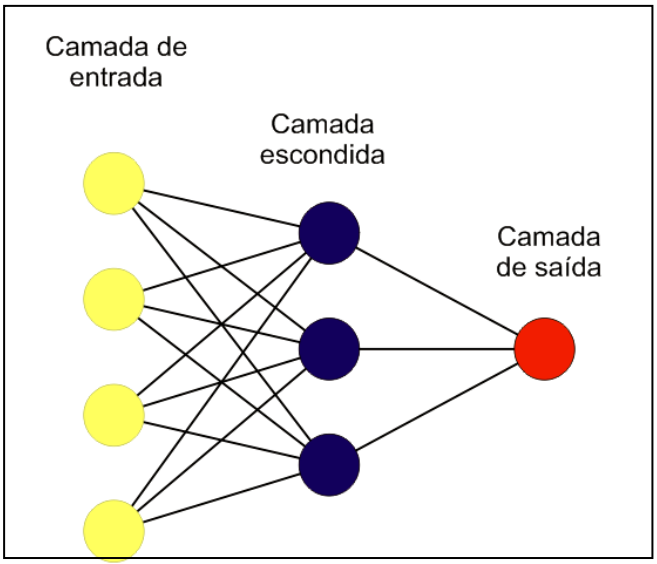


**FIGURA 4.12 –** REDE NEURAL DE MÚLTIPLAS CAMADAS
**FONTE:** O AUTOR (2012)

Muitos são os fatores que diferenciam uma RNA de outra, entre eles temos o número de camadas, quantidades de neurônios em cada camada, função de transferência, algoritmo de aprendizado, entre outros tantos fatores que podem afetar diretamente o aprendizado da RNA.

Uma leitura aprofundada dessa técnica, mostrando a rigorosidade matemática, indicando outros algoritmos de aprendizagem e funções de ativação são apresentadas por Haykin (1999).

### 4.3.1 Considerações sobre a aplicação de RNA

Na aplicação realizada nos estudos de caso desta tese, o algoritmo de aprendizado utilizado é o *backpropagation* e foi implementado em Visual Basic 6.0.

As RNA possuem uma camada de entrada, uma camada escondida e uma camada de saída. Na camada de entrada o número de neurônios é igual ao número de classes $C_i$ do problema em questão. Já na camada de saída há apenas um neurônio e o resultado obtido é comparado com a classificação do padrão, ou seja, se aquele padrão foi classificado corretamente ou não, uma vez que o aprendizado é supervisionado. Para a camada escondida foram realizados testes com *n* neurônios, em que *n* variou de 0 a 20.

Os pesos iniciais, para cada simulação, foram definidos de forma aleatória, no intervalo de (-1 a 1). A função de ativação utilizada foi a sigmoidal-logística e o treinamento da rede era finalizado ao se atingir uma das seguintes três condições: 1.000 iterações; erro médio quadrático menor ou igual a $10^{-4}$; ou número de registros classificados incorretamente, igual a zero.

Desta forma, para cada uma das etapas de aplicação (1ª etapa, 2ª etapa e 3ª etapa da validação cruzada), a rede foi treinada cinco vezes, variando o conjunto de pesos iniciais. Assim, tem-se um total de 1.505 simulações (3 etapas x 5 testes x 20 quantidades de neurônios x 5 faixas de classificação + 5 testes considerando todos os dados, um para cada faixa de classificação).

Após uma RNA ser "treinada" com os dados do conjunto de treinamento e determinada a melhor arquitetura para cada uma das faixas da etiqueta de qualidade em cada etapa de aplicação, os dados do conjunto de teste foram apresentados a RNA, verificando assim o percentual de acerto.

## 4.4 SUPPORT VECTOR MACHINE

O *Support Vector Machine* (*SVM* - Máquinas de Vetor Suporte) foi proposto pelo russo Vladimir Vapnik em 1979, tendo sua utilização grande eficiência na área de reconhecimento de padrões, mostrando alguns resultados superiores a outras técnicas, como a de RNA (Sung e Mukkamala, 2003; e Ding e Dubchak, 2001).

Essa técnica foi desenvolvida para a classificação binária, ou seja, dado o conjunto de treinamento, onde as entradas são representadas por $x_i \in \Re^n$ e as respectivas saídas por $y_i \in \{0,1\}$.

Em linhas gerais, o *SVM* procura uma função (reta no caso $\Re^2$, plano no caso $\Re^3$ e hiperplano no caso de $\Re^n$, com $n \geq 4$) que possua a mesma distância para os elementos de ambas as classes controlando, assim, a capacidade da função de decisão na busca.

A figura 4.13 apresenta exemplos em $\Re^2$ com conjuntos de dados linearmente separáveis – mesma característica dos dados desta tese. O classificador $h_1$ (figura 4.13a) não separa os dois conjuntos (classes), $h_2$ (figura 4.13b) separa tais conjuntos, mas com margem mínima, e $h_3$ (figura 4.13c) separa as classes com margem máxima.

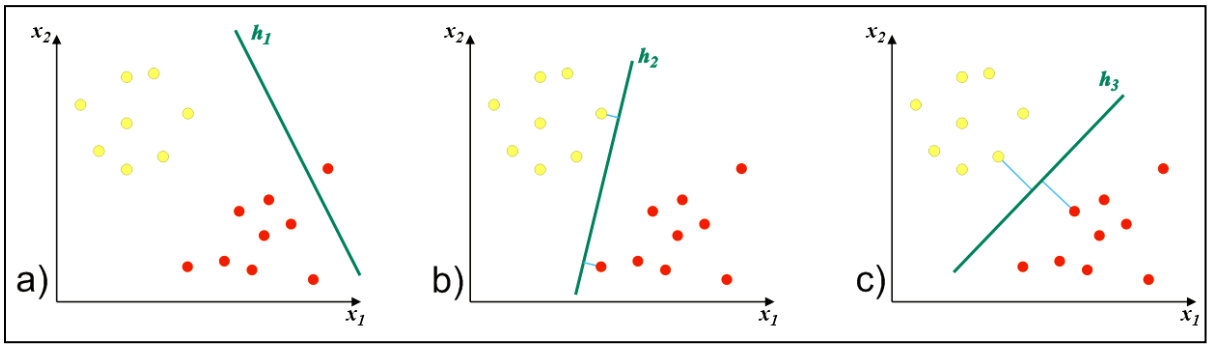


**FIGURA 4.13 –** EXEMPLO NA PROCURA DA FUNÇÃO DE DECISÃO ÓTIMA
**FONTE:** O AUTOR (2012)

Assim, os vetores de suportes (a e b na figura 4.14) são encontrados na fase de treinamento e definem a função de decisão ótima, sendo esta o classificador procurado.

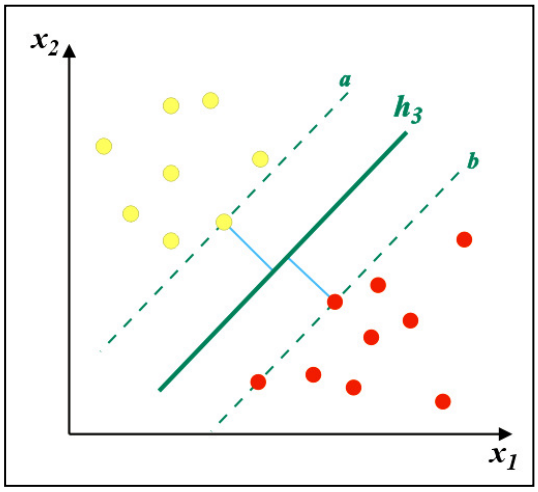


**FIGURA 4.14 –** VETORES SUPORTES E FUNÇÃO DE DECISÃO
**FONTE:** O AUTOR (2012)

No caso de conjuntos em que os dados não são separáveis linearmente, o *SVM* utiliza a função Kernel (função presente no algoritmo que calcula a função de classificação) para projetar estes dados em um espaço em que seja possível separá-los linearmente, denominado de espaço de características com uma dimensão mais elevada. Assim, o descrito acima pode ser resolvido pelo *SVM*, uma vez que a função de decisão é separável linearmente apenas no espaço de característica e não no espaço de entrada dos padrões (figura 4.15).

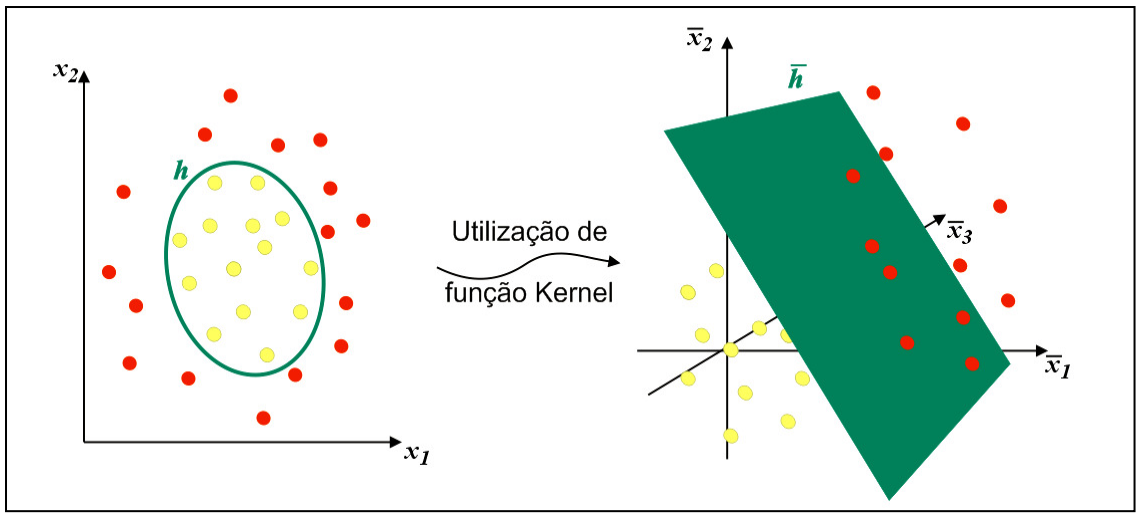


**FIGURA 4.15 –** ESPAÇO DE ENTRADA DOS PADRÕES E ESPAÇO DE CARACTERÍSTICAS
**FONTE:** O AUTOR (2012)

Dessa forma, esta subseção explicou, sem a rigorosidade matemática envolvida, como a técnica *SVM* procede na busca do classificador. Uma leitura mais aprofundada pode ser realizada em Vanipk (1995 e 1998) e Burges (1998).

### 4.4.1 Considerações sobre a aplicação de *SVM*

Na aplicação realizada nos estudos de caso desta tese foi utilizada primeiramente a função *svmtrain,* do *software Matlab* 7.9.0, com duas matrizes nos argumentos: Exemplos e Resposta. (Eq. 4.1)

$$Treino = svmtrain(Exemplos, Resposta) \quad \text{(Eq. 4.1)}$$

A matriz “Exemplos” possui em suas colunas os valores de $C_i$ e a matriz “Respostas” possui apenas uma coluna com o valor da faixa que cada padrão (linha da matriz “Exemplos”) tem como resposta.

Na sequência, foi utilizado o conjunto de teste, aqui escrito em forma de matriz denominada “NovosExemplos”, e o resultado de “Treino” com a função *svmclassify* (Eq. 4.2), com a finalidade de verificar o percentual de classificação corretas destes novos dados.

$$Classificação = svmclassify(Treino, NovosExemplos) \quad \text{(Eq. 5.2)}$$

Cabe ressaltar, que os argumentos utilizados no treinamento para a função *svmtrain* são os *defaults* do *Matlab* 7.9.0, uma vez que os conjuntos das faixas da etiqueta de qualidade são separáveis linearmente (caso $\Re^2$), por plano (caso $\Re^3$) e por hiperplano (no caso de $\Re^n$, com $n \geq 4$).

## 4.5 ALGORITMOS GENÉTICOS

Algoritmo Genético (AG ou *Genetic Algorithm*) é um algoritmo de busca baseado nos mecanismos de seleção natural e genética natural das espécies da teoria de Darwin, na qual os melhores indivíduos, os mais adaptáveis, sobrevivem (GOLDBERG, 1989). Por se basear nessa teoria os termos utilizados referem-se à terminologia biológica, entre elas têm-se: cromossomo, que se refere a cada cadeia de *bits* que representa uma solução para o problema; população é o conjunto de cromossomos no espaço de busca; geração é uma iteração completa do AG que produz uma nova população; e aptidão, que indica o quanto o indivíduo está

adaptado para prosseguir ou não com ele para a busca da solução, este grau de aptidão é chamado de *fitness*.

As principais diferenças entre o AG e os demais métodos, são: trabalha com uma população e não um único ponto na busca do melhor indivíduo; não trabalha com os parâmetros e sim com uma codificação do conjunto destes; e utiliza regras probabilísticas e não determinísticas.

O algoritmo mantém uma população de estruturas chamadas de indivíduos, as quais representam as possíveis soluções para determinado problema (GREFENSTETTE, 1986). A cada iteração os indivíduos gerados através de uma combinação, realizada por uma função de avaliação com uma estrutura de informações alteradas aleatoriamente, são avaliados quanto ao grau de adaptabilidade à população. Com base nesse processo gera-se uma nova população de possíveis soluções utilizando operadores genéticos. E assim, ao longo de cada geração pressupõe-se que os indivíduos melhor se adaptem à população e convirjam para uma boa solução do problema. De acordo com algum critério de parada (número de iterações, satisfação de restrições, e outros), o indivíduo mais apto é a solução do problema.

Os operadores genéticos frequentemente utilizados são: mutação, *crossover* e seleção. Estes são os responsáveis por todas as transformações sofridas pela população, mas possuem funções bastante distintas.

O operador genético *mutação* é a modificação da informação genética de um gene (figura 4.16), gerando um novo indivíduo (denominado de “filho”) sendo o fator fundamental para garantir a biodiversidade, permitindo que o espaço de busca seja explorado a partir de uma parte de sua extensão. Esse operador tem o papel fundamental de diminuir a probabilidade de uma convergência prematura da solução, que ocorre quando a população se estabiliza com uma adaptação pouco adequada.

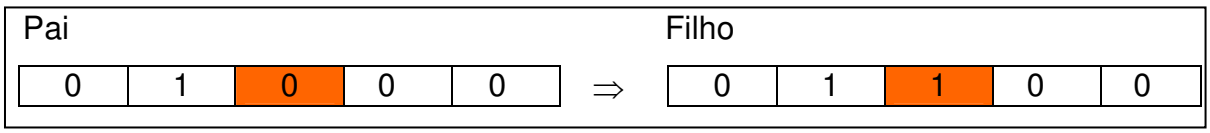


**FIGURA 4.16 –** OPERADOR GENÉTICO MUTAÇÃO
**FONTE:** O AUTOR (2012)

O operador genético *crossover* cria novos indivíduos para a população através da recombinação de partes diferentes de dois indivíduos selecionados (“pais”; figura 4.17). Primeiramente escolhe-se um ponto de *crossover* aleatoriamente (para *crossover* simples de um ponto) e dois novos indivíduos são gerados (filhos) - uma variação bastante usual desse operador é utilizar dois ou mais pontos de *crossover*. Seja qual for o esquema, deve ser mantida a integridade dos genes representados pelo cromossomo. Esse operador gera novos indivíduos com partes de outros dois denominados (Pais).

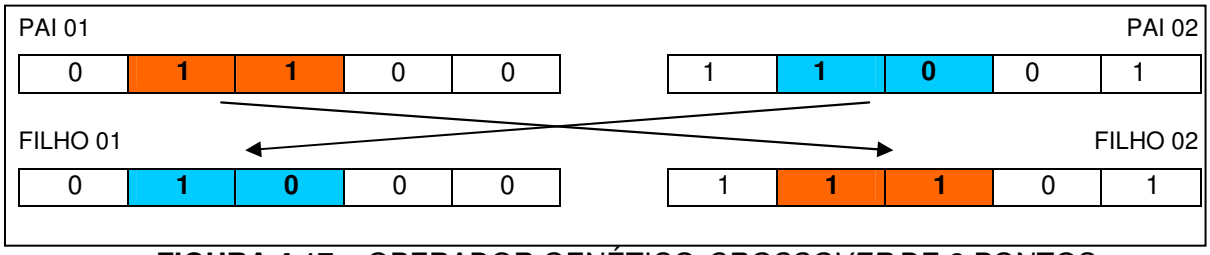


**FIGURA 4.17 –** OPERADOR GENÉTICO *CROSSOVER* DE 2 PONTOS
**FONTE:** O AUTOR (2012)

Os operadores genéticos citados influenciam diretamente no desempenho do AG, pois quanto maior for a taxa de *crossover*, uma maior diversidade de indivíduos será gerada. No entanto, um número elevado dessa taxa poderá ocasionar a perda de estruturas importantes na busca da solução; em contrapartida se o número é baixo, o AG torna-se lento. Se a taxa de mutação for alta, há uma busca aleatória, mas se for baixa não previne que a população fique estagnada.

Com relação à seleção, muitos autores, como Holland (1992) e Hoffmeister e Bäck (1990), não a consideram como um operador genético, mas sim como uma parte integrante do processo. Dentre os métodos de seleção do indivíduo para a próxima geração se destacam *Roulett Wheel*, Seleção Salvacionista e Seleção Baseada em *Rank*.

A seleção dos indivíduos para a aplicação dos operadores irá identificar os indivíduos mais adaptados, assim, a escolha do método é de extrema importância. Além disso, a porcentagem dos indivíduos que será substituída na próxima geração é importante, pois um valor alto altera em demasiado a população e um valor baixo deixa a população estagnada.

Aplicar os operadores e forma de seleção não é tarefa tão fácil no AG. O ponto mais difícil é a percepção ao modelar um problema para esta metaheurística que influencia diretamente em seu comportamento. Também o tamanho da

população é fundamental na determinação do espaço de busca, uma vez que quanto maior a população, mais pontos existirão no espaço de busca e maior será o custo computacional; se a população é pequena, oferece pouca cobertura do espaço de busca, causando uma queda no desempenho. Portanto, saber equilibrar o tamanho da população é essencial para o bom desempenho do AG.

Os trabalhos mais relevantes e, consequentemente, indicados para uma leitura mais profunda sobre a teoria de AG são Holland (1992) e Goldberg (1989).

### 4.5.1 Considerações sobre a aplicação de AG

Na aplicação realizada nos estudos de caso desta tese, o AG foi utilizado com a finalidade de determinar um hiperplano de $R^n$ (reta, se $n=2$; plano, se $n=3$; e hiperplano de $R^n$, se $n \geq 4$) de tal forma que em cada um dos subespaços determinado pelo hiperplano contenha apenas um dos conjuntos de cada etapa de aplicação, conforme a metodologia apresentada na seção 4.2.

O valor da função *fitness* é proveniente de um algoritmo que determina pontos que definam tal hiperplano, no qual as coordenadas de cada ponto são alelos dos indivíduos.

Cada indivíduo é composto de $n^2$ alelos com valores pertencentes ao conjunto dos números reais, onde $n$ é o número de classes $C_i$. Assim, os $n$ primeiros alelos representam as coordenadas de um ponto denominado $P_1$, os próximos $n$ alelos são as coordenadas do ponto $P_2$, assim sucessivamente, em que os $n$ últimos alelos são as coordenadas do ponto $P_n$. Há também o cálculo do *fitness* que leva em consideração a diferença das distâncias entre dois pontos (em conjuntos diferentes) mais próximos do hiperplano determinado. Quanto maior for a diferença entre as distâncias, maior será a penalidade aplicada no *fitness*.

Assim, a figura 4.18 apresenta este algoritmo, no qual $X$ é um vetor em que cada coordenada representa um alelo do indivíduo da população, $CL1$ e $CL2$ são os conjuntos para treinamento e *k* um elemento pertencente a $CL1 \cup CL2$.

*Defina* $P_1 = [X(1)\ X(2)\ \ldots\ X(n)]$; $P_2 = [X(n+1)\ X(n+2)\ \ldots\ X(n+n)]$; ....

$P_n = [X(n*(n-1)+1)\ X(n*(n-1)+2)\ \ldots\ X(n^2)]$.

*Determine a equação do hiperplano* $\alpha$ *que contém* $P_1, P_2 \ldots P_n$.

*Para cada elemento k*

*Substitua as variáveis da equação do hiperplano* $\alpha$ *pelos valores de k, obtendo a variável Valor.*

*Calcule a distância euclidiana de k a* $\alpha$*, obtendo a variável Dist.*

*Se k*$\in$ *CL1, então*

*Se Valor<0, então correto=correto+1;*

*Se Dist01>Dist, então Dist01 =Dist;*

*Se k*$\in$ *CL2, então*

*Se Valor>0, então correto=correto+1;*

*Se Dist02>Dist, então Dist02 =Dist;*

*z1= correto / número de exemplos k;*

*z2=modulo (Dist1 – Dist02)*penalidade;*

*Fitness de X = z1-z2;*

**FIGURA 4.18 –** PSEUDOCÓDIGO PARA CÁLCULO DO *FITNESS*
**FONTE:** O AUTOR (2012)

Para aplicar o AG foi utilizada a penalidade de 0,1 e a *toolbox* do *Matblab* 7.9.0: *gatool.* Os argumentos para o treinamento foram os *defaults* que obtiveram os melhores resultados em ambos estudos de caso, sendo alguns expostos a seguir: *Population type: Double Vector; Population size:* 20*; Fitness scaling: rank; Selection function: Stochastic uniform; Crossover fraction:* 0,8*; Crossover function: Scattered; Migration – direction: Forward; Stopping criteria – Generations:* 100*; Stopping criteria – Stall generations:* 50*;* e *Stopping criteria – Function tolerance:* 1e-6*.*

## 4.6 TÉCNICAS QUE UTILIZAM DISTÂNCIA EUCLIDIANA

São realizados quatro testes que utilizam a distância euclidiana, como se segue: somatório das distâncias do novo elemento aos elementos de cada faixa da etiqueta de qualidade; distância do novo elemento ao ponto central de cada faixa da etiqueta de qualidade; *k*-vizinhos mais próximos; e distância do novo elemento a média dos elementos de cada faixa da etiqueta de qualidade. Todos os testes foram realizados no *software* Excel.

### 4.6.1 Somatório das distâncias do novo elemento aos elementos de cada faixa da etiqueta de qualidade

Nesta aplicação para cada uma das seis faixas da etiqueta de qualidade foi realizado o somatório das distâncias do novo elemento (ponto de teste ou novo padrão) aos elementos que compõe a faixa. Com isso foram obtidos seis somatórios, sendo que o menor deles determina a classificação do novo padrão.

A técnica foi utilizada pelo fato de termos para cada faixa da etiqueta um conjunto com bom comportamento (exemplos: em $\Re^2$, um retângulo; em $\Re^3$ um paralelepípedo) e todos com mesma quantidade de elementos (seção 4.2).

### 4.6.2 Distância do novo elemento ao ponto central de cada faixa da etiqueta de qualidade

O segundo teste determina um elemento $P$, denominado de ponto central, tal que o somatório das distâncias de $P$ aos elementos de cada faixa é mínimo.

Para determinar o $P$ de cada uma das faixas de classificação da etiqueta, foi formulado o modelo matemático e resolvido no *software* Lingo 6.0. Na sequência foi calculada a distância de cada um dos seis $P$ ao novo elemento, onde a menor distância determina a classe mais próxima, ou seja, determina a classificação de $P$.

A figura 4.19, mostra que determinado o ponto central dos três conjuntos, representada pelos elementos coloridos, o elemento de teste está mais próximo do ponto central vermelho (a circunferência de centro no elemento teste contém apenas este ponto, os demais estão a uma distância maior).

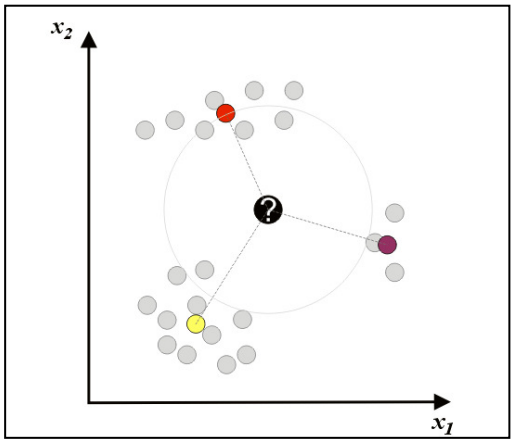


**FIGURA 4.19 –** REPRESENTAÇÃO DA SEGUNDA TÉCNICA – DISTÂNCIA DO PONTO DE TESTE AO PONTO CENTRAL DE CADA CLASSE
**FONTE:** O AUTOR (2012)

### 4.6.3 *K*-vizinhos mais próximos

O terceiro teste é realizado verificando a classificação dos $k$ elementos mais próximos do elemento de teste, sendo sua classificação definida como a de maior ocorrência entre os $k$ elementos. Caso haja empate no número de ocorrências, é verificada a classificação entre os $k+l$ elementos, em que $l$ é o número de elementos necessários para o desempate.

Para definir o valor de $k$ utilizado em cada estudo de caso, foram realizados testes com os mesmos conjuntos de treinamento e de teste da validação cruzada das técnicas de RNA, *SVM* e AG. Os valores de $k$ variaram de 1 ao número de elementos de cada faixa do conjunto de treinamento, e para cada elemento $x$ do conjunto de teste foram verificados os $k$ elementos do conjunto de treinamento mais próximos de $x$. Assim, o valor de $k$ é definido como o menor número que obteve maior acerto de classificação.

A figura 4.20 apresenta a técnica descrita acima para $k=3$, na qual o resultado da classificação do ponto de teste é a classe vermelha, que possui dois elementos mais próximos, e a classe amarela, que possui apenas um elemento.

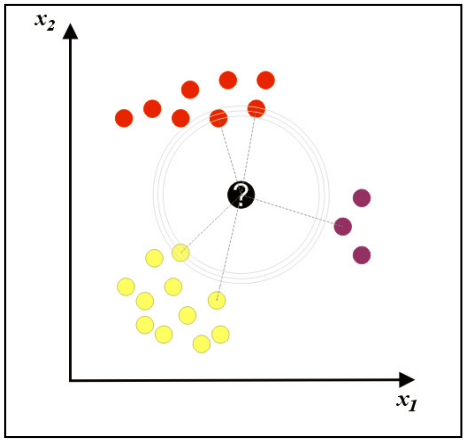


**FIGURA 4.20 –** REPRESENTAÇÃO DA TÉCNICA DOS *K*-VIZINHOS MAIS PRÓXIMOS, PARA *K=3*
**FONTE:** O AUTOR (2012)

### 4.6.4 Distância do novo elemento a média dos elementos de cada faixa da etiqueta de qualidade

O quarto teste calcula a média dos elementos de cada conjunto que representa cada faixa da etiqueta de qualidade e na sequência realiza o cálculo da distância do novo padrão a cada uma das seis médias. Assim, determinada as distâncias, o elemento teste possui a mesma classificação da média mais próxima.

Foi decidido aplicar tal técnica após a verificação que a média de cada conjunto e o ponto central do mesmo conjunto são quase coincidentes, mas a técnica que considera as médias possui tempo computacional menor.

## 4.7 TÉCNICA ESTATÍSTICA: DISTÂNCIA DE MAHALANOBIS

A distância de Mahalanobis (Mahalanobis, 1936) é uma medida de distância baseada nas correlações entre as variáveis, em que podem ser identificados e analisados padrões, ou seja, é utilizada para determinar similaridade entre uma amostra e outra.

Essa medida se distingue da distância euclidiana pelo fato de realizar as correlações entre os conjuntos de dados e não varia conforme a escala adotada. No entanto, assim como na distância euclidiana, é necessário que os elementos possuam a mesma dimensão.

Segundo Santos (2006), a distância de Mahalanobis considera o espalhamento dos padrões, formando um elipsóide, de uma determinada classe.

Dessa forma, ao utilizar esta distância em problemas de classificação deve-se realizar o cálculo do ponto de teste às $N$ classes e, para a classe que obtiver menor resultado, será atribuída à classificação do novo elemento. (Figura 4.20)

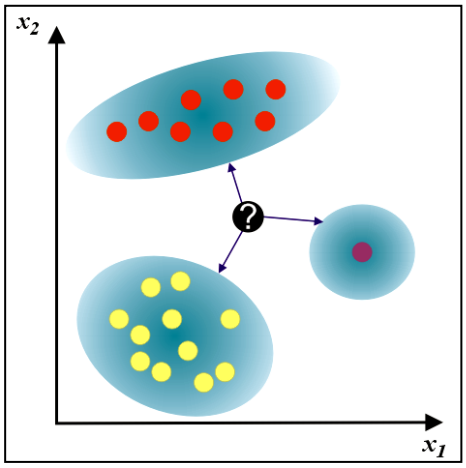


**FIGURA 4.21 –** REPRESENTAÇÃO DA DISTÂNCIA DE MAHALANOBIS DE UM PONTO A TRÊS CLASSES DISTINTAS
**FONTE:** SANTOS (2006) - MODIFICADO PELO AUTOR (2012)

A distância Mahalanobiana utiliza dois elementos de cada classe: o vetor médio dos padrões $m_j$ e a matriz de covariância $P_j$. Além disso, para seu cálculo é necessário conhecer o vetor $x$, referente ao padrão que se quer classificar (ponto teste). Sendo assim, seu cálculo é dado pela equação (Eq. 4.1) a seguir.

$$d = \sqrt{(x - m_j)^T P_j^{-1} (x - m_j)} \qquad \text{(Eq. 4.1)}$$

Para a aplicação dessa técnica foi utilizado o *software* Matblab 7.9.0, função *mahal*.

# 5 ESTUDO DE CASO 01: ÁREA ELÉTRICA

Neste capítulo são apresentadas algumas considerações sobre Qualidade de Energia Elétrica tendo como base documento da Agência Nacional de Energia Elétrica. Na sequência, é descrita a aplicação da metodologia proposta neste trabalho com o intuito de apresentar uma etiqueta de qualidade que classifique os alimentadores de uma subestação de energia elétrica em relação aos afundamentos momentâneos de tensão.

## 5.1 A QUALIDADE DE ENERGIA ELÉTRICA

As empresas dos mais diversos ramos da economia buscam a melhoria em seus serviços, pois a prestação de um serviço de alto nível é o fator crucial para a competitividade. Com isso, tem havido um crescimento no uso de equipamentos computadorizados de alta sensibilidade, bem como a discussão de Qualidade de Energia Elétrica (QEE) que estes necessitam.

Entende-se como “boa” qualidade de energia não somente a não interrupção do fornecimento de energia elétrica, mas também o fornecimento de energia cuja tensão seja adequada e com a forma de onda senoidal com poucas harmônicas, além de aspectos que envolvem valores de tensão, frequência e desequilíbrio entre fases.

Os eventos que ocorrem no sistema elétrico, devido aos mais diversos motivos, influenciam diretamente na QEE. Esses eventos podem ser de natureza acidental (queda de galhos de árvores; descargas atmosféricas) ou, então, programadas (manutenção preventiva), gerando, assim, fenômenos de ordem comumente associados à QEE.

Os danos causados pela falta da QEE são diversos, dentre os quais podem ser destacados (OLESKOVICZ *et al.*, 2006): redução do tempo de vida do equipamento; mau funcionamento de dispositivos de proteção; e interrupções na distribuição da energia, causando prejuízos às empresas e usuários de uma forma geral.

Como a qualidade de produtos e serviços prestados tem sido uma exigência crescente do consumidor, principalmente quando respaldados pelas Agências Reguladoras do Governo Federal ou Estadual, as concessionárias de energia têm buscado por soluções que possam satisfazer seus clientes quanto à melhoria da QEE e ao mesmo tempo cumprir as exigências da Agência Nacional de Energia Elétrica (ANEEL), criada em 1997.

No documento que trata da QEE, a ANEEL (2008) define terminologia e caracteriza os eventos (tensão em regime permanente, fator de potência, harmônicos, desequilíbrio de tensão, flutuação de tensão, variações de tensão de curta duração e variação de freqüência), bem como os valores de referências. Também estabelece metodologias para a verificação de indicadores quanto aos serviços prestados.

Quanto à variação de tensão de curta duração, evento da QEE utilizado no estudo de caso desta tese, a ANEEL (2008) a define como sendo os desvios significativos no valor eficaz da tensão em curtos intervalos de tempo. Dentre as classificações apresentadas, destacamos o afundamento momentâneo de tensão.

Os afundamentos momentâneo de tensão possuem duração superior ou igual a um ciclo e inferior ou igual a três segundos, com valor eficaz de tensão superior a 10% e inferior a 90%.

A ANEEL (2008) não define padrões de desempenho em relação a esses fenômenos, mas indica que “as distribuidoras devem acompanhar e disponibilizar, em bases anuais, o desempenho das barras de distribuição monitoradas”, pois estas informações podem servir como referência para o desempenho das barras das unidades consumidoras atendidas pelo Sistema de Distribuição de Alta Tensão ou pelo Sistema de Distribuição de Média Tensão com cargas sensíveis a variações de tensão de curta duração.

Ainda neste documento, há a classificação quanto à tensão verificada nas unidades consumidora em relação à tensão contratada, sendo classificada em três faixas, que para a subestação analisada no estudo de caso desta tese são: *tensão adequada* - de 95% a 100% da tensão contratada; *tensão precária* - 0,93% a 0,95% ou de 105% a 107% da tensão contratada; e *tensão crítica* - inferior a 93% ou superior a 107% da tensão contratada. Se comparados a tensão em regime

permanente todos os eventos “afundamento momentâneo de tensão” são classificados como tensão crítica. Com isso, a metodologia para a criação de etiqueta de qualidade que classifique cada alimentador de uma subestação em relação aos afundamentos de tensão apresentada neste trabalho pode preencher esta lacuna se aplicada a subestações e não a alimentadores.

Nesta tese optou-se por classificar alimentadores por dois motivos: *i)* a base de dados fornecidos pela concessionária de energia elétrica não dispunha de dados suficientes para aplicação em subestações; *ii)* temos a pretensão de utilização de dados reais e não fictícios/simulados.

## 5.2 CRIAÇÃO DA ETIQUETA DE QUALIDADE PARA CLASSIFICAR OS ALIMENTADORES DE UMA SUBESTAÇÃO EM RELAÇÃO AOS AFUNDAMENTOS MOMENTÂNEOS DE TENSÃO

Para este estudo de caso os **dados foram selecionados** de duas bases de dados de uma concessionária de energia elétrica brasileira. Os registros (eventos) da primeira base de dados (BD01 – alguns registros constam no Anexo 01) referem-se aos afundamentos de tensão e são obtidos por um dispositivo instalado na subestação que fornece, dentre outros, os 17 atributos que constam no quadro 5.1.

A segunda base de dados (BD02 – alguns registros constam no Anexo 02) registra as interrupções e são obtidas através de um *software* da própria concessionária que fornecem, dentre outros, os 29 atributos apresentados no quadro 5.2.

Os dados são provenientes de um período de quatro meses, de fevereiro a maio de 2008, no qual BD01 ficou constituída de 352 registros e a BD02, de 422 registros.

**QUADRO 5.1 –** DESCRIÇÃO DOS ATRIBUTOS DO BD01

| ATRIBUTO | | TIPO | DESCRIÇÃO |
|---|---|---|---|
| Identificação da Oscilografia | | Numérico | Enumera o registro da oscilografica, que pode conter vários registros de eventos; assim vários registros possuem o mesmo número, mas seguem ordem cronológica. |
| Nome do medidor | | Numérico | Indica a subestação a qual o registro se refere. |
| Data de início | | Data e hora | Data e hora (com milissegundos) do início do registro. |
| Tipo do evento | | Alfanumérico | Descreve os eventos ocorridos (*Sag* / Distorção Total / Desequilíbrio / *Swell*) com suas respectivas Fases (A, B ou C). |
| Data do evento | | Data e hora | Indica o final do(s) evento(s). |
| Circuito | | Numérico | Circuito em que ocorreu o evento. |
| Status do evento | | Alfanumérico | Evento que continua a ocorrer. |
| Tensão *RMS* (*Root Mean Square*) | Fase A | Numérico | Indica a tensão na respectiva fase. |
| | Fase B | | |
| | Fase C | | |
| *THD* (*Total Harmonic Distortion*) | Fase A | Numérico | Relação entre a magnitude das componentes harmônicas e a fundamental de cada fase. Ideal em zero. |
| | Fase B | | |
| | Fase C | | |
| Frequência | Fase A | Numérico | Periodicidade da componente fundamental de cada fase. Em geral 60Hz, 50Hz. |
| | Fase B | | |
| | Fase C | | |
| Desequilíbrio do circuito | | Numérico | Resultante da soma fasorial entre as fases A, B, C, idealmente nula. |

**QUADRO 5.2 –** DESCRIÇÃO DOS ATRIBUTOS DO BD02

| ATRIBUTO | | TIPO | DESCRIÇÃO |
|---|---|---|---|
| Alimentador | Cod_Alim | Numérico | Código do Alimentador onde ocorreu a interrupção. |
| | Desc_Alim | Alfanumérico | Código do Alimentador onde a interrupção foi gerada. |
| | Num_Oper_Alim | Numérico | Número do Alimentador onde ocorreu a interrupção. |
| Área Elétrica | Cod_Ael | Numérico | Código da Área Elétrica à qual a interrupção está relacionada. |
| | Desc_Ael | Alfanumérico | Descrição da área elétrica à qual a interrupção está relacionada. |
| Causa da Interrupção | Cod_Causai | Numérico | Código da Causa da interrupção de energia. |
| | Desc_Causai | Alfanumérico | Descrição da Causa da interrupção de energia |
| Área da chave | Cod_Chv | Alfanumérico | Significa a característica da área de atendimento de uma chave. |

(Continua)

**QUADRO 5.2** – DESCRIÇÃO DOS ATRIBUTOS DO BD02

(Conclusão)

| ATRIBUTO | | TIPO | DESCRIÇÃO |
|---|---|---|---|
| Componente | Cod_Crd | Numérico | Código do Componente da rede afetado pela interrupção. |
| | Desc_Crd | Alfanumérico | Descrição do Componente da rede afetado pela interrupção. |
| Condição Climática | Cod_Cclima | Numérico | Código da Condição climática registrada no momento da interrupção. |
| | Desc_Cclima | Alfanumérico | Descrição da Condição climática registrada no momento da interrupção. |
| Conjunto ANEEL | Cod_Cea | Numérico | Código do conjunto elétrico ANEEL. |
| | Desc_Cea | Alfanumérico | Descrição do conjunto elétrico ANEEL. |
| Conjunto Concessio-nária | Cod_Cec | Numérico | Código do conjunto elétrico da concessionária. |
| | Desc_Cec | Alfanumérico | Descrição do conjunto elétrico concessionária. |
| Consumidores interrompidos | Qtde_Cons_Intrp | Numérico | Quantidade de consumidores afetados pela interrupção. |
| Data de Início | Data_Inicio | Numérico | Data da interrupção. |
| | Hora_Inicio | Numérico | Horário da interrupção. |
| Duração | Duracao_Intrp | Numérico | Duração da interrupção (minutos). |
| Número da interrupção | Num_Seq_Interp_ Fico | Numérico | Número que identifica a interrupção. |
| Órgão | Num_Org8 | Numérico | Código do Órgão interno da Concessionária responsável pelo alimentador que foi interrompido. |
| | Desc_Org8 | Alfanumérico | Descrição do Órgão interno da Concessionária responsável pelo alimentador que foi interrompido. |
| Origem da Interrupção | Cod_Oint | Alfanumérico | Código da origem da interrupção. |
| Regional | Regional | Alfanumérico | Sigla da Regional à qual a interrupção está relacionada). |
| Subestação | Car_Se | Numérico | Código da subestação a qual está relacionada o alimentador interrompido. |
| | Nome_Se | Alfanumérico | Nome da subestação a qual está relacionada o alimentador interrompido. |
| Tipo | Cod_Tipo | Numérico | Código do tipo de interrupção. |
| | Desc_Tipo | Alfanumérico | Descrição do tipo de interrupção. |

Na BD01 foi realizada a **limpeza ou pré-processamento** após conversa com os engenheiros da concessária de energia elétrica e, com isso, o número de atributos da BD01 foi reduzido de 17 para 9, apresentados no quadro 5.3 em que o atributo “Nome do medidor” foi excluído, visto que a metodologia é aplicada sempre

para uma subestação e a justificativa da exclusão dos atributos "*THD*", "Frequência" e "Desequilíbrio do circuito", está no fato de que este estudo de caso considera apenas os Afundamentos Momentâneos de Tensão.

**QUADRO 5.3 –** DESCRIÇÃO DOS ATRIBUTOS DO BD01 APÓS PRÉ-PROCESSAMENTO DOS DADOS

| **ATRIBUTO** | | **TIPO** | **DESCRIÇÃO** |
|---|---|---|---|
| Identificação da Oscilografia | | Numérico | Enumera o registro da oscilografia, que pode conter vários registros de eventos; assim vários registros possuem o mesmo número, mas seguem ordem cronológica. |
| Data de início | | Data e hora | Data e hora (com milissegundos) do início do registro. |
| Tipo do evento | | Alfanumérico | Descreve os eventos ocorridos (*Sag* / Distorção Total / Desequilíbrio / *Swell*) com suas respectivas Fases (A, B ou C). |
| Data do evento | | Data e hora | Indica o final do(s) evento(s). |
| Circuito | | Numérico | Circuito em que ocorreu o evento. |
| Status do evento | | Alfanumérico | Evento que continua a ocorrer. |
| Tensão *RMS* (*Root Mean Square*) | Fase A | Numérico | Indica a tensão na respectiva fase. |
| | Fase B | | |
| | Fase C | | |

Com relação ao pré-processamento da BD02, o número de atributos foi reduzido de 29 para 12. Muitos dos atributos são do tipo "código do atributo" e "descrição do atributo" (este último refere-se ao primeiro) e outros são referentes a número do evento (protocolo), ou seja, os primeiros são atributos que fornecem a mesma informação e o último fornece um número diferente para cada evento. Além disso, tendo-se em vista o objetivo de "criar uma etiqueta de qualidade que classifique cada alimentador de uma subestação em relação aos afundamentos de tensão", foi verificado que os atributos necessários desta base de dados são apenas seis: alimentador; componente afetado; data de início; hora de início; duração; e tipo.

Em reunião com engenheiros eletricistas da concessionária, estes informaram que devem ser considerados apenas os registros em que o atributo "Tipo" seja "Acidental". Então, houve nova filtragem na BD02 e os cinco atributos a serem considerados constam no quadro 5.4. Como consequência desta nova "limpeza", tem-se que o número de registros nesta base de dados foi alterado para 181.

**QUADRO 5.4 –** DESCRIÇÃO DOS ATRIBUTOS DA BD02 APÓS PRÉ-PROCESSAMENTO DOS DADOS

| ATRIBUTO | TIPO | DESCRIÇÃO |
|---|---|---|
| Alimentador | Alfanumérico | Nome do Alimentador onde a interrupção foi gerada. |
| Componente afetado | Alfanumérico | Descrição do Componente da rede afetado pela interrupção. |
| Data de início | Numérico | Data da interrupção. |
| Hora de início | Numérico | Horário inicial da interrupção. |
| Duração | Numérico | Duração da interrupção (minutos). |

Assim, ao final desta fase do processo *KDD*, dos 17 atributos iniciais da BD01, apenas nove serão utilizados nas próximas fases. Para a BD02, dos 29 atributos apenas cinco são considerados nessa metodologia.

Ao serem analisados os registros na BD01, pôde-se verificar que para o atributo “Identificação da Oscilografia” são enumerados vários registros com mesmo valor, mas em ordem cronológica. Esses registros se referem a uma determinada configuração do aparelho para detectar o início de eventos relacionados à QEE, bem como aos eventos que estão ocorrendo naquele momento. Desta forma, foi necessária a **transformação dos dados**.

Com a finalidade de ilustrar estes registros, alguns deles são apresentados, de forma resumida, no quadro 5.5, que possui nove colunas. Na 1ª coluna, da 1ª à 3ª linhas, o atributo “Identificação da Oscilografia” possui registro “9”. Na 1ª destas três linhas, o atributo “Tipo do evento”, correspondente a “Data de início” (2ª coluna), está indicando que o registro inicia com uma *swell* (elevação de tensão) na fase A e distorção total nas três fases (A, B e C). O atributo “Status do evento” (5ª coluna) indica que, no momento “Data do Evento” (4ª coluna), está ocorrendo uma *swell* na fase A e, também, distorção nas três fases, além de um desequilíbrio de tensão.

**QUADRO 5.5 –** EXEMPLOS DE ALGUNS REGISTROS DA BD01

| ID. OSC. | DATA INICIO | TIPO DO EVENTO | DATA DO EVENTO | STATUS DO EVENTO | CIRC. | RMS FASE A | RMS FASE B | RMS FASE C |
|---|---|---|---|---|---|---|---|---|
| 9 | 2008-02-06 07:28:35.034 | Swell - Fase A; Distorção Total - Fase A; Distorção Total - Fase B; Distorção Total - Fase C; | 2008-02-06 07:28:35.199 | Swell - Fase A; Distorção Total - Fase A; Distorção Total - Fase B; Distorção Total - Fase C; Desequilíbrio de tensão | 0 | 76,4 | 63,7 | 60,1 |
| 9 | 2008-02-06 07:28:35.034 | Swell - Fase A; Desequilíbrio | 2008-02-06 07:28:35.216 | Distorção Total - Fase A; Distorção Total - Fase B; Distorção Total - Fase C | 0 | 65,5 | 65,2 | 64,9 |
| 9 | 2008-02-06 07:28:35.034 | Distorção Total - Fase A; Distorção Total - Fase B; Distorção Total - Fase C | 2008-02-06 07:28:35.232 | Swell - Fase A | 0 | 65,3 | 65,4 | 64,9 |
| 10 | 2008-02-06 20:04:14.805 | Sag – Fase C; Desequilíbrio | 2008-02-06 20:04:14.973 | Sag - Fase C; Desequilíbrio | 1 | 69,8 | 73,2 | 57,6 |
| 10 | 2008-02-06 20:04:14.805 | Swell - Fase A; Swell - Fase B; Distorção Total - Fase B; Distorção Total - Fase C | 2008-02-06 20:04:14.990 | Swell - Fase A; Swell – Fase B; Distorção Total - Fase B; Sag - Fase C; Distorção Total - Fase C; Deseq. de tensão | 1 | 82,9 | 77,1 | 35,9 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

A partir desta análise, quando um registro possui evento do tipo "*sag*" (afundamento de tensão), todos os registros com mesma "Identificação da Oscilografia" permanecem na BD01. Cada grupo de registros com mesma "Identificação da Oscilografia" é transformado em um único registro, diminuindo não somente o número de registros da BD01, mas também a quantidade de atributos, visto que "Tipo do evento" e "Status do Evento" não são mais necessários. Para o atributo "Data do evento" é considerado o registro que possui o horário "mais distante" do registrado em "Data do início", gerando, assim, um novo atributo denominado "Duração", em milissegundos, que indica o tempo para estabilização de do evento.

O conteúdo do atributo "Data do Início", contendo data e hora do registro, foi dividido em outros dois, denominados "Data Início" e "Hora Início". O mesmo ocorreu com "Data do Evento", que ficaram denominados "Data do Evento" e "Hora do Evento".

Outra transformação nos dados da BD01 se refere aos atributos que indicam a tensão remanescente em cada uma das fases, em que foi realizada a agregação por parâmetros, uma alternativa para a agregação de fases (ANEEL, 2008), e a "duração do evento é definida como a máxima duração entre os três eventos fase-neutro e o valor de magnitude que mais se distanciou da tensão de referência". Portanto, para cada grupo de registros com mesma "Identificação de Oscilografia" foi verificada a menor tensão remanescente no grupo, independente da fase. Assim, esse valor foi registrado no novo atributo "Tensão remanescente", e os atributos "Tensão RMS fase A", "Tensão RMS fase A" e "Tensão RMS fase A" foram excluídos da BD01.

Com isso, o quadro 5.6, composto de oito colunas, apresenta o quadro 5.5 reformulado, constando também os novos atributos da BD01 e excluindo os atributos não utilizados após a transformação indicada.

**QUADRO 5.6 –** ALGUNS DOS REGISTROS DA BD01 APÓS A TRANSFORMAÇÃO DOS DADOS

| ID. OSC. | DATA DO INÍCIO | HORA DO INÍCIO | DATA DO EVENTO | HORA DO EVENTO | DURAÇÃO | CIRC. | TENSÃO REMAN. |
|---|---|---|---|---|---|---|---|
| 9 | 2008-02-06 | 07:28:35.034 | 2008-02-06 | 07:28:35.232 | 218 | 0 | 60,1 |
| 10 | 2008-02-06 | 20:04:14.805 | 2008-02-06 | 20:04:14.990 | 185 | 1 | 35,9 |
| ... | ... | ... | ... | ... | ... | ... | ... |

Como já mencionado, a metodologia para a criação da etiqueta de qualidade que classifica alimentadores leva em consideração três atributos: tensão remanescente, duração e quantidade de ocorrências. Os dois primeiros atributos estão na BD01 (quadro 5.6), o terceiro é o resultado da contagem de ocorrências. No entanto, a BD01 não indica qual alimentador foi afetado pelo evento, pois os dados referentes aos alimentadores constam da BD02. Assim sendo, faz-se necessário associar os registros da BD01 com os registros da BD02 e, para isso, foi utilizado o atributo tempo. Mais especificamente foram utilizados os atributos "Data do Início" e "Hora do Início" da BD01 e "Data de Início" e "Hora de Início" da BD02.

Os engenheiros eletricistas da concessionária definiram os intervalos de associação dessas bases de dados da seguinte forma: se na BD02 um registro possui para o atributo "Componente Afetado" a informação "RA" (Religamento Automático da chave) deve-se levar em consideração um intervalo de tempo de no máximo $5min$ em relação a um registro da BD01, pois quando se tem uma chave do tipo "RA", as bases de dados registram o evento simultaneamente. Caso a informação não seja "RA", mas sim outros tipos de chaves, o intervalo a ser considerado é de até $2h$. Esta associação gerou uma nova base de dados, denominada BD03 contendo 169 registros.

O quadro 5.7, composto de 10 colunas, apresenta alguns exemplos/registros dessa associação. As informações das colunas de 1 a 5 são dados provenientes da BD01 e as das colunas 6 a 10 são as suas respectivas associações contidas na BD02. E ainda como forma de identificar os 12 alimentadores desta subestação, estes serão denominados de AA, AB, AC,..., AK, e AL.

Observa-se também no quadro 5.7, que um mesmo registro da BD01, pode ter mais de uma associação com a BD02, como no caso das três primeiras linhas deste quadro, no qual o "Identificação de Oscilografia" é 117. Isto significa que o evento "captado" na subestação, também foi "captado", ou se originou, em dois alimentadores "AC" e "AF", sendo dois registros para o "AF" com componentes afetados diferentes: "ATUAÇÃO DO RA", ou simplesmente "RA", e "ATUAÇÃO DO ELO FUSÍVEL".

Ainda analisando o quadro 5.7, tem-se que nas linhas 2 e 7 o "Componente afetado" é do tipo "RA", e verifica-se que o horário registrado na BD01, na linha 2 deste quadro, é "14:29:43" e na BD02 é "14:30", o que está de acordo com a faixa de variação de no máximo $5min$, sendo que o mesmo é observado na linha 7. Na

linha 4, o “Componente afetado” é “Condutor de AT” e a diferença entre os horários dos registros associados é de $24min$, o que respeita a faixa de variação de até $2h$.

**QUADRO 5.7 –** EXEMPLOS DOS REGISTROS BD03 (ASSOCIAÇÃO DE BD01 COM BD02)

| ID. OSC. | DATA DO INÍCIO | HORA DO INÍCIO | DURA-ÇÃO | TENSÃO REMAN. | ALIMEN-TADOR | COMPONENTE AFETADO | DATA INÍCIO | HORA INÍCIO | DURAÇÃO |
|---|---|---|---|---|---|---|---|---|---|
| 117 | 28/04/2008 | 14:28:43 | 185 | 46,3 | AC | CRUZAMENTO AEREO | 28/04/2008 | 14:30 | 135 |
| 117 | 28/04/2008 | 14:28:43 | 185 | 46,3 | AF | ATUACAO DO RA | 28/04/2008 | 14:29 | 1 |
| 117 | 28/04/2008 | 14:28:43 | 185 | 46,3 | AF | ATUACAO DO ELO FUSIVEL | 28/04/2008 | 14:42 | 41 |
| 121 | 28/04/2008 | 18:07:50 | 202 | 42,6 | AC | CONDUTOR DE AT | 28/04/2008 | 18:32 | 344 |
| 121 | 28/04/2008 | 18:07:50 | 202 | 42,6 | AI | CONDUTOR DE BT | 28/04/2008 | 18:16 | 403 |
| 136 | 02/05/2008 | 07:13:10 | 705 | 28,7 | AF | POSTE | 02/05/2008 | 07:15 | 36 |
| 139 | 08/05/2008 | 11:13:04 | 168 | 42,4 | AI | ATUACAO DO RA | 08/05/2008 | 11:14 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

Com a finalidade de simplificar a BD03, são excluídos os atributos ”Identificação de Oscilografia” (1ª coluna), “Data de início” (2ª coluna) e “Hora de Início” (3ª coluna) referentes à BD01, e os atributos “Data início” (8ª coluna), “Hora Início” (9ª coluna) e “Duração” (10ª coluna) referentes à BD02. Assim, o quadro 5.8, apresenta os atributos que são utilizados para a criação da etiqueta de QEE que classifica o alimentador em relação à subestação na qual está inserido.

**QUADRO 5.8 –** ALGUNS REGISTROS DA BD03 (ASSOCIAÇÕES DE BD01 E BD02)

| DURAÇÃO | TENSÃO REMAN. | ALIMENTADOR |
|---|---|---|
| 185 | 46,3 | AC |
| 185 | 46,3 | AF |
| 185 | 46,3 | AF |
| 202 | 42,6 | AC |
| 202 | 42,6 | AI |
| 705 | 28,7 | AF |
| 168 | 42,4 | AI |

Determinados os registros da BD03, foi construído para cada alimentador um quadro de classificação contendo os atributos: tensão remanescente, a duração e a quantidade de ocorrências.

Estes quadros possuem duas faixas para a duração: menor ou igual 500 milissegundos e maior que 500 milissegundos; e cinco intervalos de tensão remanescente: 10% a 19%, 20% a 39%, 40% a 59%, 60% a 79% e 80% a 90%. O vínculo entre a duração e a tensão remanescente pode ser melhor compreendido observando-se o quadro 5.9, no qual são apresentadas as 10 possíveis classes, aqui denominadas de $C_1$, $C_2$,... a $C_{10}$, valores baseados em Casteren *et al.* (2005).

**QUADRO 5.9 –** CLASSIFICAÇÃO CONSIDERANDO A DURACAO E A TENSÃO REMANSCENTE

| TENSÃO REMAN. | DURAÇÃO | |
|---|---|---|
| | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | $C_1$ | $C_2$ |
| 60 a 79% | $C_3$ | $C_4$ |
| 40 a 59% | $C_5$ | $C_6$ |
| 20 a 39% | $C_7$ | $C_8$ |
| 10 a 19% | $C_9$ | $C_{10}$ |

No quadro 5.9 anterior, deve-se ter em mente de que quanto menor for a duração e maior for a tensão remanescente, menos pior será a QEE daquele evento.

Tem-se, assim que a QEE dos eventos fica hierarquizada da seguinte forma: $C_1 \geq C_2 \geq ... \geq C_n$.

Para exemplificar a referida classificação, os registros do quadro 5.8 estão devidamente classificados, conforme o quadro 5.9, no quadro 5.10.

**QUADRO 5.10 –** CLASSIFICAÇÃO DOS REGISTROS DO QUADRO 5.8 CONFORME QUADRO 5.9

| **DURAÇÃO** Milissegundos | **TENSÃO REMAN.** (%) | **ALIMENTADOR** | **CLASSIFICAÇÃO DO REGISTRO** |
|---|---|---|---|
| 185 | 46,3 | AC | $C_5$ |
| 185 | 46,3 | AF | $C_5$ |
| 185 | 46,3 | AF | $C_5$ |
| 202 | 42,6 | AC | $C_5$ |
| 202 | 42,6 | AI | $C_5$ |
| 705 | 28,7 | AF | $C_8$ |
| 168 | 42,4 | AI | $C_5$ |

Ao realizar esta classificação com os 169 registros da BD03 têm-se que as quantidades de registros dos alimentadores AA e AB são apresentadas nos quadros 5.11 e 5.12. Os registros dos demais alimentadores constam no Anexo 03.

**QUADRO 5.11 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR "AA"

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 2 | 0 |
| 20 a 39% | 1 | 1 |
| 10 a 19% | 0 | 0 |

O quadro 5.11 anterior, apresenta a quantidade de registros do "AA": dois do tipo $C_5$; um do tipo $C_7$ e um do tipo $C_8$. Já o quadro 5.12, a seguir, indica que no alimentador "AB" ocorreram 20 registros $C_5$ e dois do tipo $C_7$.

**QUADRO 5.12 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR "AB"

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 20 | 0 |
| 20 a 39% | 2 | 0 |
| 10 a 19% | 0 | 0 |

Realizado o levantamento de registros por alimentador, tem-se no quadro 5.13 o total de registros classificados em cada uma das 10 faixas, sendo que na subestação analisada, apenas três destas faixas possuem registros: $C_5$, $C_7$ e $C_8$.

**QUADRO 5.13 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DA SUBESTAÇÃO ANALISADA CONSIDERANDO TODOS OS REGISTROS

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 149 | 0 |
| 20 a 39% | 15 | 5 |
| 10 a 19% | 0 | 0 |

Como a subestação analisada possui 12 alimentadores, os valores obtidos em cada classe no quadro 5.13 foram divididos por aquele número (12) obtendo-se, assim, os valores contidos no quadro 5.14 indicam a qualidade que será admitida como sendo a “qualidade média” dos alimentadores da subestação.

**QUADRO 5.14 –** CLASSIFICAÇÃO POR VOTO DOS AFUNDAMENTOS DE TENSÃO NA SUBESTAÇÃO ANALISADA

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 12,41 | 0 |
| 20 a 39% | 1,25 | 0,41 |
| 10 a 19% | 0 | 0 |

No entanto, como estes valores representam as quantidades de registros em cada classe, precisa-se arredondar os valores do quadro 5.14 obtendo-se, desta forma, o quadro 5.15.

**QUADRO 5.15 –** CLASSIFICAÇÃO POR VOTO DOS AFUNDAMENTOS DE TENSÃO NA SUBESTAÇÃO – VALORES ARREDONDADOS

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 13 | 0 |
| 20 a 39% | 2 | 1 |
| 10 a 19% | 0 | 0 |

Uma primeira classificação que pode ser realizada é verificar quais subestações estão acima da média, na média e abaixo da média. No entanto, para criar a etiqueta de QEE foi estabelecido o valor de seis faixas, sendo a “Faixa A” a de melhor qualidade e a “Faixa F” a de pior.

Para cada faixa foram multiplicados fatores, de tal forma que as faixas A e B detenham juntas 50% da faixa superior a média (25% para cada uma), a faixa C os outros 50%. As faixas D e E também possuem o mesmo comprimento da faixa C. Dessa forma, o quadro 5.15 representa a média dos alimentadores, ou seja, o limite superior da “Faixa C”.

O limite superior da “Faixa A” (quadro 5.16) é obtido multiplicando os valores do quadro 5.15 por 0,25. O limite superior da “Faixa B” (quadro 5.17) é obtido multiplicando os valores do quadro 5.15 pelo fator 0,50. Ao multiplicar os valores do quadro 5.15 por 1,5 obtém-se o limite superior da “Faixa D” (quadro 5.18). O limite superior da “Faixa E” (quadro 5.19) é obtido multiplicando pelo fator 2 os valores do quadro 5.15. Por fim, tem-se que o limite superior da “Faixa F” (quadro 5.20) é obtido verificando o maior valor apresentado em cada classe de classificação para os alimentadores analisados.

**QUADRO 5.16 –** LIMITE SUPERIOR DA “FAIXA A” DA ETIQUETA DE CLASSIFICAÇÃO DA QEE DE UM ALIMENTADOR EM RELAÇÃO À SUBESTAÇÃO

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 4 | 0 |
| 20 a 39% | 1 | 1 |
| 10 a 19% | 0 | 0 |

**QUADRO 5.17 –** LIMITE SUPERIOR DA “FAIXA B” DA ETIQUETA DE CLASSIFICAÇÃO DA QEE DE UM ALIMENTADOR EM RELAÇÃO À SUBESTAÇÃO

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 7 | 0 |
| 20 a 39% | 2 | 1 |
| 10 a 19% | 0 | 0 |

**QUADRO 5.18 –** LIMITE SUPERIOR DA "FAIXA D" DA ETIQUETA DE CLASSIFICAÇÃO DA QEE DE UM ALIMENTADOR EM RELAÇÃO À SUBESTAÇÃO

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 13 | 0 |
| 20 a 39% | 2 | 1 |
| 10 a 19% | 0 | 0 |

**QUADRO 5.19 –** LIMITE SUPERIOR DA "FAIXA E" DA ETIQUETA DE CLASSIFICAÇÃO DA QEE DE UM ALIMENTADOR EM RELAÇÃO À SUBESTAÇÃO

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 26 | 0 |
| 20 a 39% | 4 | 2 |
| 10 a 19% | 0 | 0 |

**QUADRO 5.20 –** LIMITE SUPERIOR DA "FAIXA F" DA ETIQUETA DE CLASSIFICAÇÃO DA QEE DE UM ALIMENTADOR EM RELAÇÃO À SUBESTAÇÃO

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 40 | 0 |
| 20 a 39% | 4 | 4 |
| 10 a 19% | 0 | 0 |

A figura 5.1, a seguir, apresenta a etiqueta de classificação da QEE dos alimentadores de uma subestação, de forma comparativa, com os limites de cada faixa definidos anteriormente (quadros 5.15 a 5.20).

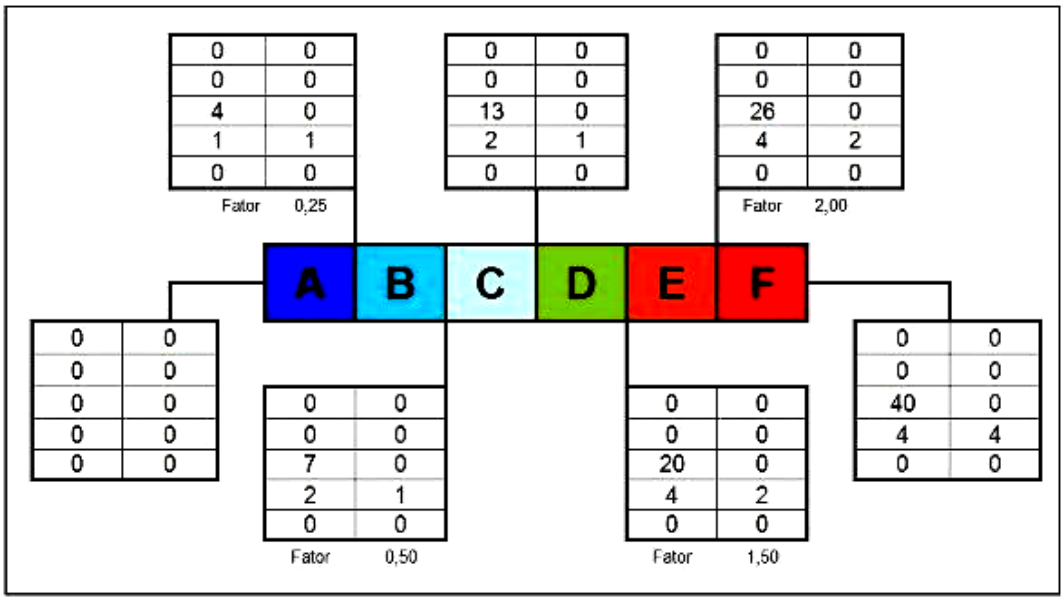


**FIGURA 5.1 –** ETIQUETA DE CLASSIFICAÇÃO DA QEE DOS ALIMENTADORES, DE FORMA COMPARATIVA, DE UMA SUBESTAÇÃO
**FONTE:** O AUTOR (2012)

Criada a etiqueta de qualidade basta verificar em qual faixa cada alimentador (quadros 5.11, 5.12 e Anexo 03) se enquadra. No entanto, esta tarefa não é tão simples, pois dos 12 alimentadores, apenas cinco se enquadram nessas faixas de valores, todos com classificação de qualidade “A”: AA, AG, AJ, AK e AL, conforme mostrado na figura 1.2, a seguir.

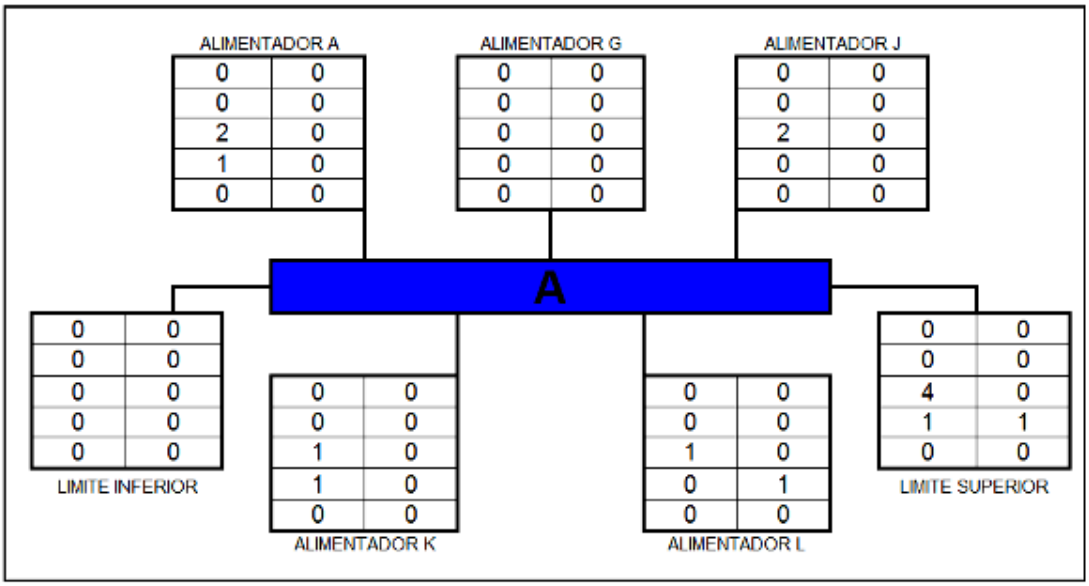


**FIGURA 5.2 –** ALIMENTADORES COM CLASSIFICAÇÃO DIRETA NA ETIQUETA DE QEE
**FONTE:** O AUTOR (2012)

A figura 5.2, anterior, mostra que os cinco alimentadores apresentam valores para $C_5$ pertencentes ao intervalo discreto [0, 4] e para $C_7$ e $C_8$, no intervalo discreto [0, 1].

Os demais alimentadores não podem ser classificados diretamente, uma vez que, por exemplo, para o alimentador AH, $C_5$ é igual a 16 (ver Anexo 03), o que indica que sua classificação seria "D". No entanto, $C_7$ e $C_8$ estão fora dos intervalos dessas classes, já que $C_7$ possui valor "1" e, por outro lado, $C_8$ possui valor "0".

Assim, qual é a classificação dos demais Alimentadores?

Para responder a esta pergunta, foram utilizadas as **técnicas de *Data Mining*** (apresentadas nas seções 4.3 a 4.7) com a finalidade de verificar suas classificações na etiqueta de qualidade.

Antes da aplicação das técnicas é apresentada a seguir a representação dos dados no espaço, com a finalidade de visualizar as faixas de classificação da etiqueta de qualidade de QEE e os alimentadores.

Tendo em vista as ilustrações apresentadas no capítulo 4 (figuras 4.3 a 4.5) e levando em consideração que o problema que está sendo analisado refere-se à figura 4.5, para cada faixa da etiqueta de classificação de QEE foram gerados, aleatoriamente, 29 registros para compor o seu conjunto de dados, totalizando 30 registros (considerando o limite superior de cada faixa).

A representação gráfica destes registros, já realizada a mudança de escala, é apresentada nas figuras 5.3 e 5.4, em que cada cor representa um conjunto de dados referente às faixas de classificação, sendo o conjunto da "Faixa A" o mais próximo da origem e o conjunto "Faixa F" o mais distante.

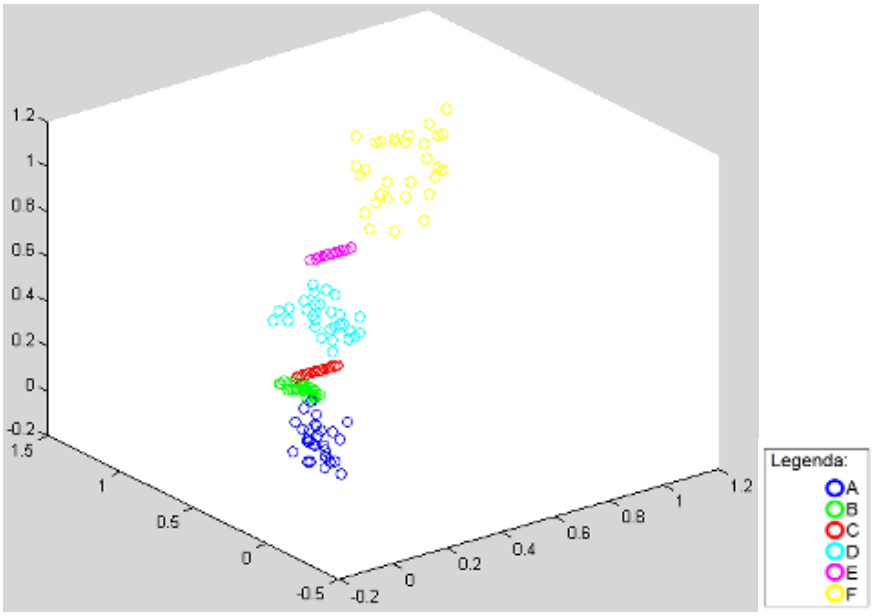


**FIGURA 5.3 –** REPRESENTAÇÃO GRÁFICA DOS REGISTROS QUE SERÃO UTILIZADOS NA APLICAÇÃO DAS TÉCNICAS DE *DATA MINING*
**FONTE:** O AUTOR (2012)

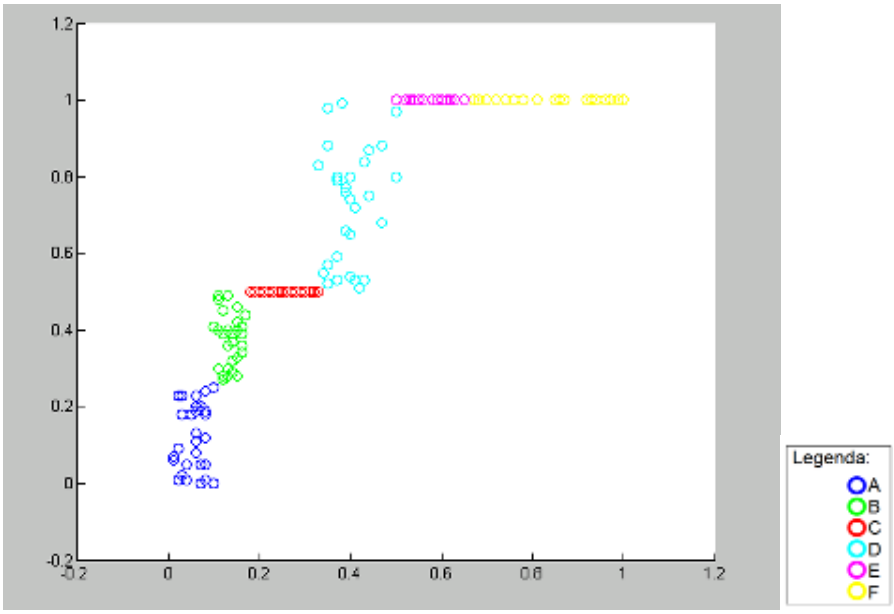


**FIGURA 5.4 –** REPRESENTAÇÃO GRÁFICA DA PROJEÇÃO NO PLANO *XY* DOS REGISTROS QUE SERÃO UTILIZADOS NA APLICAÇÃO DAS TÉCNICAS DE *DATA MINING*
**FONTE:** O AUTOR (2012)

As figuras a seguir (5.5 a 5.8) apresentam os conjuntos relativos a cada faixa de classificação (figuras 5.3 e 5.4), bem como a representação de cada um dos 12 alimentadores (representado por “+”), com a finalidade de verificar visualmente a proximidade dos alimentadores e as faixas de classificação da etiqueta de QEE.

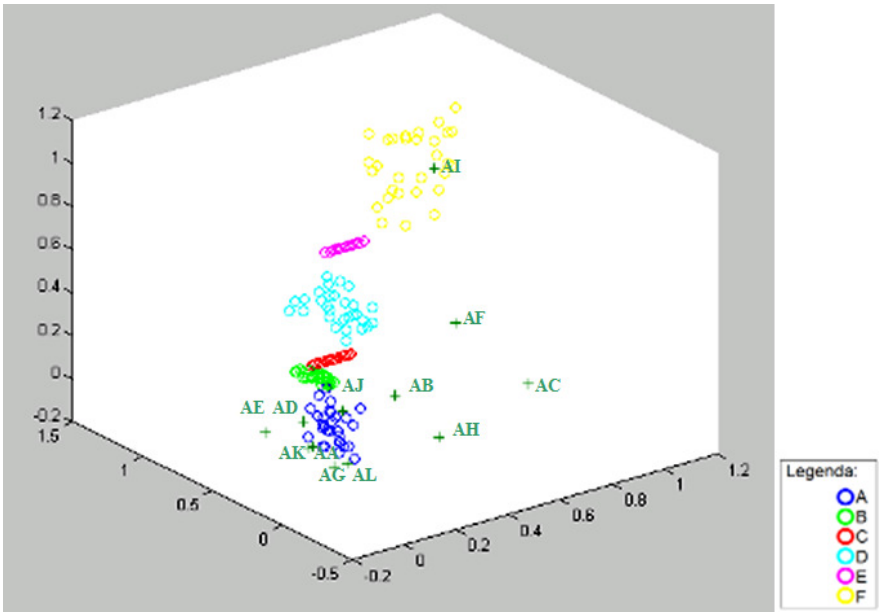


**FIGURA 5.5 –** REPRESENTAÇÃO GRÁFICA DOS REGISTROS QUE SERÃO UTILIZADOS NA APLICAÇÃO DAS TÉCNICAS DE *DATA MINING* E DOS 12 ALIMENTADORES
**FONTE:** O AUTOR (2012)

Nas figuras 5.6 a 5.8 é possível visualizar que muitos dos alimentadores estão distantes dos conjuntos de dados gerados, não sendo possível classificá-los visualmente. Também é possível verificar que os alimentadores AA, AG, AJ, AK e AL estão no conjunto da faixa A de classificação.

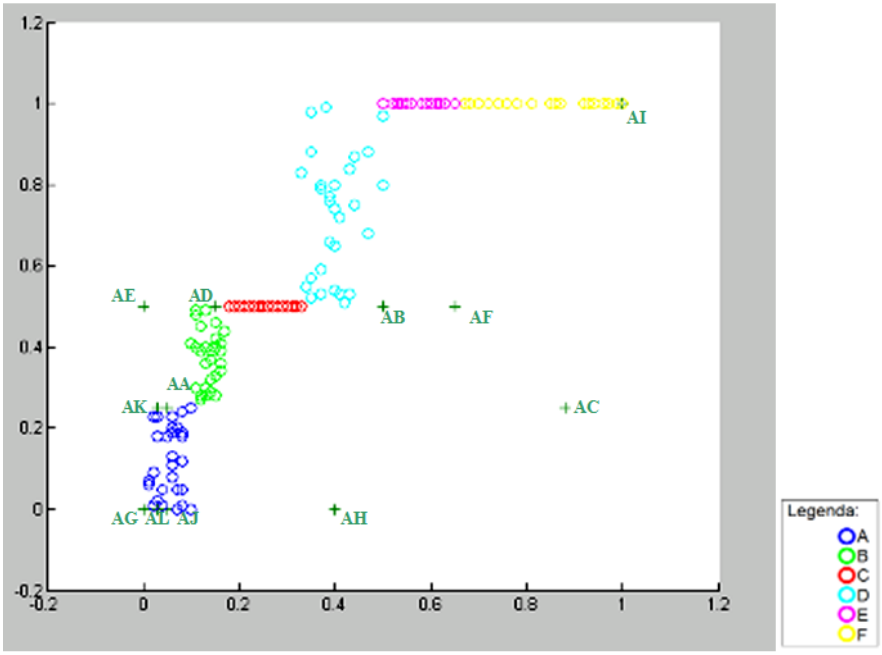


**FIGURA 5.6 –** REPRESENTAÇÃO GRÁFICA DA PROJEÇÃO NO PLANO XY DOS REGISTROS E DOS ALIMENTADORES
**FONTE:** O AUTOR (2012)

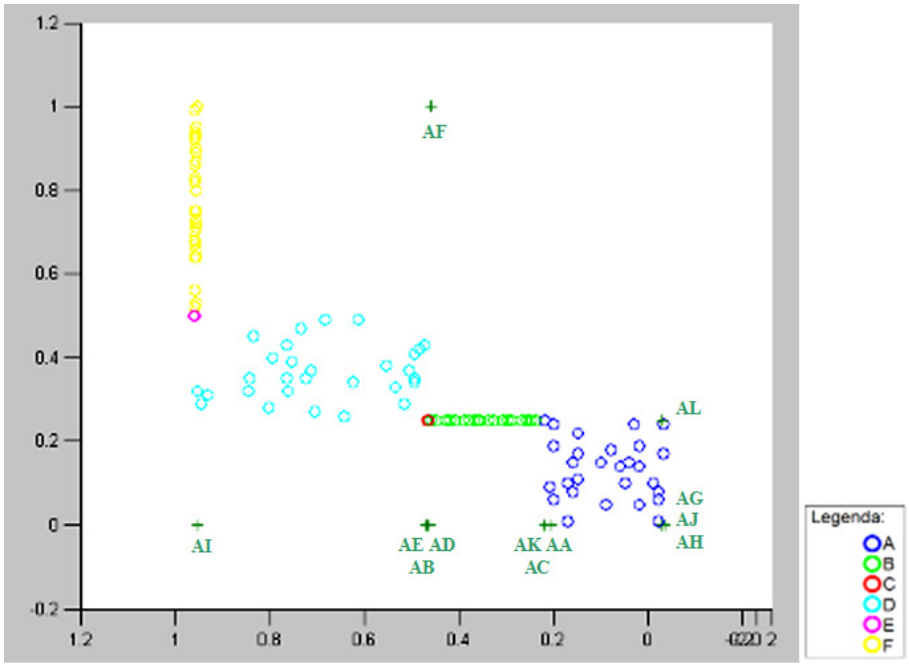


**FIGURA 5.7 –** REPRESENTAÇÃO GRÁFICA DA PROJEÇÃO NO PLANO *YZ* DOS REGISTROS E DOS ALIMENTADORES
**FONTE:** O AUTOR (2012)

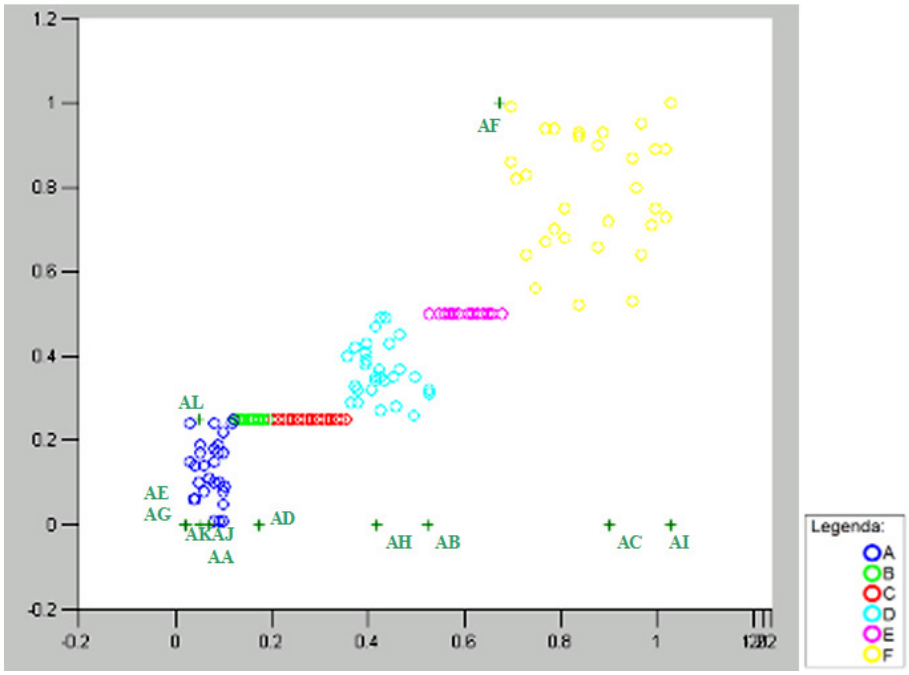


**FIGURA 5.8 –** REPRESENTAÇÃO GRÁFICA DA PROJEÇÃO NO PLANO *XZ* DOS REGISTROS E DOS ALIMENTADORES
**FONTE:** O AUTOR (2012)

Expostas essas considerações, é apresentada na próxima seção a aplicação das técnicas de *DM* a este estudo de caso, os testes e resultados obtidos em cada uma delas.

### 5.2.1 Aplicação das Técnicas de *Data Mining*

A etapa de *DM* é a mais importante do processo *KDD*, uma vez que é neste momento que se aplicam técnicas para o Reconhecimento de Padrões, seja através de procedimentos heurísticos ou metaheurísticos. Nesta pesquisa são utilizadas as metaheurísticas RNA e AG, a heurística *SVM* e outras heurísticas baseadas em distâncias (euclidiana e de Mahalanobis), descritas a seguir.

A metodologia de aplicação das técnicas a este estudo de caso está descrita no capítulo 4. Para cada faixa de classificação, há 30 exemplos compondo o conjunto de dados, sendo 29 gerados aleatoriamente respeitando os limites de cada $C_i$, mais o limite superior de cada faixa. Assim, como são seis faixas de classificação (de A a F) tem-se 180 registros, dos quais 120 são utilizados para o treinamento (ou seja, 2/3) e 60 para teste (ou seja,1/3) em cada etapa da validação cruzada.

#### *5.2.1.1 Redes Neurais*

A primeira técnica utilizada para classificar os alimentadores que não foram possíveis de serem classificados diretamente é a RNA, conforme secção 4.3.1.

Os resultados obtidos em relação ao conjunto A foram satisfatórios, uma vez que, já no treinamento, as RNA “aprenderam” corretamente todos os exemplos (registros; padrões) que foram apresentados em todas as cinco simulações para todas as topologias.

A tabela 5.1 mostra que para a 1ª etapa*,* a primeira sequência de simulações apresentou o melhor resultado com 20 neurônios na camada escondida, 431 iterações e erro quadrático médio igual a $6{,}41 \times 10^{-04}$. Na segunda sequência o melhor resultado foi obtido com 18 neurônios na camada escondida, erro quadrático médio igual a $5{,}17 \times 10^{-04}$ e 561 iterações, e assim para as demais sequências. Logo, o melhor resultado no treinamento da 1ª etapa para a “Faixa A” foi a terceira sequência, uma vez que o erro médio quadrático é o menor.

**TABELA 5.1 –** MELHORES RESULTADOS NO TREINAMENTO DA 1ª ETAPA PARA A "FAIXA A" DA ETIQUETA DE CLASSIFICAÇÃO - RNA

| Sequência de simulação | Quantidade de neurônios na camada escondida | Iterações | % de Acertos | Erro quadrático médio (aprox) |
|---|---|---|---|---|
| **1** | 20 | 431 | 100% | $6,41x10^{-04}$ |
| **2** | 18 | 561 | 100% | $5,17x10^{-04}$ |
| **3** | 20 | 530 | 100% | $4,93x10^{-04}$ |
| **4** | 17 | 387 | 100% | $6,80x10^{-04}$ |
| **5** | 12 | 144 | 100% | $1,21x10^{-03}$ |

A tabela 5.2, a seguir, apresenta o resultado da aplicação das RNA referente à "Faixa B", sendo o melhor resultado determinado na quinta sequência de simulações, com 16 neurônios na camada escondida, 642 iterações, nenhum registro classificado errado e erro médio quadrático igual a $3,68x10^{-03}$.

**TABELA 5.2 –** MELHORES RESULTADOS NO TREINAMENTO DA 1ª ETAPA PARA A "FAIXA B" DA ETIQUETA DE CLASSIFICAÇÃO - RNA

| Sequência de simulação | Quantidade de neurônios na camada escondida | Iterações | % de acertos | Erro quadrático médio (aprox) |
|---|---|---|---|---|
| **1** | 19 | 579 | 100% | $3,94x10^{-03}$ |
| **2** | 16 | 334 | 100% | $3,97x10^{-03}$ |
| **3** | 18 | 347 | 100% | $4,02x10^{-03}$ |
| **4** | 19 | 413 | 100% | $4,21x10^{-03}$ |
| **5** | 16 | 642 | 100% | $3,68x10^{-03}$ |

Na tabela 5.3 têm-se os resultados para a "Faixa C", sendo o melhor apresentado na quinta sequência, com 16 neurônios, 182 iterações e erro quadrático médio igual a $1,75x10^{-04}$.

**TABELA 5.3 –** MELHORES RESULTADOS NO TREINAMENTO DA 1ª ETAPA PARA A "FAIXA C" DA ETIQUETA DE CLASSIFICAÇÃO - RNA

| Sequência de simulação | Quantidade de neurônios na camada escondida | Iterações | % de acertos | Erro quadrático médio (aprox) |
|---|---|---|---|---|
| **1** | 12 | 176 | 100% | $2,78x10^{-04}$ |
| **2** | 13 | 232 | 100% | $2,39x10^{-04}$ |
| **3** | 6 | 486 | 100% | $3,98x10^{-04}$ |
| **4** | 16 | 965 | 100% | $2,73x10^{-04}$ |
| **5** | 16 | 182 | 100% | $1,75x10^{-04}$ |

Os resultados com os dados da “Faixa D” são apresentados na tabela 5.4 ,a seguir, e seu melhor resultado é encontrado na segunda sequência de testes, com 232 iterações, 13 neurônios na camada escondida e erro quadrático médio $1{,}39x10^{-02}$ ($1{,}3886x10^{-02}$).

**TABELA 5.4 –** MELHORES RESULTADOS NO TREINAMENTO DA 1ª ETAPA PARA A “FAIXA D” DA ETIQUETA DE CLASSIFICAÇÃO - RNA

| Sequência de simulação | Quantidade de neurônios na camada escondida | Iterações | % de Acertos | Erro quadrático médio (aprox) |
|---|---|---|---|---|
| **1** | 16 | 1000 | 99,16% | $1{,}39x10^{-02}$ |
| **2** | 18 | 1000 | 99,16% | $1{,}39x10^{-02}$ |
| **3** | 19 | 1000 | 99,16% | $1{,}40x10^{-02}$ |
| **4** | 15 | 1000 | 99,16% | $1{,}39x10^{-02}$ |
| **5** | 9 | 1000 | 99,16% | $1{,}40x10^{-02}$ |

Na tabela 5.5 têm-se os resultados para a “Faixa E” das sequências de testes. O melhor é obtido na primeira sequência de simulações, com 12 neurônios na camada escondida e erro médio quadrático igual a $3{,}67x10^{-04}$.

**TABELA 5.5 –** MELHORES RESULTADOS NO TREINAMENTO DA 1ª ETAPA PARA A “FAIXA E” DA ETIQUETA DE CLASSIFICAÇÃO - RNA

| Sequência de simulação | Quantidade de neurônios na camada escondida | Iterações | % de Acertos | Erro quadrático médio (aprox) |
|---|---|---|---|---|
| **1** | 12 | 376 | 100% | $3{,}67x10^{-04}$ |
| **2** | 12 | 307 | 100% | $6{,}01x10^{-04}$ |
| **3** | 9 | 238 | 100% | $5{,}13x10^{-04}$ |
| **4** | 20 | 175 | 100% | $4{,}73x10^{-04}$ |
| **5** | 10 | 226 | 100% | $5{,}39x10^{-04}$ |

Assim, o resultado da 1ª etapa é apresentado na tabela 5.6, a seguir. Com estes treinamentos foram realizados os testes (apresentação do conjunto de teste para cada rede que obteve melhor resultado de cada faixa da etiqueta) nesta etapa a porcentagem de acerto foi total (100%).

**TABELA 5.6 –** MELHORES RESULTADOS DA 1ª ETAPA PARA A ETIQUETA DE CLASSIFICAÇÃO - RNA

| Faixa | TREINAMENTO | | | | | TESTE |
|---|---|---|---|---|---|---|
| | Sequência de simulação | Quantidade de neurônios na camada escondida | Iterações | % de Acertos | Erro quadrático médio (aprox) | % de acertos |
| **A** | 3 | 20 | 530 | 100% | $4{,}93x10^{-04}$ | 100% |
| **B** | 5 | 16 | 642 | 100% | $3{,}68x10^{-03}$ | 100% |
| **C** | 5 | 16 | 182 | 100% | $1{,}75x10^{-04}$ | 100% |
| **D** | 2 | 18 | 1000 | 99,16% | $1{,}39x10^{-02}$ | 100% |
| **E** | 1 | 12 | 376 | 100% | $3{,}67x10^{-04}$ | 100% |

Cabe lembrar que, nessa metodologia, se um exemplo foi apresentando à primeira rede e foi classificado como “Faixa A”, este não é apresentado às próximas redes (B, C, ... F). Quando um exemplo é apresentado à rede correspondente a “Faixa E”, se a rede não “conseguir” classificá-lo como sendo desta faixa, este será automaticamente classificado como “Faixa F” (já que esta é a última faixa).

Da mesma forma foram realizadas a 2ª e a 3ª etapas, cujos resultados são apresentados nas tabelas 5.7 e 5.8, respectivamente.

**TABELA 5.7 –** MELHORES RESULTADOS DA 2ª ETAPA PARA A ETIQUETA DE CLASSIFICAÇÃO - RNA

| Faixa | TREINAMENTO | | | | | TESTE |
|---|---|---|---|---|---|---|
| | Sequência de simulação | Quantidade de neurônios na camada escondida | Iterações | % de acertos | Erro quadrático médio (aprox) | % de acertos |
| **A** | 5 | 19 | 165 | 100% | $8{,}24x10^{-05}$ | 98,33% |
| **B** | 1 | 19 | 597 | 100% | $1{,}13x10^{-03}$ | 98,33% |
| **C** | 5 | 7 | 655 | 100% | $1{,}60x10^{-04}$ | 100% |
| **D** | 1 | 11 | 544 | 100% | $5{,}85x10^{-03}$ | 100% |
| **E** | 5 | 20 | 140 | 100% | $6{,}63x10^{-04}$ | 100% |

**TABELA 5.8 –** MELHORES RESULTADOS DA 3ª ETAPA PARA A ETIQUETA DE CLASSIFICAÇÃO - RNA

| Faixa | TREINAMENTO | | | | | TESTE |
|---|---|---|---|---|---|---|
| | Sequência de simulação | Quantidade de neurônios na camada escondida | Iterações | % de acertos | Erro quadrático médio (aprox) | % acertos |
| **A** | 2 | 18 | 89 | 100% | $1{,}62E^{-03}$ | 100% |
| **B** | 1 | 18 | 1000 | 99,16% | $8{,}08E^{-03}$ | 100% |
| **C** | 5 | 10 | 327 | 100% | $3{,}29E^{-04}$ | 100% |
| **D** | 1 | 19 | 768 | 100% | $2{,}11E^{-04}$ | 98,33% |
| **E** | 4 | 10 | 164 | 100% | $5{,}24E^{-04}$ | 100% |

Através dos resultados apresentados nas tabelas 5.6 a 5.8, pode-se verificar que esses foram bastante satisfatórios através das RNA, uma vez que em apenas duas de 15 simulações as RNA não conseguiram aprender apenas um exemplo, gerando 99,89%, aproximadamente, de acerto nessa técnica.

Continuando a aplicação da técnica, em cada etapa da validação cruzada foram apresentados às redes, além dos exemplos do teste, os valores de cada alimentador. O resultado desta classificação está no quadro 5.21.

**QUADRO 5.21 –** RESULTADO DA CLASSIFICAÇÃO DOS ALIMENTADORES EM CADA ETAPA DA VALIDAÇÃO CRUZADA - RNA

| Alimentador | 1ª etapa | 2ª etapa | 3ª etapa | Classificação por voto |
|---|---|---|---|---|
| **AA** | A | A | A | A |
| **AB** | C | C | C | C |
| **AC** | D | D | D | D |
| **AD** | C | C | B | C |
| **AE** | B | B | B | B |
| **AF** | F | F | F | F |
| **AG** | A | A | A | A |
| **AH** | A | A | A | A |
| **AI** | E | E | E | E |
| **AJ** | A | A | A | A |
| **AK** | A | A | A | A |
| **AL** | A | A | A | A |

No quadro 5.21 anterior e nos demais que serão apresentados, a coluna “Classificação por voto” (última coluna) indica a classificação de maior ocorrência nas colunas anteriores. Caso não haja uma classificação com maior ocorrência, esta é definida como a *pior situação*.

Ainda referente ao quadro 5.21, apesar dos alimentadores AA, AG, AJ, AK e AL já terem as suas classificações definidas, uma vez que estes foram classificados diretamente na etiqueta de qualidade, os mesmos também foram apresentados à rede, e confirmaram as classificações já obtidas. Assim, têm-se seis alimentadores com classificação “A”, um alimentador com classificação “B”, dois alimentadores com classificação “C”, um alimentador com classificação “D”, um alimentador com classificação “E” e um alimentador com classificação “F”.

E ao finalizar a aplicação de RNA, fez-se um teste adicional: foi utilizado o conjunto de dados inicial (antes da divisão em três conjuntos para aplicação da validação cruzada) para treinar a rede e assim apresentar os dados dos alimentadores. Os melhores resultados desse treinamento estão na tabela 5.9 e a

classificação dos alimentadores segundo este treinamento encontra-se no quadro 5.22, bem como sua comparação com o resultado obtido na validação cruzada.

**TABELA 5.9 –** MELHORES RESULTADOS A ETIQUETA DE CLASSIFICAÇÃO UTILIZANDO TODOS OS DADOS NO TREINAMENTO - RNA

| Faixa | TREINAMENTO | | | | |
|---|---|---|---|---|---|
| | **Sequência de simulação** | **Quantidade de neurônios na camada escondida** | **Iterações** | **% de acertos** | **Erro quadrático médio (aprox)** |
| **A** | 5 | 6 | 325 | 100% | $2{,}09 \times 10^{-05}$ |
| **B** | 1 | 16 | 1000 | 99,44% | $5{,}42 \times 10^{-03}$ |
| **C** | 2 | 19 | 822 | 100% | $1{,}04 \times 10^{-04}$ |
| **D** | 5 | 15 | 1000 | 99,44% | $8{,}42 \times 10^{-03}$ |
| **E** | 1 | 20 | 257 | 100% | $7{,}47 \times 10^{-05}$ |

**QUADRO 5.22 –** RESULTADO DA CLASSIFICAÇÃO DOS ALIMENTADORES APÓS TREINAMENTO DA RNA COM TODOS OS EXEMPLOS

| **Alimentador** | **Treinamento com todos os dados** | **Classificação por voto *Validação cruzada*** |
|---|---|---|
| **AA** | A | A |
| **AB** | D | C |
| **AC** | D | D |
| **AD** | C | C |
| **AE** | B | B |
| **AF** | F | F |
| **AG** | A | A |
| **AH** | A | A |
| **AI** | D | E |
| **AJ** | A | A |
| **AK** | A | A |
| **AL** | A | A |

Portanto, analisando o quadro 5.22, percebe-se que há divergência em dois resultados: alimentadores “AB” e “AI”. Mas essas divergências podem ser aceitas, pois as faixas em que estes divergem são vizinhas.

### *5.2.1.2 Support Vector Machine*

A segunda técnica utilizada para classificar os alimentadores é o *SVM*, conforme seção 4.4.1.

A tabela 5.10 a seguir, apresenta a porcentagem de acerto em cada etapa da validação cruzada para cada faixa, em que se pode observar bons resultados no aprendizado, uma vez que de 15 classificações apenas quatro não conseguiram classificar um exemplo, o que gerou uma porcentagem de 98,33% nestas classificações e 99,55%, aproximadamente, nessa técnica.

**TABELA 5.10 –** PORCENTAGEM DE ACERTO EM CADA ETAPA DA VALIDAÇÃO CRUZADA NO TESTE - *SVM*

| Faixa | 1ª etapa | 2ª etapa | 3ª etapa |
|---|---|---|---|
| **A** | 100% | 98,33% | 100% |
| **B** | 100% | 98,33% | 100% |
| **C** | 100% | 98,33% | 100% |
| **D** | 100% | 100% | 98,33% |
| **E** | 100% | 100% | 100% |

Após a realização dos testes, foram apresentados ao *SVM*, os valores de cada alimentador a cada etapa da validação cruzada e seus resultados estão apresentados no quadro 5.23 a seguir.

**QUADRO 5.23 –** RESULTADO DA CLASSIFICAÇÃO DOS ALIMENTADORES EM CADA ETAPA DA VALIDAÇÃO CRUZADA - *SVM*

| Alimentador | 1ª etapa | 2ª etapa | 3ª etapa | Classificação por voto |
|---|---|---|---|---|
| **AA** | A | A | A | A |
| **AB** | C | C | C | C |
| **AC** | D | C | C | C |
| **AD** | C | C | B | C |
| **AE** | A | A | A | A |
| **AF** | F | F | F | F |
| **AG** | A | A | A | A |
| **AH** | A | A | A | A |
| **AI** | D | E | E | E |
| **AJ** | A | A | A | A |
| **AK** | A | A | A | A |
| **AL** | A | A | A | A |

O quadro 5.23 indica que a classificação por voto dos alimentadores é: sete com classificação “A”, nenhum com classificação “B”, três com classificação “C”, nenhum com classificação “D”, um com classificação “E” e um com classificação “F”.

Após o treinamento utilizando validação cruzada, houve novo treinamento considerando todos os exemplos e o resultado deste é apresentado no quadro 5.24,

bem como sua comparação com a classificação por voto das etapas da validação cruzada.

**QUADRO 5.24 –** RESULTADO DA CLASSIFICAÇÃO DOS ALIMENTADORES APÓS TREINAMENTO DO *SVM* COM TODOS OS EXEMPLOS

| Alimentador | Treinamento com todos os dados | Classificação por voto - Validação cruzada |
|---|---|---|
| **AA** | A | A |
| **AB** | C | C |
| **AC** | C | C |
| **AD** | B | C |
| **AE** | A | A |
| **AF** | F | F |
| **AG** | A | A |
| **AH** | A | A |
| **AI** | E | E |
| **AJ** | A | A |
| **AK** | A | A |
| **AL** | A | A |

No quadro 5.24 anterior é observado que há divergência de classificação em apenas no “AD”, que, na classificação por voto das etapas da validação cruzada, obteve classificação “C” e, quando utilizados todos os exemplos no treinamento, classificação “B”. Novamente, não é possível afirmar qual dos dois métodos indicou a classificação correta, mas, aqui novamente, as classificações para o “AD” ocorrem em faixas vizinhas, não tornando o resultado discrepante.

Ainda analisando o quadro 5.24, os alimentadores em que suas classificações já eram conhecidas foram classificados corretamente.

### *5.2.1.3 Algoritmo Genético*

A terceira técnica utilizada para classificar os alimentadores é o AG. A função *fitness* foi determinada conforme seção 4.5.1, que neste caso procurou um plano que separasse os conjuntos de dados, vistos que se têm apenas três classes, diferentes de zero: $C_5$, $C_7$ e $C_8$.

Os resultados da porcentagem de acerto, em cada etapa da validação cruzada nos treinamentos, estão expressos na tabela 5.11. Já o quadro 5.25 apresenta a classificação dos alimentadores obtida em cada etapa desse treinamento, com 99,11%, aproximadamente, de acerto.

**TABELA 5.11 –** PORCENTAGEM DE ACERTO EM CADA ETAPA DA VALIDAÇÃO CRUZADA NO TESTE – AG – FUNÇÃO *FITNESS*

| Faixa | 1ª etapa | 2ª etapa | 3ª etapa |
|---|---|---|---|
| **A** | 100% | 98,33% | 100% |
| **B** | 100% | 98,33% | 100% |
| **C** | 100% | 100% | 98,33% |
| **D** | 98,33% | 95% | 100% |
| **E** | 100% | 100% | 98,33% |

**QUADRO 5.25 –** RESULTADO DA CLASSIFICAÇÃO DOS ALIMENTADORES EM CADA ETAPA DA VALIDAÇÃO CRUZADA - AG – FUNÇÃO *FITNESS*

| Alimentador | 1ª etapa | 2ª etapa | 3ª etapa | Classificação por voto |
|---|---|---|---|---|
| **AA** | A | A | A | A |
| **AB** | E | C | C | C |
| **AC** | E | C | C | C |
| **AD** | B | B | C | B |
| **AE** | A | A | A | A |
| **AF** | F | F | F | F |
| **AG** | A | A | A | A |
| **AH** | E | C | C | C |
| **AI** | E | C | C | C |
| **AJ** | A | A | A | A |
| **AK** | A | A | A | A |
| **AL** | A | A | A | A |

Realizado o treinamento com todos os dados, pode-se verificar no quadro 5.26 que o resultado obtido é igual à classificação por voto obtida nas etapas da validação cruzada.

**QUADRO 5.26 –** RESULTADO DA CLASSIFICAÇÃO DOS ALIMENTADORES APÓS TREINAMENTO DA AG COM TODOS OS EXEMPLOS – AG

| Alimentador | Treinamento com todos os dados | Classificação por voto Validação cruzada |
|---|---|---|
| **AA** | A | A |
| **AB** | C | C |
| **AC** | C | C |
| **AD** | B | B |
| **AE** | A | A |
| **AF** | F | F |
| **AG** | A | A |
| **AH** | C | C |
| **AI** | C | C |
| **AJ** | A | A |
| **AK** | A | A |
| **AL** | A | A |

*5.2.1.4 Técnicas que utilizam Distância Euclidiana*

Depois de concluído os testes com a RNA, *SVM* e AG, técnicas consideradas mais elaboradas, foram realizados outros que levam em consideração a distância euclidiana entre o alimentador e os dados de cada faixa da etiqueta de qualidade. Esses testes foram realizados para verificar a eficácia de técnicas mais simples de implementação computacional, além de menor tempo de execução.

O primeiro teste calcula o somatório das distâncias do novo elemento aos elementos de cada faixa da etiqueta de qualidade (seção 4.6.1). O resultado do cálculo é apresentado No quadro 5.27, onde os valores em negrito correspondem ao menor valor do somatório obtido para cada alimentador.

**QUADRO 5.27 –** SOMATÓRIO DAS DISTÂNCIAS ENTRE CADA ALIMENTADOR E TODOS OS DADOS DAS FAIXAS DA ETIQUETA DE CLASSIFICAÇÃO

| Alimentador | Faixa A | Faixa B | Faixa C | Faixa D | Faixa E | Faixa F | Classificação |
|---|---|---|---|---|---|---|---|
| **AA** | ***5,97*** | 9,01 | 12,23 | 21,48 | 31,21 | 40,53 | A |
| **AB** | 18,09 | 13,90 | ***10,65*** | 14,15 | 21,35 | 30,08 | C |
| **AC** | 25,49 | 23,92 | ***21,69*** | 23,60 | 28,64 | 33,00 | C |
| **AD** | 12,48 | 8,62 | ***8,16*** | 15,81 | 24,67 | 35,03 | C |
| **AE** | 12,28 | ***9,54*** | 10,64 | 18,43 | 27,23 | 37,85 | B |
| **AF** | 33,65 | 27,60 | 25,53 | 22,00 | 21,39 | ***17,88*** | F |
| **AG** | ***6,22*** | 14,22 | 18,40 | 27,68 | 37,64 | 45,95 | A |
| **AH** | ***12,05*** | 15,76 | 17,41 | 24,86 | 33,94 | 40,67 | A |
| **IA** | 38,90 | 32,87 | 28,08 | 22,96 | ***19,84*** | 24,44 | E |
| **AJ** | ***5,95*** | 13,86 | 17,85 | 27,05 | 36,99 | 45,15 | A |
| **AK** | ***6,04*** | 9,21 | 12,53 | 21,79 | 31,52 | 40,89 | A |
| **AL** | ***5,88*** | 11,75 | 16,43 | 25,16 | 34,90 | 42,06 | A |

Na sequência foi aplicada outra técnica que consiste em calcular a distância do novo elemento ao ponto central de cada faixa da etiqueta de qualidade (seção 4.6.2).

A tabela 5.12 apresenta os pontos centrais obtidos para cada conjunto de dados da faixa da etiqueta de classificação.

**TABELA 5.12** – PONTO CENTRAL DE CADA CONJUNTO DE DADOS DAS FAIXAS DA ETIQUETA DE CLASSIFICAÇÃO

| Faixa | $C_5$ | $C_7$ | $C_8$ |
|---|---|---|---|
| **A** | 0,06 | 0,13 | 0,14 |
| **B** | 0,14 | 0,38 | 0,25 |
| **C** | 0,26 | 0,50 | 0,25 |
| **D** | 0,40 | 0,75 | 0,37 |
| **E** | 0,56 | 1,00 | 0,50 |
| **F** | 0,84 | 1,00 | 0,79 |

Obtidos os pontos centrais foi realizado o cálculo da distância de cada alimentador a cada um desses pontos, sendo sua classificação determinada pelo menor resultado (Quadro 5.28).

**QUADRO 5.28 –** DISTÂNCIA ENTRE CADA ALIMENTADOR E CADA PONTO CENTRAL DOS DADOS DAS FAIXAS DA ETIQUETA DE CLASSIFICAÇÃO

| Alimentador | Faixa A | Faixa B | Faixa C | Faixa D | Faixa E | Faixa F | Classifi-cação |
|---|---|---|---|---|---|---|---|
| **AA** | **0,18** | 0,30 | 0,41 | 0,71 | 1,04 | 1,35 | A |
| **AB** | 0,59 | 0,45 | **0,35** | 0,46 | 0,71 | 0,99 | C |
| **AC** | 0,84 | 0,79 | **0,71** | 0,79 | 0,96 | 1,09 | C |
| **AD** | 0,41 | 0,28 | **0,27** | 0,51 | 0,82 | 1,16 | C |
| **AE** | 0,40 | **0,31** | 0,36 | 0,60 | 0,90 | 1,26 | B |
| **AF** | 1,11 | 0,91 | 0,85 | 0,72 | 0,71 | **0,57** | F |
| **AG** | **0,20** | 0,48 | 0,62 | 0,93 | 1,25 | 1,53 | A |
| **AH** | **0,39** | 0,52 | 0,58 | 0,84 | 1,13 | 1,35 | A |
| **AI** | 1,29 | 1,09 | 0,93 | 0,75 | **0,67** | 0,81 | E |
| **AJ** | **0,19** | 0,46 | 0,60 | 0,91 | 1,23 | 1,50 | A |
| **AK** | **0,19** | 0,30 | 0,42 | 0,72 | 1,05 | 1,36 | A |
| **AL** | **0,17** | 0,40 | 0,55 | 0,84 | 1,16 | 1,40 | A |

No terceiro teste é verificada a classificação dos $k$ elementos mais próximos a cada alimentador, técnica conhecida como $k$-vizinhos mais próximos. Conforme a seção 4.6.3 foram realizados testes para definir o valor de $k$, utilizando os mesmos conjuntos, treinamento e teste, da validação cruzada das técnicas RNA, SVM e AG. A porcentagem de acerto para *k* variando de 1 a 20 está expressa na tabela 5.13, na qual para *k=1* há 100% de acerto, sendo este o escolhido.

**TABELA 5.13 –** DETERMINAÇÃO DE *K* PARA APLICAÇÃO NO ESTUDO DE CASO DA ÁREA ELÉTRICA

| $K$ | 1ª etapa | 2ª etapa | 3ª etapa | Média |
|---|---|---|---|---|
| **1** | 100% | 100% | 100% | 100% |
| **2** | 100% | 98,33% | 98,33% | 98,87% |
| **3** | 100% | 98,33% | 98,33% | 98,87% |
| **4** | 100% | 98,33% | 98,33% | 98,87% |
| **5** | 100% | 98,33% | 98,33% | 98,87% |
| **6** | 100% | 96,67% | 98,33% | 98,33% |
| **7** | 100% | 96,67% | 98,33% | 98,33% |
| **8** | 100% | 96,67% | 98,33% | 98,33% |
| **9** | 100% | 96,67% | 98,33% | 98,33% |
| **10** | 100% | 95% | 96,67% | 97,22% |
| **11** | 100% | 95% | 96,67% | 97,22% |
| **12** | 100% | 95% | 91,67% | 95,56% |
| **13** | 100% | 95% | 91,67% | 95,56% |
| **14** | 98,33% | 95% | 91,67% | 95,00% |
| **15** | 98,33% | 95% | 91,67% | 95,00% |
| **16** | 96,67% | 95% | 91,67% | 94,45% |
| **17** | 96,67% | 95% | 91,67% | 94,45% |
| **18** | 95% | 95% | 91,67% | 93,89% |
| **19** | 95% | 95% | 91,67% | 93,89% |
| **20** | 95% | 91,66% | 91,67% | 92,78% |

Analisando a tabela 5.13 anterior, percebe-se que à medida que $k$ aumenta, a porcentagem de acerto diminui. Já quadro a 5.29, a seguir, apresenta a classificação de cada alimentador, para $k=1$, em cada etapa da validação cruzada, a classificação por voto da validação cruzada e a classificação utilizando todos os dados no conjunto de treinamento.

**QUADRO 5.29 –** CLASSIFICAÇÃO DOS ALIMENTADORES – TÉCNICA DOS *K*-VIZINHOS MAIS PRÓXIMOS

| Alimentador | 1ª etapa | 2ª etapa | 3ª etapa | Classif. por voto - Validação cruzada | Todos os elementos |
|---|---|---|---|---|---|
| **AA** | A | A | A | A | A |
| **AB** | C | C | C | C | C |
| **AC** | D | D | D | D | D |
| **AD** | B | B | B | B | B |
| **AE** | A | B | B | B | B |
| **AF** | F | F | F | F | F |
| **AG** | A | A | A | A | A |
| **AH** | A | A | A | A | A |
| **AI** | F | F | F | F | F |
| **AJ** | A | A | A | A | A |
| **AK** | A | A | A | A | A |
| **AL** | A | A | A | A | A |

Analisando o quadro 5.29 anterior, tem-se que, tanto na classificação por voto obtida na validação cruzada quanto à classificação quando utilizado todos os elementos no conjunto de treinamento, o mesmo resultado. Quanto às etapas da validação cruzada, apenas para o alimentador AE houve classificações distintas, mas em faixa vizinhas.

A quarta técnica calcula a distância do novo elemento a média dos elementos de cada faixa da etiqueta de qualidade (seção 4.6.4). Essa técnica foi aplicada ao verificarmos que os pontos centrais são bastante próximos das médias dos elementos quando considerado o mesmo conjunto, como pode ser verificado na tabela 5.14.

**TABELA 5.14 –** COMPARAÇÃO DOS PONTOS CENTRAIS COM AS MÉDIAS DOS ELEMENTOS DO CONJUNTO DE DADOS DA ÁREA ELÉTRICA

| FAIXA | *PONTO CENTRAL* | | | *MÉDIA DOS ELEMENTOS DO CONJUNTO DE DADOS* | | | *DISTÂNCIA ENTRE OS PONTOS* |
|---|---|---|---|---|---|---|---|
| | $C_5$ | $C_7$ | $C_8$ | $C_5$ | $C_7$ | $C_8$ | |
| **A** | 0,06 | 0,13 | **0,14** | 0,06 | 0,13 | **0,13** | 0,01 |
| **B** | 0,14 | 0,38 | 0,25 | 0,14 | 0,38 | 0,25 | 0,00 |
| **C** | **0,26** | 0,50 | 0,25 | **0,25** | 0,50 | 0,25 | 0,01 |
| **D** | 0,40 | **0,75** | 0,37 | 0,40 | **0,74** | 0,37 | 0,01 |
| **E** | **0,56** | 1,00 | 0,50 | **0,57** | 1,00 | 0,50 | 0,01 |
| **F** | 0,84 | 1,00 | 0,79 | 0,84 | 1,00 | 0,79 | 0,00 |

Na tabela 5.14 anterior, os valores em negritos indicam as diferenças obtidas entre os pontos centrais e as médias dos elementos do conjunto de dados de cada faixa da etiqueta de classificação da qualidade. Essas diferenças, quando ocorrem, são de um centésimo. Já a última coluna da tabela indica a distância euclidiana entre ao ponto central e a correspondente média dos elementos do conjunto de dados.

O quadro 5.30 a seguir, apresenta o resultado da aplicação dessa técnica, no qual os valores em negritos indicam a menor distância.

**QUADRO 5.30 –** DISTÂNCIA ENTRE CADA ALIMENTADOR A CADA MÉDIA DOS DADOS DAS FAIXAS DA ETIQUETA DE CLASSIFICAÇÃO

| Alimentador | Faixa A | Faixa B | Faixa C | Faixa D | Faixa E | Faixa F | Classifi-cação |
|---|---|---|---|---|---|---|---|
| **AA** | **0,18** | 0,30 | 0,41 | 0,71 | 1,04 | 1,35 | A |
| **AB** | 0,59 | 0,45 | **0,35** | 0,45 | 0,71 | 0,99 | C |
| **AC** | 0,84 | 0,79 | **0,72** | 0,78 | 0,95 | 1,09 | C |
| **AD** | 0,40 | 0,28 | **0,27** | 0,51 | 0,82 | 1,16 | C |
| **AE** | 0,40 | **0,31** | 0,35 | 0,60 | 0,91 | 1,26 | B |
| **AF** | 1,11 | 0,91 | 0,85 | 0,72 | **0,71** | 0,57 | F |
| **AG** | **0,19** | 0,48 | 0,61 | 0,92 | 1,25 | 1,53 | A |
| **AH** | **0,39** | 0,52 | 0,58 | 0,83 | 1,13 | 1,35 | A |
| **AI** | 1,29 | 1,09 | 0,94 | 0,75 | **0,66** | 0,81 | E |
| **AJ** | **0,18** | 0,46 | 0,59 | 0,90 | 1,23 | 1,50 | A |
| **AK** | **0,18** | 0,30 | 0,42 | 0,72 | 1,05 | 1,36 | A |
| **AL** | **0,18** | 0,40 | 0,55 | 0,84 | 1,16 | 1,40 | A |

#### *5.2.1.5 Técnica estatística: Distância de Mahalanobis*

Para finalizar os testes para este 1º estudo de caso, foi aplicada a técnica estatística: distância de Mahalanobis (seção 4.7). Para isso foi efetuado o cálculo da distância de Mahalanobis dos alimentadores a cada conjunto de dados referentes às faixas de classificação. No entanto, ao se calcular este tipo de distância do conjunto B a cada alimentador, não foi possível obter resultado, pois a matriz covariância do conjunto B é uma matriz singular, não sendo possível determinar sua inversa. Assim, não se têm resultados desta técnica para esse 1º estudo de caso.

Finalizados os testes, a próxima seção apresenta a análise obtida com estes resultados.

### 5.2.2 Análise dos resultados

A análise dos resultados ocorre em três etapas, sendo a primeira a comparação das técnicas mais sofisticadas (RNA, *SVM* e AG), a segunda compara as técnicas que levam em consideração distâncias (euclidiana e de mahalanobis) e a terceira compara todas as técnicas. Nesta última análise apenas os testes com todos os dados no treinamento das técnicas RNA, *SVM* e AG são considerados.

*5.2.2.1 Comparação das classificações obtidas nas técnicas RNA, SVM e AG*

Na primeira analise é realizada a comparação das classificações obtidas nas técnicas RNA, *SVM* e AG. O quadro 5.31 apresenta o resultado obtido por estas, sendo que para cada uma delas é apresentado o resultado da validação cruzada e o resultado quando utilizados todos os dados no treinamento.

**QUADRO 5.31 –** COMPARAÇÃO DAS CLASSIFICAÇÕES OBTIDAS PELAS TÉCNICAS RNA, *SVM* E AG

| Alimentador | RNA | | *SVM* | | AG | |
|---|---|---|---|---|---|---|
| | TT | VC | TT | VC | TT | VC |
| **AA** | A | A | A | A | A | A |
| **AB** | D | C | C | C | C | C |
| **AC** | D | D | C | C | C | C |
| **AD** | C | C | B | C | B | B |
| **AE** | B | B | A | A | A | A |
| **AF** | F | F | F | F | F | F |
| **AG** | A | A | A | A | A | A |
| **AH** | A | A | A | A | C | C |
| **AI** | D | E | E | E | C | C |
| **AJ** | A | A | A | A | A | A |
| **AK** | A | A | A | A | A | A |
| **AL** | A | A | A | A | A | A |

**TT:** Treinamento com todos os dados; **VC:** classificação por voto considerando a *validação cruzada* em cada técnica;

De uma maneira geral, com muita cautela, pode-se afirmar que cada alimentador apresentou a classificação descrita no quadro 5.32, e a segunda coluna considera todas as colunas, dois a sete, do quadro 5.31. Já a terceira coluna do quadro 5.32 considera apenas as colunas dois, quatro e seis do quadro 5.31.

**QUADRO 5.32 –** CLASSIFICAÇÃO - COMPARAÇÃO DAS TÉCNICAS RNA, *SVM* E AG

| Alimentador | Classificação considerando a *validação cruzada* | Classificação sem considerar a *validação cruzada* |
|---|---|---|
| **AA** | A | A |
| **AB** | C | C |
| **AC** | C | C |
| **AD** | C | B |
| **AE** | A | A |
| **AF** | F | F |
| **AG** | A | A |
| **AH** | A | A |
| **AI** | E | E |
| **AJ** | A | A |
| **AK** | A | A |
| **AL** | A | A |

Comparando as duas últimas colunas do quadro 5.32, tem-se que apenas o alimentador AD teve classificações diferentes.

Como será realizada a comparação dessas técnicas com as que consideram distâncias (euclidiana e de mahalanobis – não utilizam validação cruzada, exceto para $k$-vizinhos mais próximos na determinação do $k$), aqui é admitido que a última coluna do quadro 5.32 apresenta a classificação correta dos alimentadores.

Comparando os resultados apresentados nos quadros 5.31 e 5.32, tem-se que a técnica que classificou corretamente todos os alimentadores foi a *SVM*. A segunda técnica com melhor desempenho é a RNA, que classificou cinco alimentadores (AB, AC, AD, AE e AI) diferentes daqueles que estão sendo admitidos como corretos, mas com classificações são em faixas vizinhas, por exemplo, AC foi classificado como “Faixa D”, se está admitindo que seja “Faixa C”.

Com relação ao AG, este apresentou duas classificações diferentes daquelas que estão sendo admitidas como corretas: ambos alimentadores, AH e AI, tiveram as classificações “Faixa C” no AG, ao invés de “Faixa A” e “Faixa E”, respectivamente. Pelo fato das classificações serem em faixas não vizinhas, admite-se que esta técnica, com a função *fitness* desenvolvida, obteve pouca qualidade para uma generalização. Provavelmente, um estudo mais aprofundado na obtenção da função *fitness* proporcionará uma melhor classificação.

#### *5.2.2.2 Comparação das classificações obtidas nas técnicas que consideram distâncias*

Na segunda análise é realizada a comparação das classificações obtidas nas técnicas que levam em consideração distâncias. O quadro 5.33 a seguir, apresenta os resultados obtidos por estas técnicas.

**QUADRO 5.33 –** COMPARAÇÃO DAS CLASSIFICAÇÕES OBTIDAS PELAS TÉCNICAS QUE LEVAM EM CONSIDERAÇÃO "DISTÂNCIA"

| Alimentador | Somatório das distâncias aos elementos do conjunto | Distância ao ponto central do conjunto | 3-vizinhos mais próximos | Distância à média do conjunto |
|---|---|---|---|---|
| **AA** | A | A | A | A |
| **AB** | C | C | C | C |
| **AC** | C | C | D | C |
| **AD** | C | C | B | C |
| **AE** | B | B | B | B |
| **AF** | F | F | F | F |
| **AG** | A | A | A | A |
| **AH** | A | A | A | A |
| **AI** | E | E | F | E |
| **AJ** | A | A | A | A |
| **AK** | A | A | A | A |
| **AL** | A | A | A | A |

Analisando o quadro 5.33 anterior, tem-se que os alimentadores AC, AD e AI não possuem a mesma classificação em todas as técnicas. No entanto, as classificações obtidas estão em faixas de classificação vizinhas da etiqueta de qualidade. Também é verificado que as duas primeiras técnicas e a quarta determinam a mesma classificação para este teste.

### *5.2.2.3 Comparação das classificações obtidas em todas as técnicas*

Na terceira análise é realizada comparação das classificações obtidas em todas as técnicas. Assim, o quadro 5.34, a seguir, apresenta estes resultados, bem como a classificação por voto destes.

**QUADRO 5.34 –** COMPARAÇÃO DAS CLASSIFICAÇÕES OBTIDAS PELAS TÉCNICAS

| Alimentador | RNA | *SVM* | AG | Distância | | | 3-vizinhos mais próximos | Classificação por voto |
|---|---|---|---|---|---|---|---|---|
| | | | | a todos os elementos | ao ponto central | à média | | |
| **AA** | A | A | A | A | A | A | A | **A** |
| **AB** | D | C | C | C | C | C | C | **C** |
| **AC** | D | C | C | C | C | C | D | **C** |
| **AD** | C | B | B | C | C | C | B | **C** |
| **AE** | B | A | A | B | B | B | B | **B** |
| **AF** | F | F | F | F | F | F | F | **F** |
| **AG** | A | A | A | A | A | A | A | **A** |
| **AH** | A | A | C | A | A | A | A | **A** |
| **AI** | D | E | C | E | E | E | F | **E** |
| **AJ** | A | A | A | A | A | A | A | **A** |
| **AK** | A | A | A | A | A | A | A | **A** |
| **AL** | A | A | A | A | A | A | A | **A** |

No quadro 5.34 anterior, há seis alimentadores (AA, AF, AG, AJ, AK e AL) que obtiveram a mesma classificação em todas as técnicas e, ainda, comparando a última coluna desse quadro com as demais, tem-se que as técnicas que levam em consideração “somatório das distâncias euclidiana do novo elemento aos elementos de cada faixa da etiqueta de qualidade”, “distância euclidiana do novo elemento ao ponto central de cada faixa da etiqueta de qualidade” e “distância euclidiana do novo elemento a média dos elementos de cada faixa da etiqueta de qualidade” são as técnicas com resultados idênticos a classificação por voto realizada com todas as técnicas aplicadas.

A figura 5.9 apresenta a classificação de cada alimentador, conforme quadro 5.34, através da etiqueta de classificação, de forma comparativa.

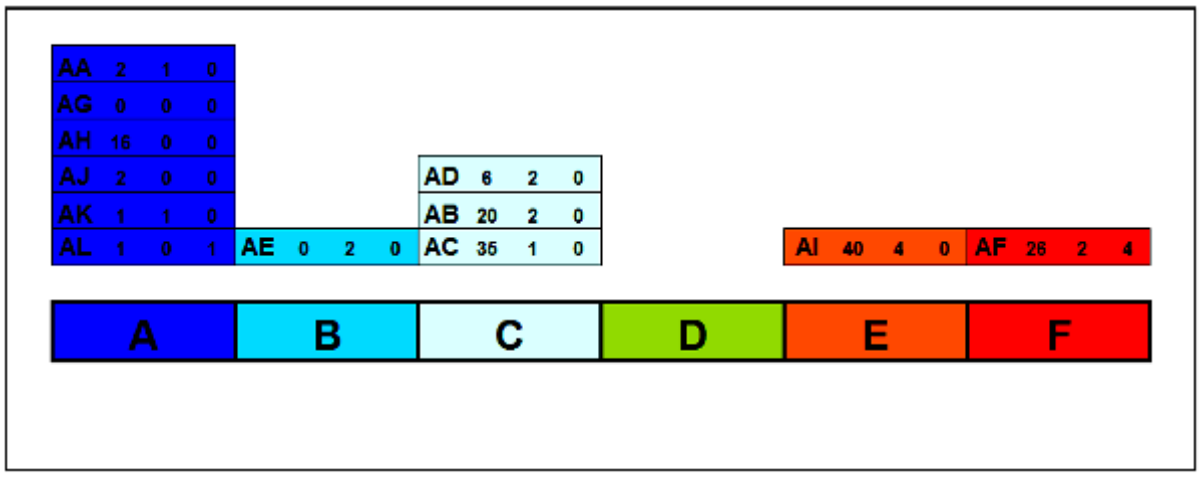


**FIGURA 5.9 –** ETIQUETA DE CLASSIFICAÇÃO DA QEE DOS ALIMENTADORES, DE FORMA COMPARATIVA
**FONTE:** O AUTOR (2012)

Os valores apresentados para cada alimentador expressam a ocorrência de eventos em cada uma das classes $C_5$, $C_7$ e $C_8$, nesta ordem. A etiqueta apresenta conhecimento não explícito quando analisado esses valores, como, por exemplo, a classificação dos alimentadores AI e AF, pois os valores de $C_5$ e $C_7$ de AI são maiores que os de AF, o que poderia indicar menor qualidade para AI em relação à AF, mas como $C_8$ possui valor menor para AI, as técnicas aplicadas indicaram que a AF possui qualidade menor que o alimentador AI.

Assim, a metodologia aqui proposta e aplicada nesse estudo de caso, no contexto *KDD*, revelou conhecimentos não explícitos nos bases de dados da concessionária de energia elétrica.

# 6 ESTUDO DE CASO 02: ÁREA EDUCACIONAL

Neste capítulo é apresentada a concepção de qualidade na educação brasileira e, na sequência, é descrita a aplicação da metodologia proposta neste trabalho com o intuito de apresentar uma etiqueta de qualidade educacional em relação ao desempenho dos estudantes na Prova Brasil.

## 6.1 O CUSTO-ALUNO QUALIDADE INICIAL - CAQi

Como já comentado o conceito de qualidade pode ter muitos significados e depende de onde seu uso é empregado. Segundo Carreira e Pinto (2007), na educação este conceito está relacionado à concepção de educação de quem o define. Fica claro que tal conceito nessa área possui diferentes significados, uma vez que há diversas concepções de educação, sendo muitas discordantes em muitos pontos.

No entanto, a Constituição Federal Brasileira (CFB) e a Lei de Diretrizes e Bases da Educação Nacional (LDB) asseguram que o ensino ofertado deve ter um padrão mínimo de qualidade. Além disto, a LDB afirma que o não cumprimento de uma qualidade mínina na educação fere o direito à aprendizagem dos alunos previsto na CFB. Já a Câmera de Educação Básica (CEB) aponta que os repasses de recursos e assistências técnicas para cumprimento deste direito é de obrigação da União.

Na busca de se estabelecer quais são os padrões mínimos e atributos relacionados à qualidade de educação, a CEB, em seu parecer 8/2010, indica o Custo-aluno Qualidade Inicial (CAQi) como um instrumento possível de apresentar com clareza os insumos necessários para garantir tal padrão. Assim, o CAQi deve ser tratado como “uma opção estabelecida para tornar viável o passo inicial rumo à qualidade, daí a designação Custo Aluno Qualidade Inicial” (BRASIL, 2010).

O CAQi teve origem na Campanha Nacional pelo Direito à Educação e, em 2008, passou a ser considerado pelo Conselho Nacional de Educação como “uma estratégia de política pública para a educação brasileira, no sentido de vencer as históricas desigualdades de ofertas educacionais em nosso país” (BRASIL, 2010). Esse conselho “entende que a adoção do CAQi representa um passo decisivo no enfrentamento dessas diferenças e, portanto, na busca de uma maior equalização

de oportunidades educacionais para todos" (BRASIL, 2010), ou seja, o conceito de qualidade aqui empregado está diretamente relacionado à perspectiva democrática e de qualidade social.

Ao apresentar o CAQi, Carreira e Pinto (2007, p. 77-78) assumem que os valores apresentados para cada etapa e modalidade de ensino estabelecem um padrão mínimo para qualidade de educação, e que este tende a crescer à medida que a exigência ou a qualidade aumenta, ou seja, é um processo dinâmico. Além disso, os valores apresentados são com base nos atributos indispensáveis ao desenvolvimento dos processos de ensino e aprendizagem, dentre eles: remuneração dos profissionais do magistério e demais profissionais da educação, infraestrutura e qualificação docente definidos pelo Plano Nacional de Educação.

A CEB destaca alguns dos fatores presentes no CAQi que estão fortemente relacionados à qualidade de educação, como o tamanho da unidade educacional, a quantidade de alunos por turma, o tempo diário de permanência do aluno na unidade educacional (parcial ou integral) e a valorização dos profissionais do magistério (formação inicial e continuada e plano de cargos e carreira).

Assim, os resultados esperados na educação estão intimamente relacionados aos recursos disponibilizados para tal, pois são estes que geram boa infraestrutura de trabalho, gestão de ensino adequada e a valorização do profissional da educação.

O parecer da CEB finaliza suas considerações indicando que há grandes desafios a serem vencidos em relação à qualidade de educação (BRASIL, 2007): universalizar o acesso desde a Pré-Escola ao Ensino Médio; reduzir a diferença entre escolas no que diz respeito às condições de infraestrutura; implantar Planos de Cargos e Carreira, o piso nacional salarial para os profissionais da educação e a hora-atividade para o docente; promover formação inicial e continuada adequada aos docentes; assegurar que os Estados, Distrito Federal e Municípios alcancem, nos próximos dez anos, um Índice de Desenvolvimento da Educação Básica (IDEB) de 6,0; melhorar a gestão educacional, tanto da escola quanto dos sistemas educacionais; e proporcionar financiamento adequado e compatível com as exigências da sociedade contemporânea.

O CAQi apresenta diversos atributos que compõem a qualidade na escola, e afirma ainda que os resultados esperados na educação estão intimamente relacionados aos recursos disponibilizados para tal. Dentre os resultados esperados

temos o desempenho escolar, que compõem muitos dos índices educacionais atuais, por exemplo, o IDEB e Programa Internacional de Avaliação de Alunos (*Programme for International Student Assessment* ou *PISA*), que utilizam alguma forma de avaliação (prova) em sua composição.

Portanto, este estudo de caso, se limitará a analisar o desempenho da escola, consequentemente, dos alunos, na Prova Brasil que compõe o IDEB.

## 6.2 CRIAÇÃO DA ETIQUETA DE QUALIDADE EDUCACIONAL EM RELAÇÃO AO DESEMPENHO NA PROVA BRASIL

O Índice de Desenvolvimento e Educação (IDEB), criado em 2007, é calculado levando em consideração os resultados de avaliações do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), taxas de aprovação/reprovação e evasão escolar, tanto em escolas públicas, quanto em escolas particulares (INEP, 2011). Com isso, espera-se que uma escola com alto índice no IDEB transpareça que seus alunos aprendem os conteúdos, não possuem reprovações e ainda frequentam a sala de aula.

O IDEB possui um índice para cada segmento da Educação Básica, ou seja, há um indicador para os anos iniciais do Ensino Fundamental, outro para os anos finais do Ensino Fundamental e outro para o Ensino Médio. Isso fragmenta a análise da escola e, por conseqüência, não verifica a qualidade total da instituição quando esta oferta mais que um destes segmentos.

A cada dois anos, um novo índice é divulgado e todos podem ter acesso no *site* do Ministério da Educação. Neste sistema de avaliação, um dos instrumentos é a Prova Brasil, aplicada somente a alunos do 5º ano (anos iniciais) e 9º ano (anos finais) do Ensino Fundamental das escolas públicas que possuem no mínimo 20 alunos matriculados nestes anos. Como este índice leva em consideração a aprovação e evasão escolar, além da nota em uma prova específica, ocorre com frequência que escolas com melhores notas na Prova Brasil têm índices inferiores do IDEB, em relação a outras.

Neste trabalho não se pretende questionar os fatores que levaram a criação do IDEB com estas considerações, mas sim propor a aplicação de uma metodologia para criação de etiqueta de qualidade, considerando apenas o desempenho escolar dos alunos. Neste estudo de caso na área educacional, serão investigadas escolas

que ofertem o Ensino Fundamental (1º ao 9º ano) em uma mesma região, de forma comparativa, utilizando apenas as notas da Prova Brasil nas duas áreas do conhecimento que esta avalia: Língua Portuguesa (notas com escala de 0 a 350) e Matemática (notas com escala de 0 a 425).

Para este estudo de caso foi selecionado o município de Araucária, região metropolitana de Curitiba/Pr, que em 2009, ano de aplicação da Prova Brasil, possuía 17 escolas municipais de Ensino Fundamental (anos iniciais e anos finais – 1º ao 9º ano). E após a coleta dos dados no site de INEP, foram **selecionados os dados** apenas da Prova Brasil. (Tabela 6.1)

**TABELA 6.1 –** NOTAS DA PROVA BRASIL

| | **Anos Iniciais** | | **Anos Finais** | |
|---|---|---|---|---|
| **Escola** | **Língua Portuguesa** | **Matemática** | **Língua Portuguesa** | **Matemática** |
| **E1** | 199,05 | 219,16 | 250,40 | 258,33 |
| **E2** | 176,19 | 204,38 | 246,03 | 243,44 |
| **E3** | 195,01 | 206,72 | 238,16 | 243,29 |
| **E4** | 192,45 | 215,27 | 247,60 | 249,31 |
| **E5** | 190,40 | 218,36 | 251,58 | 258,46 |
| **E6** | 194,40 | 214,96 | 239,08 | 244,28 |
| **E7** | 197,18 | 218,81 | 227,29 | 235,91 |
| **E8** | 183,41 | 202,93 | 219,90 | 229,18 |
| **E9** | 185,14 | 212,60 | 255,67 | 257,05 |
| **E10** | 194,20 | 214,98 | 237,33 | 252,94 |
| **E11** | 183,44 | 206,16 | 238,11 | 240,13 |
| **E12** | 174,40 | 199,76 | 240,94 | 242,30 |
| **E13** | 180,53 | 205,80 | 247,05 | 250,21 |
| **E14** | 183,24 | 229,39 | 252,05 | 267,19 |
| **E15** | 174,47 | 189,87 | 262,62 | 259,40 |
| **E16** | 201,41 | 238,69 | 260,56 | 262,36 |
| **E17** | 198,51 | 217,58 | 279,54 | 284,39 |

**FONTE:** INEP, 2011

Analisando a tabela 6.1 anterior, não é possível verificar a escola que mais se destaca em relação às notas, uma vez que, por exemplo, a escola E17 possui as melhores notas nos anos finais, mas não é a melhor escola nos anos inicias, pois a que possui as melhores notas nestes anos é a escola E16. Da mesma forma, não é possível indicar a pior escola, pois as notas mais baixas nos anos iniciais são das escolas E15 e E12, mas nos anos finais as notas mais baixas são da escola E8.

Assim, nessa metodologia as escolas são classificadas de forma comparativa, indicando sua qualidade em relação ao desempenho na Prova Brasil em uma escala de seis níveis (A, B, C, D, E e F). Como a criação de tal etiqueta

possui em seu contexto o processo *KDD*, sendo a próxima etapa deste processo o **pré-processamento dos dados**, em que são realizadas limpeza nos dados quando for necessária, nesse estudo de caso esta etapa não é necessária uma vez que os dados foram selecionados conforme o objetivo da aplicação.

A seguir é construída a etiqueta da qualidade educacional em relação ao desempenho na Prova Brasil, de forma comparativa, e para isso os dados foram organizados em quadros individuais com quatro classes de classificação: $C_1$ – notas de língua portuguesa nos anos inicias; $C_2$ – notas de matemática nos anos inicias; $C_3$ – notas de língua portuguesa nos anos finais; e $C_4$ – notas de matemática nos anos finais.

Assim, temos que $174{,}40 \leq C_1 \leq 201{,}41$; $189{,}87 \leq C_2 \leq 238{,}69$; $219{,}90 \leq C_3 \leq 279{,}54$; e $229{,}18 \leq C_4 \leq 284{,}39$.

Como exemplo, o quadro 6.1 apresenta os dados da escola E1, onde $C_1 = 199{,}05$, $C_2 = 219{,}16$, $C_3 = 250{,}40$ e $C_4 = 258{,}33$.

**QUADRO 6.1 –** NOTAS DA PROVA BRASIL DA ESCOLA E1

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 199,05 | 219,16 |
| Anos Finais | 250,40 | 258,33 |

Com base na tabela 6.1 foram elaborados quadros como o quadro 6.1 para cada escola, sendo que estes encontram-se no Anexo 04. Já o quadro a seguir apresenta o valor médio para cada classe de classificação, ou seja, para cada classe $C_i$ foi realizada a média em relação a todas as escolas.

**QUADRO 6.2 –** MÉDIA DAS NOTAS DA PROVA BRASIL DA REGIÃO ESCOLHIDA

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 188,44 | 212,67 |
| Anos Finais | 246,70 | 251,66 |

Definidos os valores médios foi estabelecido o valor das seis faixas da etiqueta, sendo a “Faixa A” a de melhor qualidade e a “Faixa F” a pior, em que os valores das etiquetas variam conforme os limites de cada $C_i$, $i=1...4$, apresentados na figura 6.1.

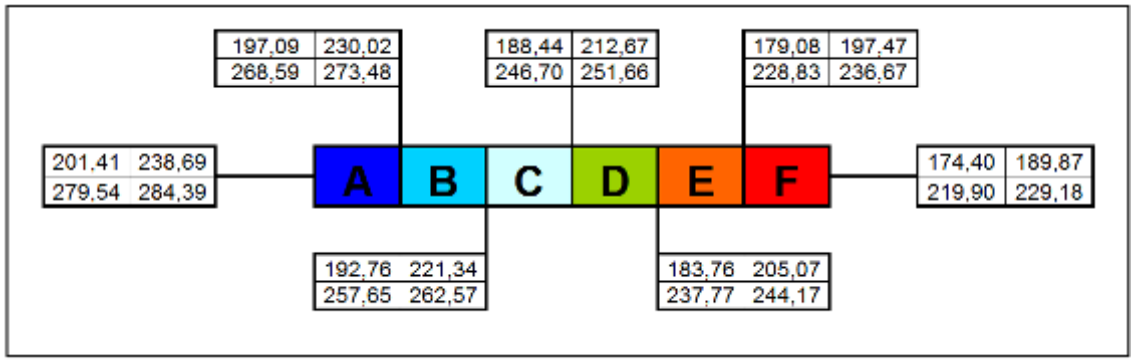


**FIGURA 6.1 –** ETIQUETA DE CLASSIFICAÇÃO DA QUALIDADE EDUCACIONAL, DE FORMA COMPARATIVA
**FONTE:** O AUTOR (2012)

O limite superior de cada faixa de classificação (Figura 6.1) foi definido da seguinte forma: o $lim\ sup\ A$ e o $lim\ sup\ B$ foram determinados tal que, para cada $C_i$, tem-se $(lim\ inf\ A - lim\ sup\ A) = (lim\ sup\ A - lim\ sup\ B) = (lim\ sup\ B - lim\ sup\ C)$. O mesmo ocorre com as faixas D, E e F: $(lim\ sup\ C - lim\ sup\ D) = (lim\ sup\ D - lim\ sup\ E) = = (lim\ sup\ E - lim\ sup\ F)$.

Assim, criada a EQ em relação ao desempenho na Prova Brasil, fica evidente que se deve verificar em qual faixa cada escola (quadro 6.1, e Anexo 04) se enquadra. No entanto, das 17 escolas analisadas, apenas uma delas se enquadra diretamente em alguma faixa da etiqueta de qualidade educacional, a E5 classificada como “Faixa C”.

A figura 6.2, mostra que a escola E5 possui $188,44 \leq C1=190,40 \leq 192,76$; $212,67 \leq C2=218,36 \leq 221,34$; $246,70 \leq C3=251,58 \leq 257,65$; e $251,66 \leq C4=258,46 \leq 262,57$.

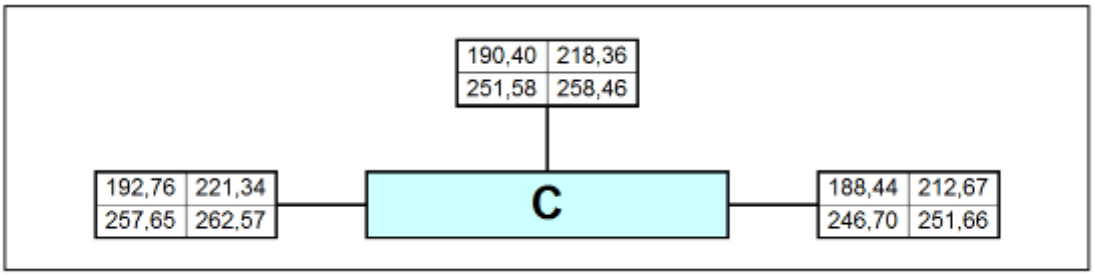


**FIGURA 6.2 –** ESCOLA E5 CLASSIFICADA DIRETAMENTE NA ETIQUETA DE QUALIDADE EDUCACIONAL
**FONTE:** O AUTOR (2012)

As demais escolas não podem ser classificadas diretamente, uma vez que, por exemplo, para a escola E4, os valores $C_1$, $C_2$ e $C_3$ pertencem à “Faixa C” da etiqueta de qualidade, mas $C_4$ pertence a “Faixa B”.

Assim, qual é a classificação das demais escolas?

Para responder a essa pergunta, foram utilizadas as **técnicas de *Data Mining*** (apresentadas nas seções 4.3 a 4.7) com a finalidade de verificar suas classificações na etiqueta de qualidade. No entanto, para a aplicação das técnicas é necessário realizar a mudança de escala (**transformação nos dados**) nos dados extraídos do site do INEP, como se segue.

Analisando a segunda coluna da tabela 6.1, os valores pertencem ao intervalo contínuo [174,40; 201,41]. Já a terceira coluna possui valores pertencentes ao intervalo contínuo [189,87; 238,69]. A quarta coluna com valores em [219,90; 279,54] e, por fim, a quinta coluna possui os valores no intervalo contínuo [229,18; 284,39]. Assim, primeiro foi subtraído de cada coluna o menor valor entre todas as linhas, por exemplo, na segunda coluna foi subtraído 174,40 de todas as linhas desta coluna. Com isso, os novos valores pertencentes aos seguintes intervalos, [0; 27,01], [0; 48,82], [0; 59,64] e [0; 55,21], respectivamente.

Na sequência, foi determinado o maior valor de cada coluna e todos os elementos desta foram divididos por esse valor, finalizando desta forma a mudança de escala dos dados, presentes na tabela 6.2 a seguir, que agora estão todos no intervalo [0, 1].

**TABELA 6.2 –** NOTAS DA PROVA BRASIL APÓS TRANSFORMAÇÃO (MUDANÇA DE ESCALA)

| | Anos Iniciais | | Anos Finais | |
|---|---|---|---|---|
| **Escola** | **Língua Portuguesa** | **Matemática** | **Língua Portuguesa** | **Matemática** |
| **E1** | 0,91 | 0,60 | 0,51 | 0,53 |
| **E2** | 0,07 | 0,30 | 0,44 | 0,26 |
| **E3** | 0,76 | 0,35 | 0,31 | 0,26 |
| **E4** | 0,67 | 0,52 | 0,46 | 0,36 |
| **E5** | 0,59 | 0,58 | 0,53 | 0,53 |
| **E6** | 0,74 | 0,51 | 0,32 | 0,27 |
| **E7** | 0,84 | 0,59 | 0,12 | 0,12 |
| **E8** | 0,33 | 0,27 | 0,00 | 0,00 |
| **E9** | 0,40 | 0,47 | 0,60 | 0,50 |
| **E10** | 0,73 | 0,51 | 0,29 | 0,43 |
| **E11** | 0,33 | 0,33 | 0,31 | 0,20 |
| **E12** | 0,00 | 0,20 | 0,35 | 0,24 |
| **E13** | 0,23 | 0,33 | 0,46 | 0,38 |
| **E14** | 0,33 | 0,81 | 0,54 | 0,69 |
| **E15** | 0,00 | 0,00 | 0,72 | 0,55 |
| **E16** | 1,00 | 1,00 | 0,68 | 0,60 |
| **E17** | 0,89 | 0,57 | 1,00 | 1,00 |

Para que os estudos de caso apresentados neste trabalho tenham o mesmo objetivo, ou seja, elementos (alimentadores e escolas) com os valores mais próximos de zero devem possuir classificação mais próxima da “Faixa A”, todos os valores da tabela 6.2 anterior foram subtraídos de “1”, gerando assim a tabela 6.3, uma vez que este problema é de maximização.

**TABELA 6.3 –** NOTAS DA PROVA BRASIL – TERCEIRA TRANSFORMAÇÃO

| | Anos Iniciais | | Anos Finais | |
|---|---|---|---|---|
| **Escola** | **Língua Portuguesa** | **Matemática** | **Língua Portuguesa** | **Matemática** |
| **E1** | 0,09 | 0,40 | 0,49 | 0,47 |
| **E2** | 0,93 | 0,70 | 0,56 | 0,74 |
| **E3** | 0,24 | 0,65 | 0,69 | 0,74 |
| **E4** | 0,33 | 0,48 | 0,54 | 0,64 |
| **E5** | 0,41 | 0,42 | 0,47 | 0,47 |
| **E6** | 0,26 | 0,49 | 0,68 | 0,73 |
| **E7** | 0,16 | 0,41 | 0,88 | 0,88 |
| **E8** | 0,67 | 0,73 | 1,00 | 1,00 |
| **E9** | 0,60 | 0,53 | 0,40 | 0,50 |
| **E10** | 0,27 | 0,49 | 0,71 | 0,57 |
| **E11** | 0,67 | 0,67 | 0,69 | 0,80 |
| **E12** | 1,00 | 0,80 | 0,65 | 0,76 |
| **E13** | 0,77 | 0,67 | 0,54 | 0,62 |
| **E14** | 0,67 | 0,19 | 0,46 | 0,31 |
| **E15** | 1,00 | 1,00 | 0,28 | 0,45 |
| **E16** | 0,00 | 0,00 | 0,32 | 0,40 |
| **E17** | 0,11 | 0,43 | 0,00 | 0,00 |

Posto isso, na seção 6.1 são aplicadas as técnicas de *DM* com a finalidade de classificar as escolas.

### 6.2.1 Aplicação das Técnicas de *Data Mining*

A metodologia de aplicação das técnicas de *DM* a este estudo de caso é descrita no capítulo 4. Para cada faixa de classificação há 60 exemplos compondo o conjunto de dados, sendo 59 gerados aleatoriamente respeitando os limites de cada $C_i$, mais o limite superior de cada faixa. Assim, como são seis faixas de classificação (de A a F) tem-se 360 registros, dos quais 240 são utilizados para o treinamento (ou seja, 2/3) e 120 para testes (ou seja,1/3) em cada etapa da validação cruzada.

#### *6.2.1.1 Redes Neurais*

Nesta aplicação (conforme seção 4.3.1) em todas as etapas as RNA aprenderam todos os exemplos corretamente. Quanto ao conjunto de teste, a 1ª e 2ª etapa aprenderam todos os exemplos, mas na 3ª etapa houve quatro exemplos classificados incorretamente, gerando assim, aproximadamente, 96,67% de acerto nessa etapa da técnica.

De maneira geral, as etapas de aplicação mostraram o grande potencial da técnica na generalização, uma vez que a média entre as três etapas é, aproximadamente, 99,63% de acertos.

Na sequência da aplicação da técnica foram apresentados às RNA os dados referentes a cada escola e esta classificação, em cada etapa de aplicação, é apresentada no quadro 6.3.

Também foi realizada a simulação utilizando todos os exemplos, ou seja, não houve a separação destes em subconjuntos. Nesta simulação todos os exemplos foram aprendidos corretamente. Quando os dados das escolas foram apresentados, a RNA obteve as classificações expostas no quadro 6.3, a seguir.

**QUADRO 6.3 –** RESULTADO DA CLASSIFICAÇÃO DAS ESCOLAS COM A APLICAÇÃO DAS RNA

| Escola | Validação Cruzada | | | | Todos os exemplos no treinamento |
|---|---|---|---|---|---|
| | 1ª etapa | 2ª etapa | 3ª etapa | Classificação por voto | |
| **E1** | C | C | B | C | C |
| **E2** | E | E | E | E | E |
| **E3** | D | D | D | D | D |
| **E4** | C | C | C | C | C |
| **E5** | C | C | C | C | C |
| **E6** | C | C | D | C | C |
| **E7** | D | D | D | D | C |
| **E8** | E | E | E | E | E |
| **E9** | C | C | C | C | C |
| **E10** | C | C | C | C | C |
| **E11** | E | E | D | E | E |
| **E12** | F | F | E | F | F |
| **E13** | D | D | D | D | D |
| **E14** | E | C | C | C | C |
| **E15** | F | D | E | F | D |
| **E16** | B | B | B | B | B |
| **E17** | A | A | A | A | A |

Ao analisar o quadro 6.3 verifica-se que as classificações das últimas duas colunas são iguais, exceto para as escolas E7 e E15. Cabe ressaltar que a escola E5, a única que já se sabia sua classificação, é classificada corretamente.

### *6.2.1.2 Support Vector Machine*

Nesta aplicação (conforme seção 4.4.1) em cada etapa (1ª etapa, 2ª etapa e 3ª etapa da validação cruzada, seção 4.2) o *SVM* apresentou 100% de acerto. Também foram apresentados os dados referentes às escolas a cada etapa, sendo que a classificação obtida está no quadro 6.4. E também, na sequência, houve novo treinamento considerando todos os exemplos e o resultado desta classificação consta no quadro 6.4, a seguir.

**QUADRO 6.4 –** RESULTADO DA CLASSIFICAÇÃO DAS ESCOLAS COM A APLICAÇÃO DO *SVM*

| Escola | Validação Cruzada | | | | Todos os exemplos no treinamento |
|---|---|---|---|---|---|
| | 1ª etapa | 2ª etapa | 3ª etapa | Classificação por voto | |
| **E1** | B | B | B | B | B |
| **E2** | E | E | E | E | E |
| **E3** | D | D | D | D | D |
| **E4** | C | C | C | C | C |
| **E5** | C | C | C | C | C |
| **E6** | D | D | D | D | D |
| **E7** | D | D | D | D | D |
| **E8** | F | F | F | F | F |
| **E9** | C | C | C | C | C |
| **E10** | C | C | C | C | C |
| **E11** | E | E | E | E | E |
| **E12** | E | E | E | E | E |
| **E13** | D | D | D | D | D |
| **E14** | C | C | C | C | C |
| **E15** | D | D | E | D | D |
| **E16** | A | B | A | A | A |
| **E17** | A | A | A | A | A |

No quadro 6.4 pode-se verificar que as classificações das últimas duas colunas são iguais. Apenas a 3ª etapa para a escola E15 e a 2ª etapa para a escola E16 são diferentes das demais colunas, considerando a mesma linha.

Esta técnica também classificou correta a escola E5 (escola com classificação direta na etiqueta de qualidade).

### *6.2.1.3 Algoritmo Genético*

A terceira técnica utilizada para classificar as escolas é o AG. A função fitness foi determinada conforme seção 4.5.1, que neste caso procurou um hiperplano que separasse os conjuntos de dados, visto que se têm quatro classes: $C_1$, $C_2$, $C_3$ e $C_4$.

Ao realizar cada etapa em busca da classificação, a 1ª e 3ª etapas classificaram todos os exemplos do teste corretamente, e na 2ª etapa houve, aproximadamente, 98,33% de acerto. A classificação de cada escola obtida nessas etapas está exposta no quadro 6.5, bem como o resultado da classificação quando foi aplicado o AG utilizando todos os exemplos no treinamento.

**QUADRO 6.5 –** RESULTADO DA CLASSIFICAÇÃO DAS ESCOLAS COM A APLICAÇÃO DO AG

| Escola | Validação Cruzada | | | | Todos os exemplos no treinamento |
|---|---|---|---|---|---|
| | 1ª etapa | 2ª etapa | 3ª etapa | Classificação por voto | |
| **E1** | C | B | C | C | B |
| **E2** | E | F | E | F | E |
| **E3** | C | D | D | D | D |
| **E4** | C | C | C | C | C |
| **E5** | C | C | C | C | C |
| **E6** | C | D | D | D | D |
| **E7** | C | D | D | D | B |
| **E8** | F | E | E | E | F |
| **E9** | D | C | C | C | C |
| **E10** | C | D | D | D | D |
| **E11** | D | D | E | D | E |
| **E12** | E | F | F | F | E |
| **E13** | E | E | D | E | D |
| **E14** | B | C | A | C | C |
| **E15** | E | F | D | F | D |
| **E16** | B | B | B | B | B |
| **E17** | A | A | B | A | A |

Ao comparar o resultado da classificação por voto da validação cruzada com o treinamento utilizando todos os dados, o quadro 6.5 mostra muitos resultados distintos. Das 17 escolas apenas oito possuem a mesma classificação nas duas últimas colunas, nas demais, tem-se que em sete as classificações são distintas em faixas vizinhas e duas em faixas não vizinhas (escolas E7 e E15).

### *6.2.1.4 Técnicas que utilizam distância euclidiana*

Concluído os testes com a RNA, *SVM* e AG, técnicas consideradas sofisticadas, foram realizados outros testes que levam em consideração a distância euclidiana da escola aos dados de cada faixa da etiqueta de qualidade.

Assim, o primeiro teste calcula o somatório das distâncias do novo elemento aos elementos de cada faixa da etiqueta de qualidade (seção 4.6.1). O resultado destes cálculos e sua classificação são apresentados no quadro 6.6, no qual cada valor em negrito corresponde ao menor valor do somatório obtido para cada escola.

Observa-se no quadro 6.6 que nenhuma das escolas possui classificação F. A escola que mais se aproximou desta faixa de classificação foi a E8 (classificada com qualidade E).

**QUADRO 6.6 –** RESULTADO DA CLASSIFICAÇÃO DAS ESCOLAS COM A APLICAÇÃO SOMATÓRIO DAS DISTÂNCIAS DO NOVO ELEMENTO AOS ELEMENTOS DE CADA FAIXA DA ETIQUETA DE QUALIDADE

| Escola | Faixa A | Faixa B | Faixa C | Faixa D | Faixa E | Faixa F | Classi-ficação |
|---|---|---|---|---|---|---|---|
| **E1** | 37,77 | 21,26 | ***19,31*** | 34,35 | 51,76 | 70,01 | C |
| **E2** | 78,95 | 58,18 | 39,45 | 24,31 | ***18,85*** | 28,46 | E |
| **E3** | 63,57 | 43,54 | 25,38 | ***20,75*** | 31,31 | 47,35 | D |
| **E4** | 50,26 | 29,22 | ***10,72*** | 17,24 | 34,65 | 52,98 | C |
| **E5** | 42,15 | 20,95 | ***6,44*** | 21,64 | 39,61 | 58,13 | C |
| **E6** | 57,85 | 37,84 | 20,40 | ***20,40*** | 34,18 | 51,10 | D |
| **E7** | 69,32 | 51,26 | 36,13 | ***33,67*** | 41,92 | 55,18 | D |
| **E8** | 92,74 | 71,29 | 49,69 | 31,89 | ***19,54*** | 19,97 | E |
| **E9** | 50,89 | 30,38 | ***15,46*** | 18,46 | 33,30 | 50,93 | C |
| **E10** | 53,39 | 33,78 | ***17,22*** | 20,89 | 35,91 | 53,31 | C |
| **E11** | 74,17 | 52,47 | 31,03 | 12,37 | ***10,19*** | 27,04 | E |
| **E12** | 86,65 | 65,78 | 46,43 | 29,64 | ***18,45*** | 22,18 | E |
| **E13** | 68,15 | 47,25 | 28,40 | ***15,18*** | 19,01 | 34,52 | D |
| **E14** | 44,18 | 29,06 | ***26,25*** | 35,64 | 49,43 | 65,64 | C |
| **E15** | 80,85 | 63,55 | 50,99 | 42,86 | ***41,73*** | 48,93 | E |
| **E16** | 24,67 | ***23,59*** | 37,70 | 55,59 | 73,53 | 91,82 | B |
| **E17** | ***22,04*** | 28,07 | 45,21 | 62,26 | 79,12 | 97,14 | A |

Para a aplicação da segunda técnica foi necessário determinar o ponto central (tabela 6.4) de cada faixa da etiqueta de qualidade (seção 4.6.2).

**TABELA 6.4 –** PONTO CENTRAL DE CADA CONJUNTO DE DADOS DAS FAIXAS DA ETIQUETA DE QUALIDADE

| Faixa | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|
| **A** | 0,08 | 0,10 | 0,10 | 0,09 |
| **B** | 0,25 | 0,27 | 0,27 | 0,32 |
| **C** | 0,39 | 0,45 | 0,48 | 0,51 |
| **D** | 0,55 | 0,61 | 0,62 | 0,67 |
| **E** | 0,73 | 0,77 | 0,78 | 0,79 |
| **F** | 0,91 | 0,92 | 0,93 | 0,93 |

Na sequência foi realizado o cálculo da distância de cada escola a cada ponto central, sendo sua classificação determinada pelo menor resultado exposto no quadro 6.7.

**QUADRO 6.7 –** DISTÂNCIA ENTRE CADA ESCOLA E CADA PONTO CENTRAL DOS DADOS DAS FAIXAS DA ETIQUETA DE QUALIDADE

| Escola | Faixa A | Faixa B | Faixa C | Faixa D | Faixa E | Faixa F | Classi-ficação |
|---|---|---|---|---|---|---|---|
| **E1** | 0,62 | 0,34 | **0,31** | 0,56 | 0,86 | 1,16 | C |
| **E2** | 1,31 | 0,95 | 0,64 | 0,40 | **0,31** | 0,47 | E |
| **E3** | 1,05 | 0,71 | 0,40 | **0,33** | 0,51 | 0,78 | D |
| **E4** | 0,84 | 0,48 | **0,16** | 0,27 | 0,57 | 0,88 | C |
| **E5** | 0,70 | 0,33 | **0,05** | 0,34 | 0,65 | 0,96 | C |
| **E6** | 0,96 | 0,62 | 0,32 | **0,32** | 0,56 | 0,84 | D |
| **E7** | 1,16 | 0,84 | 0,59 | **0,55** | 0,69 | 0,91 | D |
| **E8** | 1,54 | 1,18 | 0,82 | 0,53 | 0,32 | **0,32** | F |
| **E9** | 0,84 | 0,49 | **0,24** | 0,29 | 0,55 | 0,84 | C |
| **E10** | 0,89 | 0,55 | **0,27** | 0,33 | 0,59 | 0,88 | C |
| **E11** | 1,23 | 0,86 | 0,50 | 0,20 | **0,15** | 0,44 | E |
| **E12** | 1,44 | 1,09 | 0,77 | 0,50 | **0,30** | 0,36 | E |
| **E13** | 1,13 | 0,77 | 0,46 | **0,25** | 0,31 | 0,57 | D |
| **E14** | 0,73 | 0,47 | **0,43** | 0,59 | 0,82 | 1,09 | C |
| **E15** | 1,35 | 1,05 | 0,85 | 0,72 | **0,70** | 0,82 | E |
| **E16** | 0,40 | **0,38** | 0,63 | 0,92 | 1,22 | 1,53 | B |
| **E17** | **0,36** | 0,47 | 0,75 | 1,03 | 1,32 | 1,62 | A |

No quadro 6.7, acima, quando para duas faixas os valores são iguais, optou-se por classificar a escola com a pior colocação.

No terceiro teste é verificada a classificação dos $k$ elementos mais próximos de cada escola, técnica conhecida como $k$-vizinhos mais próximos. Conforme a seção 4.6.3, foram realizados testes para definir o valor de $k$ utilizando os mesmos conjuntos de treinamento e testes da validação cruzada das técnicas RNA, SVM e AG. Nesse estudo de caso, foram realizados testes com *k* variando de 1 a 40 (sendo 40 o número de elementos de cada faixa no conjunto de treinamento), em que se obteve 100% de acerto para todos os *k*.

Assim, optou-se em utilizar $k=1$ por ser o mais simples computacionalmente. O quadro a 6.8, a seguir, apresenta a classificação de cada escola em cada etapa da validação cruzada, a classificação por voto da validação cruzada e a classificação utilizando todos os dados no conjunto de treinamento.

**QUADRO 6.8 –** CLASSIFICAÇÃO DAS ESCOLAS – TÉCNICA DOS *K*-VIZINHOS MAIS PRÓXIMOS

| Escola | 1ª etapa | 2ª etapa | 3ª etapa | Classif. por voto - Validação cruzada | Todos os elementos |
|---|---|---|---|---|---|
| **E1** | B | B | B | B | B |
| **E2** | E | E | E | E | E |
| **E3** | D | D | D | D | D |
| **E4** | C | C | C | C | C |
| **E5** | C | C | C | C | C |
| **E6** | D | D | D | D | D |
| **E7** | D | D | D | D | D |
| **E8** | F | F | F | F | F |
| **E9** | C | D | C | C | C |
| **E10** | C | D | C | C | C |
| **E11** | E | E | D | E | E |
| **E12** | E | E | E | E | E |
| **E13** | D | D | D | D | D |
| **E14** | C | C | C | C | C |
| **E15** | E | D | D | D | D |
| **E16** | B | B | B | B | B |
| **E17** | A | A | A | A | A |

Analisando o quadro 6.8 anterior, tanto a classificação por voto obtida na validação cruzada quanto à classificação quando utilizado todos os elementos no conjunto de treinamento possuem a mesma classificação. Quanto às etapas da validação cruzada, as escolas E9, E10, E11 e E15 possuem classificações distintas, mas em faixa vizinhas.

A quarta técnica foi aplicada, pois foi verificado que os pontos centrais dos conjuntos são bastante próximos das médias dos elementos dos conjuntos, como pode ser verificado nas tabelas 6.5

**TABELA 6.5 –** COMPARAÇÃO DOS PONTOS CENTRAIS COM AS MÉDIAS DOS ELEMENTOS DO CONJUNTO DE DADOS DA ÁREA EDUCACIONAL

| FAIXA | *PONTO CENTRAL* | | | | *MÉDIA DOS ELEMENTOS DO CONJUNTO DE DADOS* | | | | *DISTÂNCIA ENTRE OS PONTOS* |
|---|---|---|---|---|---|---|---|---|---|
| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | |
| **A** | **0,08** | 0,10 | 0,10 | 0,09 | **0,09** | 0,10 | 0,10 | 0,09 | 0,01 |
| **B** | 0,25 | 0,27 | 0,27 | **0,32** | 0,25 | 0,27 | 0,27 | **0,31** | 0,01 |
| **C** | 0,39 | 0,45 | **0,48** | 0,51 | 0,39 | 0,45 | **0,47** | 0,51 | 0,01 |
| **D** | **0,55** | 0,61 | **0,62** | 0,67 | **0,56** | 0,61 | **0,63** | 0,67 | 0,01 |
| **E** | 0,73 | 0,77 | **0,78** | 0,79 | 0,73 | 0,77 | **0,77** | 0,79 | 0,01 |
| **F** | 0,91 | 0,92 | 0,93 | 0,93 | 0,91 | 0,92 | 0,93 | 0,93 | 0,00 |

Na tabela 6.5 anterior, os valores em negritos indicam as diferenças obtidas entre os pontos centrais e as médias dos elementos do conjunto de dados de cada faixa da etiqueta de classificação da qualidade. Essas diferenças, quando ocorrem, são de no máximo um centésimo. Já a última coluna da tabela indica a distância euclidiana entre ao ponto central e a correspondente média dos elementos do conjunto de dados.

Assim, esta técnica calcula a distância do novo elemento à média dos elementos de cada faixa da etiqueta de qualidade (seção 4.6.4). O quadro 6.9, a seguir, apresenta o resultado da aplicação desta técnica.

**QUADRO 6.9 –** DISTÂNCIA ENTRE CADA ESCOLA E A MÉDIA DOS DADOS DAS FAIXAS DA ETIQUETA DE QUALIDADE

| Escola | Faixa A | Faixa B | Faixa C | Faixa D | Faixa E | Faixa F | Classi-ficação |
|---|---|---|---|---|---|---|---|
| **E1** | 0,62 | 0,34 | **0,31** | 0,57 | 0,85 | 1,16 | C |
| **E2** | 1,30 | 0,96 | 0,64 | 0,39 | **0,30** | 0,47 | E |
| **E3** | 1,05 | 0,71 | 0,40 | **0,34** | 0,51 | 0,78 | D |
| **E4** | 0,84 | 0,48 | **0,16** | 0,28 | 0,57 | 0,88 | C |
| **E5** | 0,70 | 0,34 | **0,05** | 0,35 | 0,65 | 0,96 | C |
| **E6** | 0,96 | 0,63 | 0,33 | **0,33** | 0,56 | 0,84 | D |
| **E7** | 1,15 | 0,85 | 0,60 | **0,55** | 0,69 | 0,91 | D |
| **E8** | 1,54 | 1,18 | 0,82 | 0,52 | 0,32 | **0,32** | F |
| **E9** | 0,84 | 0,49 | **0,24** | 0,30 | 0,54 | 0,84 | C |
| **E10** | 0,89 | 0,56 | **0,28** | 0,34 | 0,58 | 0,88 | C |
| **E11** | 1,23 | 0,87 | 0,51 | 0,19 | **0,14** | 0,44 | E |
| **E12** | 1,44 | 1,09 | 0,77 | 0,49 | **0,30** | 0,36 | E |
| **E13** | 1,12 | 0,77 | 0,46 | **0,24** | 0,31 | 0,57 | D |
| **E14** | 0,72 | 0,47 | **0,43** | 0,59 | 0,82 | 1,09 | C |
| **E15** | 1,34 | 1,06 | 0,85 | 0,72 | **0,69** | 0,82 | E |
| **E16** | 0,40 | **0,38** | 0,62 | 0,92 | 1,22 | 1,53 | B |
| **E17** | **0,36** | 0,46 | 0,75 | 1,04 | 1,31 | 1,62 | A |

No quadro 6.9 anterior, quando para duas faixas os valores são iguais, optou-se por classificar a escola com a pior colocação.

### *6.2.1.5 Técnica estatística: Distância de Mahalanobis*

Para o cálculo da distância de Mahalanobis das escolas a cada conjunto de dados referentes às faixas de classificação, foi utilizada a ferramenta *mahal* do *Matlab* 7.9.0. O quadro 6.10 apresenta o resultado destes cálculos bem como a classificação de cada escola.

**QUADRO 6.10 –** DISTÂNCIA DE MAHALANOBIS ENTRE AS ESCOLAS E CADA CONJUNTO DE DADOS DAS FAIXAS DA ETIQUETA DE CLASSIFICAÇÃO DA QUALIDADE EDUCACIONAL

| Escola | Faixa A | Faixa B | Faixa C | Faixa D | Faixa E | Faixa F | Classi-ficação |
|---|---|---|---|---|---|---|---|
| **E1** | 13,32 | ***6,21*** | 8,01 | 11,85 | 24,29 | 25,86 | B |
| **E2** | 27,82 | 22,68 | 16,02 | ***7,97*** | 7,04 | 11,25 | D |
| **E3** | 21,79 | 13,90 | ***8,02*** | 8,15 | 12,61 | 16,92 | C |
| **E4** | 17,42 | 9,72 | ***3,19*** | 5,94 | 17,05 | 19,05 | C |
| **E5** | 14,63 | 7,29 | ***1,10*** | 7,61 | 20,11 | 22,14 | C |
| **E6** | 20,46 | 11,88 | ***6,32*** | 6,79 | 16,20 | 17,77 | C |
| **E7** | 25,50 | 15,37 | ***11,28*** | 11,58 | 19,78 | 18,60 | C |
| **E8** | 32,35 | 24,47 | 16,10 | 12,85 | 8,65 | ***6,82*** | F |
| **E9** | 17,73 | 11,89 | ***5,95*** | 7,16 | 16,64 | 19,86 | C |
| **E10** | 18,72 | 10,60 | ***5,32*** | 6,31 | 16,48 | 19,64 | C |
| **E11** | 25,64 | 19,21 | 11,06 | ***4,37*** | 4,89 | 9,60 | D |
| **E12** | 30,68 | 25,67 | 18,81 | 9,79 | ***6,64*** | 8,74 | E |
| **E13** | 23,80 | 18,27 | 11,45 | ***4,99*** | 8,98 | 13,75 | D |
| **E14** | 16,95 | ***9,08*** | 10,38 | 14,90 | 26,44 | 27,26 | B |
| **E15** | 29,02 | 26,19 | 20,32 | ***14,29*** | 18,42 | 20,39 | D |
| **E16** | 9,66 | ***8,86*** | 14,40 | 18,44 | 38,46 | 33,68 | B |
| **E17** | ***7,16*** | 8,23 | 14,47 | 25,12 | 36,86 | 38,95 | A |

### 6.2.2 Análise dos resultados obtidos

A análise dos resultados ocorre em três etapas, sendo a primeira a comparação das técnicas mais sofisticadas (RNA, *SVM* e AG), a segunda compara as técnicas que levam em consideração distâncias (euclidiana e de mahalanobis) e a terceira compara todas as técnicas. Nesta última análise apenas os testes com todos os dados no treinamento das técnicas RNA, *SVM* e AG são considerados.

#### *6.2.2.1 Comparação das classificações obtidas nas técnicas RNA, SVM e AG*

Para a primeira análise são consideradas as técnicas mais complexas (RNA, AG e *SVM*). O quadro 6.11 apresenta o resultado de cada técnica, sendo que para cada uma é apresentado o resultado da validação cruzada e o teste no qual foram considerados todos os elementos no conjunto de treinamento.

**QUADRO 6.11 –** COMPARAÇÃO DAS CLASSIFICAÇÕES OBTIDAS PELAS TÉCNICAS RNA, *SVM* E AG

| Escola | RNA | | *SVM* | | AG | |
|---|---|---|---|---|---|---|
| | TT | VC | TT | VC | TT | VC |
| **E1** | C | C | B | B | C | B |
| **E2** | E | E | E | E | F | E |
| **E3** | D | D | D | D | D | D |
| **E4** | C | C | C | C | C | C |
| **E5** | C | C | C | C | C | C |
| **E6** | C | C | D | D | D | D |
| **E7** | D | C | D | D | D | B |
| **E8** | E | E | F | F | E | F |
| **E9** | C | C | C | C | C | C |
| **E10** | C | C | C | C | D | D |
| **E11** | E | E | E | E | D | E |
| **E12** | F | F | E | E | F | E |
| **E13** | D | D | D | D | E | D |
| **E14** | C | C | C | C | C | C |
| **E15** | F | D | D | D | F | D |
| **E16** | B | B | A | A | B | B |
| **E17** | A | A | A | A | A | A |

**TT:** Treinamento com todos os dados; **VC:** classificação por voto considerando a validação cruzada *em cada técnica*;

Analisando o quadro 6.11 anterior, verifica-se que a escola E5, que já era conhecida sua colocação, foi classificada corretamente nas três técnicas. Das três técnicas, apenas o *SVM* apresenta a média da validação cruzada e o treinamento com todos os dados com classificação iguais para todas as escolas. As RNA diferem na classificação da E15 e o AG difere em várias escolas: E1, E2, E7, E8, E11, E12, E13 e E15.

Comparando as classificações obtidas, tem-se que as três técnicas tiveram os mesmos resultados para cinco escolas (colunas TT do quadro 6.11). Analisando as técnicas duas a duas tem-se que as RNA e *SVM* apresentam 11 escolas (E2, E3, E4, E5, E7, E9, E10, E11, E13, E14 e E17) com classificações iguais. Já as RNA e AG possuem 12 escolas (E1, E3, E4, E5, E7, E8, E9, E12, E14, E15, E16 e E17) com mesma classificação. Por fim, o *SVM* e AG possuem oito escolas (E3, E4, E5, E6, E7, E9, E14 e E17) com mesma classificação.

De maneira geral, com muita cautela, já que foi realizada apenas uma rodada de testes para cada técnica, pode-se afirmar que os resultados foram satisfatórios e as técnicas possuem resultados aceitáveis, uma vez que as classificações, exceto para a escola E15 (técnicas RNA e AG), apresentam classificações em faixas

vizinhas. Assim, o resultado da maior ocorrência de classificação de cada escola está descrito no quadro 6.12.

**QUADRO 6.12 –** CLASSIFICAÇÃO COMPARANDO AS TÉCNICAS RNA, *SVM* E AG

| Escola | Classificação considerando a validação cruzada | Classificação sem considerar a validação cruzada |
|---|---|---|
| **E1** | C | B |
| **E2** | E | E |
| **E3** | D | D |
| **E4** | C | C |
| **E5** | C | C |
| **E6** | D | D |
| **E7** | D | D |
| **E8** | F | F |
| **E9** | C | C |
| **E10** | C | C |
| **E11** | E | E |
| **E12** | F | E |
| **E13** | D | D |
| **E14** | C | C |
| **E15** | D | D |
| **E16** | B | B |
| **E17** | A | A |

Adotando a classificação da última coluna do quadro anterior como a adequada, nenhuma técnica obteve a mesma classificação: o *SVM* diferiu deste resultado apenas na escola E16, em faixa vizinha; as RNA possuem cinco escolas com resultados diferentes, também em faixas vizinhas: E1, E6, E7, E8 e E12; e o AG possui duas escolas (E7 e E10) com classificação não coincidente, sendo que para a escola E7 as classificações não são em faixas vizinhas (B – na técnica AG; D – na classificação adotada como correta).

#### *6.2.2.2 Comparação das classificações obtidas nas técnicas que consideram distâncias*

A segunda análise para este estudo de caso refere-se à comparação das técnicas que utilizam distâncias entre os dados das escolas aos conjuntos de dados de cada faixa da etiqueta de qualidade. O quadro 6.13, a seguir, apresenta o resultado obtido nessas técnicas.

**QUADRO 6.13 –** COMPARAÇÃO DAS CLASSIFICAÇÕES OBTIDAS PELAS TÉCNICAS QUE LEVAM EM CONSIDERAÇÃO DISTÂNCIA EUCLIDIANA E/OU DE MALAHANOBIS

| Escola | Distância Euclidiana | | | | Distância de Mahalanobis |
|---|---|---|---|---|---|
| | Distância aos elementos do conjunto | Distância ao ponto central do conjunto | 3-vizinhos mais próximos | Distância à média do conjunto | |
| **E1** | C | C | B | C | B |
| **E2** | E | E | E | E | D |
| **E3** | D | D | D | D | C |
| **E4** | C | C | C | C | C |
| **E5** | C | C | C | C | C |
| **E6** | D | D | D | D | C |
| **E7** | D | D | D | D | C |
| **E8** | E | F | F | F | F |
| **E9** | C | C | C | C | C |
| **E10** | C | C | C | C | C |
| **E11** | E | E | E | E | D |
| **E12** | E | E | E | E | E |
| **E13** | D | D | D | D | D |
| **E14** | C | C | C | C | B |
| **E15** | E | E | D | E | D |
| **E16** | B | B | B | B | B |
| **E17** | A | A | A | A | A |

No quadro 6.13 anterior, nove escolas (E3, E4, E5, E9, E10, E12, E13, E16 e E17) possuem a mesma classificação em todas as técnicas. As demais escolas possuem classificações em faixas vizinhas.

Realizando a comparação entre a classificação por voto que considera a distância euclidiana e a classificação com distância de Mahalanobis (quadro 6.14), têm-se nove escolas (E4, E5, E8, E9, E10, E12, E13, E16 e E17) com mesma classificação, e as demais possuem classificações em faixas vizinhas.

**QUADRO 6.14 –** COMPARAÇÃO DAS CLASSIFICAÇÕES OBTIDAS PELAS TÉCNICAS QUE UTILIZAM DISTÂNCIAS

| Escola | Classificação por voto – distância euclidiana | Classificação – distância Mahalanobis |
|---|---|---|
| **E1** | C | B |
| **E2** | E | D |
| **E3** | D | C |
| **E4** | C | C |
| **E5** | C | C |
| **E6** | D | C |
| **E7** | D | C |
| **E8** | F | F |
| **E9** | C | C |
| **E10** | C | C |
| **E11** | E | D |
| **E12** | E | E |
| **E13** | D | D |
| **E14** | C | B |
| **E15** | E | D |
| **E16** | B | B |
| **E17** | A | A |

*6.2.2.3 Comparação das classificações obtidas em todas as técnicas*

Na terceira análise é realizada comparação das classificações obtidas em todas as técnicas. Assim, o quadro 6.15 apresenta a classificação por voto destes resultados.

Este quadro mostra que a escola E1 obteve classificação “B” em quatro técnicas e nas outras quatro classificações “C”. Assim, decidiu-se por classificá-la como “C”, por representar a pior classificação.

**QUADRO 6.15 –** COMPARAÇÃO DAS CLASSIFICAÇÕES OBTIDAS PELAS TÉCNICAS

| Escola | RNA | SVM | AG | Distância Euclidiana | | | | Distância de Mahalanobis | Classificação por voto |
|---|---|---|---|---|---|---|---|---|---|
| | | | | a todos os elementos | ao ponto central do conjunto | 3-vizinhos mais próximos | a média do conjunto | | |
| **E1** | C | B | B | C | C | B | C | B | C |
| **E2** | E | E | E | E | E | E | E | D | E |
| **E3** | D | D | D | D | D | D | D | C | D |
| **E4** | C | C | C | C | C | C | C | C | C |
| **E5** | C | C | C | C | C | C | C | C | C |
| **E6** | C | D | D | D | D | D | D | C | D |
| **E7** | C | D | B | D | D | D | D | C | D |
| **E8** | E | F | F | E | F | F | F | F | F |
| **E9** | C | C | C | C | C | C | C | C | C |
| **E10** | C | C | D | C | C | C | C | C | C |
| **E11** | E | E | E | E | E | E | E | D | E |
| **E12** | F | E | E | E | E | E | E | E | E |
| **E13** | D | D | D | D | D | D | D | D | D |
| **E14** | C | C | C | C | C | C | C | B | C |
| **E15** | D | D | D | E | E | D | E | D | D |
| **E16** | B | A | B | B | B | B | B | B | B |
| **E17** | A | A | A | A | A | A | A | A | A |

Admitindo que o resultado apresentado na coluna “classificação por voto” seja o adequado para este estudo de caso, os resultados obtidos em cada técnica diferem desta última coluna com as seguintes quantidades: RNA com quatro escolas (E6, E7, E8 e E12); *SVM* com duas escolas (E1 e E16); AG com três escolas (E1, E7 e E10); somatório das distâncias do novo elemento aos elementos de cada faixa da etiqueta de qualidade, com duas escolas (E8 e E17); distância do novo elemento ao ponto central de cada faixa da etiqueta de qualidade, com uma escola (E15); $k$-vizinhos mais próximos, com uma escola (E1); distância do novo elemento a média dos elementos de cada faixa da etiqueta de qualidade, com uma escola (E15); e distância de Mahalanobis, com sete escolas (E1, E2, E3, E6, E7, E11 e E14).

Assim, nenhuma técnica possui o mesmo resultado que a classificação por votos. As técnicas que mais se aproximaram a este resultado foram “distância do novo elemento ao ponto central de cada faixa da etiqueta de qualidade”, “k-vizinhos mais próximos” e “distância do novo elemento a média dos elementos de cada faixa da etiqueta de qualidade”.

A figura 6.3, a seguir, apresenta a classificação das escolas, conforme última coluna do quadro 6.15, através da etiqueta de classificação da qualidade educacional, de forma comparativa, em relação ao desempenho na Prova Brasil.

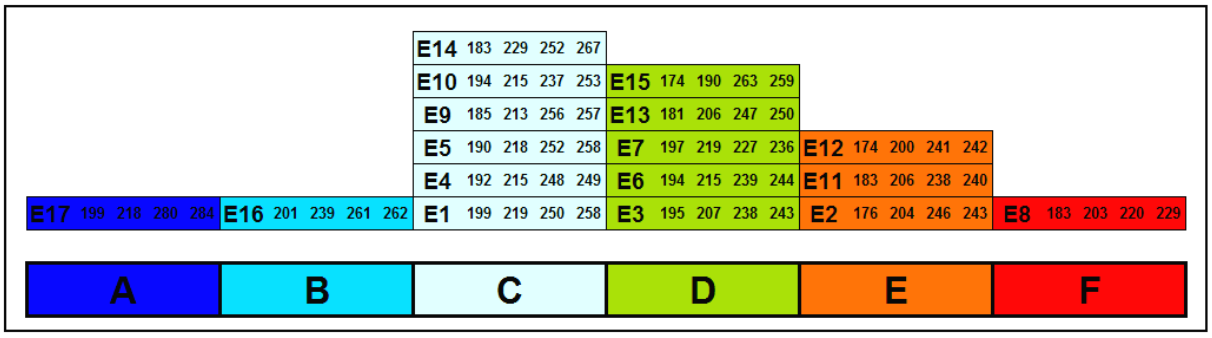


**FIGURA 6.3 –** CLASSIFICAÇÃO DA QUALIDADE EDUCACIONAL, DE FORMA COMPARATIVA
**FONTE:** O AUTOR (2012)

Os valores apresentados para cada escola expressam as quatro classes, sendo, nesta ordem, $C_1$ – notas de língua portuguesa nos anos inicias; $C_2$ – notas de matemática nos anos inicias; $C_3$ – notas de língua portuguesa nos anos finais; e $C_4$ – notas de matemática nos anos finais.

Retomando os questionamentos apresentados no início deste estudo de caso, onde afirmávamos não ser possível indicar a melhor escola apenas analisando o desempenho das escolas na Prova Brasil, visto que E17 possui as melhores notas nos anos finais e E16 nos anos iniciais, a etiqueta de qualidade indica que a melhor é E17, com classificação “A”, seguida pela escola E16, com classificação “B”. O mesmo ocorria em relação às escolas com pior desempenho na Prova Brasil, em que as escolas E15 e E12 possuem as piores notas nos anos iniciais e E8 nos anos finais, com isso, a etiqueta de qualidade indica que E15 possui classificação “D”, E12 classificação “E” e E8 classificação “F”.

Logo, a aplicação desta metodologia para a criação de etiqueta de qualidade educacional, de forma comparativa, em relação ao desempenho na Prova Brasil, apresentou conhecimento não explícito, quando analisada as notas nessa avaliação, mostrando a importância do processo *KDD* em bases de dados educacionais.

# 7 CONSIDERAÇÕES FINAIS E SUGESTÕES PARA TRABALHOS FUTUROS

Como comentado na introdução desta tese, esta pesquisa surgiu da participação em projeto de pesquisa registrado junto à ANEEL, firmado entre a UFPR/PPGMNE e uma concessionária de energia elétrica brasileira.

A inspiração para esta tese ocorreu ao analisar dois trabalhos, mais precisamente o de Casteren *et al.* (2005). Estes autores propõem a criação de etiqueta para classificar a Qualidade de Energia Elétrica, considerando afundamentos de tensão. Mas, alguns procedimentos apresentados não são explícitos e parece não haver na literatura respostas para as seguintes perguntas:

- Como utilizar dados reais para a criação da etiqueta visto que Casteren *et al.* (2005) utilizam apenas dados fictícios?
- Como definir uma referência "normal" para a etiqueta de qualidade?
- Como classificar um elemento na etiqueta de qualidade apresentada por Casteren *et al.* (2005), que não se enquadra diretamente em nenhuma faixa de classificação?

Com a metodologia apresentada nesta tese, aplicada a duas áreas distintas, esta ausência de respostas é solucionada, destacando assim a contribuição inédita deste trabalho.

Para responder a primeira pergunta utiliza-se o processo *KDD* que busca descobrir conhecimento em bases de dados históricas, ou seja, em dados reais.

A segunda pergunta é respondida ao determinar o limite superior da faixa "C" da etiqueta de qualidade, através da média dos quadros individuais de cada elemento em que se quer saber a qualidade, daí vem o nome "de forma comparativa", pois é analisado um grupo/região, ou seja, obtêm-se resultados relativos e não absolutos. É também devido a estes quadros individuais, onde constam as classes que definimos de $C_i$, que a metodologia proposta é considerada versátil.

Para responder a última pergunta, a que se refere à classificação de um elemento que não foi possível classificar diretamente nas faixas de classificação da etiqueta, utiliza-se diversas técnicas da Pesquisa Operacional, sendo duas metaheurísticas (Redes Neurais e Algoritmos Genéticos) e seis métodos heurísticos

(*Support Vector Machine*, somatório das distâncias euclidiana do novo elemento aos elementos de cada faixa da etiqueta de qualidade, distância euclidiana do novo elemento ao ponto central de cada faixa da etiqueta de qualidade, $k$-vizinhos mais próximos, distância euclidiana do novo elemento a média dos elementos de cada faixa da etiqueta de qualidade e a distância de Mahalanobis), todas relacionadas à classificação de padrões.

Quanto às aplicações da metodologia nos dois estudos de casos, uma na área elétrica e outra na área educacional, é evidente que a utilização do processo *KDD* direcionou para a descoberta de conhecimentos não explícitos e apresentou tais resultados de forma simples.

A simplicidade na apresentação dos resultados é visualmente verificada pela formatação da etiqueta, uma vez que a classificação de cada elemento é indicada em uma escala de seis níveis (“A” a “F”), com diferenciação de cores.

Quanto aos resultados obtidos com as aplicações das técnicas, podem-se constatar através das seções 5.2 e 6.2 que a técnica *SVM* é a que apresentou resultados mais adequados entre as técnicas mais sofisticadas, seguida pela técnica RNA.

Em relação às técnicas mais simples destacam-se em ambos estudos de casos, as técnicas “distância euclidiana do novo elemento ao ponto central de cada faixa da etiqueta de qualidade” e “distância euclidiana do novo elemento a média dos elementos de cada faixa da etiqueta de qualidade”, que no primeiro estudo de caso (área elétrica) apresentaram o mesmo resultado que a “classificação por voto” e no segundo estudo de caso se diferenciaram em apenas uma escola.

Assim, por ser também um dos objetivos deste trabalho indicar uma técnica para aplicações futuras da metodologia aqui proposta para criação da etiqueta de qualidade, de forma comparativa, no contexto *KDD*, é indicada a técnica “*distância euclidiana do novo elemento a média dos elementos de cada faixa da etiqueta de qualidade*”. Nesta indicação considera-se a técnica que obteve resultado mais próximo a “classificação por voto”, entre todas as técnicas em ambos estudos de casos, e também o fato de ser uma técnica simples para implementação.

Com relação à utilização do processo *KDD* na metodologia proposta, este se mostrou importante, principalmente na aplicação à área elétrica devido às transformações dos registros e as associações entre as bases de dados (BD01 e BD02), já que para a aplicação das técnicas não foram utilizados os atributos brutos provenientes destas bases, mas sim, a quantificação destes.

Com os resultados obtidos na classificação dos alimentadores em relação aos afundamentos de tensão, a concessionária poderá analisar o comportamento destes, verificando os classificados como sendo de baixa qualidade e aplicar medidas de mitigação para que os mesmos tenham sua classificação “alavancada”.

Se a metodologia proposta for aplicada a dados anuais de subestações de distribuição de energia elétrica, a etiqueta de qualidade pode ser uma alternativa a exigência da ANEEL, uma vez que esta agência não define padrões de desempenho em relação aos afundamentos de tensão, mas afirma que “as distribuidoras devem acompanhar e disponibilizar, em bases anuais, o desempenho das barras de distribuição monitoradas”, informações que podem servir como referência para as unidades consumidoras atendidas pelo Sistema de Distribuição de Alta Tensão ou pelo Sistema de Distribuição de Média Tensão com cargas sensíveis a variações de tensão de curta duração.

Com relação ao desempenho das escolas na Prova Brasil, a criação da etiqueta de qualidade revela conhecimentos não verificados quando se analisa apenas o índice IDEB (quadro 7.1). Um exemplo é o fato da escola E12 possuir maior IDEB, em ambos níveis de ensino (Fundamental anos iniciais e Fundamental anos finais), que a escola E2, mas considerando apenas as notas da Prova Brasil e aplicando a metodologia aqui proposta, a classificação é exatamente a oposta: a escola E2 possui classificação “E” e a escola E12 fica classificada com qualidade “F”. O mesmo ocorre com as escolas E17 e E14.

**QUADRO 7.1** – IDEB DAS ESCOLAS ANALISADAS

| ESCOLA | IDEB | | MÉDIA IDEB | CLASSIFICAÇÃO ETIQUETA QUALIDADE |
|---|---|---|---|---|
| | ANOS INICIAIS | ANOS FINAIS | | |
| **E1** | 4,90 | 4,00 | 4,45 | C |
| **E2** | 4,20 | 3,10 | 3,65 | E |
| **E3** | 4,60 | 3,20 | 3,90 | D |
| **E4** | 4,50 | 3,70 | 4,10 | D |
| **E5** | 4,90 | 4,20 | 4,55 | C |
| **E6** | 5,00 | 3,90 | 4,45 | D |
| **E7** | 5,30 | 3,40 | 4,35 | D |
| **E8** | 4,40 | 3,80 | 4,10 | F |
| **E9** | 5,00 | 4,30 | 4,65 | D |
| **E10** | 4,70 | 4,00 | 4,35 | D |
| **E11** | 4,50 | 3,60 | 4,05 | E |
| **E12** | 4,60 | 3,20 | 3,90 | F |
| **E13** | 4,20 | 4,20 | 4,20 | E |
| **E14** | 5,40 | 5,00 | 5,20 | C |
| **E15** | 4,10 | 4,30 | 4,20 | E |
| **E16** | 5,50 | 4,40 | 4,95 | B |
| **E17** | 5,30 | 5,00 | 5,15 | A |

Isso significa que o IDEB não aponta a escola onde os alunos possuem maior desempenho escolar, o que é considerado por boa parte da população o critério para definir uma “boa escola”, principalmente de Ensino Médio, em que pais esperam a aprovação dos filhos em vestibulares, ou seja, em provas que medem desempenho escolar.

Com a etiqueta de qualidade de desempenho na Prova Brasil, é possível verificar o aprendizado do aluno no Ensino Fundamental (1º ao 9º ano) em comparação com as demais e ainda classificar a escola, não considerando evasão e reprovação escolar.

No entanto, para os pesquisadores dessa área está claro que a “qualidade educacional” depende de outros fatores, como, por exemplo, os itens apresentados pelo Custo-Aluno Educação: tamanho da escola, relação “alunos x turma” e “alunos x professor”, formação inicial e continuada dos docentes, gestão escolar, valorização do profissional da educação, entre outros tantos. No entanto, este trabalho leva em consideração que o desempenho em avaliação, como a Prova Brasil, reflete esses interesses, pois a avaliação é fundamental para o processo de ensino e aprendizagem, e é por meio dela que a comunidade escolar pode buscar formas de melhorar a qualidade da educação.

Utilizando a metodologia aqui proposta, as mantenedoras podem criar a etiqueta de qualidade a partir de notas obtidas em outras avaliações oficiais do governo federal/estadual/municipal ou criar uma avaliação específica com base nos dados que julgarem importantes.

## 7.1 SUGESTÕES DE TRABALHOS FUTUROS

Os problemas aqui abordados estão relacionados às áreas bastante distintas: elétrica e educacional; sugere-se a aplicação desta metodologia em outras áreas como, por exemplo, saúde, ciências da terra e engenharias.

Também indica-se que sejam utilizadas outras técnicas a fim de compará-las com as utilizadas nesta tese, uma vez que os dados necessários para a criação da etiqueta de qualidade nos dois estudos de caso estão descritos no decorrer dos capítulos e/ou nos anexos.

A forma como são determinadas as faixas de classificação, que no primeiro estudo de caso são os fatores multiplicativos 0,25, 0,50, 1,50 e 2, e no segundo são faixas com mesmo comprimento, pode ser objeto de estudo de trabalhos futuros na busca da definição destes para as diversas áreas.

Ainda, tendo a etiqueta de qualidade gerada pode-se dar continuidade a este trabalho, buscando identificar o “porquê” de tal classificação dos elementos. No caso da área elétrica, por que o alimentador AF tem qualidade “F” (seção 5.2.2)? Para responder a esta pergunta, uma alternativa é aplicar novamente o processo *KDD*, mas agora com a finalidade de extrair regras dos registros desse alimentador, em que devem ser considerados outros atributos, como equipamentos, clima, manutenção e outros que devem ser discutidos com um especialista da área.

No caso da área educacional, o desempenho em avaliações reflete a falta ou presença dos itens e relações apontadas no CAQi. Sugere-se a comparação dos resultados obtidos com a etiqueta de qualidade aqui proposta com o CAQi, explicitando relações existentes entre eles.

Por fim, pode-se verificar, através de coleta de dados adicionais de cada escola, como, por exemplo, formação do professor, contexto social dos alunos, projeto pedagógico, infraestrutura das escolas e outros, como estes dados estão relacionados com as faixas de classificação da etiqueta de qualidade, buscando a

extração de regras utilizando o processo *KDD* para verificar, por exemplo, o “porquê” da escola E8 ter qualidade “F”.

Com o descrito nessa seção, vê-se a possibilidade de continuidade desta pesquisa, pois há muito ainda a ser explorado, seja nos estudos de casos realizados ou em outras áreas, além da aplicação de outras técnicas de classificação de padrões.

# REFERÊNCIAS


ADEPOJU, G. A.; OGUNJUYIGBE, S. O. A.; ALAWODE, K. O. **Application of Neural Network to Load Forecasting in Nigerian Electrical Power System.** The Pacific Journal of Science and Technology. Spring. V. 8. 2007, p. 68-72.

ALVES, R. P.; FALSARELLA, O. M. **Modelo conceitual de inteligência organizacional aplicada à função manutenção.** Revista Gestão & Produção, vol. 16, no.2, São Carlos, Apr./June, 2009.

ANEEL Agência Nacional de Energia Elétrica. **Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional – PRODIST**: *Módulo 8 – Qualidade da Energia Elétrica.* 2008.

ARORA, R.; BHATIA, R. **Optimization of Automation in Fuzzy Decision Rules.** *In: 2012 Second International Conference on Advanced Computing & Communication Technologies.* Haryana – Índia, janeiro, 2012.

ATAMANI, B.; BELDJILALI, B. **Knowledge Discovery in Database: Induction Graph and Cellular Automaton.** Computing and Informatics, Vol. 26, p. 171–197, 2007.

BANG, J.; DHOLAKIA, N.; HAMEL, L.; SHIN, S. **Customer Relationship Management and Knowledge Discovery in Databases.** *In: Encyclopedia of Information Science and Technology.* 2d ed., p. 902-907, 2009.

BITTAR, O. J. N. V. **Indicadores de qualidade e quantidade em saúde.** Revista Administração em Saúde, v.. 3, n. 12 – Jul-Set, 2001.

BONCHI, F.; FERRARI, E.; JIANG, W.; MALIN, B. **Privacy, Security, and Trust in KDD**. Springer-Verlag Berlin Heidelberg, 2009.

BRACHMAN, R. J.; ANAND, T.. **The Process of Knowledge Discovery in Databases: A First Sketch.** KDD Workshop 1994: 1-12

BRADWAJ, B. K.; PAL, S. **Mining Educational Data to Analyze Students´Performance.** International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

BRASIL. Constituição (1988). **Constituição da República Federativa do Brasil**. Brasília-DF, Senado, 1998.

BRASIL. Ministério da Educação. Conselho Nacional de Educação (CNE). **Parecer CNE/CEB n. 8/2010.** *Estabelece normas para a aplicação do inciso IX do artigo 4º da Lei n. 9.394/96 (LDB), que trata dos padrões mínimos de qualidade de ensino para a Educação Básica pública*. Brasília, DF: MEC/CNE, 5 maio 2010.

BURGES, C. J. C. **A Tutorial on Support Vector Machines for Pattern Recognition.** Data Mining and Knowledge Discovery, v. 2, p. 121-168, 1998.

CACIOTTA, M.; GIARNETTI, S.; LECCESE, F.. **Hybrid Neural Network System for Electric Load Forecasting of Telecomunication Station.** XIX IMEKO World Congress - Fundamental and Applied Metrology. Lisboa, Portugal. 2009, p. 657-661

CARREIRA, D.; PINTO, J. M. R. **Custo aluno-qualidade inicial, rumo à educação pública de qualidade no Brasil.** *In: Campanha Nacional pelo Direito à Educação.* Ed. Global, São Paulo/SP, 2007.

CARVALHO, A. A. A. **Indicadores de Qualidade de Sites Educativos.** Cadernos do Sistema de Avaliação, Certificação e Apoio à Utilização de Software para a Educação e a Formação, n. 2, Ministério da Educação, p. 55-78, 2006.

CASTEREN, J. F. L. van.; ENSLIN, L. H. R.; HULSHORST, W. T. J.; KILNG, W.L.; HAMOEN, M. D.; COBBEN, J. F. G. **A customer oriented approach to the classification of voltage dips.** In: The18th International Conference and exhibition on Electricity Distribuion - CIRED 2005.

CHAER, G. M.; TÓTOLA, M. R. **Impacto do manejo de resíduos orgânicos durante a reforma de plantios de eucalipto sobre indicadores de qualidade do solo.** Revista Brasileira de Ciências do Solo, v. 31, p. 1381-1396, 2007.

COBBEN, J. F. G.; CASTEREN, J. F. L. **Classification Methodologies for Power Quality.** Electrical Power Quality & Utilization Magazine. V. 2. 2006.

CÔRTES, S. da C.; PROCARO, R. M.; LIFSCHITZ, S.. **Mineração de Dados – Funcionalidades, Técnicas e Abordagens.** Série Monografias em Ciências da Computação. PUC, Rio de Janeiro, 2002.

DEPONTI, C. M.; ECKERT, C.; AZAMBUJA, J. L. B. **Estratégia para construção de indicadores para avaliação da sustentabilidade e monitoramento de sistemas.** Revista Agroecologia e Desenvolvimento Rural Sustentável. Porto Alegre, v.3, n.4, out/dez 2002.

DING, C. H.; DUBCHAK, I. **Multi-class protein fold recognition using support vector machines and neural networks.** Bioinformatics, v. 17, n. 4, p. 349 - 358, 2001.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases.** The American Association for Artigicial Intelligence Magazine, pp. 37-54. 1996

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in Knowledge Discovery & Data Mining**. 1 ed. American Association for Artificial Intelligence, Menlo Park, Califórnia, 1996.

FELDENS, M. A.; CASTILHO, J. M. V. **Data Mining with the Combinatorial Rule Model: an application in a health-care relational database.** *In: XXIII Conferência Latino Americana de Informática - CLEI.* Valparaiso, Chile: Universidad Tecnica Federico Santa Maria, 1997.

FERREIRA, A. B. H. **Novo Dicionário da Língua Portuguesa.** 2. ed. Rio de Janeiro: Nova Fronteira, 2001.

FIALHO, J. S.; GOMES, V. F. F.; OLIVEIRA, T. S.; SILVA JUNIOR, J. M. T. **Indicadores da qualidade do solo em áreas sob vegetação natural e cultivo de bananeiras na Chapada do Apodi-CE**. Revista Ciência Agronômica, v.37, n.3, p.250-257, 2006.

FIDALSKI, J.; TORMENA, C. A.; SCAPIM, C. A. **Espacialização vertical e horizontal dos indicadores de qualidade para um latossolo vermelho cultivado com citros.** Revista Brasileira de Ciências do Solo, v. 31, p. 9.-19, 2007.

FIGUEIREDO NETO, L. F.; SAUER, L.; BORGES, G. R. C.; BELIZARIO, J. B. **Método servqual: um estudo de satisfação em uma escola de idiomas.** *In: XIII Simpósio de Engenharia de Produção*, Bauru-SP, 2006.

FOGEL, D.B. **An introduction to simulated evolutionary computation.** IEEE Transactions on Neural Networks. v. 5, p 3-14. 1994

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. **Knowledge Discovery in Databases - An Overview.** *In: Knowledge Discovery in Databases 1991*, pp. 1--30.

GARVIN, D. A. **Gerenciando a qualidade.** Rio de Janeiro: Qualitymark, 1992.

GOLDBERG, D. E. **Genetic algorithms in search, optmization, and machine learning.** Addison-Wesley Publishing Company, Inc. Massachusetts, 1989.

GRAELLS, P. M. **Criterios para la clasificación y evaluación de espacios web de interés educativo.** Revista Educar, Barcelona – Espanha. v. 25, p.95-11, 1999.

GREFENSTETTE, J. J. **Optmizacion of Control Parameters for Genetic Algorithms.** IEEE Transactions on systems, man and cybernetics, v. 16, p. 122-128, 1986.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques.** 2a ed. Morgan Kauffmann Publishers, 2006

HARTZ, Z. M. A.; CHAMPAGNE, F.; LEAL, M. C.; CONTRANDRIOPOULOS, A. **Mortalidade infantil "evitável" em duas cidades do Nordeste do Brasil: indicador de qualidade do sistema local de saúde.** Revista Saúde Pública, v. 30, n. 4, p. 310-8, 1996.

HAYKIN, S. Neural Networks – **A Comprehensive Foundation.** 2.nd., Prentice Hall, New Jersey, 1999.

HEBB, D. O. **The organization of behavior.** New York: Wiley, 1949.

HOFFMEISTER, S., BÄCK, T. **Genetic algorithms and evolution strategie: similarities and differences.** In: SCHWESEL; H. P.; MANNER, R. (EDS.), 1(ST) WORSHOP PARALLEL PROBLEM SOLVING FROM NATURE - PPSN (1990: Dortmund, Germany). p. 455-469, 1990.

HOLLAND, J. H. **Adaptacion in natural and artificial systems.** 2ed. Cambridge, USA: Mit Press, p. 211, 1992

KURCGANT, P.; TRONCHIN, D. M. R.; MELLEIRO, M. M. **A construção de indicadores de qualidade para a avaliação de recursos humanos nos serviços de enfermagem: pressupostos teóricos.** Revista Acta Paulista de Enfermagem, v. 19, n. 1, p.88-91, 2006.

LI, S.; KUO, S. **Knowledge discovery in financial investment for forecasting and trading strategy through wavelet-based SOM networks.** Expert Systems with Applications, vol, 34, p. 935–951, February, 2008.

LIU, L.; QI, H.; LI, D. **A research for Data Mining Technology baseado n Fuzzy Neural Network.** Advanced Materials Research, vols 433-440, p. 2509-2512. 2012.

LIU, S.; TIAN, X.; ZHANG, Z. **Process planning knowledge discovery in the process database.** *In: International Conference on Computer Application and System Modeling*, v. 11, p. 370-373, Taiyuan, 2010

MAHALANOBIS, P. C. **On the generalised distance in statistics.** *In: Proceedings National Institute of Science,* India, v., p. 49-55, 1936.

MAHAPATRA, S. S. **A neural network approach for assessing quality in technical education: an empirical study**. *International Journal of Productivity and Quality Management,* v. 2, n. 3, p. 287-306, 2007.

MAO, S.; WAN, W.; WANG, Y.; WANG, Z.; YU, H. **The application of an improved BP artificial neural network in distributed data mining.** *In: IET International Conference on Smart and Sustainable City*, Shanghai – China, julho, 2011.

McCULLOCH, W. S.; PITTS, w. **A logical calculus of the ideas immanent in nervous activity.** Bulletin Mathematical Biology, v. 5, n. 4. 1943. p.115-133.

MICHALEWICZ, Z.; SCHOEMAUER, M. **Evolutionay algorthms for constrained parameter optmization problems.** Evolutionary Computation. v. 4, p 1-32. 1996.

MITCHELL, T. **Machine Learning.** McGraw Hill, 1997.

NAUMANN, F., ROLKER, C. **Assessment Methods for Information Quality Criteria.** *International Conference on Information Quality*, Cambrige, USA, 2000.

OLESKOVICZ, M.; COURY, D. V. ; CARNEIRO, A. A. F. M.; ARRUDA, E. F.; DELMONT, O.; SOUZA, S. A. **Estudo comparativo de ferramentas modernas de análise aplicadas à qualidade da energia elétrica.** Revista Controle & Automação Vol. 17 no 3. Julho, agosto e setembro 2006.

OLIVEIRA, R. P.; ARAUJO, G. C. **Qualidade do ensino: uma nova dimensão da luta pelo direito à educação.** Revista Brasileira de Educação, n. 28, 2005.

PALADINI, E.P. **Gestão da Qualidade no Processo: A qualidade na produção de bens e serviços.** São Paulo – SP, Ed. Atlas, 1995.

RIBEIRO, V. M.; RIBEIRO, V. M.; GUSMÃO, J. B. **Indicadores de qualidade para a mobilização da escola.** Revista Cadernos de Pesquisa, v. 35, n. 124, p. 227-251, 2005.

SANTOS, R. **Princípios e aplicações de mineração de dados**. Technical report, INPE, 2006.

SILVA, K. M.; SOUZA, B. A.; BRITO, N. S. D.; DANTAS, K. M. C.; COSTA, F. B.; SILVA, S. S. B. **Detecção e classificação de faltas a partir da análise de registros oscilográficos via redes neurais artificiais e Transformada Wavelet.** Revista Controle & Automação Vol. 18 no 2. Maio e junho 2007.

STEINER, M. T. A. **Uma Metodologia para o Reconhecimento de Padrões Multivariados com Resposta Dicotômica.** 158 folhas. Tese de Doutorado em Engenharia de Produção – UFSC, Florianópolis, SC, 1995

STEINER, M. T. A.; SOMA, N. Y.; SHIMIZU, T.; NIEVOLA, J. C.; STEINER NETO, P. J. **Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados.** Revista Gestão & Produção, v.13, n.2, p.325-337, mai.-ago. 2006

SUNG, A. H.; MUKKAMALA, S. **Identifying important features for intrusion detection using support vector machines and neural networks.** *In: Symposium on Applications and the Internet*, 2003. p. 209 - 216, 2003.

THURAISINGHAM, B.; KHAN, L.; CLIFTON, C.; MAURER, J.; CERUTI, M. **Dependable Real-time Data Mining.** 8º IEEE International Symposium on Object-Oriented Real-Time Distributed Computing (ISORC'05), 2005.

TRINDADE, R. M. **Sistema Digital de Detecção e Classificação de Eventos de Qualidade de Energia.** Juiz de Fora, 2005. Dissertacão (Mestrado em Engenharia Elétrica – Universidade Federal de Juiz de Fora).

TRONCHONI, A. B.; PRETTO, C. O.; ROSA, A.; LEMOS, F. A. B. **Descoberta de conhecimento em base de dados de eventos de desligamentos de empresas de distribuição.** Revista da Sociedade Brasileira de Automação: Controle & Automação, vol.21, no.2, Campinas, Mar./Apr., 2010.

VAPNIK, V. **Statistical Learning Theory**. John Wiley and Sons, Inc., New York, 1998.

VAPNIK, V. **The nature of statistical learning theory.** Springer-Verlag, New Yourk, 1995.

VERGUEIRO, W.; CARVALHO, T. **Definição de indicadores de qualidade: a visão dos administradores e clientes de bibliotecas universitárias.** Revista Perspectiva em ciência da informação, Belo Horizonte, v. 6, n. 1, p. 27-40, jan./jun, 2001.

WEISS, S.; INDURKHYA, N. **Predictive Data Mining: a pratical guide**. Morgan Kauffmann Publishers, Inc. 1998.

WITTEN, I. H., FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques.** 2a ed. Morgan Kauffmann Publishers, 2005.

| ID. OSC. | NOME MED. | DATA INICIO | TIPO TRIGGER |
|---|---|---|---|
| 9 | CIC | 2008-02-06 07:28:35.034 | Swell - Fase A / Distorção Total - Fase A / Distorção Total - Fase B |
| 9 | CIC | 2008-02-06 07:28:35.034 | Swell - Fase A / Desequilíbrio |
| 9 | CIC | 2008-02-06 07:28:35.034 | Distorção Total - Fase A / Distorção Total - Fase B / Distorção Total - Fase C |
| 10 | CIC | 2008-02-06 20:04:14.805 | Sag – Fase C / Desequilíbrio |
| 10 | CIC | 2008-02-06 20:04:14.805 | Swell - Fase A / Swell - Fase B / Distorção Total - Fase C |

| DATA TRIGGER | CIRCUITO | STATUS TRIGGER |
|---|---|---|
| 2008-02-06 07:28:35.199 | 0 | Swell - Fase A / Distorção Total - Fase A / Distorção Total - Fase C / Desequilíbrio |
| 2008-02-06 07:28:35.216 | 0 | Distorção Total - Fase A / Distorção Total - Fase B / Distorção Total - Fase C |
| 2008-02-06 07:28:35.232 | 0 | Trigger Usuário |
| 2008-02-06 20:04:14.973 | 1 | Sag - Fase C / Desequilíbrio |
| 2008-02-06 20:04:14.99 | 1 | Swell - Fase A / Swell – Fase B / Distorção Total - Fase B / Sag - Fase C / Deseq. |

| RMS FASE A | RMS FASE B | RMS FASE C | THD A | THD B | THD C | FREQ. A | FREQ. B | FREQ. C | DES. CIRC. |
|---|---|---|---|---|---|---|---|---|---|
| 76,4 | 63,7 | 60,1 | 161,4 | 167,6 | 226,7 | 59,9 | 60,04 | 59,95 | 154,63 |
| 65,5 | 65,2 | 64,9 | 153,9 | 179 | 177,6 | 59,95 | 59,95 | 59,85 | 5,31 |
| 65,3 | 65,4 | 64,9 | 17,3 | 15,6 | 19,9 | 59,95 | 59,95 | 59,95 | 5,19 |
| 69,8 | 73,2 | 57,6 | 34,5 | 29,5 | 38,9 | 60 | 60 | 60 | 138,33 |
| 82,9 | 77,1 | 35,9 | 126,3 | 126,2 | 251,2 | 59,9 | 60,18 | 59,95 | 327,67 |

| COD_ALIM | DESC_ALIM | NUM_OPER_ALIM | COD_AEL | DESC_AEL | COD_CAUSA | DESC_CAUSA |
|---|---|---|---|---|---|---|
| 216550050 | Bela Vista | 21655050 | 5 | AT (34,5 kV) | 82 | Não Identificada |
| 759850010 | Golondrina | 75985010 | 5 | AT (34,5 kV) | 82 | Não Identificada |
| 826640005 | N. Prata Iguacu | 72380031 | 5 | AT (34,5 kV) | 82 | Não Identificada |
| 895880002 | Presidente Kennedy | 75977017 | 3 | AT (13,8 kV) | 43 | Melhoria e/ou Amplia. |

| COD_CHV | COD_CRD | DESC_CRD | C_CLIMA | D_CLIMA | COD_CEA | DESC_CEA |
|---|---|---|---|---|---|---|
| 89180BVT20 | 92 | Atuacao do RA | 1 | Normal | 12280 | Marechal Cândido Rondon |
| 89180VIS10 | 92 | Atuacao do RA | 1 | Normal | 12280 | Marechal Cândido Rondon |
| 83240DV031 | 92 | Atuacao do RA | 1 | Normal | 12264 | Dois Vizinhos |
| 8324002784 | 89 | Rede de Distribuição | 1 | Normal | 12264 | Dois Vizinhos |

| COD_CEC | DESC_CEC | QTD_CONS | DATA_INIC. | HORA_INIC | DUR_ INTRP | NUM_SEQ_INTERP |
|---|---|---|---|---|---|---|
| 844 | 34.5 Ra Alto Alegre | 571 | 20080201 | 307 | 0 | 12040957378 |
| 649 | 34.5 Ra Golondrina (E-500) | 221 | 20080201 | 320 | 0 | 12040957381 |
| 129 | 34.5 Nova Prata Do Iguacu | 1327 | 20080201 | 545 | 0 | 12040957382 |
| 524 | Vere | 33 | 20080201 | 835 | 92 | 12040957415 |

| NUM_ORG8 | DESC_ORG8 | COD_OINT | REGIONAL | CAR_SE | NOME_SE | COD_TIPO | DESC_TIPO |
|---|---|---|---|---|---|---|---|
| 18224 | ST Cont. Qualidade Toledo | P | SDO | 71470 | SE Guaira | 1 | Acidental |
| 18224 | ST Cont. Qualidade Toledo | P | SDO | 75985 | EC Vista Alegre | 1 | Acidental |
| 18223 | ST Cont. Qualidade Beltrão | P | SDO | 72380 | SE Dois Vizinhos | 1 | Acidental |
| 18223 | ST Cont. Qualidade Beltrão | P | SDO | 75977 | SE Vere | 2 | Programada |

# ANEXO 03 – QUANTIDADES DE REGISTROS DOS ALIMENTADORES

Nos quadros a seguir, são apresentadas as quantidades de registros ocorridos nos alimentadores AC, AD, AE, EF, AG, AH, AI. AJ, AK e AL de acordo com a duração e tensão remanescente para a posterior classificação da QEE de cada um deles.

**QUADRO A3.1 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR "AC"

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 35 | 0 |
| 20 a 39% | 1 | 0 |
| 10 a 19% | 0 | 0 |

**QUADRO A3.2 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR "AD"

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 6 | 0 |
| 20 a 39% | 2 | 5 |
| 10 a 19% | 0 | 0 |

**QUADRO A3.3 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR "AE"

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 0 | 0 |
| 20 a 39% | 2 | 0 |
| 10 a 19% | 0 | 0 |

**QUADRO A3.4 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR "AF"

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 26 | 0 |
| 20 a 39% | 2 | 4 |
| 10 a 19% | 0 | 0 |

**QUADRO A3.5 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR "AG"

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 0 | 0 |
| 20 a 39% | 0 | 0 |
| 10 a 19% | 0 | 0 |

**QUADRO A3.6 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR "AH"

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 16 | 0 |
| 20 a 39% | 1 | 0 |
| 10 a 19% | 0 | 0 |

**QUADRO A3.7 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR "AI"

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 40 | 0 |
| 20 a 39% | 4 | 0 |
| 10 a 19% | 0 | 0 |

**QUADRO A3.8 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR "AJ"

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 2 | 0 |
| 20 a 39% | 0 | 0 |
| 10 a 19% | 0 | 0 |

**QUADRO A3.9 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR "AK"

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 1 | 0 |
| 20 a 39% | 1 | 0 |
| 10 a 19% | 0 | 0 |

**QUADRO A3.10 –** CLASSIFICAÇÃO DOS AFUNDAMENTOS DE TENSÃO DO ALIMENTADOR “AL”

| **TENSÃO** | **DURAÇÃO** | |
|---|---|---|
| **REMAN.** | ≤ 500 milissegundos | > 500 milissegundos |
| 80 a 90% | 0 | 0 |
| 60 a 79% | 0 | 0 |
| 40 a 59% | 1 | 0 |
| 20 a 39% | 0 | 1 |
| 10 a 19% | 0 | 0 |

# ANEXO 04 – NOTAS NA PROVA BRASIL DAS ESCOLAS ANALISADAS

Nos quadros a seguir, são apresentadas as notas na Prova Brasil das escolas E2 a E17, indicada por segmento do Ensino Fundamental (anos iniciais e anos finais) e disciplinas (língua portuguesa e matemática).

**QUADRO A4.1 –** NOTAS DA PROVA BRASIL DA ESCOLA E2

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 176,19 | 204,38 |
| Anos Finais | 246,03 | 243,44 |

**QUADRO A4.2 –** NOTAS DA PROVA BRASIL DA ESCOLA E3

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 195,01 | 206,72 |
| Anos Finais | 238,16 | 243,29 |

**QUADRO A4.3 –** NOTAS DA PROVA BRASIL DA ESCOLA E4

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 192,45 | 215,27 |
| Anos Finais | 247,60 | 249,31 |

**QUADRO A4.4 –** NOTAS DA PROVA BRASIL DA ESCOLA E5

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 190,40 | 218,36 |
| Anos Finais | 251,58 | 258,46 |

**QUADRO A4.5 –** NOTAS DA PROVA BRASIL DA ESCOLA E6

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 194,40 | 214,96 |
| Anos Finais | 239,08 | 244,28 |

**QUADRO A4.6 –** NOTAS DA PROVA BRASIL DA ESCOLA E7

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 197,18 | 218,81 |
| Anos Finais | 227,29 | 235,91 |

**QUADRO A4.7 –** NOTAS DA PROVA BRASIL DA ESCOLA E8

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 183,41 | 202,93 |
| Anos Finais | 219,90 | 229,18 |

**QUADRO A4.8 –** NOTAS DA PROVA BRASIL DA ESCOLA E9

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 185,14 | 212,60 |
| Anos Finais | 255,67 | 257,05 |

**QUADRO A4.9 –** NOTAS DA PROVA BRASIL DA ESCOLA E10

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 194,20 | 214,98 |
| Anos Finais | 237,33 | 252,94 |

**QUADRO A4.10 –** NOTAS DA PROVA BRASIL DA ESCOLA E11

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 183,44 | 206,16 |
| Anos Finais | 238,11 | 240,13 |

**QUADRO A4.11 –** NOTAS DA PROVA BRASIL DA ESCOLA E12

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 174,40 | 199,76 |
| Anos Finais | 240,94 | 242,30 |

**QUADRO A4.12 –** NOTAS DA PROVA BRASIL DA ESCOLA E13

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 180,53 | 205,80 |
| Anos Finais | 247,05 | 250,21 |

**QUADRO A4.13 –** NOTAS DA PROVA BRASIL DA ESCOLA E14

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 183,24 | 229,39 |
| Anos Finais | 252,05 | 267,19 |

**QUADRO A4.14 –** NOTAS DA PROVA BRASIL DA ESCOLA E15

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 174,47 | 189,87 |
| Anos Finais | 262,62 | 259,40 |

**QUADRO A4.15 –** NOTAS DA PROVA BRASIL DA ESCOLA E16

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 201,41 | 238,69 |
| Anos Finais | 260,56 | 262,36 |

**QUADRO A4.16 –** NOTAS DA PROVA BRASIL DA ESCOLA E17

| **Nível de Ensino** | **Área do conhecimento** | |
|---|---|---|
| | Língua Portuguesa | Matemática |
| Anos Iniciais | 198,51 | 217,58 |
| Anos Finais | 279,54 | 284,39 |