

**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA
E TECNOLOGIA FLUMINENSE**

**PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS APLICADOS À
ENGENHARIA E GESTÃO**

Carlos Adriano Arruda Teixeira

**AVALIAÇÃO DO EFEITO DO BRANQUEAMENTO DO POLÍMERO
FLUORETO DE POLIVINILIDENO (PVDF) ATRAVÉS DA
DESCOBERTA DE CONHECIMENTO EM BANCOS DE IMAGENS**

Campos dos Goytacazes/RJ

2019

**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E
TECNOLOGIA FLUMINENSE**

**PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS APLICADOS À
ENGENHARIA E GESTÃO**

CARLOS ADRIANO ARRUDA TEIXEIRA

**AVALIAÇÃO DO EFEITO DO BRANQUEAMENTO DO POLÍMERO
FLUORETO DE POLIVINILIDENO (PVDF) ATRAVÉS DA DESCOBERTA
DE CONHECIMENTO EM BANCOS DE IMAGENS**

**D.Sc. Simone Vasconcelos Silva
(Orientador)**

**D.Sc. Antônio José da Silva Neto
(Co-orientador)**

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação do Instituto Federal de Educação, Ciência e Tecnologia Fluminense, no Curso de Mestrado Profissional em Sistemas Aplicados à Engenharia e Gestão (MPSAEG), como parte dos requisitos necessários à obtenção do título de Mestre em Sistemas Aplicados à Engenharia e Gestão.

Campos dos Goytacazes/RJ

2019

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E
TECNOLOGIA FLUMINENSE

PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS APLICADOS À
ENGENHARIA E GESTÃO

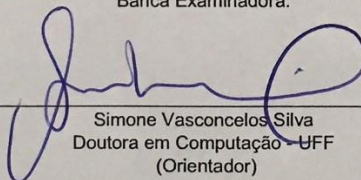
Carlos Adriano Arruda Teixeira

AVALIAÇÃO DO EFEITO DO BRANQUEAMENTO DO POLÍMERO
FLUORETO DE POLIVINILIDENO (PVDF) ATRAVÉS DA DESCOBERTA
DE CONHECIMENTO EM BANCOS DE IMAGENS

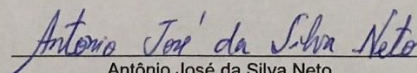
Dissertação de Mestrado apresentada ao Programa de Pós-Graduação do Instituto Federal de Educação, Ciência e Tecnologia Fluminense, no Curso de Mestrado Profissional em Sistemas Aplicados à Engenharia e Gestão (MPSAEG), como parte dos requisitos necessários à obtenção do título de Mestre em Sistemas Aplicados à Engenharia e Gestão.

Dissertação apresentada em 26 de Setembro de 2019.

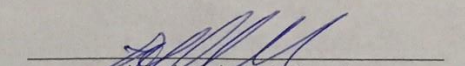
Banca Examinadora:



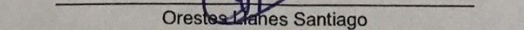
Simone Vasconcelos Silva
Doutora em Computação - UFF
(Orientador)



Antônio José da Silva Neto
Doutor em Engenharia Mecânica - IPRJ/UERJ
(Coorientador)



João José de Assis Rangel
Doutor em Engenharia e Ciência dos Materiais - UENF



Orestes Llanes Santiago
Doutor em Controle Descontínuo de Sistemas Não Lineares -
Universidade de Los Andes Venezuela

Dedico este trabalho, em primeiro lugar, à Deus, que me deu forças para superar todas as dificuldades, à minha esposa e filho por tudo que representam em minha vida e por serem os meus maiores motivadores em todos os momentos, em principal os momentos de maiores incertezas. Dedico este trabalho também ao meu pai, mãe e irmãos, que me ensinaram o real valor da família e a sempre seguir o caminho da dignidade e da honestidade, não importando os obstáculos encontradas pelo trajeto, mas sim a grandeza de se alcançar onde quer chegar de forma justa e honrosa.

AGRADECIMENTOS

Não seria possível concluir este trabalho sem o precioso apoio de algumas pessoas, em especial de meus familiares, alguns colegas de trabalho e de minha orientadora. Por isso é importante explicitar que este trabalho é fruto não apenas do empenho do autor, mas também da colaboração destas pessoas de diferentes maneiras, porém substancialmente importantes em igual teor.

Por este motivo, gostaria de expressar minha enorme gratidão à minha orientadora, professora doutora Simone Vasconcelos Silva, que tanto se empenhou para viabilizar o estudo da forma adequada, com muita paciência e dedicação. Muito obrigado pela forma como conduziu esta orientação, empregando seus comentários, observações e correções de forma extremamente profissional, sem em nenhum momento me desmotivar, mesmo diante dos momentos mais críticos ao longo do desenvolvimento deste trabalho;

Gostaria também de expressar a importância da minha esposa e filho, que são o alicerce da minha vida, que me guiam, me incentivam e me dão forças para nunca desistir, não importando quão grande for o obstáculo. Agradeço também a meus pais, mãe e a meus quatro irmãos que sempre contribuíram, mesmo que emocionalmente em algumas ocasiões, de maneira a sempre reforçar a importância de concluir este trabalho e como isso mudaria a minha vida positivamente;

Expresso também minha gratidão aos meus colegas de trabalho, que me auxiliaram no entendimento do problema, nas propostas de solução e análise de resultados deste trabalho. A estes agradeço por transformarem as discussões técnicas relacionadas a elaboração desta pesquisa sempre de forma amistosa e interessante, agregando muito valor para este trabalho.

Por fim agradeço a NOV FLEXIBLES EQUIPAMENTOS E SERVICOS LTDA, que viabilizou o estudo ao permitir que fossem explorados recursos internos da empresa. Não seria possível avançar com o trabalho sem a contribuição das imagens e dados dos produtos fabricados e tampouco sem o suporte técnico de alguns profissionais que atuam na área científica correlata a esta pesquisa.

RESUMO

Para tornar possível a exploração de óleo e gás natural em águas marítimas é necessário desenvolvimento de alta tecnologia para dar suporte aos projetos de perfuração e exploração de poços submarinos. Um dos equipamentos que viabiliza a exploração *offshore* é o duto flexível. As linhas de dutos flexíveis possuem um complexo e meticuloso processo produtivo, devido à tecnologia de ponta envolvida em seu método de construção. Considerando o alto custo e a complexidade do processo de fabricação de dutos flexíveis, é objetivo deste trabalho analisar o efeito do aparecimento de manchas identificadas como branqueamento cuja ocorrência é observada no material polimérico conhecido como *fluoreto de Polivinilideno* (PVDF) que é utilizado na manufatura de camadas de revestimento interno em linhas flexíveis. A metodologia proposta consiste na elaboração de um modelo de predição, com base em classificação, que avalie o branqueamento através da extração de conhecimento oriundos de um banco de imagens. Amostras e produções comerciais são analisadas neste trabalho, provenientes da colaboração de uma indústria do setor. Através da metodologia proposta são identificados os parâmetros mais relevantes associados ao problema foco de estudo passíveis de estabelecerem regras de classificação. Como resultado, algumas relações entre atributos, frutos da extração de dados em imagens, e características físicas do produto, foram identificadas através de árvores de decisão. Conclui-se que tais relações proveem suporte para a tomada de decisão quando feita por um especialista de domínio.

Palavras-chave: Duto Flexível, PVDF, Branqueamento, Processos de Manufatura, Mineração de dados

ABSTRACT

To make possible the exploration of oil and natural gas in maritime waters, it is necessary to develop high technology to support the projects of drilling and exploration of submarine wells. One of the equipment that enables offshore exploration is the flexible pipes. Flexible pipes have a complex and meticulous production process, due to the cutting-edge technology involved in their construction method. Considering the high cost and complexity of the flexible pipe manufacturing process, the objective of this work is to analyze the effect of the occurrence of whitening stress identified as bleaching observed in the polymeric material known as Polyfluoride of Vinylidene (PVDF) which is used in the manufacture of inner liner layers. in flexible pipes. The proposed methodology is to develop a prediction model based on classification, which evaluates the whitening bleaching by extracting knowledge from an image database. Samples and commercial products are analyzed in this dissertation, through the collaborative participation of an Industry from this sector. Through the proposed methodology, the most relevant parameters associated with the study problem that can define classification rules are identified. As a result, some relationships between some characters, fruits of image data extraction and product characteristics were identified through decision trees. The conclusion is that these relationships provides support for decision making when performed by a domain expert.

Keywords: Flexible Pipe, PVDF, Image Processing, Whitening, Manufacturing Processes, data mining

DISSERTAÇÃO

Esta dissertação é baseada em estudos conduzidos pela empresa NOV FLEXIBLES EQUIPAMENTOS E SERVICOS LTDA, integrante do grupo National Oilwell Varco – SPS (Subsea Production Systems) e, portanto, o material aqui publicado não deve ser utilizado para fins comerciais, mas sim apenas como pesquisa científica.

DOCUMENTOS NOV FLEXIBLES

Os experimentos de laboratório apresentados como parte do presente projeto foram documentados em relatórios técnicos internos emitidos como documentos NOVFlexibles e são parte do material científico que é de propriedade intelectual desta companhia. Logo, esses documentos não estão disponíveis para o público. No entanto, alguns dos resultados experimentais são apresentados nos capítulos 3 e 4.

LISTA DE ABREVIACOES

AM	Aprendizado de Mquina
ARFF	Attribute-Relation File Format. Formato do arquivo relacionado ao atributo
BD	Banco de dados
CBF	Consistency-Based Filter
CFS	Correlation-based Feature Selection
CSE	Classifier Subset Evaluator
<i>InfoGain</i>	Information-Gain Attribute Ranking
J48	rvore de deciso do WEKA adaptada do algoritmo C4.5
KDD	Knowledge Discovery Databases
MBR	Mnimo raio de curvatura
MD	Minerao de dados
PVDF	Fluoreto de polivinilideno
ROI	Region of interest
SA	Seleo de Atributos
TA	Transformao de Atributos
WEKA	Software de minerao de dados
PA-11	Poliamida 11
PA-12	Poliamida 12
HDPE	Polietileno de alta densidade
XLPE/PEX	Polietileno reticulado

LISTA DE FIGURAS

Figura 1 – Camadas que compõem um duto flexível	20
Figura 2 – Identificação da camada de estanqueidade em uma linha de duto flexível.....	23
Figura 3 – Barreira de Pressão de PVDF em um tubo Flexível Dissecado.	25
Figura 4 - Esboço de uma estrutura de duto flexível típico unbonded. A camada metálica mais interna, conhecida como carcaça é carregada pela pressão hidrostática externamente e pela pressão do fluido internamente, que é contido pelo revestimento interno.	25
Figura 5 – Matriz de combinação de cores.....	29
Figura 6 – Representação de uma imagem digital bidimensional	30
Figura 7 - Imagem em cinzentos e representação em linguagem computacional	30
Figura 8 - Atividades do pré-processamento.	33
Figura 9 - Processo de descoberta de conhecimento em bancos de dados.....	35
Figura 10 - Interface típica do WEKA	41
Figura 11 – Micro vazios à partir de microscopia da camada de barreira extrudada em PVDF com diferentes graus de deformação.	46
Figura 12 – Branqueamento do revestimento interno de uma linha flexível.....	47
Figura 13 – Ocorrência de branqueamento durante o enrolamento do duto no carretel.	48
Figura 14 – Aparecimento de branqueamento ao longo de todo o comprimento do duto na camada mais próxima ao núcleo do carretel.....	48
Figura 15 – Variação de passo em uma carcaça de duto flexível	49
Figura 16 – Branqueamento formado em regiões onde houve variação no passo da carcaça..	50
Figura 17 – Branqueamento localizado onde houve redução de espessura do material.....	51
Figura 18 - Característica opaca e translúcida do PVDF.....	51
Figura 19 – Ocorrências de branqueamento ao longo de diferentes comprimentos na circunferência do duto.	52
Figura 20 – Possíveis padrões geométricos para branqueamento no PVDF	53
Figura 21 – Amostras recolhidas de camada de barreira extrudada em PVDF.....	55
Figura 22 - Resultado de testes para amostras não-deformadas e pré-deformadas (com branqueamento) com espessura total da camada de barreira, que deformou até 3,5% e então mantida por 4 minutos.	56
Figura 23 - Resultado de testes para 2 amostras: não-deformadas e pré-deformadas (com branqueamento), com espessura total da camada de barreira, que deformou até 5,25% e então mantida por 4 minutos.	56

Figura 24 - Resultado de testes para amostras não-deformadas e pré-deformadas (com branqueamento) com espessura total da camada de barreira, que deformou até a ruptura.	57
Figura 25 - Características visuais do branqueamento da amostra #01: camada em PVDF manufaturado sobre carcaça (acima); perfil da amostra removida (abaixo).	58
Figura 26 - Características visuais do branqueamento da amostra #02: camada em PVDF manufaturado sobre carcaça (acima); perfil da amostra removida (abaixo).	59
Figura 27 – Exemplo de imagem sendo calibrada para um valor de diâmetro físico conhecido (à esquerda) incluindo apresentação de escala em milímetros (à direita).	61
Figura 28 – Imagem original à esquerda e imagem convertida para escala de cinzas 8-bit à direita.	62
Figura 29 – Exemplo do arquivo resultante da aplicação de filtros de imagens.	63
Figura 30 – Extração das informações das imagens através do software FIJI.	64
Figura 31 – Identificação de alguns atributos.	66
Figura 32 – Estrutura do conjunto de dados do arquivo .arff utilizado no trabalho.	67
Figura 33 – Configuração do modelo de experimento	68
Figura 34 – Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A1, para todos os atributos.	72
Figura 35 – Distribuição de resultados por grupos de atributos	75
Figura 36 – Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A3, grupo TT.	78
Figura 37 – Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A9, grupo TT.	78
Figura 38 - Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A1, grupo TT, com BD ajustado para atributo PC com valor acima de 16.38mm apenas.	79
Figura 39 - Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A1, grupo IM.	80
Figura 40 – <i>Dataset (test_data.arff)</i> criado para avaliação das amostras calculadas na avaliação físico-mecânica.	81
Figura 41 - Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A3, grupo GTT2.	82
Figura 42 – Resultado da predição do modelo refinado para as amostras 01 e 02 para o caso A3, grupo GTT2.	83
Figura 43 - Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A3, grupo GTT2 com aplicação de filtro <i>Resample</i> .(1 semente aleatória).	84

Figura 44 - Árvore de decisão gerada pelo algoritmo J48 configurado para o caso , grupo GTT2, com aplicação de filtro <i>Resample</i> .(8 sementes aleatórias).	84
Figura 45 – Número de aparições de cada atributo nas principais análises	88

LISTA DE TABELAS

Tabela 1 – Composição característica de um duto flexível e principal função de cada camada.	20
Tabela 2 - Quantidade de atributos selecionados por cada um dos algoritmos de seleção de atributos.	38
Tabela 3 - Média de erro para cada conjunto de dados e algoritmos considerados (método J48).	38
Tabela 4 – Comportamento mecânico do PVDF em diferentes temperaturas.....	54
Tabela 5 – Especificações do teste de tração	55
Tabela 6 - Valores de deformação máxima na região mais externa do duto sob flexão no carretel onde se observou aparecimento de branqueamento.	59
Tabela 7 - Características extraídas das imagens para o banco de dados	65
Tabela 8 – Avaliação dos algoritmos (70% treinamento, 30% teste).	70
Tabela 9 – Diferentes condições de configurações do algoritmo J48.....	71
Tabela 10 – Definição dos atributos de maior relevância por tipo de algoritmo avaliador de atributos	73
Tabela 11 – Análise dos grupos de atributos	74
Tabela 12 – Principais regras extraídas da subdivisão das árvores geradas para os casos A1, A3 e A9, grupo TT.....	77
Tabela 13 - Principais regras extraídas da subdivisão das árvores geradas para o caso A1, grupo TT, com atributo PC ajustado.	79
Tabela 14 – Análise dos atributos mais relevantes da Figura 41.....	82
Tabela 15 – Análise dos atributos mais relevantes da Figura 43.....	85
Tabela 16 – Análise dos atributos mais relevantes da Figura 44.....	85
Tabela 17 – Configuração A3 para os atributos diâmetro interno (DI), largura da mancha (W) e área da mancha (A) para o modelo refinado.....	85
Tabela 18 - Configuração A3 para os atributos relação pixels por mm ² (PI), mínimo de intensidade de cinza (MI), diâmetro interno (DI) e diâmetro externo (OD) para o modelo refinado.....	86
Tabela 19 - Configuração A3 para os atributos largura da mancha (W), máxima intensidade de cinza (MA) e Relação pixels por mm ² (PI) para o modelo refinado	86
Tabela 20 – Principais regras de classificação condicionadas à aprovação para dutos de até 6”.	89

Tabela 21 – Principais regras de classificação condicionadas à aprovação para dutos maiores ou iguais a 8”	89
Tabela 22 – Regras gerais de classificação condicionadas à aprovação.....	89
Tabela 23 - Modulo de Young em diferentes temperaturas	95
Tabela 24 - Coeficiente de expansão térmica	95
Tabela 25 - Outras propriedades relevantes.....	95
Tabela 26 – Temperatura de serviço	95

SUMÁRIO

CAPÍTULO 1 - INTRODUÇÃO	16
1.1. OBJETIVOS	17
1.1.1. Objetivo principal	17
1.1.2. Objetivos Específicos	17
1.2. JUSTIFICATIVAS	18
1.3. ESTRUTURA DO TRABALHO	18
CAPÍTULO 2 - REVISÃO BIBLIOGRÁFICA	20
2.1. DUTO FLEXÍVEL	20
2.1.1. Processo de manufatura	22
2.1.2. Revestimento interno (<i>inner Liner</i>)	23
2.1.3. PDVF	24
2.1.4. PVDF aplicado a dutos flexíveis	24
2.1.5. Branqueamento do PVDF (<i>Stress whitening</i>).....	26
2.2. MODELO DE IMAGEM DIGITAL	28
2.3. MINERAÇÃO DE DADOS	31
2.3.1. Conceitos preliminares	31
2.3.2. Banco de dados	32
2.3.3. Processo de descoberta de conhecimento em bancos de dados.....	34
2.3.4. Aprendizado de máquina	36
2.3.5. Estimativa, ruído e desvio	36
2.3.6. Seleção de atributos	37
2.3.7. Principais avaliadores de atributos	37
2.3.8. Tipos de aprendizado de máquina: preditivo e descritivo	38
2.3.9. Classificação	39
2.4. WEKA	40
2.4.1. J48.....	42
2.4.2. Avaliação dos modelos	43
CAPÍTULO 3 - METODOLOGIA	45
3.1. ETAPA II - ANÁLISE DO PROBLEMA	46
3.1.1. Avaliação mecânica do problema.....	53
3.2. ETAPA III - SELEÇÃO, PRÉ-PROCESSAMENTO E TRANSFORMAÇÃO DAS IMAGENS	60
3.3. ETAPA IV - APLICAÇÃO DAS TÉCNICAS DO KDD	64

3.3.1.	KDD fase I – Seleção, pré-processamento e transformação	64
3.3.2.	KDD Fase II – Mineração de dados	67
3.3.3.	KDD fase III - Avaliação do desempenho.....	69
CAPÍTULO 4 - RESULTADOS E DISCUSSÕES		70
4.1.	AVALIAÇÃO DOS ATRIBUTOS E SUAS INFLUÊNCIAS.....	73
4.2.	EXTRAÇÃO DAS REGRAS E INTERPRETAÇÃO.....	75
4.3.	VALIDAÇÃO DO MODELO	80
4.3.1.	Refinamento do modelo	81
4.3.2.	Análise dos atributos após implementação do filtro <i>Resample</i>	83
4.4.	DISCUSSÃO DOS RESULTADOS.....	86
CAPÍTULO 5 - CONCLUSÃO		89
REFERÊNCIAS.....		92
APÊNDICE A – CARACTERÍSTICAS DO PVDF		95
APÊNDICE B - ÁRVORES DE DECISÃO GERADAS NO FORMATO TEXTO		96
APÊNDICE C – ALGORITMOS GERADOS PELO WEKA.....		97
APÊNDICE D - ARQUIVO .ARFF UTILIZADO NO TRABALHO		100
APÊNDICE E – IMAGENS DAS AMOSTRAS #01 E #02.....		102

CAPÍTULO 1 - INTRODUÇÃO

Conforme dados da Agência Nacional de Petróleo, Gás Natural e Biocombustíveis (ANP, 2017), a produção de petróleo no Brasil atingiu 2,5 milhões de barris por dia, enquanto a produção de gás natural atingiu 103 milhões de metros cúbicos por dia (m³/d). Neste contexto, 95% do total do petróleo e 83% do de gás natural produzido no país foi oriundo dos campos marítimos, a partir de seus 746 poços. Aproximadamente, 47,5% da produção total já era oriundo dos campos do pré-sal.

Segundo Martins (2011), um sistema para a exploração de petróleo *offshore* consiste em um conjunto de equipamentos de uma especificidade complexa, os quais podem ser resumidos basicamente em quatro grupos: as linhas, os poços, os equipamentos submarinos e as unidades de produção.

As linhas flexíveis correspondem aos dutos de transferência do produto do poço até a unidade de produção, os poços provêm o acesso ao reservatório subterrâneo de hidrocarboneto, os equipamentos submarinos auxiliam, monitoram e controlam o processo de elevação dos fluidos e as unidades de produção (plataformas, FSO ou FPSO) compõem a poção mais visível do sistema *offshore*, situando-se na superfície, em locais de diferentes profundidades de lâmina d'água (BOSCHEE, 2012).

A avaliação do comportamento mecânico de polímeros para diferentes aplicações vem despertando grande interesse e neste contexto, a presente dissertação terá como foco a análise do efeito do branqueamento (*stress whitening*) em material polimérico Fluoreto de Polivinilideno (PVDF) presente na camada de revestimento interno (*inner liner*) de dutos flexíveis (ABREU, 2012).

Durante processos de manufatura, desvios de qualidade decorrentes de variações de parâmetro de processo podem implicar na perda total ou parcial de uma produção. Esses tipos de perdas são gravemente prejudiciais para um negócio uma vez que retrabalhos custam caro, impactam no prazo e podem depreciar o valor de um produto no mercado.

Piccirilli (2013), demonstra que é possível extrair informações úteis de bancos de dados e que uma análise consistente pode ser decisiva na detecção de irregularidades por possuir caráter multidisciplinar contendo técnicas de banco de dados, estatística e aprendizado de máquina, apresentando grandes utilidades na descoberta de regras ou padrões, na previsão de futuras tendências e comportamento de grupos similares.

A motivação desta pesquisa se dá através da possibilidade de estabelecer parâmetros quantitativos, por meio da extração de conhecimento de imagens para dar auxílio à tomada de

decisão durante processos de manufatura. A base deste trabalho é propor um método base para elaborar um modelo, a partir da técnica de classificação por mineração de dados, que seja capaz de identificar as principais características que podem levar uma avaliação a ser conduzida de forma assertiva durante o aparecimento de branqueamento no PVDF durante o processo de extrusão em uma fábrica.

O presente trabalho representa um fato inovador ao abordar o contexto de desvios de fabricação sobre a ótica de descoberta de conhecimento em banco de dados. Os desafios correntes da indústria em garantir a qualidade de seus produtos justificam a importância do estudo ao abordar o processo de seleção de atributos para propor melhoras no processo de identificação e avaliação de anomalias. Assim como em (PICCIRILLI, 2013), a qualidade do resultado do modelo está condicionada à disposição das informações e técnicas empregadas para obter o modelo de classificação.

1.1. OBJETIVOS

1.1.1. Objetivo principal

O objetivo principal deste trabalho é a extração de conhecimentos, através da mineração de dados, de um banco de imagens visando possibilitar a classificação das informações para auxiliar na tomada de decisão em caso de ocorrência do efeito de “branqueamento” nos revestimentos internos de linhas flexíveis durante o processo de fabricação.

1.1.2. Objetivos Específicos

Para alcançar a proposta deste trabalho, o objetivo principal pode ser dividido nos seguintes objetivos específicos:

- Estabelecer um conjunto de características de imagens que possam gerar conhecimento útil, onde dados são extraídos de cada imagem para compor um banco de dados, o qual será utilizado na mineração;
- Utilização da técnica de mineração, classificação, através de algoritmos de árvore de decisão, aplicada ao banco de dados obtido das imagens;
- Extração dos principais conhecimentos oriundos do modelo construído e validação dos mesmos junto aos especialistas do domínio.

1.2. JUSTIFICATIVAS

Linhas flexíveis são dutos compostos por várias camadas sobrepostas de materiais diferentes, desprovidas de qualquer tipo de união, sendo promotoras do transporte do petróleo nas plataformas de produção. Estima-se que, do custo agregado para o desenvolvimento de um campo petrolífero submarino, 25% advenha dos dutos flexíveis (MARTINS, 2011).

Devido ao aumento da utilização dos tubos flexíveis no ambiente marinho, há uma crescente investigação sobre o comportamento destas estruturas flexíveis. Embora já se tenha evoluído de forma importante quanto a uma configuração otimizada dessas estruturas, durante sua fabricação, alguns pontos ainda se configuram como obstáculos, necessitando ainda de estudos (SILVA, 2010).

Ainda segundo Silva (2010), a estrutura do duto flexível precisa ter um controle de qualidade rigoroso e intenso, por isso, as indústrias fabricantes deste tipo de produto tem um grande desafio para garantir a qualidade de construção conforme projeto de estrutura, de forma a garantir que todas as especificações do produto sejam atendidas em plenitude.

O problema de branqueamento dos dutos flexíveis durante o processo de fabricação gera diversos transtornos e prejuízos para as indústrias, onde muitas das vezes este problema ocasiona a perda de todo o material. Logo, pesquisas que possam contribuir para solucionar ou minimizar este problema é de grande utilidade para a indústria de dutos flexíveis.

1.3. ESTRUTURA DO TRABALHO

Este trabalho encontra-se dividido nos seguintes capítulos:

- Capítulo 1 – é apresentada uma introdução sobre o assunto;
- Capítulo 2 – uma revisão bibliográfica é realizada sobre temas relacionados às estruturas de dutos flexíveis e suas propriedades e modos de falha, branqueamento em PVDF, análise de imagens digitais e também sobre técnicas de descoberta de conhecimento em bancos de dados e mineração de dados;
- Capítulo 3 – apresenta a metodologia proposta pelo autor para estudar a problemática. A análise das características mais relevantes das manchas é apresentada e discutida e uma sequência de ações é estabelecida através da metodologia KDD com intuito viabilizar classificações com base aprendizado supervisionado e que podem vir a gerar e critérios de aceitação;

- Capítulo 4 – apresenta uma discussão sobre os resultados gerados, sobre a revisão do modelo e sobre a viabilidade de implementar o modelo proposto em uma situação real. Esta discussão embasa a conclusão que encerra o estudo;
- Capítulo 5 – a conclusão acerca da confiabilidade do modelo proposto e do potencial classificatório elaborado é apresentada, assim como as dificuldades encontradas ao longo do desenvolvimento do trabalho e algumas sugestões de trabalhos futuros.

CAPÍTULO 2 - REVISÃO BIBLIOGRÁFICA

2.1. DUTO FLEXÍVEL

A classificação dos dutos flexíveis pode ser dada a partir da sua construção tubular ou de sua função. No contexto deste trabalho o tipo de duto flexível que será abordado é conhecido como *unbonded pipe*, ou seja, linhas flexíveis de camadas sobrepostas não adesivas. Essa nomenclatura é derivada de seu modelo de construção, que por sua vez é composto de camadas poliméricas e metálicas intercaladas, propiciando movimentação entre essas camadas, provendo a adequada flexibilidade duto (BOSCHEE, 2012).

Na Tabela 1, são apresentadas as camadas típicas que compõem um duto flexível do tipo *unbonded pipe*, assim como as suas características principais.

Tabela 1 – Composição característica de um duto flexível e principal função de cada camada.
Fonte: Própria (2018)

Camada	Principal Função
Carcaça	Prover resistência ao Colapso
Barreira de pressão	Prover Estanqueidade do fluido interno
Armadura de pressão	Prover resistência à Pressão interna, colapso, esmagamento
Armadura de tração	Prover resistência à Cargas axiais de tração
Anti-desgaste	Reduzir o atrito entre camadas metálicas
Capa externa	Prover Estanqueidade do fluido externo

Nesse trabalho, foi dado ênfase apenas às características dos dutos flexíveis do tipo *unbonded pipe*. A Figura 1 ilustra as camadas que compõem um duto flexível típico.



Figura 1 – Camadas que compõem um duto flexível
Fonte: SILVA (2010)

As camadas que compõem uma estrutura de duto flexível são descritas por Boschee (2012) como dispostas e brevemente descritas da seguinte forma:

- Carcaça Intertravada (*carcass*) - Camada mais interna, normalmente feita de aços inoxidáveis especiais que tem como função resistir à pressão externa, prevenindo, assim, o colapso do duto;
- Barreira de Pressão, revestimento interno, camada de barreira ou simplesmente camada de Estanqueidade (*inner liner*) - Camada formada por um material polimérico responsável pela estanqueidade do fluido a ser transportado. Os materiais utilizados levam em consideração a temperatura e composição química do fluido transportado. Normalmente são utilizados PA-11 ou PA-12 (Poliamida) e PVDF (Fluoreto de polivinilideno). A camada será abordada neste trabalho simplesmente como revestimento interno;
- Armadura de Pressão (*pressure armour*) - Formada de aço carbono de baixa liga. Tem a função de auxiliar na resistência à pressão interna e externa;
- Camada anti-desgaste (*anti-wear*) - composto, na maioria das vezes, pelos mesmos polímeros utilizados na camada de estanqueidade e tem como objetivo reduzir o atrito entre as camadas;
- Armadura de Tração (*tensile armour*) - Geralmente composta por várias camadas de aços dispostas em forma de helicóide com a função de prover resistência à tração;
- Revestimento externo (*outer sheath*) - Construídas a partir de materiais poliméricos como HDPE, PVDF e outros, com a função de proteger todas as camadas internas da exposição ao meio externo.

Ainda de acordo com Boschee (2012), os dutos podem ser classificados por aplicação, da seguinte forma:

- Aplicação estática (*flowline*) - O duto flexível não é exposto a cargas e flexões cíclicas significativas durante a operação normal. Ele é dimensionado para operar inteiramente, ou parcialmente, depositado ou enterrado no leito marinho. Segundo Boschee (2012), aplicações estáticas são aquelas em que o componente não sofre carregamento dinâmico durante sua operação, ou aquelas situações em que os carregamentos dinâmicos são de ordem muito baixa, não oferecendo qualquer risco de falha por fadiga ao duto.
- Aplicação Dinâmica (*Riser*) - Os Dutos Flexíveis são expostos a cargas e flexões cíclicas significativas durante a operação normal. Ele é dimensionado especialmente

para resistir a um grande número de ciclos de dobramento e torções. Aplicações dinâmicas são todas aquelas em que há um carregamento cíclico de tal ordem a ser capaz de gerar falha por fadiga em um duto flexível.

2.1.1. Processo de manufatura

A fabricação de um duto flexível é composta por vários processos, onde cada processo basicamente é relacionado a uma das camadas que compõem o produto final. Em cada processo existe, uma série de fatores que podem contribuir para que haja um desvio de qualidade na produção, gerando um desvio de fabricação, o qual deve ser rapidamente identificado e avaliado. Na identificação, verifica-se se a ocorrência realmente configura-se como um desvio e, em seguida, avalia-se quais ações devem ser tomadas e quem será o responsável por conduzir as mesmas e/ou a quem recorrer (TEE et al., 2013).

Considerando que, a manufatura de um produto altamente tecnológico possui alto custo agregado aos seus processos de fabricação, entende-se que o consumo de tempo para tomada de decisões pode ser excessivamente oneroso e causar grandes prejuízos ao negócio.

Para garantir e controlar a estabilidade do processo, evitando elevado consumo de tempo durante a tomada de decisão, diferentes critérios de aceitação devem ser avaliados, dentre eles podem ser citados as tolerâncias geométricas de fabricação, assim como os padrões de caracterização de anomalias (marcas, pigmentos, ranhuras, entre outros).

Segundo a norma (API-17J, 2009), durante a fabricação de uma linha de duto flexível deve-se realizar a inspeção visual ao longo de todo o processo, que normalmente é feito em pontos estratégicos, através de um ou mais operadores qualificados. Além disso, toda a matéria-prima utilizada deve ser inspecionada, tanto durante o recebimento, quanto durante o carregamento do material nas bobinas de processo (ZHANG et al., 2003). Tais procedimentos tem o propósito de suprimir a ocorrência de não conformidades durante a fabricação dos produtos.

Contudo, durante o processo de manufatura pode ser necessário executar a tratativa de eventuais não conformidades em uma linha de produção, onde duas medidas devem ser imediatamente adotadas: i) investigar a causa raiz da ocorrência e atuar para mitigar a fonte do problema; ii) avaliar a falha no produto e tomar uma decisão sobre como proceder a partir de uma conclusão (BOSCHEE, 2012).

Desta forma, é objetivo da indústria de dutos flexíveis buscar a contínua melhoria dos processos de manufatura, melhorando a relação custo x benefício e aumentar a satisfação do

cliente. Para isso, faz-se necessário um rigoroso controle de qualidade durante a fabricação dos dutos.

2.1.2. Revestimento interno (*inner Liner*)

O revestimento interno de uma linha flexível, também chamada de barreira de pressão ou camada de estanqueidade, traduções do termo em inglês *inner Liner*, será a camada de interesse deste trabalho (Figura 2). O revestimento interno é a segunda camada de um duto flexível, feita de material polimérico que tem como função assegurar a estanqueidade do fluido além de transferir esforços mais internos para outras camadas mais externas (ZHANG ET AL., 2003). Por ser uma camada termoplástica, a barreira de pressão exerce uma função importante no isolamento térmico do fluido. Ajuda a garantir que o gradiente de temperatura entre os meios internos e externos do tubo não interfiram na viscosidade do fluido, assim garantindo uma melhor condução do mesmo (SOARES, 2015).

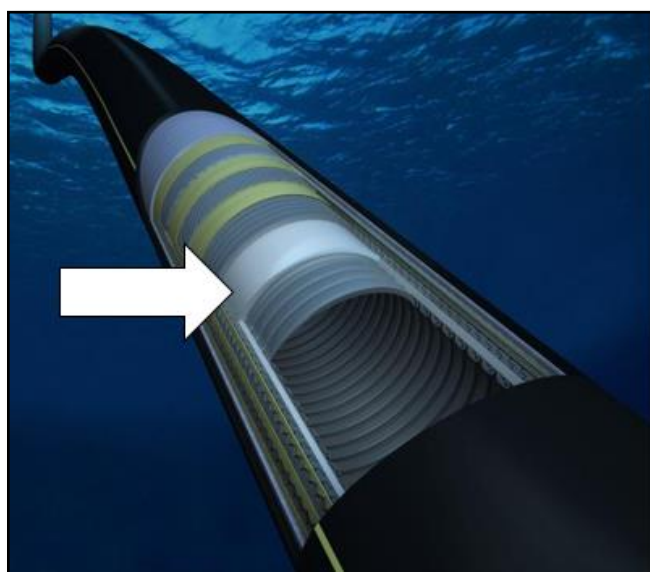


Figura 2 – Identificação da camada de estanqueidade em uma linha de duto flexível.

Fonte: Adaptado de site Web (<https://www.geoilandgas.com/subsea-offshore/flexible-pipe-solutions/flexible-pipe-solutions>), acesso em: 22-MAR-2018

Os materiais poliméricos são selecionados por apresentarem em sua composição química elementos capazes de ter boa resistência à corrosão, assim como também excelentes propriedades mecânicas a fim de suportar as condições de operações impostas pelos carregamentos à alta pressão e elevadas temperaturas (BOSCHEE, 2012). Os principais materiais poliméricos empregados nessa camada, assim como as aplicações recomendadas são:

- Polietileno de alta densidade – HDPE (água / gás e óleo dependendo da temperatura de pressão);

- Polietileno reticulado – XLPE ou PEX (água / gás e óleo dependendo da temperatura de pressão);
- Poliamida 11 – PA-11 ou nylon 11 (gás e óleo com limitações dependendo de temperatura, pressão, presença de água e seu pH);
- Poliamida 12 – PA-12 ou nylon 12 (gás e óleo com limitações dependendo de temperatura, pressão, presença de água e seu pH);
- *Polifluoreto de vinilideno* - PVDF (gás e óleo, altas pressões e temperaturas).

Este trabalho estará voltado para análise específica do material *Polifluoreto de Vinilideno*, comumente conhecido como PVDF e será abordado mais detalhadamente na Seção 2.1.3.

2.1.3. PDVF

Os materiais poliméricos utilizados nas linhas flexíveis são selecionados de acordo com as características da linha, como pressão, temperatura e composição química do fluido e de acordo com as funções de cada camada onde serão utilizados (MOTTA, 2017).

O Fluoreto de Polivinilideno (PVDF) é um polímero semicristalino termoplástico da classe dos fluorados. Possui grande estabilidade química, sendo resistente a solventes orgânicos, além de elevada resistência mecânica e térmica, podendo atuar em ambientes com até 130 °C de temperatura e 15000 psi de pressão (GOSI DE AQUINO, 2015). Uma das mais importantes características deste material é o fato de não sofrer hidrólise ao entrar em contato com a água do mar. Portanto este polímero é muito utilizado como material de vedação de diversos equipamento submarinos inclusive condicionado a elevadas temperaturas de operação (MOTTA, 2017).

O PVDF é um polímero linear formado pela reação de adição do monômero fluoreto de vinilideno ($\text{CH}_2 = \text{CF}_2$), tendo em média um peso molecular da ordem de 105g/mol, correspondendo a 2000 unidades repetitivas. Essa disposição espacial simétrica dos átomos de hidrogênio e flúor ao longo da cadeia, confere polaridade que influencia as suas propriedades. Sua resistência química está diretamente relacionada à forte ligação formada entre carbono e flúor no polímero

2.1.4. PVDF aplicado a dutos flexíveis

Devido às boas propriedades mecânicas e químicas do PVDF, a sua fácil extrusão e principalmente ao fato desse material não sofrer hidrólise, o PVDF é muito aplicado na indústria

de óleo e gás como material de revestimento impermeável nas barreiras de pressão de tubos flexíveis, Figura 3.



Figura 3 – Barreira de Pressão de PVDF em um tubo Flexível Dissecado.
Fonte: Abreu (2012)

Durante esse processo de manufatura, o PVDF ainda em estado líquido penetra na região onde existem os vales na superfície da carcaça, que ocorre na região do intertravamento do perfil dessa camada e, ao solidificar-se, realiza a completa vedação do duto flexível.



Figura 4 - Esboço de uma estrutura de duto flexível típico unbonded. A camada metálica mais interna, conhecida como carcaça é carregada pela pressão hidrostática externamente e pela pressão do fluido internamente, que é contido pelo revestimento interno.
Fonte: Adaptado de Abreu (2012)

Gosi de Aquino (2015), explica que devido ao advento da exploração de petróleo e gás natural na costa brasileira, novos campos vêm sendo descobertos em laminais d' água cada vez

mais profundas. Nesse ambiente, os hidrocarbonetos saem dos poços com alta temperatura e pressão, exigindo que os equipamentos submarinos utilizados para a sua produção sejam projetados de maneira a resistir circunstâncias severas.

Tee et al. (2013) apresentam um refinado estudo de seleção e especificação dos materiais de cada uma das camadas do tubo flexível, bem como uma análise do comportamento mecânico dos mesmos submetidos a ambientes severos, e explicitam ainda que este entendimento é de vital importância para a avaliação de seu funcionamento sob condições de operação.

2.1.5. Branqueamento do PVDF (*Stress whitening*)

Para Abreu (2012), Laiarinandrasana et al. (2009) e de SOUSA (2011), o branqueamento, conhecido na literatura internacional como *stress whitening*, é um fenômeno definido como a mudança na opacidade de um material polimérico devido a nucleação e/ou crescimento de micro vazios, resultando em aumento de volume e possível modificação de propriedades mecânicas. Sua formação pode surgir do processamento do material em algumas condições de temperaturas ou tensões mecânicas que atuam nos polímeros.

Castagnet et al. (2007) e Jerjen et al. (2013) mencionam que a deformação de termoplásticos durante o processamento pode induzir uma morfologia com uma variedade de defeitos específicos, incluindo branqueamento, cavitação ou porosidade. Tais defeitos ou descontinuidades podem ser prejudiciais à funcionalidade, ou pelo menos reduzir o tempo de vida útil esperado das peças.

Durante a moldagem por injeção, o material é altamente deformado (tensionado) no processo de fabricação. Para peças de fluoreto de polivinilideno (PVDF), o branqueamento em condições complexas de tensão e temperatura pode estar relacionado a dois mecanismos microestruturais, a nucleação e o crescimento de micro vazios. Portanto, a detecção e avaliação de defeitos microscópicos e descontinuidades são importantes para garantir a qualidade da produção.

Jerjen et al. (2013) propõem testes não destrutivos com métodos avançados de raios-X para fornecer informações sobre poros e descontinuidades semelhantes em peças termoplásticas moldadas por injeção ou, no caso de porosidade funcional, na distribuição e tamanho dos poros. No entanto, métodos convencionais de teste de raios-X não podem ser aplicados economicamente em uma linha de produção para controle de qualidade de 100%.

A compreensão do fenômeno de branqueamento e seu efeito sobre o comportamento mecânico do PVDF é de grande importância para engenharia uma vez que o uso deste material

vem amplamente sendo utilizado no mercado de óleo e gás em exploração em águas ultraprofundas (MOTTA, 2017).

Alguns dos principais fatores foram identificados na literatura como agentes causadores do fenômeno de branqueamento localizados nas camadas de revestimento de PDVF em dutos flexíveis.

- Tração excessiva do material: Castagnet et al. (2007) definem branqueamento do PVDF como um micro-mecanismo de dano conhecido como cavitação e é frequentemente relatado para micro e macro escalas de tensões em polímeros semicristalinos estirados. Neste caso, micro-vazios são nucleados dentro da fase amorfa. Em tensão, a cavitação começa nas camadas amorfas entre os cristais de lamelas perpendiculares ao eixo de estresse. À medida que o alongamento dos esferulitos aumenta, a cavitação se espalha progressivamente do mais inclinado para as camadas amorfas do eixo de tensão. Abreu (2012a) executou testes de tração em corpos de provas de PVDF e observou que o efeito de branqueamento ocorre quando há o alinhamento das cadeias do polímero levando a uma mudança na opacidade do material, primeiro de forma homogênea, ao longo de toda a área útil e, em seguida, com a formação de pescoço, esse branqueamento se torna mais intenso e concentra-se ao redor da região de estricção.
- Fadiga: Abreu (2012a) ainda efetuou uma análise do efeito do branqueamento sob carregamento cíclico demonstrando que o branqueamento também ocorre quando por fadiga e por envelhecimento do material exposto a hidrocarbonetos crus, como o petróleo. Motta (2017) apresentou um estudo que descreve o comportamento do PVDF para carregamentos monótonos e cíclicos e que simula o fenômeno de fluência e relaxação do material.
- Envelhecimento: O branqueamento pode aumentar a taxa de permeação do gás de operação, reduzir a vida útil da camada de estanqueidade e potencialmente levar o duto flexível à falha de operação (JERJEN et al., 2013). Chaves (2017) concluiu em seu trabalho que o tempo de envelhecimento influenciou na redução do módulo de elasticidade do PVDF. Foi evidenciado em seus estudos que o efeito de branqueamento isoladamente praticamente não influenciou no módulo de elasticidade do material. Entretanto, a associação do processo de envelhecimento com a geração de branqueamento ocasionou uma progressão de danos a resistência

mecânica do PVDF, não importando em qual ordem a geração de branqueamento aconteceu, se antes ou após o envelhecimento.

- **Fabricação:** Ainda segundo Motta (2017), o fato pode estar relacionado ao processo de fabricação da barreira de pressão de um tubo flexível, pois tensões são geradas durante o próprio processo de extrusão do polímero. Isso é explicado em estudos realizados por Motta, onde constataram que algumas barreiras de pressão de PVDF apresentam características de branqueamento logo após a sua fabricação. Tal efeito foi relacionado com a nucleação e crescimentos de microvazios devido choque de resfriamento e a aplicação de cargas mecânicas.

Tendo em vista os riscos associados ao branqueamento, é fundamental controlar a ocorrência deste tipo de característica durante o processo de fabricação, a fim de garantir a qualidade do produto conforme desempenho esperado (LAIARINANDRASANA et al., 2009b).

2.2. MODELO DE IMAGEM DIGITAL

Para Gonzalez et al. (2001), as técnicas de processamento de imagens, desenvolvidas ao longo do tempo, apresentam duas facetas principais: uma de melhoramento da qualidade das informações, facilitando a observação humana; a outra de permitir a extração de dados, de forma automática ou semiautomática, por uma máquina. Com o aprimoramento computacional (aquisição, armazenamento, capacidade de processamento, comunicação e exibição) e o desenvolvimento de novos algoritmos, o tratamento digital de imagens se faz presente em novas áreas da ciência, tornando-se, hoje em dia, muitas vezes, procedimento padrão e até mesmo transparente ao usuário.

O uso de imagens digitais iniciou-se na década de 20 do século passado como forma de reduzir o tempo do transporte de imagens entre os jornais de Londres e Nova York. As imagens eram digitalizadas e enviadas por um cabo submarino, reduzindo o tempo de envio de mais de uma semana para menos de três horas (IGNACIO, 2013). Isso fez com que o interesse na digitalização de imagens aumentasse e chegasse ao ponto em que estamos hoje. A este processo de transformar uma imagem real capturada por equipamentos específicos em uma imagem digital deu-se o nome de Processamento Digital de Imagens (PDI).

Segundo Queiroz e Gomes (2006), objetos que emitem luz visível são percebidos em função da soma das cores espectrais emitidas. Tal processo de formação é denominado aditivo. O processo aditivo pode ser interpretado como uma combinação variável em proporção de

componentes monocromáticas nas faixas espectrais associadas às sensações de cor verde, vermelho e azul, as quais são responsáveis pela formação de todas as demais sensações de cores registradas pelo olho humano. Assim, as cores verde, vermelho e azul são ditas cores primárias. Ainda segundo Queiroz e Gomes (2006), este processo de geração suscitou a concepção de um modelo cromático denominado RGB (*Red, Green e Blue*), para o qual a Comissão Internacional de Iluminação (CIE) estabeleceu as faixas de comprimento de onda das cores primárias. A combinação dessas cores, duas a duas e em igual intensidade, produz as cores secundárias, Ciano, Magenta e Amarelo.

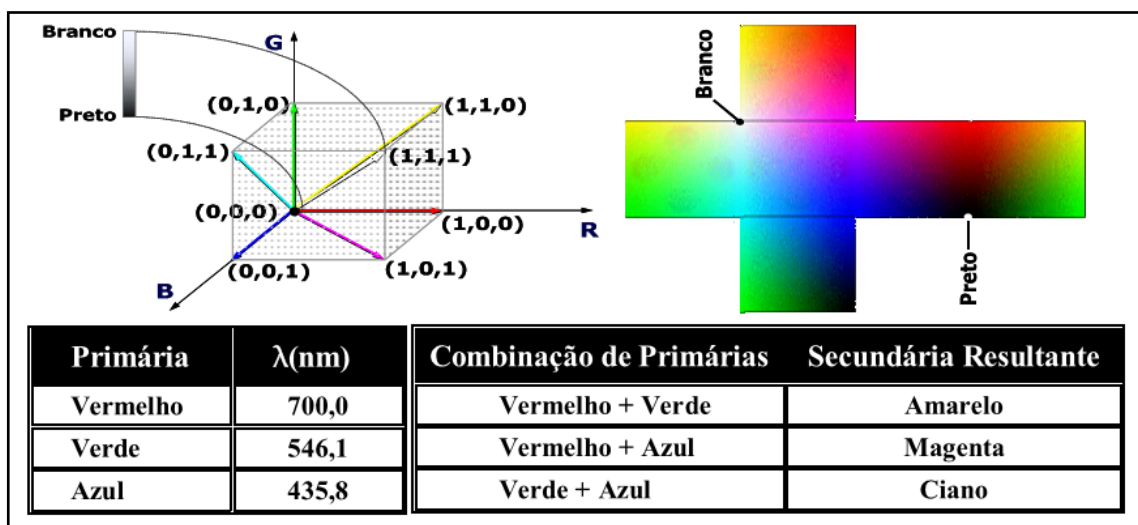


Figura 5 – Matriz de combinação de cores.
Fonte: Adaptado de Queiroz e Gomes (2006).

Queiroz e Gomes (2006) definem uma imagem monocromática como uma função bidimensional contínua $f(x,y)$, na qual x e y são coordenadas espaciais e o valor de f em qualquer ponto (x,y) é proporcional à intensidade luminosa (brilho ou nível de cinza) no ponto considerado. Como os computadores não são capazes de processar imagens contínuas, mas apenas vetores (*arrays*) de números digitais, é necessário representar imagens como arranjos bidimensionais de pontos. Cada ponto na grade bidimensional que representa a imagem digital é denominado elemento de imagem ou pixel.

Na Figura 6, apresenta-se a notação matricial usual para a localização de um pixel no arranjo de pixels de uma imagem bidimensional. O primeiro índice denota a posição da linha, m , na qual o pixel se encontra, enquanto o segundo, n , denota a posição da coluna. Se a imagem digital contiver M linhas e N colunas, o índice m variará de 0 a $M-1$, enquanto n variará de 0 a

N-1. Observe-se o sentido de leitura (varredura) e a convenção usualmente adotada na representação espacial de uma imagem digital (GONZALEZ at al., 2001).

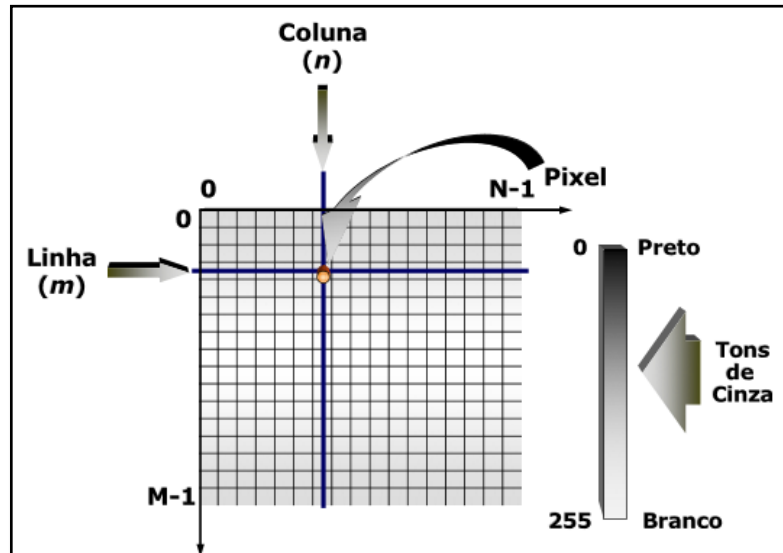


Figura 6 – Representação de uma imagem digital bidimensional
Fonte: Adaptado de Queiroz e Gomes (2006).

Para representar imagens em preto-e-branco (e com tons intermédios de cinza) basta associar a cada pixel um valor numérico numa escala de tons. Se quisermos, por exemplo, uma escala com 256 tons de cinzento (do preto absoluto ao branco absoluto) podemos associar a cada pixel um valor inteiro não-negativo de um byte. Convencionamos que o valor 0 corresponde ao preto e 255 corresponde ao branco; os valores intermédios correspondem aos tons intermediários chamados de cinzentos.

A Figura 7 ilustra uma possível codificação de uma imagem em linguagem computacional. Note que são declarados os valores dos pixels como para representar valores entre 0 e 255.

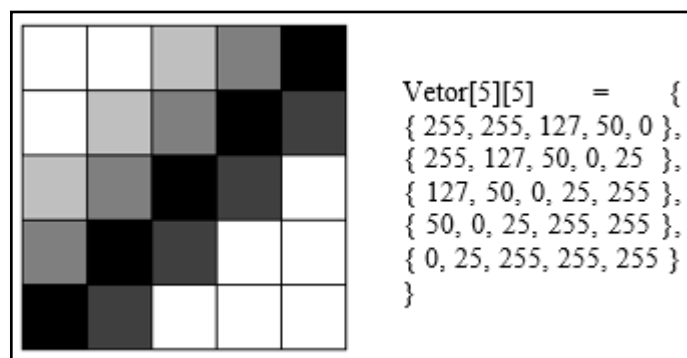


Figura 7 - Imagem em cinzentos e representação em linguagem computacional
Fonte: Adaptado de Queiroz e Gomes (2006).

2.3. MINERAÇÃO DE DADOS

O objetivo de um processo da mineração de dados (MD) é extrair conhecimento a partir de informações “ocultas” nos dados, que sejam úteis nas tomadas de decisões, utilizando métodos de diferentes áreas científicas, como Estatística, Banco de Dados, Inteligência Artificial, Aprendizado de Máquina e Reconhecimento de Padrões, caracterizando esse processo como interdisciplinar (WITTEN et al., 2016). As etapas desse processo são compostas pelo pré-processamento, construção de padrões e modelos e pós-processamento de dados, que envolvem, respectivamente, o tratamento dos dados de forma a qualificá-los para a entrada dos algoritmos de mineração de dados, a execução dos algoritmos para extração de padrões, e a compreensão das saídas para geração de conhecimento.

Para Witten et al. (2016), na MD, a principal atividade envolve algoritmos de extração de padrões a partir da definição de uma tarefa de análise. As tarefas podem ser preditivas ou descritivas. As tarefas preditivas consistem na generalização de exemplos ou experiências passadas com respostas conhecidas, relacionadas às classes dos exemplos. Se a classe dos exemplos pertence a um conjunto discreto e finito de valores, a tarefa é denominada classificação; e se pertence a um conjunto numérico de valores, é denominada regressão. Já as tarefas descritivas consistem na identificação de comportamentos intrínsecos do conjunto de dados, sem que haja classes definidas dos exemplos. As tarefas descritivas podem ser de construção de regras de associação ou de agrupamento de dados

2.3.1. Conceitos preliminares

Esta seção destina-se à definição dos termos técnicos adotados ao decorrer da dissertação. Os termos abordam notações matemáticas necessárias para a descrições dos algoritmos de classificação. Além disso, é apresentada a definição de mineração de dados, o contexto em que é aplicada e a exemplificação de uma tarefa de classificação.

A noção de encontrar padrões úteis em grandes volumes de dados pode ser conhecida por diversos nomes, tais como mineração de dados, extração de conhecimento, descoberta de informações, arqueologia de dados e processamento de padrões de dados dentre outros. O termo mineração de dados é o mais usado por profissionais da computação, estatísticos e analistas de dados.

Segundo Han et al. (2016), a fase de mineração de dados é uma fase do processo de Descoberta de Conhecimento em Banco de dados (KDD). Esta etapa é responsável pela

aplicação dos algoritmos que são capazes de identificar e extrair padrões relevantes presente nos dados.

O processo KDD é interdisciplinar, tanto em sua aplicação, quanto das suas fundamentações teóricas. O processo pode ser aplicado a qualquer problema de identificação de padrões em dados e contém fundamentação de diversas áreas como a banco de dados, inteligência artificial, estatística, probabilidade e visualização de dados.

2.3.2. Banco de dados

Parmezan (2012) explica que um banco de dados é composto por um conjunto de amostras descritas em um conjunto de variáveis. As classes representam o rótulo de cada amostra, determinando o estado de uma amostra. As características, também denominadas variáveis, são específicas em cada banco de dados. Seguem alguns exemplos: os componentes químicos de alimentos ou outros produtos, perguntas de um questionário aplicado a um conjunto de pessoas ou corporações, assim como estados ou posições de uma molécula perante uma reação.

Parmezan (2012) fala ainda que conhecer o tipo dos dados com o qual se irá trabalhar também é fundamental para a escolha do(s) método(s) mais adequado(s) de MD. Pode-se categorizar os dados em dois tipos: quantitativos e qualitativos. Os dados quantitativos são representados por valores numéricos. Eles ainda podem ser discretos e contínuos. Já os dados qualitativos contêm os valores nominais e ordinais (categóricos). Em geral, antes de se aplicar os algoritmos de mineração é necessário explorar, conhecer e preparar os dados.

Com um entendimento inicial da base de dados definido, faz-se necessário estudar mais profundamente tais dados, afim de avaliar fatores que possam comprometer a qualidade da mineração de dados, tais como: valores em branco ou nulos, valores viciados, variáveis duplicadas ou com interdependência, entre outras. À medida em que problemas vão sendo encontrados e o entendimento vai sendo obtido, ocorre a preparação dos dados para que os algoritmos de mineração possam ser aplicados.

O processo de preparação dos dados para a mineração, também chamado de pré-processamento, pode ser observado através da Figura 8. Segundo Han et al. (2016), o processo consiste principalmente em:

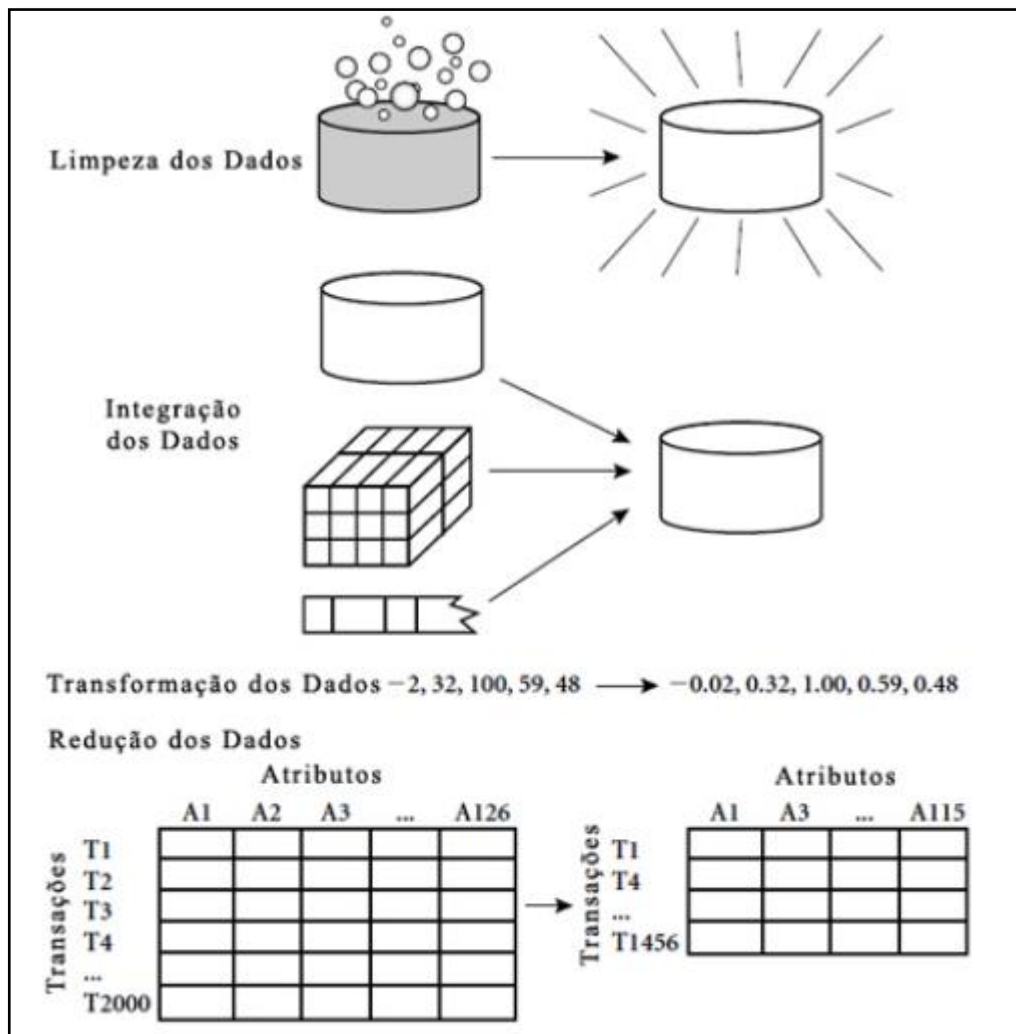


Figura 8 - Atividades do pré-processamento.
Fonte: HAN et al. (2016)

Limpeza dos dados: Dados são usualmente encontrados com diversas inconsistências. Falhas como registros incompletos, valores errados e dados inconsistentes são citados por Camilo (2009) como potências problemas encontrados nos bancos de dados. A etapa de limpeza dos dados visa eliminar estes problemas de modo que eles não influam no resultado dos algoritmos usados. HAN et al. (2016), demonstram que dada a complexidade deste processo, existe grande necessidade de utilização de métodos específicos para executar a limpeza de dados.

Integração dos dados: Uma vez que dados podem ser extraídos de qualquer fonte, como por exemplo videos, imagens, arquivos, documentos, etc., existe a necessidade da integração destes dados com intuito de estabelecer um repositório único e consistente. Para isto, uma análise detalhada dos dados é necessária para que redundâncias, dependências por meio das variáveis e valores conflitantes, possam ser identificados.

Transformação dos dados: Esta etapa do processo é muito importante uma vez que os dados a serem explorados normalmente não estão dispostos da forma como eles serão trabalhados. Dependendo do tipo de algoritmo a ser utilizado nas etapas seguintes, o repositório deve ser adequado ao modelo para que o algoritmo possa funcionar adequadamente. Valores binários de classificação, representações numéricas, texto, são possíveis formas de representar parâmetros em um conjunto de dados (*dataset*). Os dados devem ao final do processo de transformação, estarem ajustados para as funcionalidades a que se destina o modelo de estudo de acordo com os objetivos pretendidos.

Redução dos dados: Com frequência encontra-se bancos de dados com alto volume de informação. Para melhorar a performance dos algoritmos uma redução de dados pode potencialmente resultar em melhor tempo de análise, melhor taxa de acerto e apresentar regras de classificação mais pertinentes. É possível converter os dados originais de um banco de dados em um conjunto reduzido de informações, ainda sim mantendo preservada a sua capacidade de extrair conhecimento. É comum que encontrar bancos de dados reduzidos de seu tamanho original, uma vez que o próprio modelo precisa ser testado.

Segundo Camilo (2009), dado este fato, os bancos de dados podem ser representados em três conjuntos distintos:

- 1) Conjunto de Treinamento (*Training Set*): conjunto de registros usados no qual o modelo é desenvolvido;
- 2) Conjunto de Testes (*Test Set*): conjunto de registros usados para testar o modelo construído;
- 3) Conjunto de Validação (*Validation Set*): conjunto de registros usados para validar o modelo construído;

Ainda segundo Camilo (2009), essa divisão de conjuntos é importante para que o modelo possa ser treinando sob uma condição independente, onde os testes de validação não sejam feitos a partir de “dados viciados” que já foram considerados para o treinamento e, portanto, ao serem submetidos a outros testes possam apresentar inconsistências. Este efeito é chamado de efeito Bias. A medida que se aumenta a precisão do modelo para um conjunto de dados específico, perde-se a precisão para outros conjuntos.

2.3.3. Processo de descoberta de conhecimento em bancos de dados

Conforme a definição de Fayyad et al. (1996), a mineração de dados faz parte do processo de descoberta de conhecimento em banco de dados, *Knowledge Discovery Databases*

(KDD). Este processo visa à extração de padrões e informações em bases de dados que são previamente desconhecidas e que podem ser potencialmente úteis. A Figura 1.1 a seguir apresenta as fases do processo de KDD.

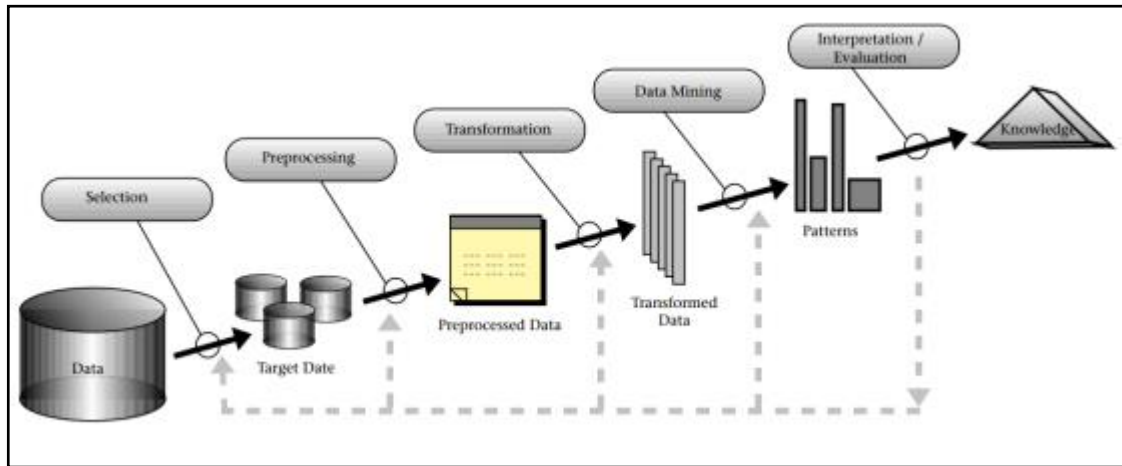


Figura 9 - Processo de descoberta de conhecimento em bancos de dados.
Fonte: Fayyad et al. (1996)

De forma resumida, é descrito a seguir as etapas que consistem o processo de descoberta de conhecimento em banco de dados (KDD).

Seleção: este primeiro passo tem como objetivo entender o domínio de aplicação dos dados a serem analisados, identificar as possibilidades de padrões que podem existir e selecionar um conjunto de dados que podem produzir o resultado esperado.

Pré-processamento: no segundo passo realiza-se uma série de procedimentos que caracterizam o pré-processamento dos dados. Essa etapa inclui as possíveis ações de remoção de ruídos, o tratamento de valores ausentes em algumas amostras, normalização e o balanceamento das classes.

Transformação: O terceiro passo consiste na redução ou transformação dos dados. Este processo ocorre com as técnicas de seleção e extração de variáveis, descritas na seção 2.3.2. Nessa etapa, espera-se selecionar ou transformar as variáveis que contribuem suficientemente para a representação dos dados de acordo com as classes.

Mineração de dados: É onde ocorre a mineração de dados especificamente. Nesta etapa, os algoritmos são aplicados com intuito de identificar os padrões e relacionamentos existentes nos dados.

Interpretação dos padrões: Basicamente a interpretação dos resultados gerados e podem ser avaliados por um especialista de domínio capacitado. Os padrões gerados pelos algoritmos de mineração de dados são identificados como pertinentes ou não propiciando uma

revisita ao modelo de MD se for necessário realizar ajustes de parâmetros dos algoritmos ou aprimorar os padrões e relacionamentos encontrados.

Ainda segundo Fayyad et al. (1996), os objetivos do KDD resumem-se em dois tópicos: i) verificação, limitando-se a verificar a hipótese do usuário; e ii) descoberta, em que há a descoberta de novos padrões. A descoberta de padrões é subdividida em PREDIÇÃO e DESCRIÇÃO. A predição utiliza-se dos padrões encontrados para predizer novas amostras. A descrição explica os padrões encontrados de forma que o usuário possa compreendê-los.

2.3.4. Aprendizado de máquina

Para atingir o objetivo do KDD, os algoritmos realizam um aprendizado de máquina (AM). Este termo define o processo que um algoritmo realiza para aprender rapidamente padrões complexos que regem os dados (CAMILO, 2009).

O aprendizado de máquina supervisionado é caracterizado pela presença de um professor, que auxilia no processo de aprendizagem. Este professor é representado pelas classes dos dados. A partir de amostras com classes pré-definidas, o algoritmo aprende como as classes são organizadas perante suas variáveis. Posteriormente, o conhecimento obtido é utilizado para predizer classes de amostras não rotuladas.

O aprendizado de máquina não supervisionado não possui a presença de um professor. Em conjunto de dados não rotulados, o aprendizado não supervisionado analisa as variáveis do conjunto de dados e define grupos que estão presentes nos dados.

2.3.5. Estimativa, ruído e desvio

A estimativa é usada para definir um valor para alguma variável contínua desconhecida como, por exemplo, altura de uma pessoa ou saldo de cartão de crédito. Ela trabalha com resultados contínuos. Pode ser usada para executar uma tarefa de classificação, convencendo-se que diferentes faixas de valores contínuos correspondem a diferentes classes (FAYYAD et al., 1996). Como exemplos de tarefas de estimativa têm-se: estimar o número de filhos em uma família, estimar a renda total de uma família, estimar o valor em tempo de vida de um cliente, estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de um conjunto de diagnósticos médicos ou prever a demanda de um consumidor para um novo produto, ainda segundo Fayyad et al. (1996).

O “ruído” representa conteúdo nas bases de dados que pode prejudicar a qualidade da informação extraída, a partir de qualquer método, seja ele tradicional ou baseado em estratégias mais elaboradas. Destacam-se como ruídos: valores fora do domínio, ausência de valores,

inconsistências etc. É importante lembrar que o mundo real é ruidoso, ou seja, se uma base de dados representa uma abstração deste mundo real, esta será ruidosa a despeito dos esforços despendidos para a sua modelagem e respectiva população. Cabe aos profissionais da área de tecnologia da informação minimizar o impacto negativo que estes ruídos possam representar nas informações extraídas e disponibilizadas aos gestores.

A tarefa de redução de desvio tem por objetivo descobrir um conjunto de valores que não seguem padrões definidos. Para esta tarefa é necessário adotar padrões antecipadamente. Pode-se usar esta tarefa para identificar fraudes baseadas em elementos que estão fora dos padrões ou são exceções às regras.

2.3.6. Seleção de atributos

Parmezan et al. (2012), explica que existem diferentes definições na literatura para determinar a importância de um atributo. Algumas definições consideram a importância do atributo em relação à classe, enquanto outras trabalham com a seleção de atributos (SA) não supervisionada. Essa necessidade de estimativa da importância de atributos é comum, tanto à avaliação individual quanto à avaliação de subconjuntos de atributos. Por meio das medidas de importância é possível, por exemplo, avaliar se os atributos selecionados auxiliam a melhorar o desempenho preditivo do classificador, ou auxiliam a simplificar o modelo construído de modo que ele seja mais compreensível.

É importante ressaltar que as medidas de precisão são, geralmente, dependentes do algoritmo de aprendizado considerado, pois os subconjuntos de atributos selecionados são importantes em relação ao desempenho preditivo do modelo induzido por um determinado algoritmo. Diferentemente, as medidas de separabilidade de classes/grupos são independentes do algoritmo de aprendizado escolhido para a posterior construção do modelo.

Os experimentos apresentados neste trabalho foram realizados utilizando seis algoritmos para a tarefa de SA, os quais são baseados nas principais medidas de importância de atributos dispostas nas categorias clássica, consistência e precisão, e dois algoritmos para indução de modelos.

2.3.7. Principais avaliadores de atributos

Em seu trabalho, Parmezan et al. (2012) analisam o comportamento dos principais algoritmos de seleção de atributos disponibilizados no WEKA:

- Avaliação de correlação de atributos (*Correlate Attribute* - CFS)
- Avaliador de Subconjuntos (CSE)

- Avaliador de recursos de taxa de ganho (*Gain Ratio*)
- Filtro de classificação de ganho de informações (*Info gain*)
- Avaliador de recursos OneR (*OneR attribute*)
- Filtro de classificação do relevo F (*Relief attribute*)
- Filtro de classificação de incerteza simétrica (*Symmetrical Uncert*)

Através do método de árvores de decisão, a partir do algoritmo J48, Parmezan et al. (2012) obtém resultados para diferentes *datasets* disponíveis pelo próprio WEKA. A Tabela 2 apresenta o resultado da redução de tamanho do conjunto de atributos a partir da aplicação de cada algoritmo de SA enquanto a

Tabela 3 apresenta o erro médio de classificações através da aplicação do método de árvores de classificação J48, considerando a aplicação de cada algoritmo de SA abordado.

Tabela 2 - Quantidade de atributos selecionados por cada um dos algoritmos de seleção de atributos.
Fonte: Adaptado de (PARMEZAN et al., 2012)

Conjunto de dados	Original	CBF	CFS	CSE	CSE	InfoGain	ReliefF
BreastCancer	9	8	5	5	8	3	3
Bupa	6	1	1	4	5	2	2
Haberman	3	2	2	3	3	1	1
Hepatitis	19	12	10	6	12	6	6
Hungarian	13	10	6	8	13	4	4
LungCancer	56	4	8	2	4	17	17
Pima	8	8	4	6	8	2	2

Tabela 3 - Média de erro para cada conjunto de dados e algoritmos considerados (método J48).
Adaptado de (PARMEZAN et al., 2012)

Conjunto de dados original	Original	CBF	CFS	CSE(J)	InfoGain	ReliefF
BreastCancer	25,72	27,49	27,06	25,37	28,14	26,67
Bupa	34,16	37,76	37,76	32,48	37,50	31,93
Haberman	27,84	27,97	27,97	27,84	28,79	26,47
Hepatitis	20,78	20,26	19,73	17,26	16,15	15,78
Hungarian	19,78	19,60	21,00	22,33	20,56	21,09
LungCancer	20,75	16,42	16,42	15,75	17,42	19,33
Pima	25,51	25,51	25,62	24,86	25,35	25,35

Em Parmezan et al. (2012) esta análise é abordada em maiores detalhes.

2.3.8. Tipos de aprendizado de máquina: preditivo e descritivo

Segundo Carvalho et al. (2014), o aprendizado preditivo analisa os dados que representam eventos passados buscando relações entre estes que permitam prever situações em novos dados futuros, tais como a classificação para previsões de valores discretos e a

regressão para predições de valores contínuos. Já o aprendizado descritivo efetua a busca pela compreensão dos acontecimentos do presente, para tomadas de decisão instantaneamente.

Como exemplo de aprendizado preditivo, pode-se identificar potenciais clientes interessados em algum novo produto a ser divulgado. Como exemplo de aprendizado descritivo, pode-se citar a associação entre eventos demandados por pacientes, que podem indicar possíveis relações de causa-efeito ao ser complementada com a respectivo espaço de tempo entre estes eventos associados.

Carvalho et al. (2014), ainda diz que em geral, a natureza do problema em questão define a natureza do aprendizado a ser adotado na experimentação de mineração de dados, ou seja, em geral os problemas são preditivos ou descritivos. Porém muitas vezes as duas estratégias podem ser adotadas de forma complementar. Por exemplo, o sistema aprende sobre as preferências históricas de compras e avaliações e a partir deste aprendizado realiza recomendações de novos produtos. O aprendizado para esta recomendação pode ser obtido pela hibridização destes dois tipos de aprendizados.

2.3.9. Classificação

As tarefas de classificação se apresentam cada vez mais diversificada, dado a vasta área de aplicação da mineração de dados. Tais tarefas de classificação já demandam desde o início do processo de extração de conhecimento, que seja definido o método a ser utilizado, uma vez que a variação de resultados pode depender do método.

A classificação pode ser considerada como uma tarefa de aprendizado que mapeia dados de entrada, ou conjuntos de dados de entrada, em um número finito de classes. Nela, cada exemplo pertence a uma classe, entre um conjunto pré-definido de classes. O objetivo de um algoritmo de classificação é encontrar alguma correlação entre os atributos e uma classe, de modo que o processo de classificação possa usá-la para predizer a classe de um exemplo novo e desconhecido.

O processo consiste em obter um modelo baseado em um conjunto de exemplos que descrevem uma função desconhecida. Esse modelo é utilizado posteriormente para fornecer o valor de atributos de novos exemplos. Para Parmezan et al. (2012), exemplos de tarefas de classificação podem ser: separar pedidos de créditos em baixo, médio e alto risco ou identificar a forma de tratamento mais adequado para um paciente, baseando-se em classes de pacientes que respondem bem a determinado tipo de tratamento médico.

Para Larose (2006), resumidamente, as principais regras de classificação que são abordadas no WEKA podem ser identificadas e descritas como:

- Rede neural (Multilayer Perceptron) é uma tentativa de reproduzir o aprendizado não linear que ocorre na rede de neurônios do sistema neural biológico. É utilizada para tarefas que buscam estimação ou predição. Apesar da estrutura de um neurônio ser relativamente simples, uma rede pode apresentar uma conjuntura densa conectada por inúmeros neurônios com o intuito de executar tarefas complexas como classificação e reconhecimento de padrões;
- Regressão logística - é uma técnica de modelagem preditiva classificada como método bivariado. Em alguns casos o objetivo de interesse é utilizar o valor de uma variável para determinar o valor de outras variáveis em determinada linha do tempo;
- Naive Bayes - na abordagem bayesiana os parâmetros são considerados variáveis aleatórias e os dados são considerados conhecidos. Os parâmetros são considerados provenientes de uma distribuição dos possíveis valores, e os bayesianos procuram os dados observados para fornecer informações sobre valores prováveis de parâmetros;
- Árvore de decisão é um fluxo em uma estrutura de árvores, em que cada nó representa um teste de um dos atributos, cada ramo representa o resultado desse teste e cada folha possui uma classe. A partir do campo escolhido como dado de saída, que será exibido nas folhas da árvore, o campo mais significativo é colocado como nó raiz (topo) e, partindo daí, de forma recursiva, novos campos são colocados como nós, conectando a raiz às folhas com os próximos campos mais significativos. A técnica de árvore de decisão, devido à sua boa visualização, é comumente utilizada para tarefas de classificação, pois sua construção é intuitiva e de fácil entendimento por humanos (LAROSE, 2006). Outros fatores que corroboram para ampla utilização dessa técnica, além de possuir um bom desempenho, é a possibilidade de trabalhar com dados multidimensionais e não requerer conhecimento prévio de domínio (HAN ET AL., 2016).

2.4. WEKA

A escolha da ferramenta analisada, deu-se não apenas pelo fato de ser bastante conhecida na área, mas também por ter uma variedade de funcionalidades em comum e possuir versões de acesso livre, conforme já abordado em (IVANCIUC, 2008).

WITTEN et al. (2016) descreve o WEKA como uma coleção de algoritmos de aprendizado de máquina e ferramentas úteis para o processamento de dados. Desenvolvida na Universidade de Waikato, Nova Zelândia, de forma que fosse possível testar diferentes técnicas em diferentes bancos de dados rapidamente. Em relação a MD, o WEKA possui ferramentas

para tarefas de regressão, classificação, agrupamento (*clustering*), associação e seleção de atributos. Esta ferramenta foi escolhida para a realização das tarefas de MD envolvendo a classificação de dados. A facilidade de uso, vasta documentação disponível, concentração de algoritmos, gratuidade e possibilidade de conexão direta com banco de dados foram essenciais para a escolha desse software.

WEKA é um projeto *open source* que significa *Waikato Environment for Knowledge Analysis*. Foi criado como um projeto de *Machine Learning* pela universidade de Waikato na Nova Zelândia. O projeto tem o objetivo de disseminar técnicas de *Machine Learning* através da disponibilização do *software* para utilização de pesquisadores, alunos e para resolver problemas reais da indústria além de contribuir com a ciência pelo mundo.

O grande diferencial do WEKA além de todo o seu arsenal de métodos e algoritmos é a sua interface gráfica (GUI – *Graphical User Interface*) que torna as tarefas de mineração de dados extremamente fáceis e rápidas. Através da interface é possível consultar dados em sistemas de bancos de dados, executar métodos de processamento de dados, executar e configurar parâmetros dos algoritmos e visualizar os resultados através de gráficos. Tudo isso sem precisar escrever comandos ou programar.

O WEKA tem funcionalidades para manipulação de bases de dados (pre-processamento), interface para visualização de dados, e ainda disponível diversos algoritmos de *machine learning* e *Data Mining*. Isso facilita muito a vida dos seus usuários que não tem que dominar diversas ferramentas para fazer seu trabalho.

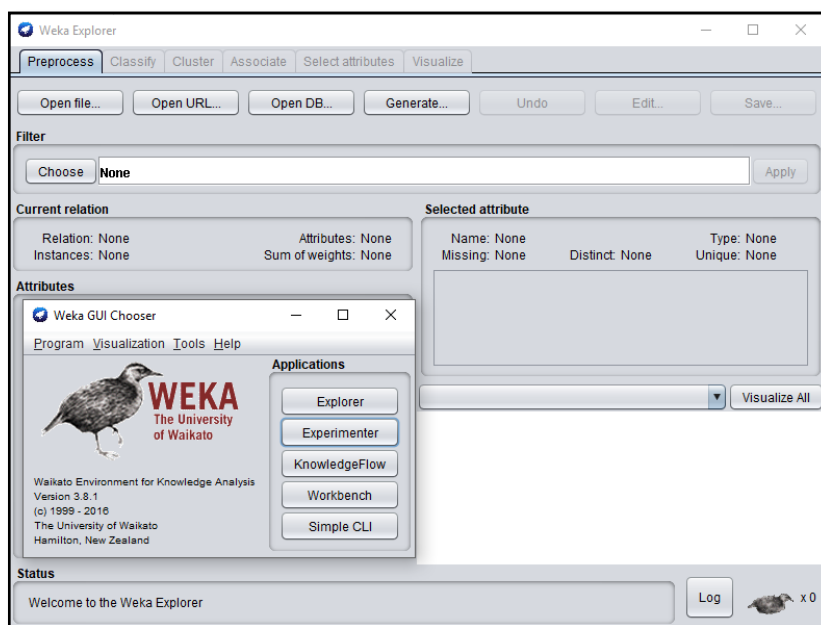


Figura 10 - Interface típica do WEKA
Fonte: WEKA (2018)

Os resultados da aplicação de cada metodologia serão avaliados através de alguns *outputs* apresentados pela ferramenta WEKA descritos a seguir conforme (WITTEN et al., 2016):

- **TP rate:** taxa de verdadeiros positivos (instâncias classificadas corretamente como uma determinada classe). Quanto mais próximo de 1 TP rate for, mais próximo está de 100% do acerto nas classificações;
- **FP rate:** taxa de falsos positivos (casos classificados falsamente como uma determinada classe). É o oposto do TP rate e quanto mais próximo de 0, mais próximo está de 100% do acerto nas classificações;
- **Precision:** proporção de instâncias que são verdadeiramente de uma classe dividida pelo total de instâncias classificadas como essa classe;
- **MCC:** é usado no aprendizado de máquina como uma medida da qualidade das classificações binárias (duas classes). Ele leva em conta os verdadeiros e falsos positivos e negativos e é geralmente considerado como uma medida balanceada que pode ser usada mesmo se as classes forem de tamanhos muito diferentes;
- **Class:** é a classificação final do objeto, no caso deste estudo, foi definido como amostras aprovadas ou reprovadas. Estas classificações serão os resultados das previsões apresentadas mais adiante.

2.4.1. J48

Avaliando o trabalho de Salzberg (1994), entende-se que o método de árvores de decisão construídas por meio do algoritmo J48, é uma versão otimizada do algoritmo C4.5 de QUINLAN (1996) implementada no pacote de software do WEKA. Justifica-se essa escolha devido ao algoritmo ser o mais utilizado e conhecido na sua categoria, sendo amplamente incorporado em ferramentas de mineração de dados.

Utilizando o método “dividir para conquistar”, o C4.5 é chamado com três parâmetros: “D”, partição de dados que inicialmente é compreendida por todos os atributos do conjunto de treino com suas respectivas classes; “Lista_Atributos”, valores que descrevem os atributos da partição “D” e “Método_Seleção_Atributo”, que especifica o procedimento heurístico usado para revelar o atributo que melhor define a amostra.

Dessa maneira, em cada nó da árvore, o algoritmo escolhe através do seu Método_Seleção_Atributo, o atributo que mais particiona o seu conjunto de amostra “D” em novos subconjuntos que tenderão a uma categoria ou outra (HAN et al., 2016).

A estrutura final da Árvore de Decisão é formada por uma sequência hierárquica em forma de árvore invertida da raiz para as folhas. As folhas das árvores representam regras que explicam o conhecimento embutido nos dados de forma humanamente compreensível por regras “se-então”

Algumas medidas de ajustes podem ser adequadas ao algoritmo J48, como por exemplo a aplicação ou não do MDL. Davies et al. (2002) explica que medidas baseadas em MDL usam técnicas de codificação para definir a “melhor” árvore de decisão como aquela que requer menos número de bits para (1) codificar a árvore e (2) codificar as exceções para a árvore (ou seja, casos que não são classificados corretamente pela árvore). A ideia principal é que a mais simples de soluções é a preferida.

2.4.2. Avaliação dos modelos

Para avaliar o desempenho da previsão dos classificadores, isto é, estimar o erro de generalização do classificador de árvore de decisão. (IVANCIUC, 2008) demonstra que o MCC pode ser determinado partir de quatro diferentes medidas. A sensibilidade (também chamada *Recall*, ou seja, a fração de exemplos positivos que são previsto corretamente) é dado por $TP / (TP + FN)$, onde TP é o número de verdadeiros positivos e FN é o número de falsos negativos ou sub-previsões. A especificidade (também chamada precisão, ou seja, a fração de exemplos negativos previstos corretamente) é fornecida por $TN / (TN + FP)$, onde TN é o número verdadeiro negativos e FP é o número de falsos positivos ou previsões exageradas. A precisão geral da previsão é dado por $TP + TN / (TP + TN + FP + FN)$ e, finalmente, o Coeficiente de Correlação de Matthews (MCC) é definido por Matthews (1975) como:

$$MCC = \frac{TP \times TN - FP \times FN}{[(TP + FP) * (FN + TN) * (FP + TN) * (TP + FN)]^{\frac{1}{2}}} \quad \text{Eq. 05}$$

De acordo com (BOUGHORBEL; JARRAY; EL-ANBARI, 2017), a métrica MCC é relevante para avaliar modelos de classificação devido aos seguintes fatores:

- O MCC pode ser calculado usando a matriz de confusão.
- O cálculo da métrica MCC utiliza as quatro grandezas (TP, TN, FP e FN), o que dá um melhor resumo do desempenho dos algoritmos de classificação.

- O MCC não é definido se qualquer uma das quantidades $TP + FN$, $TP + FP$, $TN + FP$ ou $TN + FN$ for zero.

O MCC foi escolhido como a métrica eletiva na iniciativa liderada pelo FDA dos EUA, MAQC-II, que visa alcançar um consenso sobre as melhores práticas para o desenvolvimento e validação de modelos preditivos para medicina personalizada (MAQC CONSORTIUM, 2010). Com base na confiabilidade deste estudo, a medida que será adotada como referência para avaliação de desempenho será também a métrica (MCC). O MCC retorna um intervalo de -1 a 1, onde -1 indica um classificador binário completamente errado ou aleatório, enquanto 1 indica um classificador binário completamente correto (MATTHEWS, 1975).

CAPÍTULO 3 - METODOLOGIA

A metodologia proposta nesta dissertação está dividida em cinco etapas, tais como a seguir:

- Etapa I – Revisão Bibliográfica: na primeira etapa uma revisão bibliográfica foi realizada sobre as estruturas de dutos flexíveis e suas propriedades, análise de imagens digitais e também sobre técnicas de descoberta de conhecimento e mineração de dados. A revisão bibliográfica pôde ser previamente observada na seção CAPÍTULO 2 - desta dissertação. Através da revisão realizada, os principais autores identificados com publicações de conteúdo pertinente às propostas deste trabalho foram identificados e avaliados. Os principais tópicos pesquisados foram:
 - (i) Dutos Flexíveis: Uma introdução sobre o equipamento submarino foi realizada expondo as camadas que compõem a estrutura tubular flexível e identificando a camada de barreira de pressão que é foco deste trabalho.
 - (ii) Branqueamento do PVDF: Uma revisão dos principais trabalhos publicados sobre o branqueamento foi realizada, apontando as análises e resultados realizados nos últimos anos.
 - (iii) Mineração de dados: Uma revisão sobre mineração de dados, com ênfase para as etapas do KDD e extração de conhecimento em banco de dados foi realizada, direcionando para as técnicas e algoritmos relevantes que serão explorados durante este estudo.
- Etapa II – Análise do Problema: nesta etapa, o problema foco do estudo proposto é descrito como um desvio de qualidade do produto que é identificado como a ocorrência de manchas branqueadas (ou simplesmente branqueamento) no material polimérico PVDF, observado durante o processo fabril. Nesta etapa as condições que podem contribuir para a ocorrência do desvio são discutidas, assim como é apresentado um estudo mais elaborado sobre as propriedades mecânicas do material polimérico especificamente para aplicação em linhas flexíveis;
- Etapa III – Criação do banco de dados, extração de dados e definição de atributos: apresenta um estudo com base no conjunto de imagens disponível para este trabalho que foram catalogadas durante diversas fabricações comerciais. As imagens cedidas para este estudo foram armazenadas e as principais características das manchas branqueadas

foram exploradas. Uma análise das características mais relevantes das manchas é apresentada e discutida;

- Etapa IV – Aplicação do método proposto: aborda uma técnica específica da descoberta de conhecimento em bancos de dados conhecida como KDD. Através desta técnica uma sequência de ações é estabelecida com intuito viabilizar classificações de novas imagens com base em resultados previamente conhecidos que possuem potencial de gerar critérios de classificação;
- Etapa V – Análise dos resultados: apresenta uma discussão sobre os resultados gerados, sobre a revisão do modelo e sobre a viabilidade de implementar a técnica em uma situação real. Esta discussão é apresentada no CAPÍTULO 4 - e embasa a conclusão que encerra o estudo, onde espera-se avaliar a confiabilidade do potencial classificatório do modelo elaborado.

3.1. ETAPA II - ANÁLISE DO PROBLEMA

Para Laiarinandrasana et al. (2009c) o branqueamento do PVDF é um fenômeno óptico ocasionado devido à formação de microvazios no material após deformação. Testes de fadiga e tração em regime controlado indicam que os microvazios na região de branqueamento podem esgotar o material de forma a aumentar ou causar crescimento de trincas e que seu aparecimento se dá entre deformações na ordem de 3,5% a 4%.

A Figura 11 apresenta a micrografia do PVDF em diferentes graus de deformação onde pode ser observado crescimento dos microvazios no material em questão, identificados por crateras ou “furos negros”.

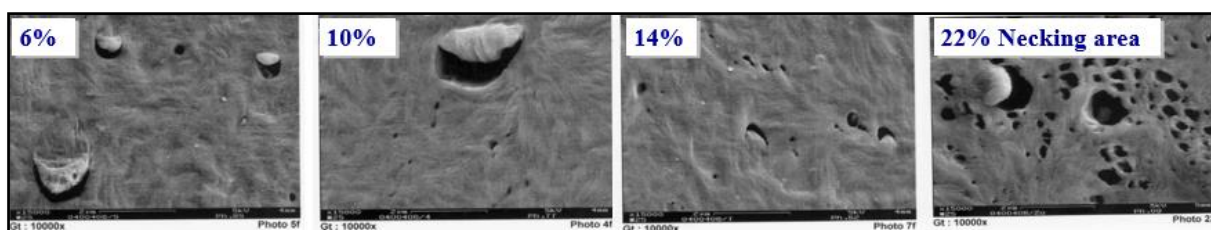


Figura 11 – Micro vazios à partir de microscopia da camada de barreira extrudada em PVDF com diferentes graus de deformação.

Fonte: (NOVFlexibles,, 2008)

Para efeitos ilustrativos, uma imagem de revestimento interno com efeito típico do branqueamento e apresentado na Figura 12.



Figura 12 – Branqueamento do revestimento interno de uma linha flexível.
Fonte: NOVFlexibles, (2008)

O alto contraste da coloração branca no polímero, ou seja, grande variação de opacidade na região onde houve estiramento de material polimérico pode indicar grande incidência de microvazios ou cavitações, gerando um ponto frágil da região, tornando-a susceptível a falhas durante dobramento do duto ou por fadiga durante operação.

Para entender o fenômeno deste contraste de coloração branca mais profundamente foram realizadas algumas dissecações em dutos já fabricados e também produções simulando condições propensas ao aparecimento de branqueamento no PVDF onde algumas possíveis causas raízes foram apontadas.

As principais causas raízes para o aparecimento de branqueamento em revestimentos extrudados com PVDF durante o processo fabril foram identificados como sendo de diferentes naturezas. Esta seção abordará algumas possíveis causas raízes previamente apontadas pelos *stakeholders* de uma empresa do setor de dutos flexíveis.

Os carretéis de produção têm um papel fundamental para evitar deformações excessivas durante a produção do *liner*. Cada projeto de duto flexível possui uma especificação de raio mínimo de flexão (MBR) definido do projeto estrutural da linha e que deve ser respeitado durante a produção. O núcleo do carretel deve corresponder à especificação sendo igual ou superior ao solicitado pelo raio mínimo de flexão da camada de estanqueidade. Na Figura 13 é possível observar que o branqueamento começa a aparecer no ponto de dobra da linha enquanto

é bobinada para o carretel de produção e a Figura 14 apresenta um efeito generalizado de branqueamento ao longo de todo o comprimento em contato com o núcleo do carretel.



Figura 13 – Ocorrência de branqueamento durante o enrolamento do duto no carretel.
Fonte: Imagem cedida por NOVFlexibles



Figura 14 – Aparecimento de branqueamento ao longo de todo o comprimento do duto na camada mais próxima ao núcleo do carretel.
Fonte: Imagem cedida por NOVFlexibles

A carcaça é a primeira camada de um duto flexível e também a camada anterior à extrusão da camada de estanqueidade. O material extrudado pode ingressar na carcaça através

do efeito chamado fluência entre espaçamentos e quando o material se conforma nestas lacunas, uma movimentação desta região pode forçar o material extrudado induzindo-o a uma estricção localizada onde é potencial região de aparecimento de branqueamento. Como pode ser observado pela Figura 15, uma variação de passo irregular é detectada. Esta variação pode ser dar por vários motivos, entre as quais podem ser aperto excessivo dos enrolamentos da carcaça, variação de parâmetros do processo, falhas no equipamento de tracionamento, etc.



Figura 15 – Variação de passo em uma carcaça de duto flexível
Fonte: Imagem cedida por NOVFlexibles

A região da carcaça que apresentou maior variação de passo com abertura próximo da máxima possível, resultou em um branqueamento excessivo com posterior falha da camada ainda durante o enrolamento do duto no carretel de manufatura. Grandes comprimentos de duto também podem ser impactados por branqueamento generalizado, como é o caso observado na Figura 16. Em casos como este não é recomendável a aceitação do produto, uma vez que se entende que a qualidade da produção já está comprometida como um todo.



Figura 16 – Branqueamento formado em regiões onde houve variação no passo da carcaça.
Fonte: Imagem cedida por NOVFlexibles

A espessura do material tem papel fundamental no cálculo de resistência da estrutura. Uma camada mais fina pode permitir maior flexibilidade à camada, além de reduzir o custo do produto, porém oferecerá menor resistência aos esforços solicitados pelo projeto do duto e aos carregamentos sofridos pelos processos da produção.

Durante o processo de extrusão da camada de estanqueidade a espessura da camada deve ser controlada a fim de não variar para uma gama fora das tolerâncias de fabricação especificadas com base nos cálculos de projeto. É apresentado na Figura 17 uma ocorrência de branqueamento na região onde houve uma variação de espessura do material polimérico extrudado.

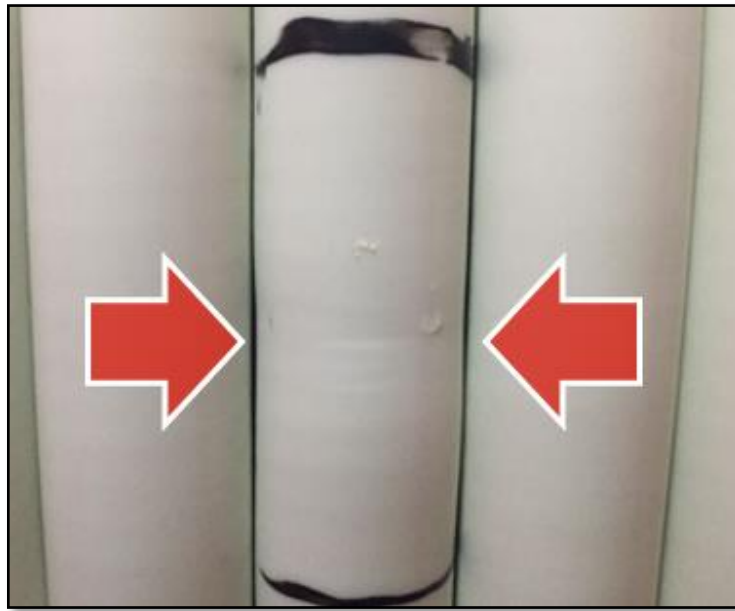


Figura 17 – Branqueamento localizado onde houve redução de espessura do material
Fonte: Imagem cedida por NOVFlexibles

Como outras possíveis causas que possam gerar ou contribuir para o aparecimento de branqueamento são apontadas emendas de fitas da camada intermediária entra a carcaça e a camada de estanqueidade durante a troca de bobinas, problemas com arrefecimento da camada durante conformação, etc.

O PVDF possui uma característica translúcida e também opaca, como pode ser observado na Figura 18. Tal característica será explorada ao longo do desenvolvimento deste estudo com intuito de definir o limiar da aceitação da variação na opacidade do material através do contraste da coloração branca.

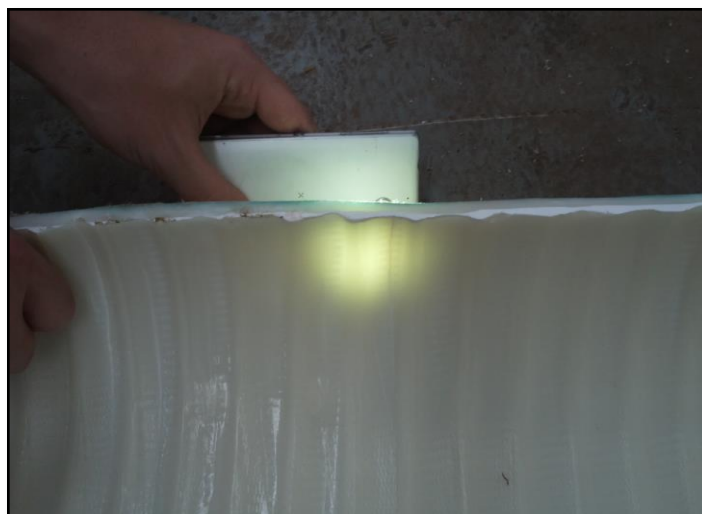


Figura 18 - Característica opaca e translúcida do PVDF
Fonte: Imagem cedida por NOVFlexibles

É propósito deste trabalho definir um parâmetro comparativo com base nas imagens registradas antes de ensaios destrutivos, para posterior avaliação a partir dos resultados dos ensaios. Os resultados dos ensaios podem ser associados a padrões visuais aceitáveis ou não, uma vez que apresentem variação suficientemente perceptível para atribuir um critério com base nestes padrões.

O efeito de branqueamento em *liners* de PVDF aparece basicamente através da circunferência do duto. Entretanto, ao analisar mais profundamente, pode ser observado que o material pode sofrer uma estrição expandindo a região de branqueamento no sentido longitudinal ao duto.

Branqueamento concentrado em seções menores que 90° podem potencialmente ser mais susceptíveis a falha devido a maior estiramento local, embora estudos ainda não tenham chegado a uma conclusão que efetivamente comprove tal afirmativa. Da mesma forma, o comprimento de circunferência afetado em comprimento de arco maiores (próximas à de 180°) não necessariamente tem impacto mais severo na propriedade estrutural da camada, pois um maior arco de propagação apresenta menor concentração de micro vazios comparado com ocorrências de cerca de 90° . A Figura 19 apresenta um comparativo de ocorrências em 180° e 90° ao longo da circunferência do duto em regiões muito próximas de uma mesma produção.

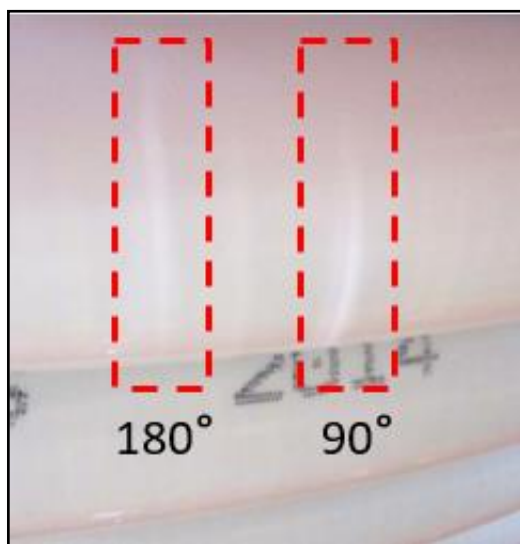


Figura 19 – Ocorrências de branqueamento ao longo de diferentes comprimentos na circunferência do duto.
Fonte: Imagem cedida por NOVFlexibles

Ainda observando a Figura 19, é possível distinguir a intensidade do efeito do branqueamento nas duas ocorrências destacadas em pontilhado vermelho. Esta diferença na tonalidade de branqueamento será abordada no decorrer do estudo proposto neste trabalho,

como principal recurso para avaliar a qualidade do produto em relação aos critérios de aceitação.

Além da análise de diferença entre tons de coloração que toma como base a variação de opacidade do material, também se faz necessário uma abordagem que relacione a geometria e extensão da região afetada. É possível observar diferentes geometrias, assim como diferentes comprimentos de arco nas ocorrências de branqueamento que são observadas em produções comerciais. A Figura 20 ilustra possíveis formas que o branqueamento pode assumir tanto em relação ao comprimento de arco no sentido radial do duto quanto em relação a propagação axial.

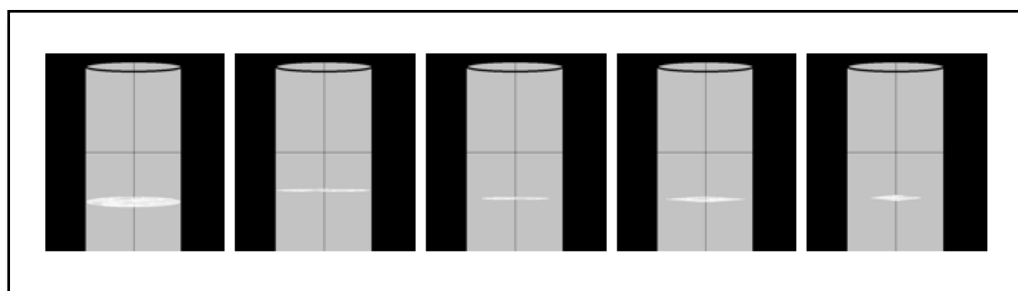


Figura 20 – Possíveis padrões geométricos para branqueamento no PVDF
Fonte: Elaboração própria (2018)

Através de uma análise da imagem, será realizada uma extração das características geométricas de cada mancha típica do efeito de branqueamento encontrada no banco de imagens disponibilizado pela empresa NOVFlexibles. Características a serem estimadas digitalmente são basicamente largura [mm], altura [mm], perímetro [mm], área [mm²].

A geometria da mancha branqueada será estimada com base em um valor de referência conhecido que será utilizado para calibrar cada imagem individualmente. Além dos parâmetros geométricos, outros parâmetros relevantes a serem extraídos das imagens também serão abordados mais adiante.

3.1.1. Avaliação mecânica do problema

A camada polimérica de PVDF avaliada neste estudo considera os efeitos de deformação pré-carregados no material após o processo de extrusão da camada em produção. Para avaliar tal efeito é considerado os princípios da lei de Hooke, considerando o método da anisotropia elástica induzida por tensões (RYCHLEWSKI, 1984).

A partir do estudo detalhado em NOVFlexibles (2014a), a deformação residual se dá com base nas forças de retração que atuam no revestimento interno quando o mesmo é resfriado a partir da temperatura de solidificação.

A norma API-17J (2009) reconhece que há um limite para a deformação do polímero antes que ele altere suas propriedades mecânicas originais e, portanto, se torne suscetível a ruptura. Os critérios de projeto definidos para as camadas de polímero usadas como revestimento interno (barreira de pressão) para PVDF são, deformação máxima de 7% para dutos estáticos ou de condição de armazenamento e de 3,5% para dutos com operação dinâmica.

Devido à observação do aparecimento de branqueamento em dutos fletidos durante o processo de fabricação, uma deformação local é considerada. A Tabela 4 apresenta o resultado de deformação e utilização e resistência à tração calculados em diferentes temperaturas conforme elaborado por NOVFlexibles (2014b). Para efeitos de consideração sobre as condições de climáticas no Brasil, onde foram executados os processos de manufatura avaliados, será tomado como referência a temperatura na faixa de 23°C para as análises posteriores. Pode-se observar que a deformação percentual calculada devido à conformação do material para a camada de revestimento interno já é em torno de 2,3%.

Tabela 4 – Comportamento mecânico do PVDF em diferentes temperaturas¹
Fonte: (NOVFlexibles, 2014b)

Temperatura [°C]	Deformação [%]	Utilização [%]	Tensão (Módulo de Young) [MPa]
150	0	0	0
130	0,4	0,1	1
23	2,3	0,7	26
0	3,7	5,1	56
-20	3,0	19	62
-40	3,4	28	136

Para efeitos de análise apenas, a deformação local será somente considerada onde tiver sido detectado aparecimento de branqueamentos. Um limite médio de deformação local de 5,25% é estabelecido para qualquer aplicação de dutos, seguindo as diretrizes advindas da norma API-17J (2009) para o material PVDF com respeito à projetos de camada de barreira em linhas flexíveis. A base para esta premissa tem sido principalmente a detecção visual de ocorrência do branqueamento após a flexão dos dutos nos carretéis de manufatura, juntamente com a preocupação das consequências que o branqueamento pode causar, como principalmente a propagação de trincas.

¹ As propriedades físicas e mecânicas do material PVDF estão disponíveis no Apêndice A.

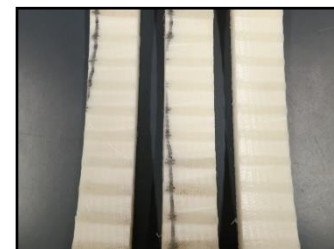
Duas amostras de camadas de barreira extrudadas em PVDF foram selecionadas para serem investigadas neste trabalho, uma vez que o material já se encontra disponível em NOVFlexibles (2014b). Detalhes das amostras podem ser observados na Figura 21.



(a) Amostra da esquerda (1) foi fletida no carretel, a da direita (2) sem sofrer flexão alguma devido ao dobramento no carretel



(b) As amostras da direita para a esquerda são tensionadas na máquina de tração a 0%, 3,5% e 5,25%, respectivamente, sem deformação inicial no carretel.



(c) As amostras da direita para a esquerda são deformadas na máquina a tração a 0%, 3,5% e 5,25%, respectivamente, pré-deformadas no carretel.

Figura 21 – Amostras recolhidas de camada de barreira extrudada em PVDF
Fonte: NOVFlexibles, (2014b)

A integridade mecânica e o efeito do branqueamento observado nas camadas de barreira de PVDF foram avaliados por meio de análise de tração segundo especificações padrões de amostragem e especificação de teste conforme ISO 37 (2017) aplicado para a situação relevante de tração e deformação a serem avaliadas. Condições do teste são detalhadas na Tabela 5.

Tabela 5 – Especificações do teste de tração

Parâmetro de teste	Descrição
Instrumento de teste	Máquina de tração Instron equipada com célula de carga de 250 kN.
Espécimes de teste	Barras com espessura total da camada de barreira, Tipo I.
Velocidade de teste	1 mm / min. para espessura total e 35 mm / min. para 2 mm de espessura.
Esquema do ensaio	3,5%, 5,25% com tempo de parada de 4 minutos e deformação até a ruptura.

As propriedades mecânicas dessas amostras foram avaliadas de acordo com D20-ASTM COMMITTEE (2014) e ISO 37 (2017) e então comparadas com as especificações de propriedades mecânicas calculadas, conforme já observado na Tabela 4. Os resultados dos ensaios são mostrados nas Figura 22 (amostra 1) e Figura 23 (amostra 2).

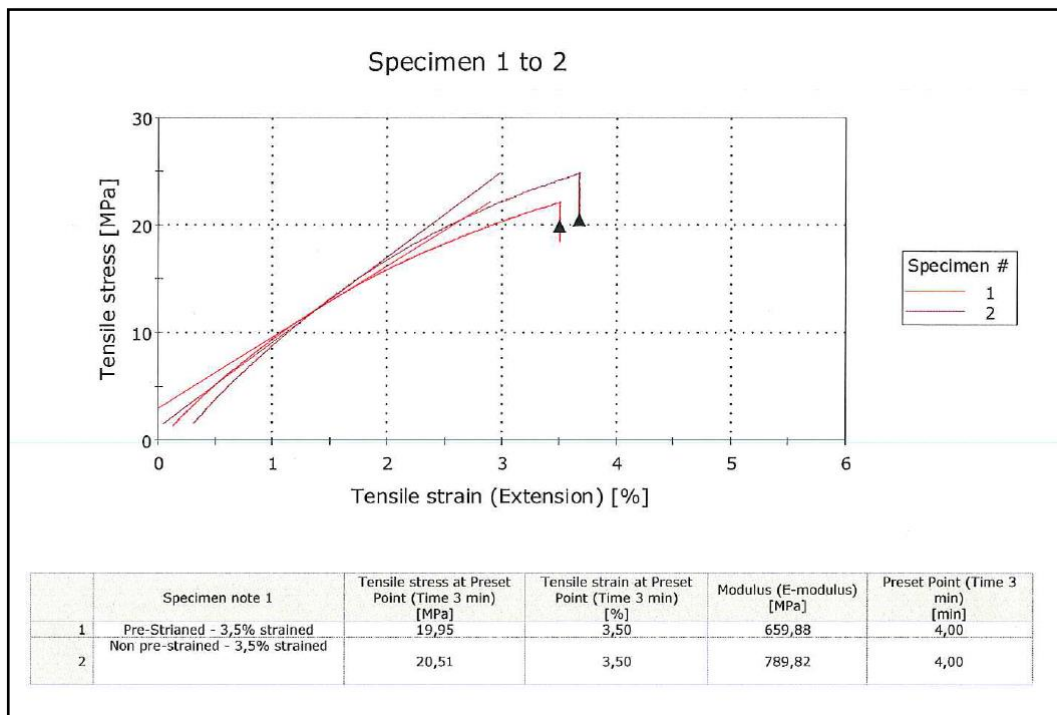


Figura 22 - Resultado de testes para amostras não-deformadas e pré-deformadas (com branqueamento) com espessura total da camada de barreira, que deformou até 3,5% e então mantida por 4 minutos.

Fonte: NOVFlexibles (2014b)

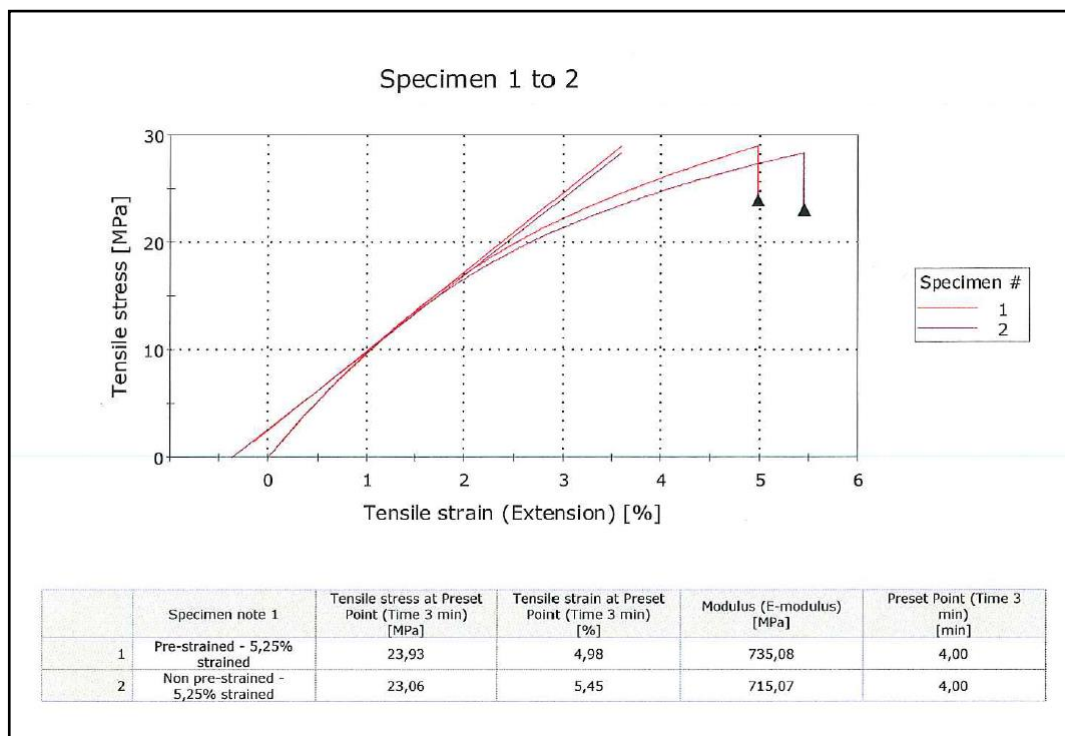


Figura 23 - Resultado de testes para 2 amostras: não-deformadas e pré-deformadas (com branqueamento), com espessura total da camada de barreira, que deformou até 5,25% e então mantida por 4 minutos.

Fonte: NOVFlexibles (2014b)

No gráfico da Figura 22 é possível notar que a amostra pré-deformada em 3,25% demandou esforço de tração relativamente menor do que para a amostra 2 para um mesmo valor alvo de deformação sob tração, ou seja, a amostra 1 demandou menor esforço para atingir a mesma deformação, indicando uma sutil fragilização do material neste ponto.

Já no gráfico da Figura 23 observa-se o oposto. A amostra pré-deformada em 5,25% sofre maior carregamento em tração e deforma menos quando comparado à amostra 2. Isso é uma indicação de que o polímero não sofre dano físico significativo ao ser sobrecarregado no limiar do branqueamento, demonstrando ser de fato um valor bastante tolerável.

Adicionalmente, para identificar em que nível de tensão o branqueamento começa a desenvolver trincas, algumas das amostras de polímero com espessura total também foram forçadas até a ruptura. O resultado pode ser observado na Figura 24.

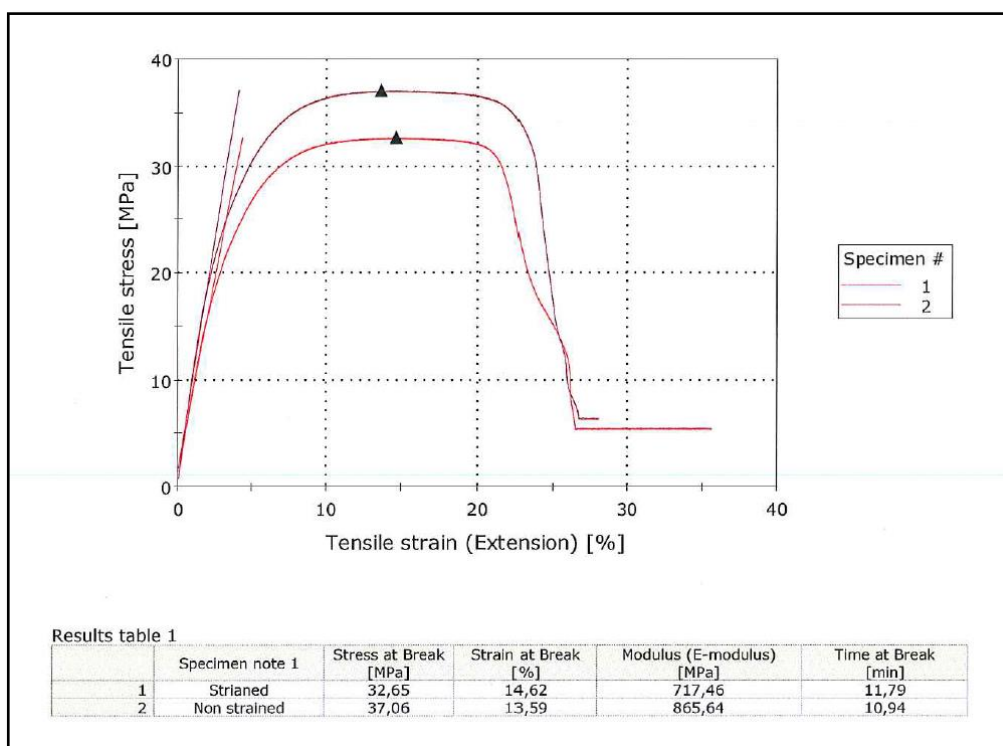


Figura 24 - Resultado de testes para amostras não-deformadas e pré-deformadas (com branqueamento) com espessura total da camada de barreira, que deformou até a ruptura.

Fonte: NOVFlexibles (2014b)

Para estabelecer um padrão associativo de referência, as características de branqueamento observadas durante a flexão do duto nos carretéis de manufatura são definidas. NOVFlexibles (2014c) avaliou algumas amostras com dados pertinentes a este trabalho. O comprimento de arco afetado pelo branqueamento também é medido para 2 amostras. Para a amostra #01, um valor de 281 mm foi medido como comprimento de arco na região mais

visualmente afetada pelo branqueamento. A Figura 25 apresenta os aspectos visuais da amostra #01 após o corte, na região interna da amostra. Da mesma maneira, um comprimento de arco medido de 163mm para a amostra #02.

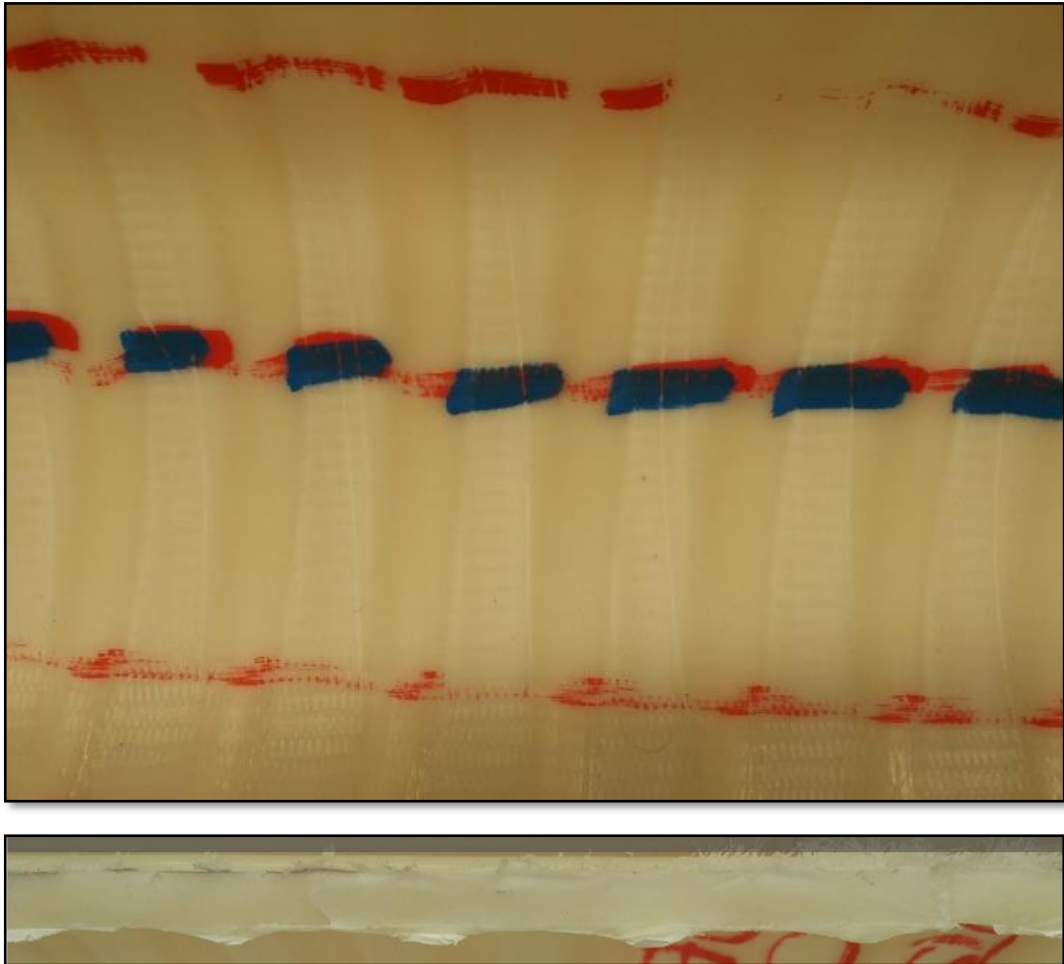


Figura 25 - Características visuais do branqueamento da amostra #01: camada em PVDF manufacturado sobre carcaça (acima); perfil da amostra removida (abaixo).

Fonte: NOVFlexibles (2014c)



Figura 26 - Características visuais do branqueamento da amostra #02: camada em PVDF manufaturado sobre carcaça (acima); perfil da amostra removida (abaixo).

Fonte: NOVFlexibles (2014c)

A Tabela 6 apresenta um quadro comparativo para os valores de deformação medidos para as amostras #01 e #02.

Tabela 6 - Valores de deformação máxima na região mais externa do duto sob flexão no carretel onde se observou aparecimento de branqueamento.

Fonte: Adaptado de NOVFlexibles

Identificação da amostra	#01	#02
Diâmetro externo da camada de estanqueidade [mm]	252,7	250,8
Deformação total [%]	7,81	6,85
Deformação do branqueamento [%]	5,51	4,55
Comprimento de arco [mm]	281	163

As deformações totais para as amostras #01 e #02 são calculadas. A deformação no branqueamento é igual à deformação total subtraída da deformação inicial de 2,3% resultante do processo de extrusão. Pode-se observar que a amostra #02 teve uma deformação ligeiramente

menor quando em comparação com a amostra #01 e o comprimento de arco medido também apresentou uma diferença relevante. Com base no critério de 5,25% apenas, a amostra #02 seria aceitável. Este é um bom ponto de partida para as análises que serão realizadas nas seções seguintes e esta abordagem será considerada como referência para validação do modelo.

3.2. ETAPA III - SELEÇÃO, PRÉ-PROCESSAMENTO E TRANSFORMAÇÃO DAS IMAGENS

Por se tratar de um estudo onde o recurso base a ser analisado são imagens fotográficas, todo o espaço amostral se dará com bases nestas imagens e na preparação das mesmas. Durante a etapa de preparação de imagens foram realizadas várias atividades afim de estabelecer um conjunto de características que possam ser processadas futuramente durante a etapa de mineração de dados. A sequência de atividades principais que foram realizadas durante este estudo são apresentadas a seguir.

Um banco de dados criado a partir de 25 ocorrências gerou um espaço amostral de cerca de 60 imagens de desvios de fabricação que apresentaram manchas brancas no PVDF que foram aprovadas ou reprovadas pelos critérios da empresa. Dentre as imagens disponíveis, 19 são classificadas como aprovadas e 41 como reprovadas.

Após a avaliação detalhada dos especialistas, incluindo decisão final sobre a possibilidade de aceitação do branqueamento na camada de extrusão, as ocorrências foram divididas em dois grupos: aprovados (AP) ou reprovados (RP). O critério de avaliação dos especialistas é com base em critérios internos da NOVFlexibles e, portanto, não será discutido neste estudo.

As imagens originais foram identificadas com variadas resoluções distintas, pois cada observador registrava com aparelhos fotográficos individuais sem uma configuração padronizada. De uma maneira geral a maioria das imagens possuía resolução definida em 1224x1632 pixels, porém um redimensionamento foi realizado de forma a reduzir o poder computacional necessário para processamento de dados e facilitar o entendimento dos resultados. Todas as imagens deste trabalho tiveram redimensionamento definido por padrão como 640x480 pixels.

A calibração da imagem digital com relação a uma escala real é uma etapa essencial para fazer medições mais precisas nas estruturas apresentadas nas imagens. O processo é realizado por meio de uma associação do número de pixels na imagem com medidas reais do mundo físico (p.ex. cm, μm).

Para executar a metodologia proposta, o software FIJI (2017) foi utilizado para executar calibração da imagem digital por ser um programa de acesso aberto e livre, criado para auxiliar pesquisas em ciências biológicas, mas que atende perfeitamente a demanda deste estudo. A Figura 27 ilustra o processo de calibração, onde uma reta é traçada de um lado ao outro do duto, definindo uma distância conhecida conforme referência real que é associada com uma referência em valor de pixels da imagem.

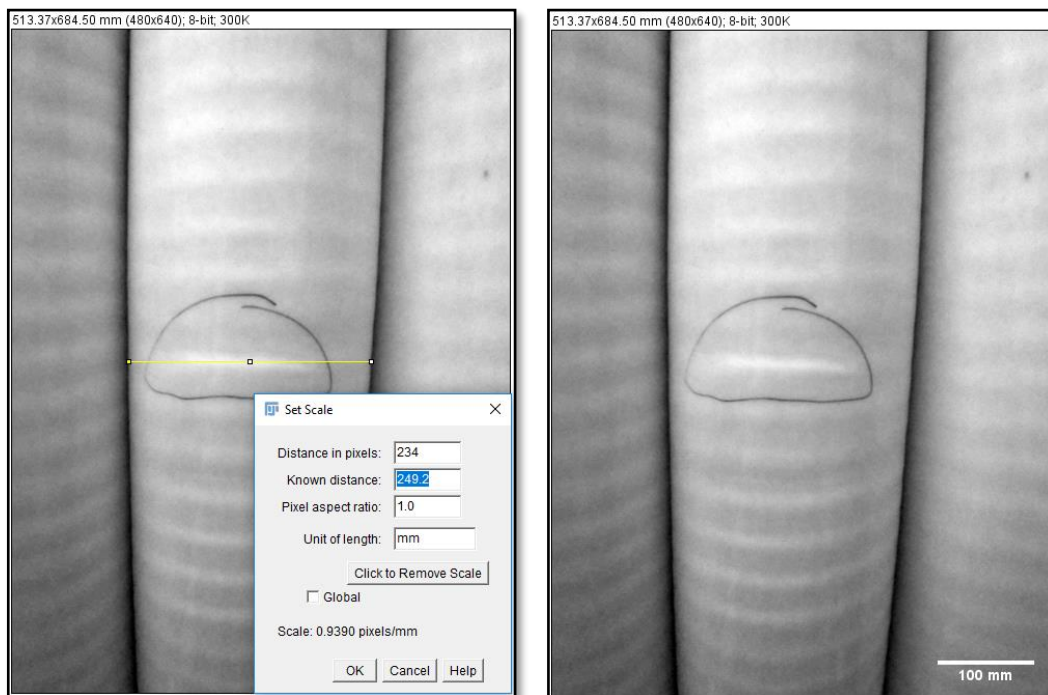


Figura 27 – Exemplo de imagem sendo calibrada para um valor de diâmetro físico conhecido (à esquerda) incluindo apresentação de escala em milímetros (à direita).

Fonte: Imagem cedida por NOVFlexibles, adaptada pelo autor

Todas as imagens foram calibradas definindo como distância horizontal da estrutura sendo igual ao seu diâmetro externo nominal de projeto (considerando carcaça e revestimento interno apenas) e assim, uma relação pixel/mm é estabelecida para cada imagem disponível.

Uma vez que a detecção de um branqueamento não é necessariamente algo objetivo, ou seja, influência da luz externa, alterações de pigmentos do material polimérico, manchas, ondulações, sombreados, etc, podem influenciar na identificação do fenômeno, todas as regiões de interesse são advindas de registros de análise de oriundas dos especialistas da NOVFlexibles que constataram que o branqueamento de fato ocorreu. Os contornos das manchas foram igualmente identificados mediante ao entendimento de um especialista em desvios de manufatura da fábrica de flexíveis em questão e, portanto, este não será um tópico abordado no presente estudo.

Durante a etapa de processamento das imagens vários testes foram executados afim de estabelecer as técnicas que trariam resultados mais coerentes com a proposta do trabalho. Todavia, apenas recursos mais relevantes e que foram de fato implementados nas imagens serão apresentados. Alguns dos recursos mais importantes utilizados durante as etapas deste trabalho foram a conversão das imagens para escala de cinza (8-bit) e aplicação de filtros: Suavização (*Smooth*), Melhoria do contraste (*Enhance contrast*), Saturação dos pixels (0,3%, normalizado e com equalização de histograma) e Aguçamento (*Sharpen*).

A etapa mais básica na fase do pré-processamento da imagem é convertê-la em escala de cinza. Este recurso permite avaliar cada pixel em um escalar e desta forma é possível avaliar as diferentes intensidades de tom de branco em cada região de uma imagem monocromática.

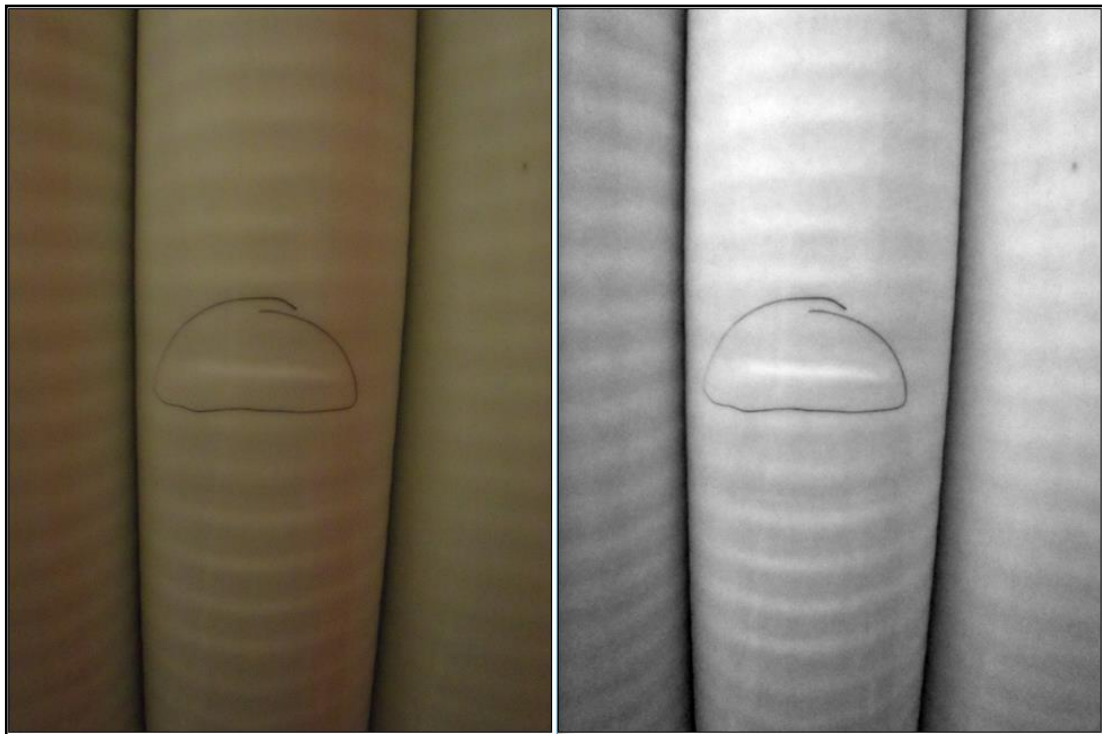


Figura 28 – Imagem original à esquerda e imagem convertida para escala de cinzas 8-bit à direita.
Fonte: Imagem cedida por NOVFlexibles, adaptada pelo autor

Alterações de aspecto e distorções resultantes do redimensionamento serão desconsideradas da análise. A região de interesse, que pode ser identificada pelos tons mais claros em maior contraste, será selecionada individualmente uma a uma para cada imagem do banco de dados para então preparar para extrair dados posteriormente. Como resultado da aplicação das técnicas mencionadas obtém-se resultado conforme apresentado na Figura 29 a seguir.

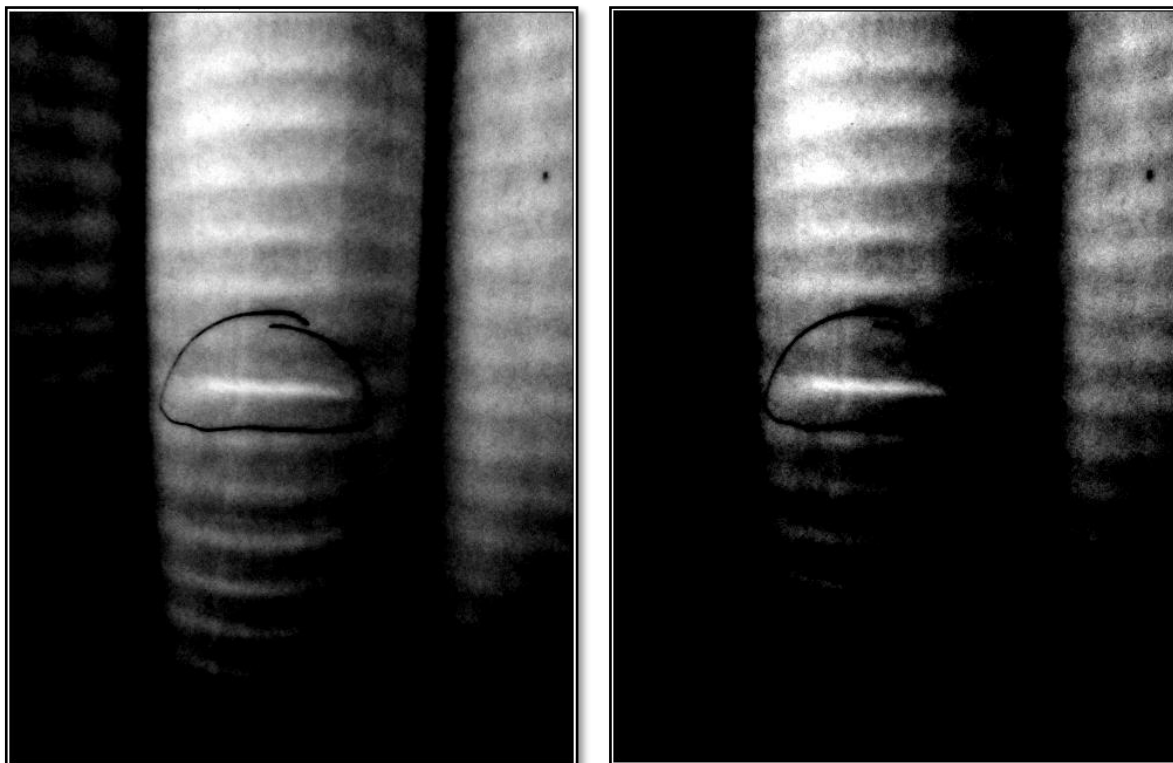


Figura 29 – Exemplo do arquivo resultante da aplicação de filtros de imagens.
Fonte: Imagem cedida por NOVFlexibles, adaptada pelo autor

É possível notar que as características do efeito de branqueamento são realçadas e destacadas na imagem a ponto de simplificar ao máximo o trabalho do agente humano. Desta forma é possível explorar com mais facilidade os parâmetros relacionados com a mancha característica do efeito de branqueamento para a extração dos parâmetros.

As informações contidas na imagem podem ser facilmente extraídas a partir de alguns recursos da ferramenta FIJI. Através ferramenta de seleção livre e “*brush tool*” é possível selecionar a região de interesse e definir os tipos de medições que desejamos obter da imagem na região previamente delimitada (FIJI, 2017). Com a utilização deste recurso é feito um contorno sobre a região branqueada para então executar a extração dos parâmetros buscados. A própria ferramenta disponibiliza um recurso de medição dos parâmetros. Os valores foram extraídos e incluídos no banco de dados para posterior avaliação.

A Figura 30 exemplifica o método de medição de parâmetros na região de interesse das imagens para criação do banco de dados a ser minerado posteriormente.

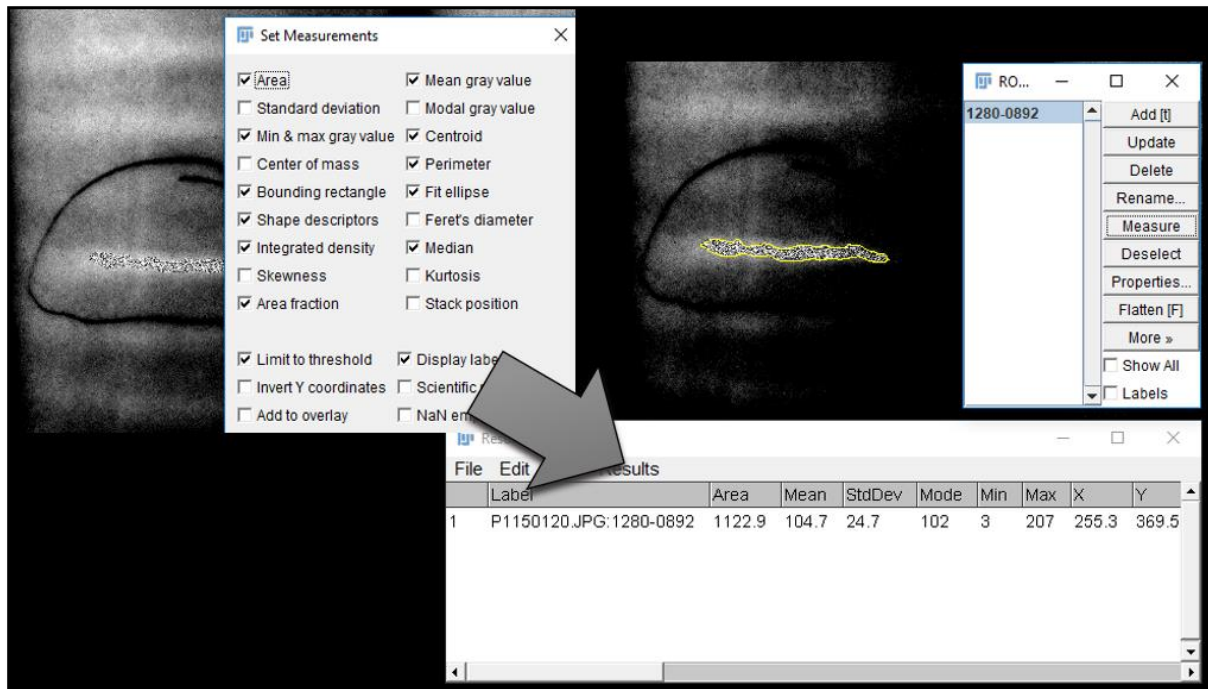


Figura 30 – Extração das informações das imagens através do software FIJI.
 Fonte: Imagem cedida por NOVFlexibles, adaptada pelo autor

3.3. ETAPA IV - APLICAÇÃO DAS TÉCNICAS DO KDD

Através do modelo KDD, pretende-se desenvolver uma estrutura de predição que alcance resultados satisfatórios na aplicação da tarefa de classificação através de um algoritmo já conhecido, não sendo objeto deste estudo, desenvolver novas técnicas ou ferramentas computacionais que auxiliem na classificação dos parâmetros de uma imagem com manchas brancas em uma imagem.

3.3.1. KDD fase I – Seleção, pré-processamento e transformação

Na aplicação do KDD, a primeira fase, possui essencialmente dois objetivos: conhecer o domínio da aplicação e a base de dados e prepará-los para a próxima-fase. Dentre as diversas tarefas que podem ser realizadas nessa fase destacam-se: preparação dos dados, limpeza dos dados, transformação de dados e atributos e a seleção dos atributos mais relevantes (SILVA et al., 2017).

Uma vez que a fonte fornecedora dos dados é também detentora do maior conhecimento sobre o produto a ser estudado, fica relativamente simplificado preencher os requisitos iniciais para execução do KDD. Para dar sequência na etapa de classificação de imagens, uma série de características das imagens foram pré-estabelecidas como relevantes para análise, porém um estudo mais detalhado sobre a efetiva contribuição de cada parâmetro para predições confiáveis

será discutido mais adiante. Abaixo é possível observar quais os parâmetros que estão sendo genericamente extraídos para cada mancha branqueada presente nas imagens isoladamente:

- Distância vertical entre cada mancha (DV);
- Altura (H) e largura (W) de cada “mancha”;
- Área da mancha (A);
- Perímetro da mancha (PE);
- Máximo (MA) e mínimo (MI) nível de tom de cinza na mancha;
- Média da intensidade de cinza na região de interesse (AV).
- Número de pixels por mm² (PI).
- Quantidade de pixels (QP) na região da mancha
- Definição do formato de borda da mancha como irregular (IR) ou regular (RE).

As características mecânicas extraídas fisicamente do produto fabricado são listadas a seguir:

- Diâmetro interno da linha (DI);
- Passo da carcaça (PC);
- Diâmetro externo real de cada tubo (OD);
- Houve redução de espessura considerável sobre na região afetada (TR)?
- Quantidade de “manchas” (N).

Uma das formas mais simples e utilizada para a representação de dados é realizada por meio de atributos e seus respectivos valores e é denominada de atributo-valor. Esse formato consiste na descrição de exemplos por atributos, de modo que cada exemplo equivale a uma linha e cada atributo a uma coluna, como ilustrado na Tabela 7.

Tabela 7 – Formato do banco de dados para registro dos atributos

EA	DI [pol]	OD [mm]	PC [mm]	N	DV	H	W	A	PE	MA	MI	Pixel/ mm	QP	IR/RE	Class.
D001															
D002															
...															
D00n															

Os atributos extraídos das imagens podem assumir valores quantitativos e qualitativos e as operações possíveis sobre eles incluem transformação de atributos (TA) e seleção de atributos (AS). A SA abrange as tarefas de construção, extração e discretização de atributos,

enquanto na MD o objetivo é determinar um subconjunto de atributos ou um conjunto ordenado de atributos, ambos, segundo algum critério ou medida de importância (PARMEZAN et al., 2012).

Juntamente a um especialista da área foi realizado um refinamento na lista de parâmetros previamente estabelecidos, de forma a evitar incluir no estudo, características que apresentam correlação de aspectos que possuem interdependência e potencialmente podem gerar uma falsa relação que leve os resultados a serem tendenciosos ou falsos. Com isso, foram excluídos das análises por não se mostrarem relevantes: espaço amostral, quantidade de manchas, distância entre branqueamentos, definição do formato de borda. Através da Figura 31 é possível identificar alguns dos atributos selecionados e que serão avaliados neste trabalho.

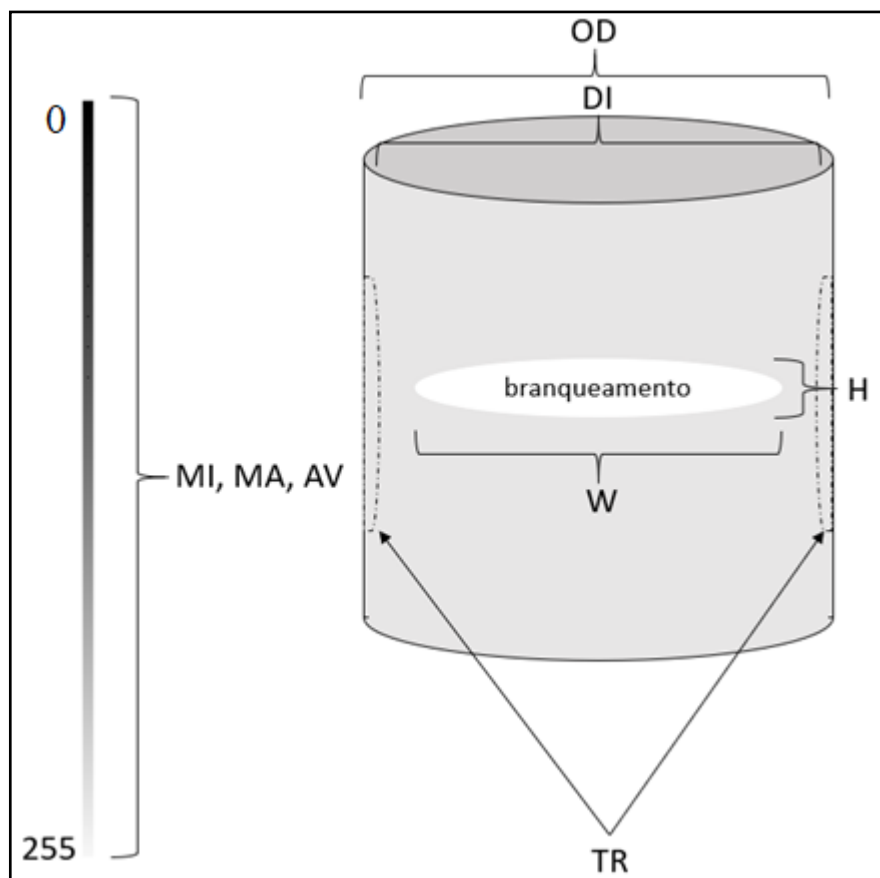


Figura 31 – Identificação de alguns atributos.
Fonte: Elaboração própria (2019)

Durante o tratamento dos dados foi necessário um ajuste para um formato padrão a ser lido e interpretado pelo *software* WEKA. As informações após uma fase de pré-processamento são armazenadas em um único arquivo do tipo texto dentro do formato *.arff* (*Attribute Relation File Format*), que é um padrão de leitura aceito por diversas ferramentas de mineração de dados,

incluindo o WEKA. A Figura 32 apresenta a estrutura do conjunto de dados utilizada neste estudo adequada para utilização no WEKA.

```
@Relation Whitenig_classifier
@attribute DI real
@attribute OD real
@attribute PC real
@attribute TR {Sim,Não}
@attribute H real
@attribute W real
@attribute A real
@attribute PE real
@attribute AV real
@attribute MA real
@attribute MI real
@attribute PI real
@attribute QP real
@attribute Class {Aprovado,Reprovado}

@data
9,262.8,22.93,Não,11.8,203.5,2139.2,418,230.1,138,109,11.59,287354.67,Reprovado
9,262.8,22.93,Não,8,154.2,1403.8,368,224.5,130,112,11.59,188569.79,Reprovado
9,262.8,22.93,Não,10.4,156.1,1589.4,325.3,237.8,140,108,11.59,213501.08,Reprovado
9,262.8,22.93,Não,19.1,76.5,1089.4,162.2,115.2,148,79,5.0686,27987.46,Reprovado
9,262.8,22.93,Não,9.6,144.8,1052,302.4,145.2,179,115,3.74,14714.96,Reprovado
8,230.8,16.38,Não,21.3,168,2284.3,360.3,166.8,190,137,2.2622,11690.02,Aprovado
8,230.8,16.38,Não,10.7,190.9,1322.2,402.4,186.5,216,154,2.4237,7767.03,Aprovado
8,230.8,16.38,Não,9.9,222.1,2129.8,405.3,183.1,224,155,2.4237,12511.13,Aprovado
8,230.8,16.38,Não,16.5,237.5,2209.1,508.2,175.3,197,138,2.4237,12976.96,Aprovado
6,190.4,28.29,Não,10.7,75.4,413.5,148.7,130.6,171,83,9.9027,40549.24,Reprovado
8,249.2,36.14,Não,11.1,110.8,740.5,245.1,159.9,173,145,2.1672,3477.95,Aprovado
8,249.2,36.16,Sim,6.5,187.9,1012.5,399.6,162.2,174,136,1.0032,1018.99,Reprovado
8,249.2,36.16,Sim,14.2,205,2220.7,486.9,154.5,178,82,1.0032,2234.94,Reprovado
8,249.2,36.16,Sim,13.7,204.4,1398,423,150.1,174,68,1.0032,1406.96,Reprovado
8,249.2,36.16,Sim,11.7,123.8,1308.2,264,103.4,118,84,3.7806,18698.02,Reprovado
8,249.2,36.16,Sim,21.8,171.1,3057.7,382.2,135.1,150,110,2.6966,22234.53,Reprovado
8,249.2,36.16,Sim,17.9,149.4,1905.7,341.6,135.8,150,109,2.6966,13857.59,Reprovado
8,249.2,36.16,Sim,14.9,133.7,1424.2,282.5,135.7,140,116,2.6966,10257.01,Reprovado
```

Figura 32 – Estrutura do conjunto de dados do arquivo .arff utilizado no trabalho.
Fonte: Elaboração própria (2019)

Uma vez que o propósito deste trabalho não é avaliar com profundidade os métodos de ranqueamento de atributos, esta não será uma área muito abordada, porém uma breve análise sobre os resultados dos métodos disponibilizados pelo software WEKA será apresentada mais adiante na seção de resultados. Os seguintes avaliadores disponíveis no WEKA foram utilizados: avaliação da correlação de atributos (*Correlate Attribute - CFS*), avaliador de Subconjuntos (*CSE*), avaliador de recursos de taxa de ganho (*Gain Ratio*), filtro de classificação de ganho de informações (*Info gain*), avaliador de recursos OneR (*OneR attribute*), filtro de classificação do relevo F (*Refief attribute*) e Filtro de classificação de incerteza simétrica (*Symmetrical Uncert*).

3.3.2. KDD Fase II – Mineração de dados

A segunda fase do KDD, extração de padrões, tem como objetivo a construção de modelos que possam representar o conhecimento embutido nos dados. Nesta fase de modelagem as técnicas (algoritmos) de seleção de atributos e mineração serão aplicadas.

As características consideradas importantes, após extraídas, são organizadas e transformadas em formato padrão (.arff) para serem interpretadas pelo software WEKA (WITTEN et al., 2016). Ao todo foram definidos 14 atributos com 60 instâncias iniciais que correspondem as extrações dos parâmetros de cada mancha de branqueamento disponível no BD. Os atributos efetivamente relevantes serão avaliados através dos algoritmos disponíveis pelo próprio software WEKA. A classificação para este espaço amostral já é conhecida para cada uma das imagens avaliadas, onde 19 instâncias são classificadas como aprovadas e 41 como reprovadas.

Algoritmos de MD, em especial os classificadores, são de caráter preditivo e nesses casos, desenvolvem aprendizado com uma rotina de treinamento para atingir o objetivo de classificação. Por essa característica peculiar, o conjunto final de dados foi subdividido em um conjunto de treinamento e um conjunto de testes, como será apresentado.

Ao todo, cinco algoritmos de classificação disponibilizados pelo software WEKA foram explorados durante a etapa de mineração de dados. Todos esses algoritmos, à exceção do CSE, foram executados em configuração de parâmetros com valores padrão do próprio WEKA seguindo dez rodadas de testes cada. Os algoritmos analisados são: Naive Bayes, Regressão Logística (*Logistic regression*), Árvore de Decisão (*Decision trees*) e Rede neural (Multilayer Perceptron).

Uma representação esquemática do processo de descoberta do conhecimento que será aplicado neste estudo é apresentado na Figura 33.

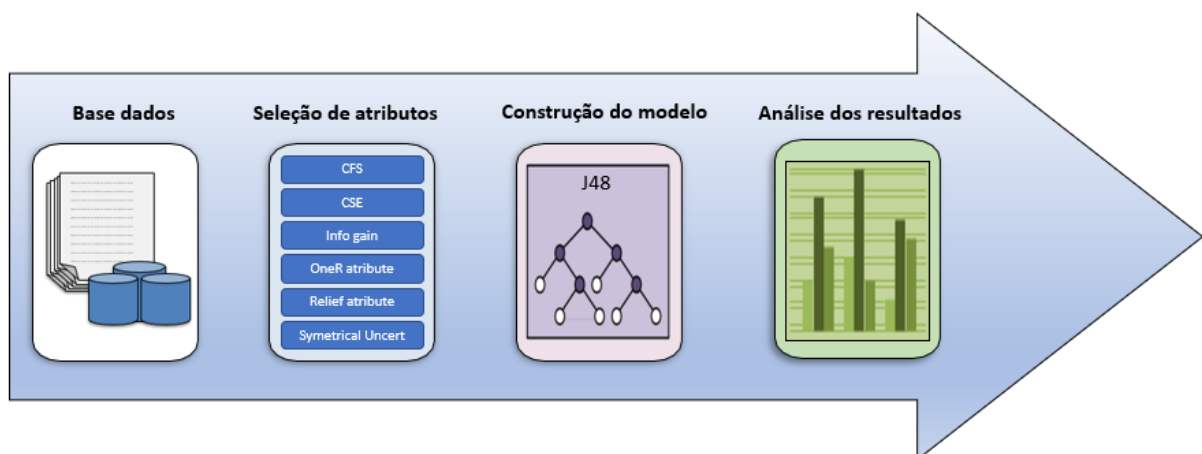


Figura 33 – Configuração do modelo de experimento
Fonte: Adaptado de Parmezan et al. (2012)

Na última fase, pós-processamento de conhecimento, o objetivo é avaliar, validar e consolidar o conhecimento extraído. A avaliação é realizada, por exemplo, com a interpretação dos resultados por meio de visualização dos padrões extraídos, remoção de padrões irrelevantes

ou redundantes e tradução de padrões úteis para formas compreensíveis para os usuários. Durante essa fase, os resultados devem ser avaliados para garantir que estes sejam estatisticamente significativos e confiáveis. Os resultados são coletados para medidas de desempenho diferentes, como estatística Kappa e MCC (Coeficiente de Correlação de Matthews).

3.3.3. KDD fase III - Avaliação do desempenho

O Ambiente Experimental do WEKA permite realizar testes estatísticos sobre as diferentes medidas de desempenho para uma melhor análise do experimento visando responder às seguintes perguntas sobre os experimentos:

- Qual algoritmo teve o melhor desempenho?
- Qual é a classificação dos algoritmos por desempenho?

O desempenho preditivo com que o classificador prediz a classe dos exemplos está relacionado a sua representação interna utilizando os atributos. Se os atributos não são capazes de representar o conhecimento implícito nos exemplos, o desempenho preditivo é deteriorado, ou seja, haverá uma grande chance da classe de novos exemplos submetidos ao classificador ser predita incorretamente (PARMEZAN et al., 2012).

CAPÍTULO 4 - RESULTADOS E DISCUSSÕES

Uma primeira análise é gerada com base nas configurações de modelo padrão disponível pelo WEKA. Para cada algoritmo uma proporção de 70% de treinamento (*training set*) e 30% de teste (*test set*) foi definido para gerar os resultados. Os dados de treinamento e de teste são selecionados aleatoriamente e adicionalmente, duas amostras previamente avaliadas serão utilizadas como referência para validar o modelo (*validation set*) ao final de todo o processo.

Um resumo dos resultados de classificação para a condição supramencionada é apresentado na Tabela 8.

Tabela 8 – Avaliação dos algoritmos (70% treinamento, 30% teste).

	Naive Bayes	Regressão Logística	Árvore de Decisão (J48)	Rede neural
Classificações corretas [%]	66,67	83,33	88,89	83,33
Classificações incorretas [%]	33,33	16,67	11,11	16,667
Estatística Kappa	0,365	0,658	0,766	0,658
MCC	0,403	0,663	0,766	0,663

Os resultados inicialmente apresentados, salvo exceção para o método de *Naive Bayes*, que apresentou uma menor capacidade de predição em relação ao resultado dos demais modelos testados, apresentaram grande similaridade. Portanto, foi decidido pelo autor que o algoritmo classificatório de árvore de decisão J48 do WEKA será utilizado para uma análise mais elaborada durante as próximas análises. Além disso, o J48 apresenta o conhecimento extraído de forma mais útil e de interpretação mais simplificada para as análises previstas, que são o propósito deste estudo.

As árvores de decisão representam uma abordagem supervisionada para classificação. A indução da árvore de decisão constrói uma estrutura semelhante a um fluxograma, na qual cada nó interno (não-folha) indica um teste em um atributo, cada ramificação corresponde a um resultado do teste e cada nó externo (folha) indica uma previsão de classe (QUINLAN, 1993). Em cada nó, o algoritmo escolhe o atributo melhor para particionar os dados em classes individuais. Quando a indução de árvore de decisão é usada para seleção de subconjunto de atributos, uma árvore é construída a partir dos dados fornecidos.

Todos os atributos que não aparecem na árvore são considerados irrelevantes. O conjunto de atributos que aparecem na árvore formam o subconjunto reduzido de atributos. Quinlan (1996) popularizou a abordagem da árvore de decisão com sua pesquisa realizada há mais de 20 anos, já apresentando uma melhoria com relação ao seu trabalho anterior (QUINLAN, 1993). A mais recente implementação de domínio público do modelo de Quinlan

é o C4.5. Porém, o pacote de classificadores do WEKA possui sua própria versão do C4.5, conhecida como J48.

O algoritmo J48 foi configurado para ajustar as árvores geradoras de dados com critérios de poda (*pruned*) sendo considerado, ou não, e então combinados com a redução, ou não, do erro de poda e a correção MDL. A Tabela 9 apresenta os resultados divididos em diferentes casos de configurações para o algoritmo J48 aplicado para todo o conjunto de atributos.

Tabela 9 – Diferentes condições de configurações do algoritmo J48

Caso	Instâncias por folha	Sem poda (<i>unpruned</i>)?	Redução de erro da poda?	Correção MDL?	MCC	Precisão clas. aprovados	Precisão clas. reprovados
A1	2	Sim	Não	Sim	0,766	0,833	0,833
A2	3	Sim	Não	Sim	0,645	0,857	0,909
A3	2	Sim	Não	Não	0,886	1	0,917
A4	2	Não	Sim	Sim	0,766	0,857	0,909
A5	3	Não	Sim	Sim	0,766	0,857	0,909
A6	2	Não	Sim	Não	0,766	0,857	0,909
A7	2	Não	Não	Sim	0,766	0,857	0,909
A8	3	Não	Não	Sim	0,645	0,833	0,833
A9	2	Não	Não	Não	0,886	1	0,917

Nota-se que as configurações que levaram aos melhores resultados de avaliação pelo indicador MCC são representados pelos casos A3 e A9, onde 17 instâncias foram classificadas corretamente de um total de 18 possíveis. As análises serão conduzidas tomando como referência as configurações dos casos A3 e A9 que apresentarem melhores resultados.

A Figura 34 apresenta a árvore gerada para o caso A3, aplicando uma configuração específica do J48 para todos os atributos do BD.

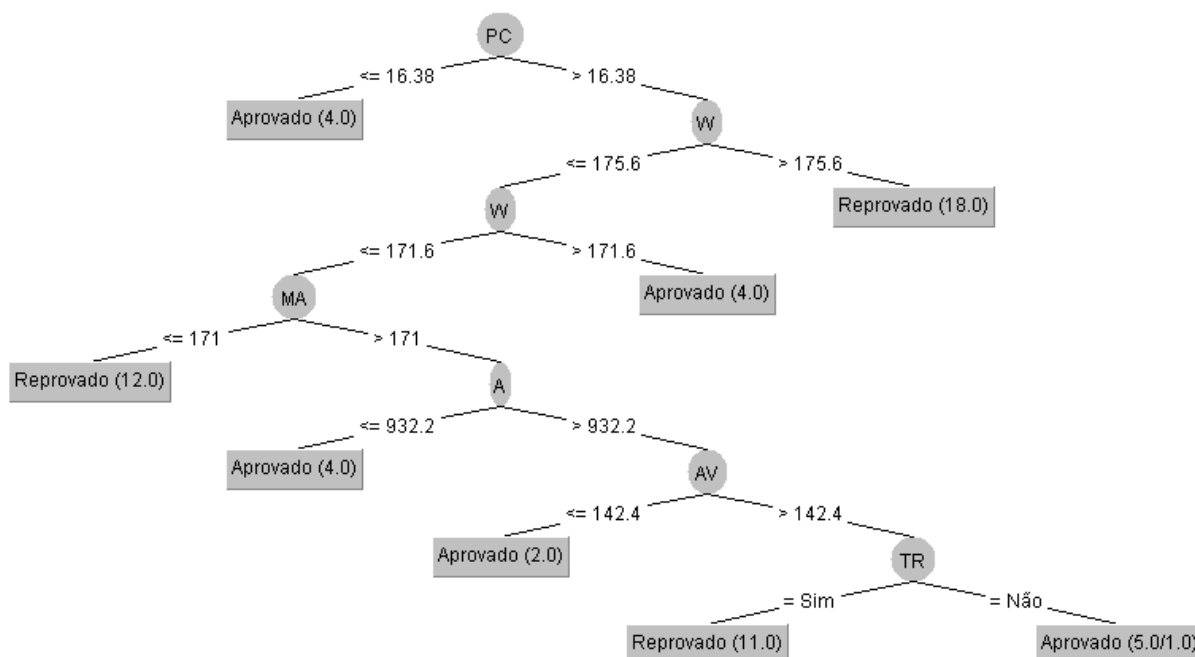


Figura 34 – Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A1, para todos os atributos.
Fonte: Elaboração própria (2019)

Analisando a Figura 34, pode-se observar que o nó raiz da árvore gerada é o passo da carcaça (PC). Este nó sozinho não traz muita informação útil neste momento, uma vez que um passo de 16,38mm é em média, considerado como pequeno e só será encontrado em carcaças de material confeccionado com baixas larguras de fita, possivelmente associados a linhas flexíveis com aplicação em lâminas d'água de baixa profundidade. Observando mais atentamente o banco de dados, é possível notar ainda que apenas um caso exemplo consta no BD com passo da carcaça menor ou igual à 16,38mm, onde 4 manchas foram extraídas para avaliação de uma mesma ocorrência e todas foram aprovadas nesta ocasião com características bem semelhantes.

Fazendo uma interpretação desta árvore, para casos onde o passo for maior que 16.36mm temos que a largura da mancha de branqueamento (W) passa a ser o subgrupo divisor associado. A largura da mancha passa a apresentar uma condição classificatória direta, onde larguras superiores a 175,8mm são consideradas reprobatórias e das instâncias avaliadas, um total de dezoito casos se enquadram nesta condição. Ainda assim, quatro casos foram classificados como aprovados com valores de W entre 171,6mm e 175,6mm, indicando que existe uma sensibilidade neste subgrupo com grande potencial de gerar erros, dada a proximidade da faixa de valores.

Mais adiante na interpretação é possível perceber uma relação muito interessante com o máximo nível de cinza extraído (MA). Sabemos que a escala de cinza 8-bits vai de 0 a 255

níveis que que 255 representa o branco máximo, logo valores acima de 171 tendem a ser muito “brancos” na imagem real, indicando que regiões com nível de escala de cinza elevados podem gerar reprovação. Tomando em consideração apenas este parâmetro, chama a atenção o fato de valores abaixo de 171 indicarem reprovação diretamente, pois logicamente poderia ser qualquer valor a partir deste limiar tendendo a tons mais escuros.

A média de intensidade de cinza (AV) também está associada com a leitura desta árvore. Um valor médio de intensidade de cinza superior à 142,4 indica que a mancha como um todo é relativamente branca e que pode haver um grande contraste na imagem entre valores de MI e MA. Finalmente, como subgrupo resultante final tem-se a redução de espessura da camada de revestimento interno (TR). Observa-se que onde houve redução de espessura foram classificadas reprovações, onze no total, e onde não houve redução de espessura foram observadas classificações aprobatórias, 5 no total, sendo que uma foi classificada erroneamente.

4.1. AVALIAÇÃO DOS ATRIBUTOS E SUAS INFLUÊNCIAS

Uma análise preliminar foi executada, onde os atributos do arquivo .arff carregados no WEKA foram testados com relação ao seu grau de influência através de vários algoritmos classificadores. Um quadro comparativo apresenta os parâmetros que mostraram maior relevância nas análises de classificação em cada método considerado, conforme Tabela 10.

Tabela 10 – Definição dos atributos de maior relevância por tipo de algoritmo avaliador de atributos

Grupo	Classificador	Maior relevância				
CF	CFS	TR	MA	OD	PI	QP
CS	CSE	PC	MA	PI	-	-
GR	Gain Ratio	PI	PC	MA	TR	-
IG	Info gain	PI	MA	PC	TR	-
OA	OneR attribute	PI	TR	OD	MA	DI
RA	Refief attribute	TR	MI	MA	AV	PC
SU	Symmetrical Uncert	PI	PC	MA	TR	-

Fazendo a intercessão com os resultados das aplicações de diferentes algoritmos de avaliação de atributos, nota-se que os parâmetros considerados mais relevantes para estabelecer relações foram MA, OD, PI, AV, TR, QP e PC, onde destes, TR, PC e PI aparecem como os de principal contribuição para classificações. Há interesse em adicionar no trabalho este grupo em específico, uma vez que não foi apontado pelos algoritmos de seleção de atributos avaliados. Este novo grupo será identificado como TT. A Tabela 11 apresenta um detalhamento sobre a assertividade dos modelos com base em cada grupo de atributos especificamente. É notável que

os grupos CS e SU apresentam resultados que devem ser avaliados detalhadamente, visto que estes obtiveram cem por cento de assertividade nas classificações.

Adicionalmente dois novos grupos foram também considerados, onde MC corresponde aos atributos puramente físico-mecânicos do duto flexível (PC, DI, OD, TR) e IM corresponde aos atributos puramente extraídos das imagens (H, W, A, MI, MA, AV, PI, PE e QP). Nestes últimos casos, o interesse é em avaliar o comportamento do modelo ao considerar características físicas e digitais isoladamente e efetuar uma comparação com os demais grupos para investigar onde as regras se distanciam ou se mesclam. Uma forma diferente de visualizar os resultados é apresentada na Figura 35.

Tabela 11 – Análise dos grupos de atributos

Grupo	Atributo raiz	N. de classificações corretas	N. de classificações incorretas	MCC
CF	PI	15	3	0.663
CS	PI	18	0	1
GR	PI	17	1	0.886
IG	PI	15	3	0.663
AO	PC	14	4	0.523
RA	PI	17	1	0.886
SU	PI	18	0	1
TT	PC	17	1	0.886
MC	PC	14	4	0.532
IM	PI	18	0	1

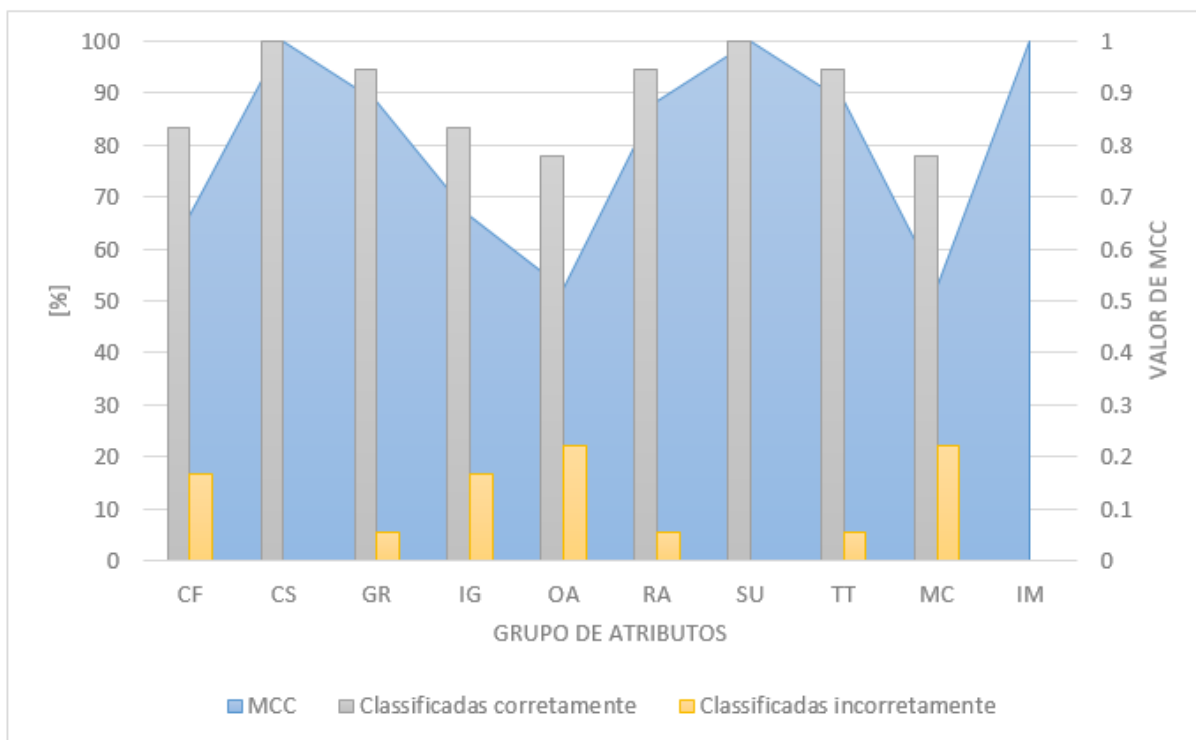


Figura 35 – Distribuição de resultados por grupos de atributos
 Fonte: Elaboração própria (2019)

4.2. EXTRAÇÃO DAS REGRAS E INTERPRETAÇÃO

Neto (2016), observou em seu trabalho que o modelo induzido pela árvore de decisão apresenta como principal vantagem sua capacidade de analisar os valores da divisão dos nós em subgrupos. A partir desses valores é possível entender a influência que um atributo exerce sobre outro e em que sentido, uma vez que interpreta-se o J48 com a orientação “*top-down*”, ou seja, de cima para baixo.

Ao contrário de outros modelos de classificação, que não apresentam as relações com visibilidade, pois possuem metodologia baseada na assertividade, o modelo de funcionamento do J48 pode ser avaliado e compreendido, disponibilizando valores gerados da base de dados com as relações sequentes ao longo da árvore, onde o especialista de domínio pode comparar os valores apresentados pelo modelo com particularidades da questão real a ser analisada e pesá-los com o embasamento de sua perícia.

Este tipo de modelo é facilmente aplicável na área industrial, uma vez que os atributos do modelo são exatamente as informações a serem ajustadas no processo fabril e conhecer um pouco mais sobre a sensibilidade de cada parâmetro em uma classificação final pode auxiliar no entendimento ainda mais profundo do processo, consequentemente implicando em melhorias e ajustes em um produto.

Alguns fatores da árvore de decisão auxiliam na capacidade interpretativa de um analista de dados, como exemplo pode-se citar a apresentação gráfica e hierárquica dos resultados, que exhibe uma estrutura intuitiva de fácil leitura, permitindo ao analista, mesmo sem experiência com algoritmos e modelos numéricos, fazer extração de informação útil e analisar a eficácia do modelo para uma possível solução de problema de classificação.

Outro fator relevante deste modelo em árvore é que se pode avaliar o peso de cada nó ao compará-lo com sua proximidade da raiz, uma vez que este é o mais relevante. Entretanto, Neto (2016) reitera que neste ponto, às vezes o atributo pode aparecer repetidamente em uma ramificação e assim, sua importância estará inicialmente associada ao seu primeiro aparecimento na estrutura da árvore, apesar disto não ser uma regra. Um atributo A pode estar mais próximo da raiz do que o atributo B e, mesmo assim, B pode ser mais relevante devido a ter classificado um número maior de registros.

Neto (2016) ainda chama a atenção para o fato de que durante o processo indutivo de uma árvore algum ramo pode trazer valor sem importância real para a classificação, logo não sendo de real relevância para o processo, mesmo que o banco de dados já tenha sido previamente tratado. Esta situação pode ocorrer devido à natureza do algoritmo de indução, pois, uma vez que um atributo é selecionado para ser o nó da árvore, cada valor deste atributo deve ser incluído. Logo, poderia haver uma situação onde o ramo da árvore tenha sido acrescentado sem agregar valor real de classificação.

Na indução de uma árvore, há a fase de crescimento da árvore seguida de uma fase de poda que contribui para melhorar o modelo resultante. O método de poda não consegue resolver o problema de fragmentação apresentado, uma vez que o único caminho para remover os valores irrelevantes (ramos) da árvore remove, também, o atributo relevante (nó).

Nas árvores geradas pelo algoritmo J48 do WEKA, os nós que geram uma classificação são seguidos por um número (às vezes dois) entre parênteses. O primeiro número indica quantas instâncias no conjunto de treinamento são corretamente classificadas por este nó. O segundo número, se existir (se não, é considerado 0,0), representa o número de instâncias classificadas incorretamente pelo nó (WITTEN et al., 2016).

A Tabela 12 apresenta as principais regras extraídas da subdivisão das árvores geradas para os casos A1, A3 e A9 aplicando os atributos contidos no grupo TT, uma vez que este grupo engloba todos os atributos que obtiveram os melhores resultados pelo classificador MCC.

A partir da leitura dos valores de cada nó divisor para definição de cada novo subgrupos, é possível perceber o peso de cada relação nas classificações geradas. Os ramos de árvore que contribuíram com relevância superior a 10% podem ser observados em destaque.

Ainda observando a Tabela 12, é possível identificar alguns padrões classificatórios. Por exemplo, o atributo AV aparece tanto para o caso A1 quanto para o caso A9 no nó 6 indicando potencial de reprovação para valores acima de 142.4 e da mesma forma o atributo MA apresenta potencial de reprovação para valores abaixo de 171. Áreas de manchas branqueadas (A) também apresentam grande potencial de reprovação quando superiores a 932.2mm².

Tabela 12 – Principais regras extraídas da subdivisão das árvores geradas para os casos A1, A3 e A9, grupo TT.

Nó	W	A	AV	MA	MI	TR	PC	Class.	Nº de ocorrências	Relevância [%]	
Caso A1	1						≤16,38	Reprovado	4	6,7	
		>175,6					>16,38	Reprovado	18	30	
	2	171,6<x					>16,38	Aprovado	4	6,7	
		≤175,6									
	3	≤171,6			≤171			>16,38	Reprovado	12	20
	4	>171,6	≤932,2		≤171			>16,38	Aprovado	4	6,7
	5	>171,6	>932,2	≤142,4	≤171			>16,38	Aprovado	2	3,4
6	>171,6	>932,2	>142,4	≤171		Sim	>16,38	Reprovado	11	18,4	
7	>171,6	>932,2	>142,4	≤171		Não	>16,38	Aprovado	5	8,4	
Caso A3	1						≤16,38	Aprovado	4	6,7	
			≤957,1				>16,38	Aprovado	3	5	
	2		>957,1		≤186	≤108		>16,38	Reprovado	3	5
	3		>957,1		>186	≤108		>16,38	Aprovado	4	6,7
	4		>957,1			108<x		>16,38	Reprovado	21	35
						≤189					
5		>957,1			>189	Sim	>16,38	Reprovado	3	5	
6		>957,1			>189	Não	>16,38	Aprovado	2	3,4	
Caso A9	1						≤16,38	Aprovado	4	6,7	
		>175,6					>16,38	Reprovado	18	30	
	3	171,6<x					>16,38	Aprovado	4	6,7	
		<175,6									
	4	<171,6			≤171			>16,38	Reprovado	12	20
	5	<171,6		≤142,4	>171			>16,38	Aprovado	2	3,4
	6	<171,6		>142,4	>171		Sim	>16,38	Reprovado	12	20
7	<171,6		>142,4	>171		Não	>16,38	Aprovado	8	13,4	

Outras árvores também foram geradas, como pode ser observado nas Figura 36 e Figura 37, onde os atributos do grupo TT foram testados variando as configurações do algoritmo J48 para os casos A3 e A9 respectivamente.

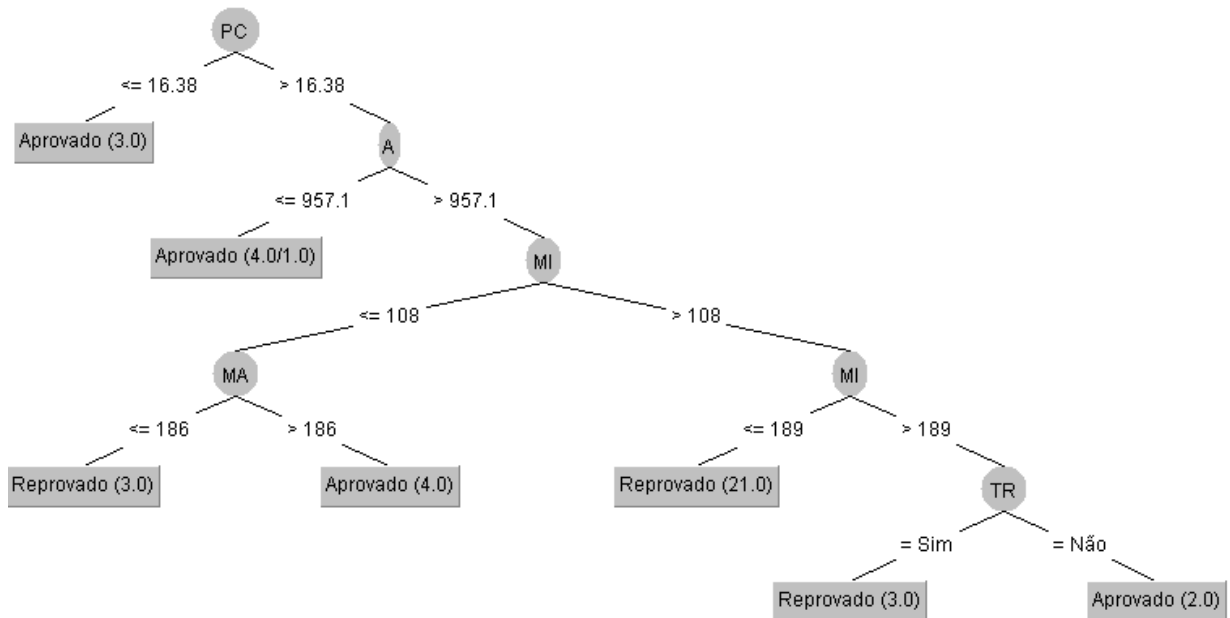


Figura 36 – Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A3, grupo TT.
Fonte: Elaboração própria (2019)

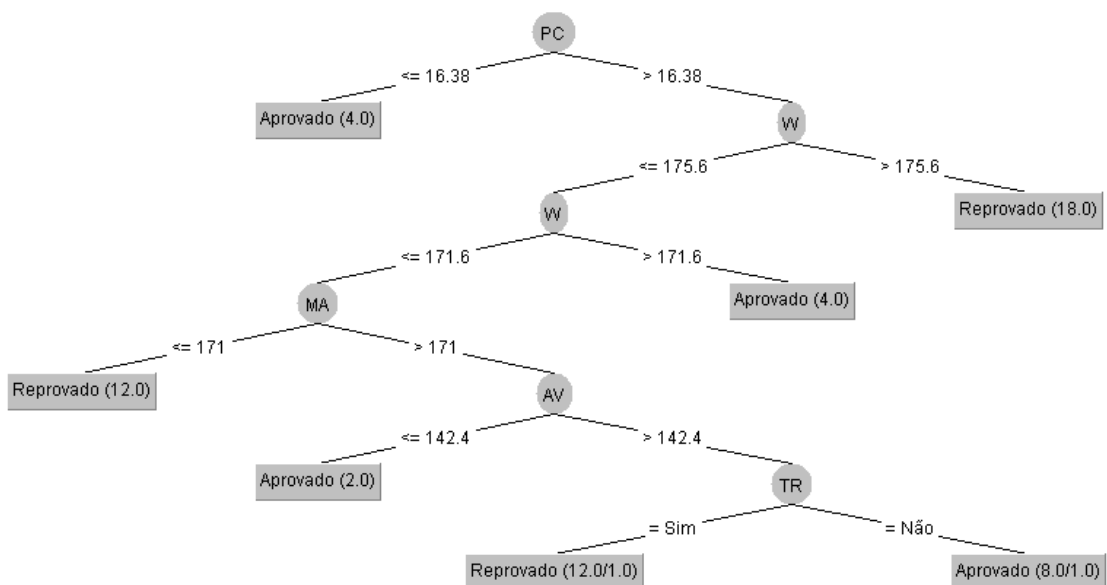


Figura 37 – Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A9, grupo TT.
Fonte: Elaboração própria (2019)

Um fator que muito chama a atenção nas três árvores geradas para o grupo TT, considerando os casos A1, A3 e A9, é que em todos os casos o passo da carcaça é a raiz da

árvore. Para entender um pouco melhor a influência do passo da carcaça nesta avaliação, uma nova análise foi executada removendo as situações onde passos menores ou iguais a 16,38mm ocorreram, uma vez que após uma verificação do BD foi constatado que todas as ocorrências com este valor de passo foram identificadas sendo de uma mesma ocorrência de desvio de manufatura e todas foram aprovadas com o mesmo embasamento do especialista. Nenhuma outra ocorrência foi observada em linha flexíveis com passos similares.

Portanto, para melhor avaliar o passo da carcaça neste modelo, uma nova árvore foi gerada, onde esta, apresentou como nó raiz a largura da mancha de branqueamento (W). Outros nós são subdivididos a partir do passo da carcaça partindo de um valor de 36,15mm, logo podemos entender que existe de fato uma relação forte do passo da carcaça com o efeito de branqueamento, demonstrando que há uma grande aderência da variação deste atributo com as características do processo de extrusão do revestimento interno. A árvore resultante para o modelo ajustado é apresentada na Figura 38.

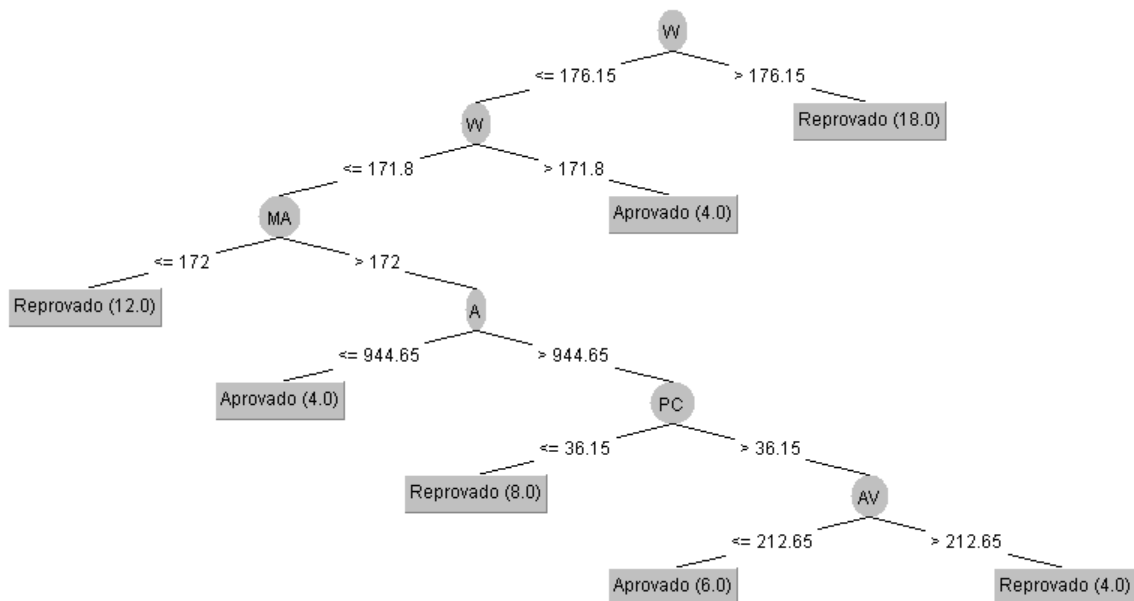


Figura 38 - Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A1, grupo TT, com BD ajustado para atributo PC com valor acima de 16,38mm apenas.

Fonte: Elaboração própria (2019)

A Tabela 13 apresenta as regras de classificação geradas para o BD ajustado.

Tabela 13 - Principais regras extraídas da subdivisão das árvores geradas para o caso A1, grupo TT, com atributo PC ajustado.

	Nó	W	A	AV	MA	PC	Classificação	Nº	Relevância [%]
Caso A1	1	$x > 176,15$					Reprovado	18	30
		$171,8 < x \leq 176,15$					Aprovado	4	6,7

2	$\leq 171,8$		≤ 172		Reprovado	12	20
3	$\leq 171,8$	$\leq 944,65$		>172	Aprovado	4	6,7
4	$\leq 171,8$	$\leq 944,65$		>172	Reprovado	8	13,4
5	$\leq 171,8$	$> 944,65$	$>212,65$	>172	Reprovado	4	6,7
6	$\leq 171,8$	$> 944,65$	$<212,65$	>172	Aprovado	6	10

Uma última análise também foi feita excluindo-se todos os atributos mecânicos associados ao produto analisado e mantendo apenas as informações extraídas da imagem, conforme especificado para o grupo IM. Como resultado a árvore gerada é apresentada na Figura 39.

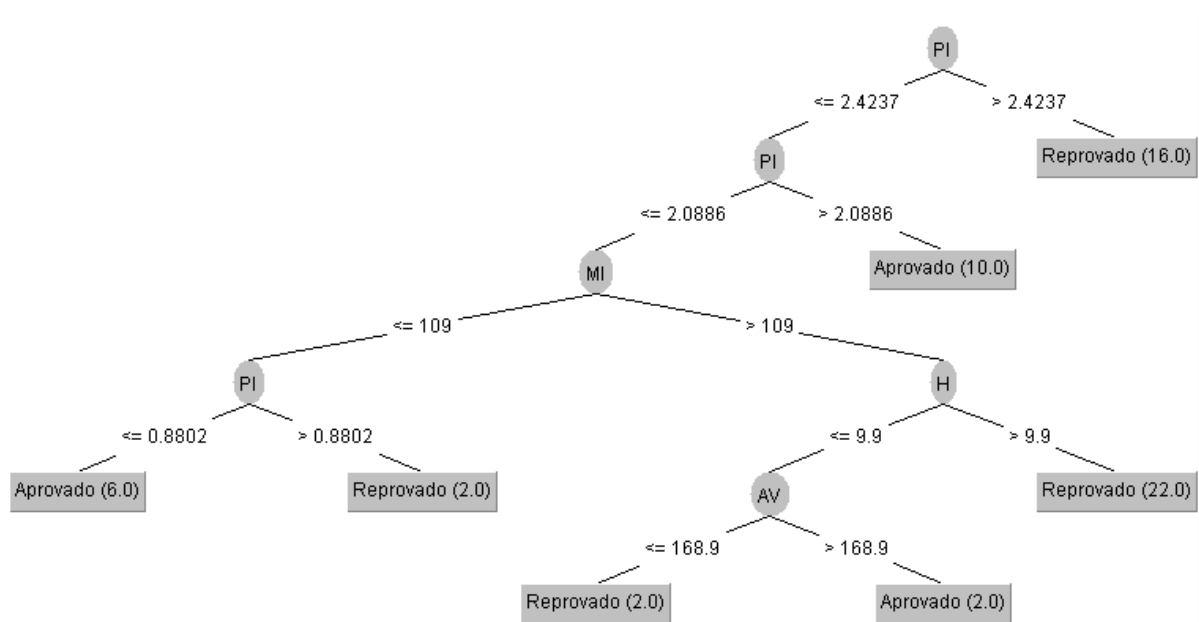


Figura 39 - Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A1, grupo IM.
Fonte: Elaboração própria (2019)

4.3. VALIDAÇÃO DO MODELO

Para validação do modelo é proposto executar um teste a partir das características extraídas das duas amostras identificadas na Seção 3.1.1. As amostras possuem características visuais bem semelhantes, no entanto, uma das amostras foi reprovada, quando avaliada numericamente pelos cálculos, e uma outra amostra foi considerada como aprovada com base nas mesmas premissas. Portanto, é conveniente fazer uma análise comparativa objetivando avaliar o modelo criado em relação ao seu potencial de predição.

Um novo arquivo com extensão *.arff* é criado com valores das duas amostras apenas. A Figura 40 apresenta o *dataset* a ser carregado no WEKA para ser testado pelo modelo gerado. O novo *dataset* não possui um valor de classificação definido, porém é adicionado um sinal de interrogação “?” ao final de cada instância, para que o WEKA preencha este valor a partir de uma predição gerada pelo modelo já criado. O modelo utiliza as mesmas regras inicialmente criadas para avaliar o novo *dataset*, chamado de *test_data.arff* e então lhe atribui uma classificação com base no que já foi aprendido.

```
@Relation Whitening_classifier
@attribute DI real
@attribute OD real
@attribute PC real
@attribute TR {Sim,Não}
@attribute H real
@attribute W real
@attribute A real
@attribute PE real
@attribute AV real
@attribute MA real
@attribute MI real
@attribute PI real
@attribute QP real
@attribute Class {Aprovado,Reprovado}

@data
8,249.2,36.14,Não,23.4,133.8,1663.6,309.7,196.7,201,111,0.4300,3910.67,?
8,249.2,36.14,Não,17.7,119.0,1818.6,289.5,145.0,165,125,0.6700,2713.11,?
```

Figura 40 – *Dataset (test_data.arff)* criado para avaliação das amostras calculadas na avaliação físico-mecânica.
Fonte: Elaboração própria (2019)

Como resultado do mecanismo de classificação, tanto a amostra #01 quanto a amostra #02 foram consideradas reprovadas. Sabe-se que esta não é uma verdade, considerando os resultados de cálculos que definiram uma classificação prévia com base nos parâmetros físico-mecânicos, pois sabe-se que a amostra #02 teria sido aprovada pelo critério de máximo 5.25% de deformação. É proposto reavaliar o modelo através de um refinamento para que outras regras de classificação mais efetivas possam ser geradas e estudadas.

4.3.1. Refinamento do modelo

Analisando as características de imagem puramente, podemos observar através da Figura 39 uma relação forte de PI, MI e H. Porém um grupo que associa tais parâmetros com os atributos mecânicos de maior relevância ainda não foi explorado. Novamente é gerada uma árvore de decisão, porém agora associando estes parâmetros de imagem aos parâmetros

mecânicos que se mostraram mais relevantes (PC e TR). A estes parâmetros foi associado um novo grupo identificado como GTT2. A Figura 41 apresenta o resultado desta nova análise para o grupo GTT2.

A Figura 41 ilustra uma árvore, indicando a representatividade da regra (i.e. redução de espessura), seguida de relações de mínimo de intensidade de cinza e por fim quantidade de pixels por unidade de área. Desta forma, é possível concluir que de fato existem relações a serem extraídas como conhecimento útil deste mecanismo. O modelo abordado tem potencial classificatório para ser explorado com mais profundidade, apesar de trazer pouca informação útil sobre o atributo MI neste momento. A Tabela 14 apresenta uma análise do comportamento dos atributos principais que geraram regras através desta árvore.

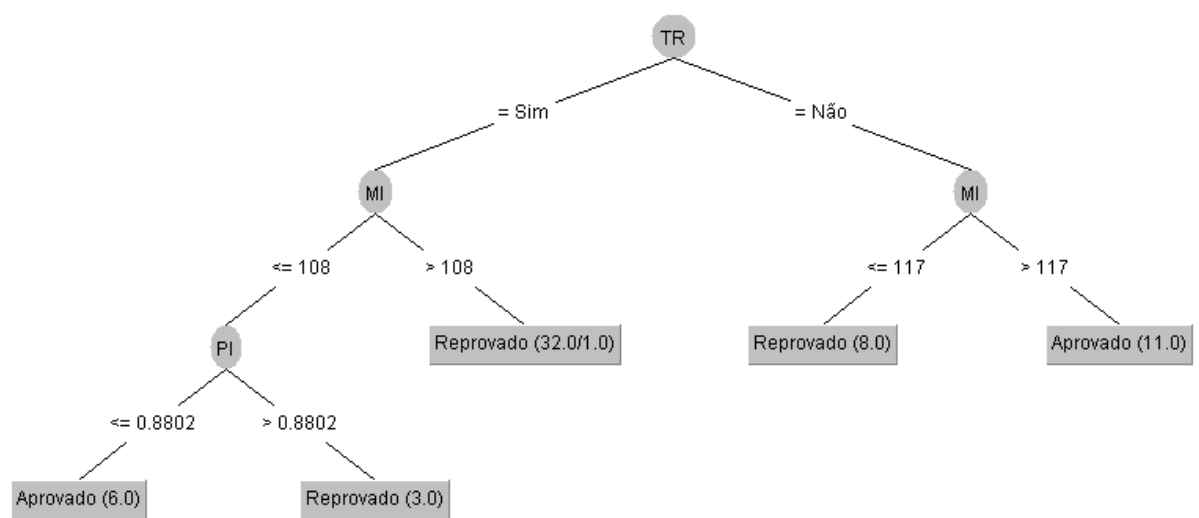


Figura 41 - Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A3, grupo GTT2.
Fonte: Elaboração própria (2019)

Tabela 14 – Análise dos atributos mais relevantes da Figura 41.

Atributo	Descrição da regra	Chance de ocorrer	Probabilidade reprovado [%]	de	Probabilidade de aprovar [%]
TR	Sim	41/60 (68,3%)	34/41 (82,9%)	7/41	(17,1%)
TR	Não	19/60 (31,7%)	8/19 (42,1%)	11/19	(57,9%)
TR (sim) + MI	Inferior a 108	9/41 (22,0%)	5/18 (27,8%)	13/18	(72,2%)
TR (sim) + MI	Maior que 108	32/41 (78,0%)	32/41 (78,0%)	3/36	(8,3%)
TR (não) + MI	Inferior a 117	46/60 (76,7%)	43/46 (93,5%)	6/46	(13,0%)
TR (não) + MI	Maior que 117	11/60 (18,3)	-	11/11	(100%)

Uma revalidação do modelo é então executada a partir do caso de configuração A3, grupo GTT2 no WEKA. O resultado pode ser visto na Figura 42. Nesta última situação, os resultados da predição do modelo foram coerentes com os resultados da análise físico-mecânica

e retornou uma predição que aponta para a mesma direção, ou seja, amostra 01 reprovada e amostra 02 aprovada.

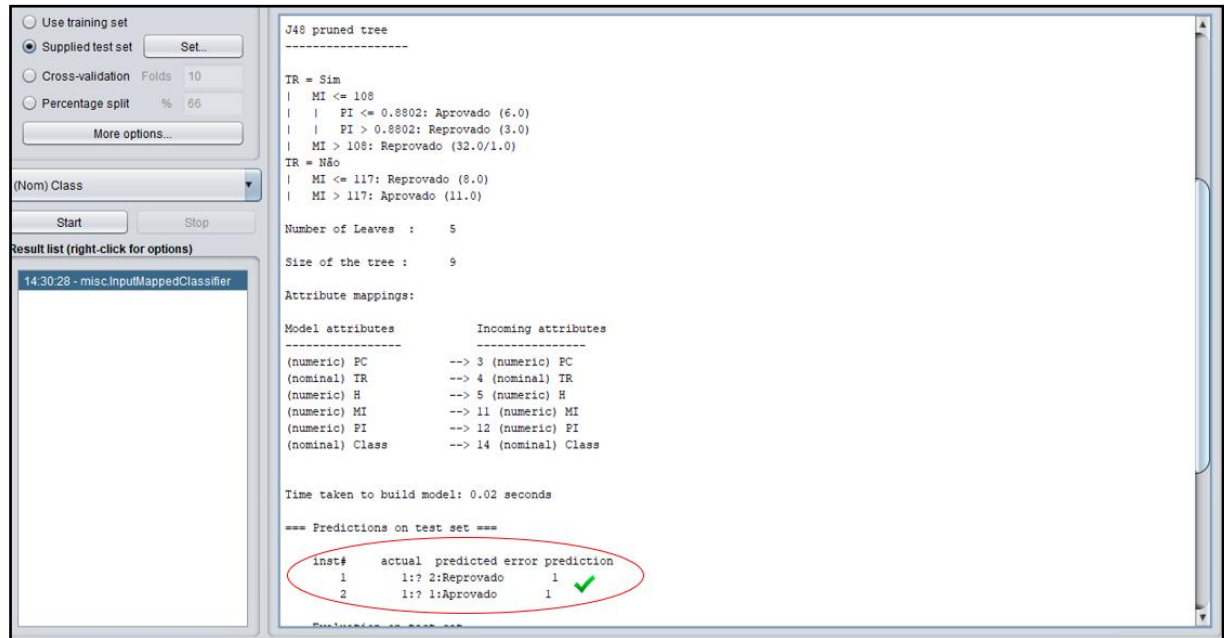


Figura 42 – Resultado da predição do modelo refinado para as amostras 01 e 02 para o caso A3, grupo GTT2.

Fonte: Elaboração própria (2019)

4.3.2. Análise dos atributos após implementação do filtro *Resample*

O filtro *resample* produz uma subamostra aleatória de um conjunto de dados usando amostragem com substituição ou sem substituição. O número de instâncias no conjunto de dados gerado é especificado para um atributo de classe nominal. Este filtro é aplicado para influenciar a distribuição da classe em direção a uma distribuição uniforme na subamostra. O parâmetro do filtro *resample* que será alterado da configuração padrão do WEKA é o *random seed number*, ou seja, número de sementes aleatórias da árvore.

A aplicação do filtro *resample* no *dataset*, já ajustado para passos de carcaça maiores que 16.38mm é realizada apenas com intuito de reavaliar as características que geraram a árvore da Figura 41. Esta análise é executada como uma reavaliação do modelo, apenas com interesse em reforçar a contribuição dos atributos identificados no refinamento do modelo, através da árvore gerada pelo caso A3, grupo GTT2. O resultado é apresentado nas Figura 43 e Figura 44, onde é de fato observado a atuação dos mesmo atributos (TR, PI e MI), porém em diferentes disposições.

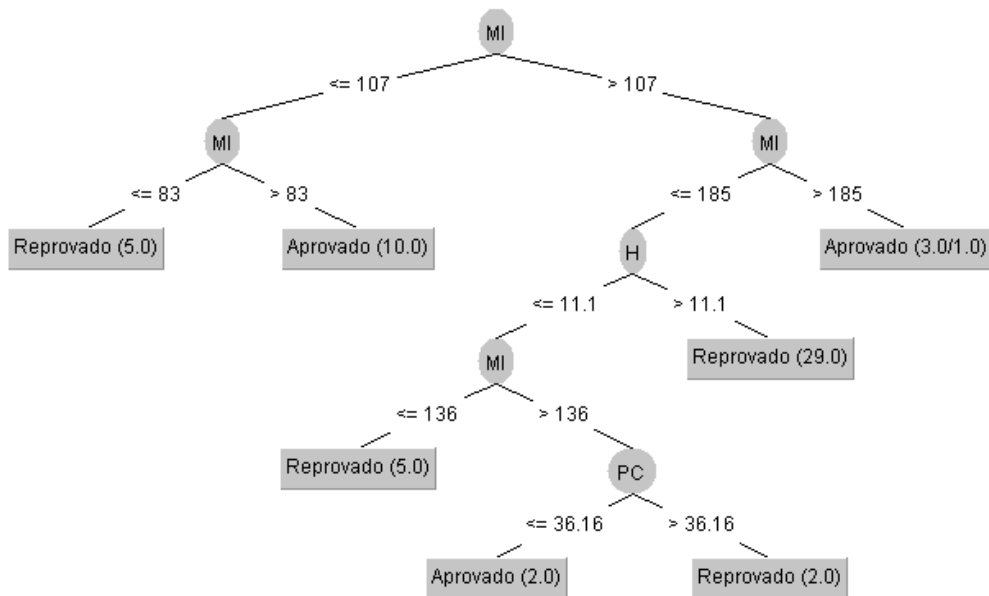


Figura 43 - Árvore de decisão gerada pelo algoritmo J48 configurado para o caso A3, grupo GTT2 com aplicação de filtro *Resample*.(1 semente aleatória).
Fonte: Elaboração própria (2019)

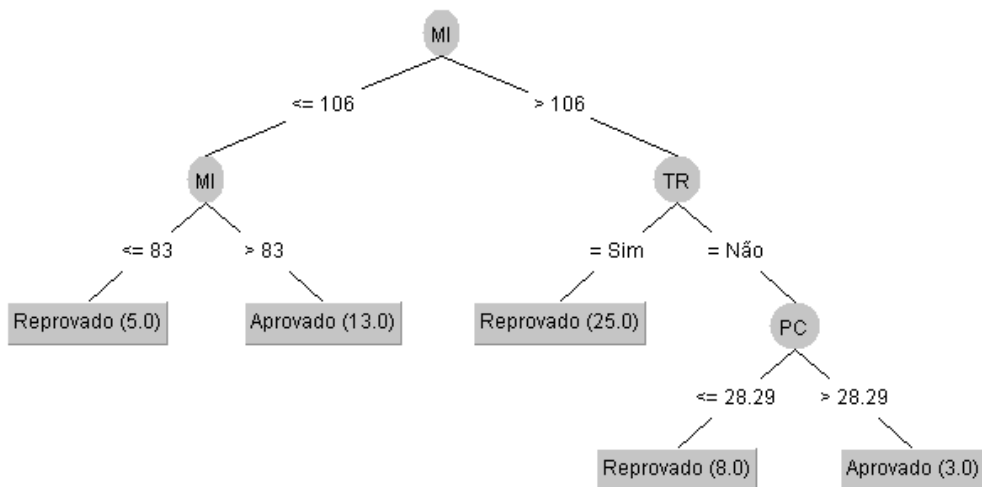


Figura 44 - Árvore de decisão gerada pelo algoritmo J48 configurado para o caso , grupo GTT2, com aplicação de filtro *Resample*.(8 sementes aleatórias).
Fonte: Elaboração própria (2019)

A Tabela 15 apresenta uma análise detalhada sobre a assertividade dos atributos da árvore de decisão apresentado na Figura 43. Já a Tabela 16 apresenta uma análise detalhada sobre a assertividade dos atributos da árvore de decisão apresentada na Figura 44.

Tabela 15 – Análise dos atributos mais relevantes da Figura 43

Atributo	Descrição da regra	Chance de ocorrer	Probabilidade reprovado [%]	de	Probabilidade de aprovar [%]
MI	Inferior a 107	15/60 (25,0%)	5/15 (33,3%)		10/15 (66,7%)
MI	Maior que 107	41/60 (68,33%)	37/41 (90,24%)		9/41 (9,8%)
MI > 107 + H	H maior que 11,1mm	38/41 (92,47%)	29/38 (76,3)		-
MI > 107 + H	H inferior a 11,1mm	9/41 (22,00%)	7/9 (77,8%)		2/9 (22,2%)

Tabela 16 – Análise dos atributos mais relevantes da Figura 44

Atributo	Descrição da regra	Chance de ocorrer	Probabilidade reprovado [%]	de	Probabilidade de aprovar [%]
MI	Inferior a 106	18/60 (30,0%)	5/18 (27,8%)		13/18 (72,2%)
MI	Maior que 106	36/60 (60,0%)	33/36 (91,7%)		3/36 (8,3%)
MI > 106 + TR	TR Sim	25/36 (69,4%)	25/25 (100,0%)		-
MI > 106 + TR	TR Não	11/36 (30,6%)	8/11 (72,7%)		2/9 (27,3%)

Nota-se a partir da Figura 45 que DI e OD não foram parâmetros considerados relevantes como atributos em nenhuma regra de classificação, no entanto é interessante fazer uma análise associando estes dois parâmetros com outras características importantes. Os diâmetros, tanto interno da linha, quanto externo da camada de revestimento interno, podem ainda trazer algum conhecimento útil para este estudo, uma vez que os cálculos de deformação são influenciados diretamente por estes parâmetros.

Para fazer uma análise específica destes parâmetros o modelo foi ajustado conforme o caso de configuração A3 para os atributos diâmetro interno (DI), largura da mancha (W) e área da mancha (A) conforme resultados apresentados na Tabela 17, assim como os parâmetros relação de pixels por mm² (PI), mínimo de intensidade de cinza (MI), diâmetro interno (DI) e diâmetro externo (OD) na Tabela 18 e por fim os parâmetros largura da mancha (W), máxima intensidade de cinza (MA) e relação pixels por mm² (PI) apresentados na Tabela 19.

Tabela 17 – Configuração A3 para os atributos diâmetro interno (DI), largura da mancha (W) e área da mancha (A) para o modelo refinado

ID	Regra	Classificação	Probabilidade de ocorrer o evento. [%]
	W	A	
Se maior que 8"	-	-	Reprovado 5/60 (8,3%)
Se menor ou igual à 8"	Maior que 79,4mm	Entre 1079,1mm ² e 1820mm ²	Aprovado 4/60 (6,7%)
Se menor ou igual à 8"	Entre 190mm e 231,7mm		Reprovado 13/60 (21,7%)

Tabela 18 - Configuração A3 para os atributos relação pixels por mm² (PI), mínimo de intensidade de cinza (MI), diâmetro interno (DI) e diâmetro externo (OD) para o modelo refinado

PI	MI	Regra		OD	Classificação	Probabilidade de ocorrer o evento. [%]
Se maior que 2,4237pi/mm ²	-	-	-	-	Reprovado	16/60 (26,7%)
Se menor ou igual à 2,4237pi/mm ²	Se maior que 109	Maior que 6"		-	Reprovado	22/60 (36,7%)
Se menor ou igual à 2,4237pi/mm ²	Se maior que 109	Menor ou igual à 6"	ou	Maior que 186,8mm	Aprovado	2/60 (3,3%)
Se menor ou igual à 2,4237pi/mm ²	Se maior que 109	Menor ou igual à 6"	ou	Menor ou igual à 186,8mm	Reprovado	2/60 (3,3%)

Tabela 19 - Configuração A3 para os atributos largura da mancha (W), máxima intensidade de cinza (MA) e Relação pixels por mm² (PI) para o modelo refinado

W	MA	Regra		PI	Classificação	Probabilidade de ocorrer o evento. [%]
Se maior que 175,6mm	-	-	-	-	Reprovado	18//60 (30,0%)
Se menor ou igual à 175,6mm	Se maior que 219			Se maior que 0,63	Reprovado	4/60 (6,7%)
Se entre 68,9mm e 175,6mm	Se entre 172 e 219			Se maior que 0,63	Aprovado	13/60 (21,7%)
-	Se menor que que 171			-	Reprovado	12/60 (20,0%)

4.4. DISCUSSÃO DOS RESULTADOS

Testes de validação dos algoritmos e escolha do melhor método: uma comparação entre algumas técnicas de classificações foi executada através da aplicação de algoritmos disponibilizados pelo WEKA, onde definiu-se fazer o uso do método de árvores de decisão (J48), uma vez que este apresentava o resultado de forma mais compatível com os propósitos deste estudo.

Os parâmetros de número de instâncias por folha, redução de poda, redução de erro da poda e correção DML foram trabalhados em diferentes combinações e então avaliados segundo avaliador de performance MCC.

Os casos que apresentaram melhor performance foram identificados e considerados nas análises por apresentarem maior confiabilidade. A seguir, algumas configurações possíveis do J48 foram testadas em combinação com grupos de atributos distintos, visando obter as melhores

avaliações de performance indicados pelo Coeficiente de Correlação de Matthews (MCC) para serem avaliadas em maior profundidade.

Através das análises executadas para caso de configuração A3, grupo GTT2 com e sem aplicação do filtro *resample*, foi constatado que valores de MI entre 107 e 116, tanto isoladamente quanto em grupo, tem mais de 90% de chance de classificar uma avaliação como reprovado, assim como valores de MI entre 83 e 106 tendem a serem aprovados (mais de 67% de chance). Essa margem causa estranheza e indica que os valores mínimos de intensidade de cinza podem ter sido extraídos com dificuldade das imagens, principalmente sabendo que este é um parâmetro sensível, que pode sofrer grande variação de valor por pixel quadrado dependendo da forma como é feita a seleção da região da mancha branqueada. Já valores de PI entre 2,08 e 2,42 tem alguma chance de obter classificações aprovadas (16,7%), mas valores superiores ou inferiores a estes apresentam chances em torno de 70% de reprovação o que torna arriscado se embasar neste parâmetro para tomar uma decisão.

Ainda analisando os resultados, o atributo H definitivamente indica reprovação quando se obtém valor superior a 11,1mm. No entanto, já a partir de 9,9mm o índice de reprovação é bem relevante, cerca de 84%. Valores de H inferiores a 9,0mm tendem a indicar classificação aprovatória e este acaba se tornando um importante referencial para investigar características que possam indicar uma possível aprovação. Os valores de H que se encontram entre 9,6mm e 12,5mm funcionam bem como caráter classificatório para obter aprovação em dutos com diâmetros internos (DI) de até 6" e diâmetro externo do revestimento (OD) entre 186,8mm e 249,2mm, caso a largura da mancha (W) fique abaixo dos 175,6mm.

Na Figura 45 é possível perceber a influência de cada atributo nas principais regras de classificação geradas a partir do refinamento do modelo. O passo da carcaça apareceu 27 vezes como nós ou sementes se mostrando um importante agente classificatório nas árvores de decisão. Outro atributo que também teve forte presença nos resultados foi "W". Este atributo foi muito relevante para estabelecer associações entre parâmetros físicos do produto com parâmetros digitais da imagem. Os atributos de máximo (MA) e mínimos (MI) de intensidade de cinza também tiveram contribuição na geração das árvores. Estes parâmetros são interessantes, pois abordam a variação da tonalidade quantitativamente e proporcionam uma análise através das características puramente oriundas das imagens.

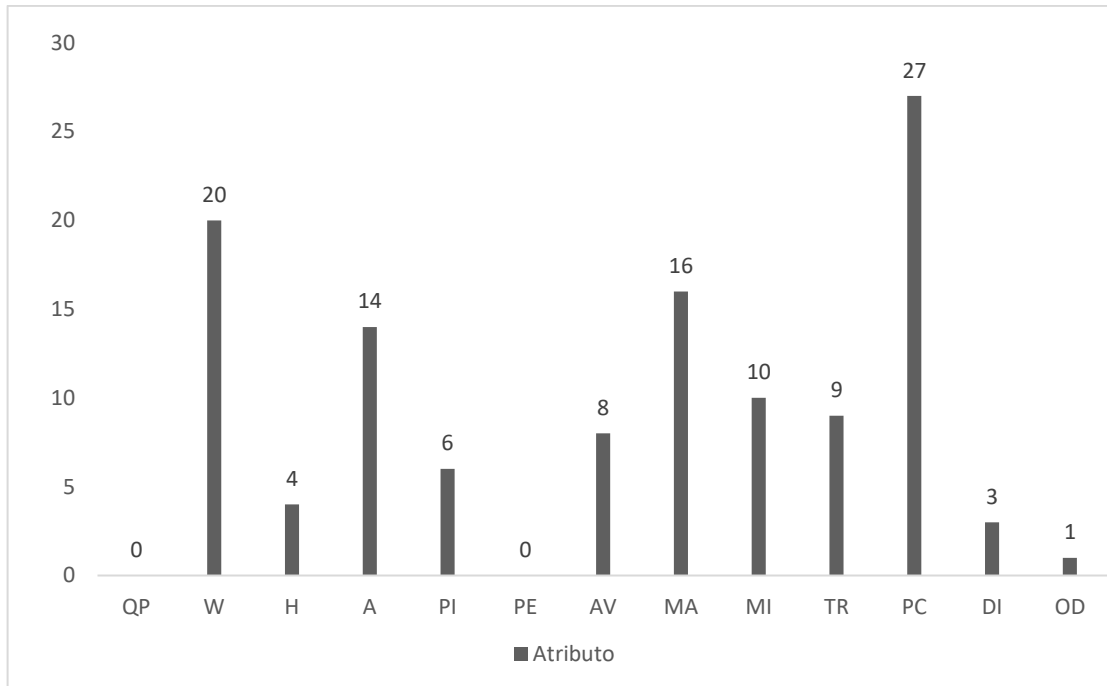


Figura 45 – Número de aparições de cada atributo nas principais análises
Fonte: Elaboração própria (2019)

As árvores geradas pelo modelo foram analisadas para os diferentes casos e combinações de atributos para as amostras de referência, #01 e #02 e geraram um resultado incompatível com a análise numérica. Logo, um refinamento do modelo foi proposto, incluindo uma abordagem específica sobre a seleção dos atributos, caso de configuração A1, grupo IM em adição aos parâmetros PC e TR que formou um novo grupo identificado como GTT2. Além deste refinamento, um teste paralelo foi executado onde um filtro conhecido como *resample* foi aplicado no *dataset* para o modelo e então uma nova árvore foi gerada. A interpretação das árvores apontou principalmente para os parâmetros MI, PI e TR apresentando forte relação. Todavia, um estudo envolvendo DI e OD também foi feito e deste, uma nova relação entre os atributos é identificada para PI, MI, DI e OD ou DI, W e A. Neste caso os mais consistentes são W, MA e PI quando se trata de referência aprobatória (21,7%) e PI, MI e DI para referência reprobatória (36,7%)

Um novo teste do modelo é feito a partir das características das amostras #01 e #02, onde os atributos TR, PI e MI apresentam as relações classificatórias. Desta vez, após o refinamento, o resultado predito é compatível com a análise numérica, encerrando assim os testes e análises propostos por este trabalho.

CAPÍTULO 5 - CONCLUSÃO

A partir de uma pesquisa bibliográfica e de trabalho de campo, um estudo foi elaborado na área industrial com foco na fabricação de dutos flexíveis, especificamente na execução do processo de extrusão da camada de estanqueidade da estrutura do duto. A metodologia proposta neste trabalho propõe avaliar uma característica específica do material PVDF que pode ocorrer durante o processo de manufatura do mesmo.

A anomalia tratada neste estudo, conhecida como branqueamento do PVDF, foi avaliada através de características de imagens disponibilizadas pela empresa NOVFlexibles seguindo o conceito de mineração de dados e do uso das técnicas de classificação definidas durante a pesquisa. A avaliação apontou para relações interessantes entre parâmetros de imagem e características físicas do produto avaliado. Estas relações puderam ser quantificadas pelos valores de subdivisão de grupos identificados pelo algoritmo J48 do WEKA.

A possibilidade de associação das relações identificadas neste estudo com as avaliações reais feitas por especialistas foi vista como de relevante potencial para estabelecer regras de classificação. Entretanto, o grau de maturidade deste estudo ainda é insuficiente para concluir com precisão qualquer tipo de regra final com base apenas nos resultados apresentados.

Um resumo das principais regras de classificação é apresentado para dutos com diâmetro interno de até 6 polegadas na Tabela 20. Para dutos com diâmetros internos maiores ou iguais a 8 polegadas, o resultado é apresentado na Tabela 21. Uma regra mais geral é apresentada na Tabela 22.

Tabela 20 – Principais regras de classificação condicionadas à aprovação para dutos de até 6”.

Descrição da regra	Probabilidade de evento [%] / Assertividade
Se MI for maior que 107 e H inferior à 11,1mm	22,2% / >100%
Se PI for menor ou igual à 2,4237pi/mm ² e MI maior que 109	3,3% / >80%

Tabela 21 – Principais regras de classificação condicionadas à aprovação para dutos maiores ou iguais a 8”.

Descrição da regra	Probabilidade de evento [%] / Assertividade
Se W for maior que 79,4mm e A estiver entre 1079,1mm ² e 1820mm ²	6,7% / >100%

Tabela 22 – Regras gerais de classificação condicionadas à aprovação.

Descrição da regra	Probabilidade de evento [%] / Assertividade
Se W estiver entre 68,9mm e 175,6mm e MA entre 172 e 219	21,7% / >80%
Se MI for maior que 109, H menor que 9,9mm e AV estiver entre 168,9 e 121,65	10,0% / >80%

Algumas dificuldades foram identificadas ao longo do desenvolvimento do presente estudo, como por exemplo o ângulo de inclinação que o duto apresenta nas imagens devido à diferença de altura entre a região de interesse e a pessoa que faz o registro das imagens. Esta variação impede que os diâmetros calibrados sejam os mesmos ao longo de todo o duto em uma imagem. Outro exemplo de dificuldade foi a dependência de um especialista em toda extração de dados.

Seria importante definir parâmetros mais objetivos para identificar as manchas de branqueamento a partir de valores chave pré-definidos pelos próprios especialistas. Além disso, o atributo PI apresenta uma relação de pixel por milímetro quadrado na imagem, que pode variar não apenas com o diâmetro do duto ou a área afetada pelo branqueamento em si, mas também conforme o zoom da foto no momento do registro. Mesmo executando a padronização de resolução nas imagens, este parâmetro poderá apresentar variação com relação a cada imagem individualmente. Como este parâmetro se encontra presente em vários resultados, é necessário ter cuidado com as associações feitas a partir deste, prezando pela coerência das relações verdadeiramente assertivas.

Pode-se finalmente concluir, como demonstrado no presente estudo, que é possível extrair uma relação “geometria da mancha branqueada *versus* geometria das camadas *versus* intensidade de tons de cinza da região”. Apesar dos resultados serem insuficientes para embasar uma relação objetiva, forte o bastante para estabelecer valor classificatório confiável, já demonstra potencial de orientação para classificar ocorrências de desvios de branqueamento, caso haja.

O objetivo inicial do trabalho de encontrar uma relação entre características da imagem com parâmetros do produto fabricado é finalmente alcançado. Estudos adicionais, no entanto, são fortemente recomendados para avaliar a análise custo-benefício do algoritmo proposto no modelo. O uso de árvores de decisão, no entanto, combinado com a experiência de especialistas do produto em questão é considerado um passo importante em direção ao aperfeiçoamento do modelo proposto neste estudo, visando obter dados ainda mais sólidos e precisos.

Como sugestão de trabalhos futuros recomenda-se:

- A execução do processamento de imagens na identificação do ROI (*region of interest*) automaticamente e executar a extração dos parâmetros relevantes sem a necessidade do agente humano. Tal recurso seria de vital valor para um aperfeiçoamento da técnica, uma vez que a extração seria executada de forma padronizada, minimizando as variações ocasionadas pelo recurso de seleção

livre da ferramenta FIJI que dependem da habilidade do usuário que ainda conta com o auxílio de um especialista;

- Padronizar o método de aquisição de imagens para viabilizar o processo de automação de extração de parâmetros e reduzir os ruídos e erros atualmente presentes no BD. Um dispositivo com controle de luminosidade apropriado e portátil pode ser construído para disponibilizar este recurso. O controle de parâmetros na aquisição das imagens tornaria a qualidade da extração das características da imagem, além de equalizadas, muito mais confiáveis;
- Ampliação do BD com novas informações de imagens oriundas de desvios em processo de fabricação de camadas de revestimento interno em PVDF. A empresa em NOFlexibles foi de vital papel para viabilizar este estudo, uma vez que cedeu as imagens de seu acervo interno de registro da qualidade para o propósito deste trabalho. Todavia, se fosse possível catalogar mais imagens, inclusive de outras empresas do setor, seria um complemento que poderia levar este trabalho para um nível de aplicabilidade muito mais próximo do real;
- Utilização de técnicas de processamento de imagens para aquisição automática das características das imagens;
- Desenvolvimento de um *software* que unifique o processo de classificação através da aplicação de um modelo mais sofisticados de predição, com base no que foi proposto neste trabalho e que possa integrar a metodologia de árvores de classificação ou outras, com a automação do processo de aquisição de imagens e extração de características. Este seria o passo mais importante para a criação de uma ferramenta inteligente, que pudesse passar por todas as etapas do processo de avaliação e dar um laudo com base em probabilidades, sobre qual é o impacto do branqueamento na qualidade do produto.

REFERÊNCIAS

ABREU, F. G. ANÁLISE DE POLIFLUORETO DE VINILIDENO (PVDF) SOB CARREGAMENTO CÍCLICO. p. 139, mar. 2012b.

ADDIEGO, F. et al. Characterization of volume strain at large deformation under uniaxial tension in high-density polyethylene. *Polymer*, v. 47, n. 12, p. 4387–4399, maio 2006.

ANP. AGÊNCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS - ANP Escritório Central. p. 263, 2017.

API-17J. API 17J: Specification for Unbonded Flexible Pipe. n. thirty edition, p. 85, 1 jan. 2009.

BESSON, J.; STEGLICH, D.; BROCKS, W. Modeling of crack growth in round bars and plane strain specimens. *International Journal of Solids and Structures*, v. 38, n. 46–47, p. 8259–8284, nov. 2001.

BOUGHORBEL, S.; JARRAY, F.; EL-ANBARI, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, v. 12, n. 6, p. e0177678, 2 jun. 2017.

CAMILO, C. O. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. p. 29, 1 ago. 2009.

CARVALHO, D. R.; DALLAGASSA, M. R. Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas. *AtoZ: novas práticas em informação e conhecimento*, v. 3, n. 2, p. 82, 31 dez. 2014.

CASTAGNET, S.; DEBURCK, Y. Relative influence of microstructure and macroscopic triaxiality on cavitation damage in a semi-crystalline polymer. *Materials Science and Engineering: A*, v. 448, n. 1–2, p. 56–66, mar. 2007.

CASTAGNET, S.; GACOUGNOLLE, J.-L.; DANG, P. Correlation between macroscopical viscoelastic behaviour and micromechanisms in strained a polyvinylidene fluoride (PVDF). *Materials Science and Engineering A*, p. 8, 2000.

CHAVES, E. É. G. EFEITO DO ENVELHECIMENTO E DA GERAÇÃO DE “WHITENING” NO COMPORTAMENTO MECÂNICO DO PVDF. p. 67, 1 out. 2017.

D20-ASTM COMMITTEE. Test Method for Tensile Properties of Plastics. [s.l.] ASTM International, 2014. Disponível em: <<http://www.astm.org/cgi-bin/resolver.cgi?D638-14>>. Acesso em: 19 ago. 2019.

DAVIES, R. H. et al. A minimum description length approach to statistical shape modeling. *IEEE Transactions on Medical Imaging*, v. 21, n. 5, p. 525–537, maio 2002.

DE SOUSA, E. E. ANÁLISE DO COMPORTAMENTO EM FADIGA DO POLIFLUORETO DE VINILIDENO (PVDF). p. 59, 1 nov. 2011.

DOS SANTOS, W. N.; IGUCHI, C. Y.; JR., R. G. DETERMINAÇÃO DAS PROPRIEDADES TÉRMICAS DO POLIFLUORETO DE VINILIDENO. *Anais do 9 Congresso Brasileiro de Polímeros*, n. 9, p. 6, 2007.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMITH, P. From Data Mining to Knowledge Discovery in Databases. p. 18, 1996.

FIJI. Fiji Is Just ImageJ. [s.l.] GNU General Public License., 2017.

GHELLERE, J. S. Detecção de objetos em imagens por meio da combinação de descritores locais e classificadores. p. 90, 12 jun. 2015.

GONZALEZ, R.; WOODS, R. *Digital Image Processing*. 2. ed. [s.l.] Prentice Hall, 2001.

HAN, J.; KAMBER, M.; PEI, J. *Data mining: Concepts and Techniques*. Third edition ed. [s.l.] Morgan Kaufmann Publishers, 2016.

IGNACIO, J. Processamento e análise digital de imagens em estudos da cinética de recristalização de ligas Al-Mg-X Instituto de pesquisas energéticas e nucleares - Autarquia associada à universidade de São Paulo, 2013.

ISO 37. RUBBER, VULCANIZED OR THERMOPLASTIC -- DETERMINATION OF TENSILE STRESS-STRAIN PROPERTIES. [s.l.] ISO/TC 45/SC 2, nov. 2017.

IVANCIUC, O. Weka Machine Learning for Predicting the Phospholipidosis Inducing Potential. *Current Topics in Medicinal Chemistry*, v. 8, n. 18, p. 1691–1709, 1 dez. 2008.

JERJEN, I. et al. Detection of stress whitening in plastics with the help of X-ray dark field imaging. *Polymer Testing*, v. 32, n. 6, p. 1094–1098, set. 2013.

LAIARINANDRASANA, L. et al. Temperature dependent mechanical behaviour of PVDF: Experiments and numerical modelling. *International Journal of Plasticity*, v. 25, n. 7, p. 1301–1324, jul. 2009b.

LAROSE, D. T. *Data Mining Methods and Models*. p. 340, 2006.

MAQC CONSORTIUM. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, v. 28, n. 8, p. 827–838, ago. 2010.

MARTINS, M. A. DE L. Otimização da configuração de risers verticais complacentes.pdf. *Structural Engineering and Mechanics*, 9 dez. 2011.

MATTHEWS, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, v. 405, n. 2, p. 442–451, out. 1975.

MONTGOMERY, D. C. *Design and analysis of experiments*. Eighth edition ed. Hoboken, NJ: John Wiley & Sons, Inc, 2013.

MOTTA, E. P. Modelagem Elasto-Viscoplástica do PVDF. p. 122, 28 ago. 2017.

NETO, C.; DE MOURA, C. Análise inteligente de dados em um banco de dados de procedimentos em cardiologia intervencionista. n. Versão corrigida. Resolução CoPGr 6018/11, de 01 de Novembro de 2011., p. 173, 2016.

NOVFLEXIBLES. Qualification report for crack growth analysis of PVDF inner liner, 2008.

NOVFLEXIBLES. Discussion of stress/strain at PVDF liner due contraction, 2014a.

NOVFLEXIBLES. Investigation of whitening effect on over all mechanical integrity of extruded PVDF inner liner, 2014b.

NOVFLEXIBLES. Tension-whitening with PVDF: Results from 14X07 trial order with 8" carcass and PVDF inner liner, 2014c.

NOVFLEXIBLES. Research on bending radius of 6" & 8" Pre-salt pipe with PVDF inner liner during fabrication processes, 2014d.

PARMEZAN, A. R. S. et al. Avaliação de Métodos para Seleção de Atributos Importantes para Aprendizado de Máquina Supervisionado no Processo de Mineração de Dados. p. 67, dez. 2012.

PICCIRILLI, T. MINERAÇÃO DE DADOS APLICADA A CLASSIFICAÇÃO DOS CONTRIBUINTES DO ISS. Goiânia: PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS, 2013.

QUEIROZ, J.; GOMES, H. Introdução ao Processamento Digital de Imagens. *Revista RITA: instruções para preparação de documentos em Word*. v. VIII, 2006.

QUINLAN, J. R. Improved Use of Continuous Attributes in C4.5. *Journal of Artificial Intelligence Research*, v. 4, p. 77–90, 1 mar. 1996.

QUINLAN, R. J. C4.5: Programs for Machine Learning by J. Ross Quinlan. *Machine Learning*, v. 16, n. 3, p. 235–240, 1993.

RHODIA, S. G. Solef® PVDF. Disponível em: <<https://www.rhodia.com.br/pt/mercados-e-produtos/catalogo-de-produtos/solef-pvdf.html>>. Acesso em: 15 jun. 2019.

RYCHLEWSKI, J. ON HOOKE'S LAW. v. 48, No.3, p. pp.303-314, 1984.

SENCADAS, V. et al. α - to β Transformation on PVDF Films Obtained by Uniaxial Stretch. Materials Science Forum, v. 514–516, p. 872–876, maio 2006.

SILVA, L. ANÁLISE DA CAMADA INTERNA DE UM TUBO FLEXÍVEL SUBMETIDO À CARREGAMENTOS EXTERNOS UNIVERSIDADE FEDERAL FLUMINENSE, 2010.

SILVA, V. et al. Mineração de Dados Utilizando Análise de Redes Social para Identificar Tendências de Participação em Aulas Presenciais. In: XXVIII SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO - SBIE (BRAZILIAN SYMPOSIUM ON COMPUTERS IN EDUCATION). Recife, Pernambuco, Brasil: 27 out. 2017 Disponível em: <<http://www.br-ie.org/pub/index.php/sbie/article/view/7677>>. Acesso em: 17 ago. 2019

SOARES, H. F. ANÁLISE DO COMPORTAMENTO MECÂNICO DO POLIFLUORETO DE VINILIDENO (PVDF) EM ELEVADAS TEMPERATURAS. p. 74, 2 set. 2015.

TEE, K. F.; KHAN, L. R.; CHEN, H.-P. Probabilistic failure analysis of underground flexible pipes. Structural Engineering and Mechanics, v. 47, n. 2, p. 167–183, 25 jul. 2013b.

VONK, C. G. A small angle X-ray scattering study of polyethylene fibres, using the two-dimensional correlation function. Colloid and Polymer Science, v. 257, n. 10, p. 1021–1032, out. 1979.

WEKA. WEKA - Waikato Environment for Knowledge Analysis. [s.l: s.n.].

WITTEN, I. H. et al. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques. v. Fourth, 2016.

WU, P. D.; VAN DER GIESSEN, E. On neck propagation in amorphous glassy polymers under plane strain tension. International Journal of Plasticity, v. 11, n. 3, p. 211–235, jan. 1995.

ZHANG, Y. et al. State of the Art Analytical Tools Improve Optimization of Unbonded Flexible Pipes for Deepwater Environments. Offshore Technology Conference. Anais... In: OFFSHORE TECHNOLOGY CONFERENCE. Houston, Texas: Offshore Technology Conference, 2003b Disponível em: <<http://www.onepetro.org/doi/10.4043/15169-MS>>. Acesso em: 17 ago. 2019

APÊNDICE A – CARACTERÍSTICAS DO PVDF

Tabela 23 - Modulo de Young em diferentes temperaturas

Temperatura [°C]	Modulo de Young [MPa]
-20	2550
0	2047
23	1583
70	636
130	67.6

Fonte: (RHODIA, 2019)

Tabela 24 - Coeficiente de expansão térmica

Temperatura [°C]	[°C ⁻¹]
-20	139×10^{-6}
+20	196×10^{-6}
+60	232×10^{-6}
+100	280×10^{-6}

Fonte: (RHODIA, 2019)

Tabela 25 - Outras propriedades relevantes

Condutividade térmica [W/(m*K)]	Resistência à tração [MPa]	Coeficiente de Poisson	Capacidade térmica [J/gK]
0.178	>25	0.40	1.23

Fonte: (RHODIA, 2019)

Tabela 26 – Temperatura de serviço

Condição de serviço	Temperatura de serviço (estado sólido) [°C]
Mínima	-30
Máxima	+130

Fonte: (RHODIA, 2019), (DOS SANTOS; IGUCHI; JR., 2007)

APÊNDICE B - ÁRVORES DE DECISÃO GERADAS NO FORMATO TEXTO

Resultados para as diferentes árvores de decisão criadas pelo algoritmo J48 em diversas condições.

Parâmetros considerados: (DI, OD, PC, TR, H, W, A, PE, AV, MA, MI, PI, QP, Class)

<pre> PC <= 16.38: 1 (4.0) PC > 16.38 PI <= 2.2793 PI <= 0.7851: 0 (15.0) PI > 0.7851 PI <= 0.8802: 1 (8.0) PI > 0.8802 PI <= 2.0886: 0 (11.0) PI > 2.0886: 1 (6.0) PI > 2.2793: 0 (16.0) Number of Leaves : 6 Size of the tree : 11 </pre>	<pre> TR = 1 MI <= 108 PI <= 0.8802: 1 (5.0) PI > 0.8802: 0 (2.0) MI > 108: 0 (28.0) TR = 0 MI <= 117: 0 (6.0) MI > 117: 1 (7.0) Number of Leaves : 5 Size of the tree : 9 </pre>
<p>Árvore gerada aplicando treinamento em todo o <i>dataset</i></p>	<p>Árvore gerada aplicando 70% treinamento, 30% teste.</p>

APÊNDICE C – ALGORITMOS GERADOS PELO WEKA

Algoritmo gerado pelo modelo inicial:

```

=== Source code ===
// Generated with Weka 3.8.3
//
// This code is public domain and comes with no warranty.
//
// Timestamp: Thu Aug 29 15:54:23 BRT 2019

package weka.classifiers;

import weka.core.Attribute;
import weka.core.Capabilities;
import weka.core.Capabilities.Capability;
import weka.core.Instance;
import weka.core.Instances;
import weka.core.RevisionUtils;
import weka.classifiers.Classifier;
import weka.classifiers.AbstractClassifier;

public class WekaWrapper
  extends AbstractClassifier {

  /**
   * Returns only the toString() method.
   *
   * @return a string describing the classifier
   */
  public String globalInfo() {
    return toString();
  }

  /**
   * Returns the capabilities of this classifier.
   *
   * @return the capabilities
   */
  public Capabilities getCapabilities() {
    weka.core.Capabilities result = new weka.core.Capabilities(this);

    result.enable(weka.core.Capabilities.Capability.NOMINAL_ATTRIBUTES);
    result.enable(weka.core.Capabilities.Capability.NUMERIC_ATTRIBUTES);
    result.enable(weka.core.Capabilities.Capability.DATE_ATTRIBUTES);
    result.enable(weka.core.Capabilities.Capability.MISSING_VALUES);
    result.enable(weka.core.Capabilities.Capability.NOMINAL_CLASS);
    result.enable(weka.core.Capabilities.Capability.MISSING_CLASS_VALUES);

    result.setMinimumNumberInstances(0);

    return result;
  }

  /**
   * only checks the data against its capabilities.
   *
   * @param i the training data
   */
  public void buildClassifier(Instances i) throws Exception {
    // can classifier handle the data?
    getCapabilities().testWithFail(i);
  }

  /**
   * Classifies the given instance.
   *
   * @param i the instance to classify
   * @return the classification result
   */
  public double classifyInstance(Instance i) throws Exception {
    Object[] s = new Object[i.numAttributes()];

```

```

for (int j = 0; j < s.length; j++) {
    if (!i.isMissing(j)) {
        if (i.attribute(j).isNominal())
            s[j] = new String(i.stringValue(j));
        else if (i.attribute(j).isNumeric())
            s[j] = new Double(i.value(j));
        }
    }

// set class value to missing
s[i.classIndex()] = null;

return WekaClassifier.classify(s);
}

/**
 * Returns the revision string.
 *
 * @return the revision
 */
public String getRevision() {
    return RevisionUtils.extract("1.0");
}

/**
 * Returns only the classnames and what classifier it is based on.
 *
 * @return a short description
 */
public String toString() {
    return "Auto-generated classifier wrapper, based on weka.classifiers.trees.J48 (generated with Weka 3.8.3).\n" +
        this.getClass().getName() + "/WekaClassifier";
}

/**
 * Runs the classifier from commandline.
 *
 * @param args the commandline arguments
 */
public static void main(String args[]) {
    runClassifier(new WekaWrapper(), args);
}
}

class WekaClassifier {

    public static double classify(Object[] i)
        throws Exception {

        double p = Double.NaN;
        p = WekaClassifier.N7aa4b5160(i);
        return p;
    }

    static double N7aa4b5160(Object []i) {
        double p = Double.NaN;
        if (i[3] == null) {
            p = 1;
        } else if (i[3].equals("1")) {
            p = WekaClassifier.N2820dc3c1(i);
        } else if (!i[3].equals("1")) {
            p = WekaClassifier.Na8a68834(i);
        }
        return p;
    }

    static double N2820dc3c1(Object []i) {
        double p = Double.NaN;
        if (i[10] == null) {
            p = 0;
        } else if (((Double) i[10]).doubleValue() <= 108.0) {
            p = 0;
        } else if (((Double) i[10]).doubleValue() > 108.0) {
            p = WekaClassifier.N708d9e892(i);
        }
        return p;
    }

    static double N708d9e892(Object []i) {

```

```
double p = Double.NaN;
if (i[1] == null) {
    p = 1;
} else if (((Double) i[1]).doubleValue() <= 249.2) {
    p = WekaClassifier.N55597b3b3(i);
} else if (((Double) i[1]).doubleValue() > 249.2) {
    p = 1;
}
return p;
}
static double N55597b3b3(Object []i) {
    double p = Double.NaN;
    if (i[1] == null) {
        p = 1;
    } else if (((Double) i[1]).doubleValue() <= 2.0886) {
        p = 1;
    } else if (((Double) i[1]).doubleValue() > 2.0886) {
        p = 1;
    }
    return p;
}
static double Na8a68834(Object []i) {
    double p = Double.NaN;
    if (i[10] == null) {
        p = 1;
    } else if (((Double) i[10]).doubleValue() <= 117.0) {
        p = 1;
    } else if (((Double) i[10]).doubleValue() > 117.0) {
        p = 0;
    }
    return p;
}
}
```

APÊNDICE D - ARQUIVO .ARFF UTILIZADO NO TRABALHO

1) *data_table_weka_current_AP_RP.arff*

@Relation Whitening_classifier

@attribute DI real
 @attribute OD real
 @attribute PC real
 @attribute TR {Sim,Não}
 @attribute H real
 @attribute W real
 @attribute A real
 @attribute PE real
 @attribute AV real
 @attribute MA real
 @attribute MI real
 @attribute PI real
 @attribute QP real
 @attribute Class {Aprovado,Reprovado}

@data

9,262.8,22.93,Não,11.8,203.5,2139.2,418,230.1,138,109,11.59,287354.67,Reprovado
 9,262.8,22.93,Não,8,154.2,1403.8,368,224.5,130,112,11.59,188569.79,Reprovado
 9,262.8,22.93,Não,10.4,156.1,1589.4,325.3,237.8,140,108,11.59,213501.08,Reprovado
 9,262.8,22.93,Não,19.1,76.5,1089.4,162.2,115.2,148,79,5.0686,27987.46,Reprovado
 9,262.8,22.93,Não,9.6,144.8,1052,302.4,145.2,179,115,3.74,14714.96,Reprovado
 8,230.8,16.38,Não,21.3,168,2284.3,360.3,166.8,190,137,2.2622,11690.02,Aprovado
 8,230.8,16.38,Não,10.7,190.9,1322.2,402.4,186.5,216,154,2.4237,7767.03,Aprovado
 8,230.8,16.38,Não,9.9,222.1,2129.8,405.3,183.1,224,155,2.4237,12511.13,Aprovado
 8,230.8,16.38,Não,16.5,237.5,2209.1,508.2,175.3,197,138,2.4237,12976.96,Aprovado
 6,190.4,28.29,Não,10.7,75.4,413.5,148.7,130.6,171,83,9.9027,40549.24,Reprovado
 8,249.2,36.14,Não,11.1,110.8,740.5,245.1,159.9,173,145,2.1672,3477.95,Aprovado
 8,249.2,36.16,Sim,6.5,187.9,1012.5,399.6,162.2,174,136,1.0032,1018.99,Reprovado
 8,249.2,36.16,Sim,14.2,205,2220.7,486.9,154.5,178,82,1.0032,2234.94,Reprovado
 8,249.2,36.16,Sim,13.7,204.4,1398,423,150.1,174,68,1.0032,1406.96,Reprovado
 8,249.2,36.16,Sim,11.7,123.8,1308.2,264,103.4,118,84,3.7806,18698.02,Reprovado
 8,249.2,36.16,Sim,21.8,171.1,3057.7,382.2,135.1,150,110,2.6966,22234.53,Reprovado
 8,249.2,36.16,Sim,17.9,149.4,1905.7,341.6,135.8,150,109,2.6966,13857.59,Reprovado
 8,249.2,36.16,Sim,14.9,131.7,1424.3,282.5,135.7,149,116,2.6966,10357.01,Reprovado
 8,256.4,47.49,Não,17.1,82,1440.7,217.2,210.2,218,200,2.2793,7484.74,Aprovado
 8,256.4,47.49,Não,22.8,122.3,1915.6,340,210,216,201,2.2793,9951.94,Aprovado
 8,256.4,47.49,Não,11.5,108.8,1093.3,258.2,205.8,214,198,2.2793,5679.92,Aprovado
 8,256.4,47.49,Não,15.9,133.9,1443.1,306.5,200.9,207,194,2.2793,7497.21,Aprovado
 8,256.4,47.49,Sim,15.1,158.8,2067.1,368,142.4,155,115,1.8137,6799.74,Reprovado
 8,249.2,36.16,Sim,16,172.2,3134,407.1,149.3,234,106,0.8802,2428.07,Aprovado
 8,249.2,36.16,Sim,16.3,175,2335,407.9,144.4,208,104,0.8802,1809.05,Aprovado
 8,249.2,36.16,Sim,15.3,175.6,2098.8,389.6,145.2,204,106,0.8802,1626.05,Aprovado
 8,249.2,36.16,Sim,13,172,1712.9,355.3,142.1,196,107,0.8802,1327.07,Aprovado
 8,249.2,36.16,Sim,13.8,164.4,2424.1,377.2,140.4,199,107,0.8802,1878.08,Aprovado
 8,249.2,36.16,Sim,13.4,148.5,1820,360.3,128.4,203,105,0.8802,1410.05,Aprovado
 6,189.2,26.72,Sim,13.7,118.3,1079.1,242,139.7,111,147,4.3716,20622.56,Reprovado
 6,189.2,26.72,Não,7,56.5,313.1,117.1,186.6,200,171,0.8626,232.97,Aprovado
 6,189.2,26.72,Não,9.3,105,932.2,236.3,178.3,192,159,0.8129,616,Aprovado
 8,249.2,36.14,Sim,15.8,194,198.9,419.9,198.9,219,171,2.0886,867.65,Reprovado
 8,249.2,36.14,Sim,12.5,68.9,1299.4,201.8,185.2,199,169,2.0886,5668.31,Reprovado
 8,249.2,36.14,Sim,11.9,79.4,957.1,173.8,173.4,186,161,0.3906,146.02,Reprovado
 8,249.2,36.14,Sim,15.4,213.7,4365.4,457.2,199.1,213,179,0.7851,2690.75,Reprovado
 8,249.2,36.14,Sim,16.6,210,3549.4,443,195.6,210,175,0.7851,2187.79,Reprovado
 8,249.2,36.14,Sim,15.8,204.7,2967.1,424.6,194.5,208,176,0.7851,1828.87,Reprovado
 8,249.2,36.14,Sim,17.5,209.7,3546.2,429.2,187.4,201,172,0.7851,2185.81,Reprovado
 8,249.2,36.14,Sim,21,213.2,3966.4,502.2,168.9,181,155,0.7851,2444.82,Reprovado
 8,249.2,36.14,Sim,19.8,210.2,3495.9,502.2,166.2,179,150,0.7851,2154.81,Reprovado
 8,249.2,36.14,Sim,14.7,205.4,2512,475.1,163.1,179,141,0.7851,1548.35,Reprovado
 8,249.2,36.14,Sim,15.1,176.7,2398.4,401.9,192.2,206,162,0.63,951.92,Reprovado
 8,249.2,36.14,Sim,11.1,176.8,2534.4,399.7,194.7,209,165,0.63,1005.9,Reprovado
 8,249.2,36.14,Sim,14.3,171.6,2259.8,375.3,198,210,169,0.63,896.91,Reprovado
 8,249.2,36.14,Sim,17.5,166,2867,360.5,201.3,214,175,0.63,1137.91,Reprovado
 8,249.2,36.14,Sim,20.7,151.3,2884.6,344.1,202.8,214,178,0.63,1144.9,Reprovado
 8,249.2,36.14,Sim,19.9,138,3207.1,312.3,205.1,216,183,0.63,1272.9,Reprovado
 8,249.2,36.14,Sim,23.8,79.4,1307.5,172.3,205.4,217,189,0.63,518.95,Reprovado
 6,186.8,23.57,Não,9.3,86.7,896.3,212.1,125.1,134,112,1.8362,3021.99,Reprovado
 6,186.8,23.57,Não,11.3,107.7,982.6,229.6,126.1,135,117,1.8362,3312.96,Reprovado

8,256.4,47.49,Sim,10.8,191.7,1438.8,409.9,214.3,227,185,2.5429,9303.77,Reprovado
 8,256.4,47.49,Sim,14.8,207.2,2029.7,416.7,216,227,183,2.5429,13124.73,Reprovado
 8,256.4,47.49,Sim,14.5,205.9,1436.5,431.4,218.1,230,183,2.5429,9288.9,Reprovado
 6,197.2,36.03,Sim,9.6,79.5,620.3,1806,181.4,191,158,2.2252,3071.42,Aprovado
 8,256.4,47.49,Sim,11.8,182.7,2205.1,400.9,213.3,223,199,5.1253,57925.11,Reprovado
 8,256.4,47.49,Sim,13.5,136.8,1834.2,316.4,225.1,236,178,2.4493,11003.5,Reprovado
 8,256.4,47.49,Sim,17.7,49.7,994.7,137,215.1,206,221,1.5601,2421.01,Reprovado
 8,256.4,47.49,Sim,29.8,99.5,1937.3,249.9,217.8,223,208,1.5601,4715.22,Reprovado
 8,256.4,47.49,Sim,27.6,102.6,1566.3,225.9,217.6,224,207,1.5601,3812.24,Reprovado

2) *test_data.arff*

@Relation Whitening_classifier_test_data

@attribute DI real

@attribute OD real

@attribute PC real

@attribute TR {Sim,Não}

@attribute H real

@attribute W real

@attribute A real

@attribute PE real

@attribute AV real

@attribute MA real

@attribute MI real

@attribute PI real

@attribute QP real

@attribute Class {Aprovado,Reprovado}

@data

8,249.2,36.14,Não,23.4,133.8,1663.6,309.7,196.7,201,111,0.4300,3910.67,?

8,249.2,36.14,Não,17.7,119.0,1818.6,289.5,145.0,165,125,0.6700,2713.11,?

APÊNDICE E – IMAGENS DAS AMOSTRAS #01 E #02

Imagem do duto que gerou a amostra #01:



Imagem do duto que gerou a amostra #02:

