UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
MESTRADO EM ENGENHARIA DE PRODUÇÃO

JOVANI TAVEIRA DE SOUZA

MÉTODOS DE SELEÇÃO DE ATRIBUTOS E ANÁLISE DE COMPONENTES PRINCIPAIS: UM ESTUDO COMPARATIVO

DISSERTAÇÃO

PONTA GROSSA
2017

JOVANI TAVEIRA DE SOUZA

MÉTODOS DE SELEÇÃO DE ATRIBUTOS E ANÁLISE DE COMPONENTES PRINCIPAIS: UM ESTUDO COMPARATIVO

Dissertação apresentada como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, do Programa de Pós-Graduação em Engenharia de Produção, da Universidade Tecnológica Federal do Paraná – Campus Ponta Grossa.

Orientador: Prof. Dr. Antonio Carlos de Francisco

Coorientadora: Prof.ª Dr.ª Maria Helene Giovanetti Canteri

PONTA GROSSA

2017

Ficha catalográfica elaborada pelo Departamento de Biblioteca
da Universidade Tecnológica Federal do Paraná, Campus Ponta Grossa
n.31/17

S729 Souza, Jovani Taveira de

Métodos de seleção de atributos e análise de componentes principais: um estudo comparativo. / Jovani Taveira de Souza. 2017.
73 f. : il. ; 30 cm.

Orientador: Prof. Dr. Antonio Carlos de Francisco
Coorientadora: Profa. Dra. Maria Helene Giovanetti Canteri

Dissertação (Mestrado em Engenharia de Produção) - Programa de Pós-Graduação em Engenharia de Produção. Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2017.

1. Câncer - Diagnóstico. 2. Expressão gênica. 3. Engenharia de produção. I. Francisco, Antonio Carlos de. II. Canteri, Maria Helene Giovanetti. III. Universidade Tecnológica Federal do Paraná. IV. Título.

CDD 670.42

Universidade Tecnológica Federal do Paraná
Campus Ponta Grossa
Diretoria de Pesquisa e Pós-Graduação
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

FOLHA DE APROVAÇÃO

Título de Dissertação N°**304/2017**

MÉTODOS DE SELEÇÃO DE ATRIBUTOS E ANÁLISE DE COMPONENTES PRINCIPAIS: UM ESTUDO COMPARATIVO

por

Jovani Taveira de Souza

Esta dissertação foi apresentada às **11h00min** de **28 de abril de 2017** como requisito parcial para a obtenção do título de MESTRE EM ENGENHARIA DE PRODUÇÃO, com área de concentração em Gestão Industrial, Programa de Pós-Graduação em Engenharia de Produção. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. Dr. Roquemar de Lima Baldan
(IFES)

Prof. Dr. Antonio Carlos de Francisco
(UTFPR) - *Orientador*

Prof. Dr. Cassiano Moro Piekarski
(UTFPR)

Visto do Coordenador:

Antonio Carlos de Francisco (UTFPR)
Coordenador do PPGEP

A FOLHA DE APROVAÇÃO ASSINADA ENCONTRA-SE ARQUIVADA NA SECRETARIA ACADÊMICA DA UTFPR – CÂMPUS PONTA GROSSA

“Nada na vida deve ser temido, somente compreendido. Agora é hora de compreender mais para temer menos”.
(Maria Curie)

AGRADECIMENTOS

Quero agradecer, em primeiro lugar, a Deus, pelas bênçãos, inspirações e forças a mim concedidas.

Aos meus pais, por toda orientação, amor e educação.

À minha namorada, por todo apoio, amor e compreensão.

Ao meu orientador, pelas oportunidades, amizades e orientações.

À minha coorientadora, pela disponibilidade, diálogos e orientações.

Aos amigos e colegas do grupo de pesquisa LESP, pelo incentivo e pelo apoio constante.

Aos professores participantes das bancas de qualificação e defesa da dissertação, pela disponibilidade e críticas construtivas.

Ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Tecnológica Federal do Paraná, incluindo docentes, servidores, alunos e colaboradores.

À CAPES pelo apoio financeiro através da bolsa de estudo para o desenvolvimento deste trabalho.

Enfim, a todos os que por algum motivo contribuíram para a realização desta pesquisa.

RESUMO

SOUZA, Jovani Taveira de. **Métodos de seleção de atributos e análise de componentes principais:** um estudo comparativo. 2017. 73 f. Dissertação (Mestrado em Engenharia de Produção) - Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2017.

A neoplasia é um grande desafio para os pesquisadores devido a sua alta complexidade. Apesar dos avanços em diagnósticos, os estudos apontam que, além da análise de dados, são necessários métodos que otimizem e auxiliem o processo de tomada de decisão. Neste sentido, a redução de dimensionalidade de dados tem contribuído significativamente, auxiliando nesse processo, devido à quantidade de genes (atributos), ser muito ampla comparada ao número de amostras (classes). Este trabalho, portanto, visa fornecer um estudo comparativo entre dois métodos de redução de dimensionalidade, aplicados em três bases de dados no domínio de expressão gênica: LungCancer-Michigan, LungCancer-Ontario e LungCancer-Harvard, todas relacionadas ao câncer de pulmão. Os métodos aplicados foram: Seleção de Atributos e Análise de Componentes Principais (PCA), ambos usados como uma etapa de pré-processamento na Mineração de Dados. Os algoritmos de classificação escolhidos foram: Naive Bayes, SVM, J48, 1-NN, 3-NN, 5-NN e 7-NN. Foi utilizado o Weka como *software* para procedimentos de análise. Uma série de experimentos foi realizada para avaliar a acurácia e aplicabilidade dos algoritmos para ambos os métodos. Como resultado, foram evidenciados avanços significativos nas taxas de acerto (acurácia) dos classificadores envolvendo os métodos empregados, utilizando como critério de avaliação a Validação Cruzada. A abordagem *Wrapper*, do método de Seleção de Atributos, obteve os melhores resultados para as três bases de dados analisadas. O método de Análise de Componentes Principais, mesmo apresentando taxa de acerto inferior, não pode ser descartado. Os algoritmos Naive Bayes, SVM e 1-NN foram os que apresentaram melhor desempenho dentre as bases. Foram denotados os atributos (genes) que apresentaram maior frequência nas bases de dados. Portanto, a partir dos subconjuntos escolhidos, estes podem ser submetidos a análises específicas, no intuito de direcionar diagnósticos mais precisos.

Palavras-chave: Redução de Dimensionalidade. Seleção de Atributos. Análise de Componentes Principais. Expressão Gênica.

ABSTRACT

SOUZA, Jovani Taveira de. **Methods of attribute selection and principal component analysis:** a comparative study. 2017. 73 f. Dissertation (Master in Production Engineering) - Federal Technology University - Parana. Ponta Grossa, 2017.

Neoplasm is a major challenge for researchers because of its high complexity. Despite advances in diagnosis, studies point out that in addition to data analysis, methods to optimize and aid the decision-making process are necessary. In this sense, the dimensionality reduction of data has contributed significantly, helping in this process, due to the large number of genes (attributes) compared to the number of samples (classes). This work, therefore, aims to provide a comparative study between two methods of dimensionality reduction, applied to three databases in the field of gene expression: LungCancer-Michigan, LungCancer-Ontario and LungCancer-Harvard, all related to lung cancer. The methods applied were: Attribute Selection and Principal Component Analysis (PCA), both used as a pre-processing step in Data Mining. The classification algorithms chosen were Naive Bayes, SVM, J48, 1-NN, 3-NN, 5-NN and 7-NN. Weka was used as a software for analyses procedures. A series of experiments was performed to evaluate the accuracy and applicability of the algorithms for both methods. As a result, significant advances in the hit rate (accuracy) of the classifiers involving the methods were evidenced, using Cross-Validation as the assessment criterion. The Wrapper approach, from the Attribute Selection method, obtained the best results for the three analyzed databases. The Principal Component Analysis method, even presenting lower hit rate, could not be ruled out. The Naive Bayes, SVM and 1-NN algorithms presented the best performance within the databases. The attributes (genes) which presented the highest frequency in the databases were denoted. Therefore, from the chosen subsets, these can be submitted to specific analyzes in order to direct more precise diagnoses.

Keywords: Dimensionality Reduction. Attribute Selection. Principal Component Analysis. Gene Expression.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS E ACRÔNIMOS

cDNA	Complementary DNA
CFS	Correlation-based Feature Selection
CSE	Consistency Subset Eval
DNA	Deoxyribonucleic Acid
DRM-F	Dimensionality Reduction Method – Framework
ISOMAP	Isometric Feature Mapping
KDD	Knowledge Discovery in Databases
KNN	K – Nearest Neighbors
LLE	Locally Linear Embedding
mRNA	Messenger Ribonucleic Acid
NCBI	National Center for Biotechnology Information
NSCLC	Non Small Cell Lung Cancer
PCA	Principal Component Analysis
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SCLC	Small Cell Lung Cancer
SVM	Support Vector Machines
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

SUMÁRIO

1 INTRODUÇÃO

Os estudos sobre neoplasia têm instigado pesquisadores na descoberta por novos métodos e maneiras de combater essa patologia. O conhecimento acerca das bases de dados que envolvem essa doença é um grande desafio, pois as informações contidas nessas bases se acumulam e aumentam exponencialmente, o que torna difícil e complexo lidar com esse enorme volume de dados (ZOU et al., 2015)

Os riscos de se desenvolver neoplasia geralmente são devidos a dois fatores: ambientais e hereditários (INCA, 2011). No que se refere às condições ambientais, pode-se afirmar que a exposição a determinados agentes ambientais pode desencadear a predisposição ao desenvolvimento do câncer. Já na questão fatores hereditários, as neoplásicas decorrem justamente por características herdadas geneticamente. Além disso, fatores genéticos podem influenciar significativamente a probabilidade de desenvolver cânceres induzidos pelo meio ambiente (KUMAR; ASTER; ABBAS, 2015).

As doenças oncológicas são constituídas por um conjunto com mais de cem patologias que possuem em comum o crescimento desordenado de células. Essa “massa” invade tecidos e órgãos e, posteriormente, pode se tornar agressiva e incontrolável, ocasionando a formação de tumores ou neoplasias malignas (INCA, 2011). Essas células cancerosas ou células neoplásicas são células anormais decorrentes de alteração no ácido desoxirribonucleico (DNA). Portanto, a compreensão dessas células torna-se necessária pois, a partir disso, será possível diagnosticar, com maior precisão, os genes susceptíveis a padrões neoplásicos.

Neste contexto, a utilização da técnica de microarranjos se destaca, em função de que determina, em uma amostra, os padrões de expressão de milhares de genes, simultaneamente (APOLLONI; LEGUIZAMON; ALBA, 2016; MACEDO, 2015). A análise dos níveis de expressão permite identificar os genes alterados de uma determinada doença (LATKOWSKI; OSOWSKI, 2015).

De acordo com os autores Khalilabad e Hassanpour (2017), os desafios enfrentados por esta técnica são a grande quantidade de informações e a necessidade de técnicas estatísticas, tecnologias de informação ou abordagens de aprendizado de máquinas que auxiliem nesse processo.

A dificuldade associada à dimensionalidade direciona a busca por uma representação adequada, menor e relevante de uma base de dados bruta, a fim de

tornar a análise de dados e reconhecimento de padrões mais fácil e eficiente (YIN; HUANG, 2010).

Neste sentido, os métodos empregados na Mineração de Dados (*Data Mining*) têm papel importante, pois exploram grandes quantidades de dados, a fim de encontrar padrões, regras ou dados ocultos em uma determinada base (KAMBER; HAN; PEI, 2012).

Data mining é uma técnica pertinente à área de pesquisa denominada de Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases* – KDD). Dentro deste contexto, a etapa de mineração de dados tem como intuito transformar os dados em conhecimento útil (FAYYAD et al.,1996).

Basicamente, *data mining* é um processo constituído por algumas etapas que podem ser divididas em: seleção; pré-processamento; mineração de dados e pós-processamento, que aplicados conjuntamente, permitem a descoberta do conhecimento (BORGES; NIEVOLA, 2012).

Ainda segundo os autores, neste âmbito, os métodos de redução de dimensionalidade são utilizados, pois permitem reduzir o volume de informações, mais precisamente o número de atributos, eliminando os dados irrelevantes e/ou redundante de uma determinada base de dados. Em mineração de dados, a redução de dimensionalidade é uma questão muito importante no processamento de dados de alta dimensão (BARTL et al., 2011; ZHANG; JIANG; YE, 2010).

Existem vários métodos de redução de dimensionalidade, dentre os quais podem ser identificados: Isomap (*Isometric Feature Mapping*), Análise de Componentes Principais (*Principal Component Analysis* - PCA), Seleção de Atributos, Kernel, Kernel PCA (combinação de PCA com Kernel), *Locally Linear Embedding* (LLE), entre outros. A adoção de alguns desses métodos neste trabalho justifica-se pela natureza do problema a analisar, além da ausência de estudos sobre o mesmo. Nesta perspectiva, foram escolhidos os métodos de Seleção de Atributos e Análise de Componentes Principais.

O método de Seleção de Atributos tem importância fundamental em identificar os atributos relevantes para uma determinada tarefa. Segundo Kira e Rendell (1992), nesse método seleciona-se um subconjunto de atributos relevantes com objetivo de melhorar o processo de aprendizagem, garantindo a qualidade dos dados. De acordo com Macedo (2015), o método irá detectar os dados de microarranjos significativos para o estudo em questão.

A utilização da Seleção de Atributos em estudos envolvendo expressão de genes pode aumentar o entendimento dos resultados produzidos, identificando a influência de cada um dos atributos selecionados (BORGES; NIEVOLA, 2012). São utilizadas duas abordagens no método de seleção de atributos a Filtro e a *Wrapper*, distintas entre si na forma em que os subconjuntos são testados.

Já, a Análise de Componentes Principais (PCA), um método de análise de dados, prevê uma sequência das melhores combinações lineares a partir dos atributos originais de um determinado conjunto, além de ser uma das técnicas mais tradicionais para redução de dimensionalidade (BARSHAN et al., 2011; YIN; HUANG, 2010).

A PCA visa à obtenção de um novo e reduzido conjunto de variáveis não correlacionadas (componentes principais), de tal forma que seja perdido o mínimo possível de informação. De acordo com Prieto-Moreno, Llanes-Santiago e García-Moreno (2015), o método proporciona uma representação dimensional inferior, porém preserva a estrutura de correlação dos dados originais.

Recentemente, houve alguns avanços significativos em bases no domínio de expressão gênica, dentre os quais o estudo de Borges e Nievola (2012), que compararam a eficiência de dois métodos de Redução de Dimensionalidade - seleção de atributos e projeção aleatória - em bases cancerígenas, referentes a linfoma e leucemia. Já Macedo (2015), propôs uma nova metodologia de redução de dimensionalidade, denominado de DRM-F (*Dimensionality Reduction Method – Framework*), e verificou sua aplicabilidade nos mesmos tipos de bases.

Ghosh e Barman (2016) aplicaram cálculo de distância euclidiana e análise de componentes principais para identificação de genes saudáveis e cancerígenos com base na composição dos aminoácidos, em bases de câncer de mama, contudo, utilizando outros padrões de comparação. Latkowski e Osowski (2015) estudaram a aplicação de métodos de mineração de dados em dados associados com o autismo, com objetivo de selecionar os genes mais importantes nesse âmbito.

Considerando outras bases, por exemplo, Sakthivel et al. (2014), comparou o método PCA com métodos não lineares: Kernel PCA, Isomap, *Locally Linear Embedding*, entre outros, que ajudaram a identificar prováveis falhas de uma bomba centrífuga utilizando sinais de vibração.

Existem trabalhos relevantes que utilizam métodos de redução de dimensionalidade, como os relatados acima, tanto em bases de expressão gênica, como em outros tipos de bases, no entanto, de acordo com as buscas realizadas na

literatura, não foi encontrado um estudo comparando esses dois métodos no domínio de expressão gênica.

Sendo assim, visando utilizar os métodos de redução de dimensionalidade, esta pesquisa pretende responder a seguinte questão:

Como realizar a identificação de atributos relevantes através dos métodos de redução de dimensionalidade em bases de dados de expressão gênica?

1.1 OBJETIVOS DA DISSERTAÇÃO

1.1.1 Objetivo Geral

Avaliar a aplicação dos métodos de Seleção de Atributos e Análise de Componentes Principais em dados de expressão gênica.

1.1.2 Objetivos Específicos

- Selecionar as bases de dados referentes ao domínio de expressão gênica;
- Aplicar os métodos de Seleção de Atributos e Análise de Componentes Principais (PCA) nas bases de dados escolhidas;
- Executar os algoritmos de classificação;
- Analisar as taxas de acerto do Método de Seleção de Atributos e do Método PCA;
- Comparar os resultados dos métodos utilizados.

Após os objetivos definidos, serão apresentadas as justificativas para o desenvolvimento da pesquisa.

1.2 JUSTIFICATIVA DA PESQUISA

Pesquisas envolvendo câncer são objetos de estudo em todo o mundo, por conta de sua alta incidência e taxa de mortalidade. De acordo com os estudos, realizada pela *World Health Organization* (Organização Mundial da Saúde), o câncer vitimou cerca de 8,8 milhões de pessoas em todo o mundo, no ano de 2015 (WHO, 2017).

Ainda segundo os estudos da WHO (2017), as causas mais comuns de morte por câncer atingiram o pulmão, com um total 1,69 milhões de mortes, seguido do fígado, com 788 mil mortes, a região colorretal, com 774 mil mortes, estômago com 754 mil mortes e mama, com 571 mil mortes.

No Brasil, as estimativas de incidência de câncer a cada ano são preocupantes. De acordo com o Instituto Nacional de Câncer Alencar Gomes da Silva - INCA (2015), as estimativas de incidência de câncer para o ano de 2016, foram de aproximadamente 295.200 casos novos de todos os tipos de câncer para o sexo masculino, com um risco estimado de 298,13 casos para cada 100 mil homens. Já o sexo feminino, apresenta estimativa de 300.870 casos novos de todos os tipos de câncer, tendo um risco estimado de 291,54 casos para cada 100 mil mulheres. Os números apresentados excluem dados de capitais brasileiras.

Uma das abordagens que auxilia na identificação do câncer é realizada através da técnica de microarranjos, que estuda milhares de genes simultaneamente (KHALILABAD; HASSANPOUR, 2017). Segundo os autores, são produzidas dezenas de imagens, cujo processamento associado à análise dos dados, são de elevada importância. Diante disso, a análise de dados de expressão gênica necessita fazer o uso de métodos eficazes que corroboram para a qualidade dos diagnósticos, assim como conhecimentos quanto à doença analisada (MACEDO, 2015).

A mineração de dados contribui com essa descoberta de conhecimento, pois através de técnicas e ferramentas, ajudam a buscar correlações importantes entre os dados (FAYYAD et al., 1996).

De acordo com Schuch et al. (2010), as técnicas de mineração de dados contribuem significativamente em empresas e organizações, em razão de sua capacidade de gerar informações e produzir o conhecimento, identificando aspectos relevantes que possam ser utilizados em nível estratégico para tomada de decisão.

Com esses métodos de otimização, a análise de expressão gênica se beneficiou, em virtude da possibilidade de estudo, através da identificação de dados característicos em diversas bases de dados, caracterizando genes que podem estar diretamente relacionados com alguma doença.

A análise de dados de expressão gênica permite a descoberta de grupos significativos de genes com funcionalidades relacionadas (NASCIMENTO; TOLEDO; CARVALHO, 2010). Os genes com tais alterações podem aumentar nosso conhecimento do mecanismo de formação da doença, permitindo prever o risco de ser afetado por tal doença (LATKOWSKI; OSOWSKI, 2015).

Sendo assim, a Seleção de Atributos e o PCA, métodos escolhidos para este estudo, podem fornecer maneiras de se obter resultados mais precisos, auxiliando na escolha dos dados realmente significativos. A comparação dos mesmos apresenta-se como um diferencial, pois auxilia na escolha do método mais apropriado para as bases de dados selecionadas.

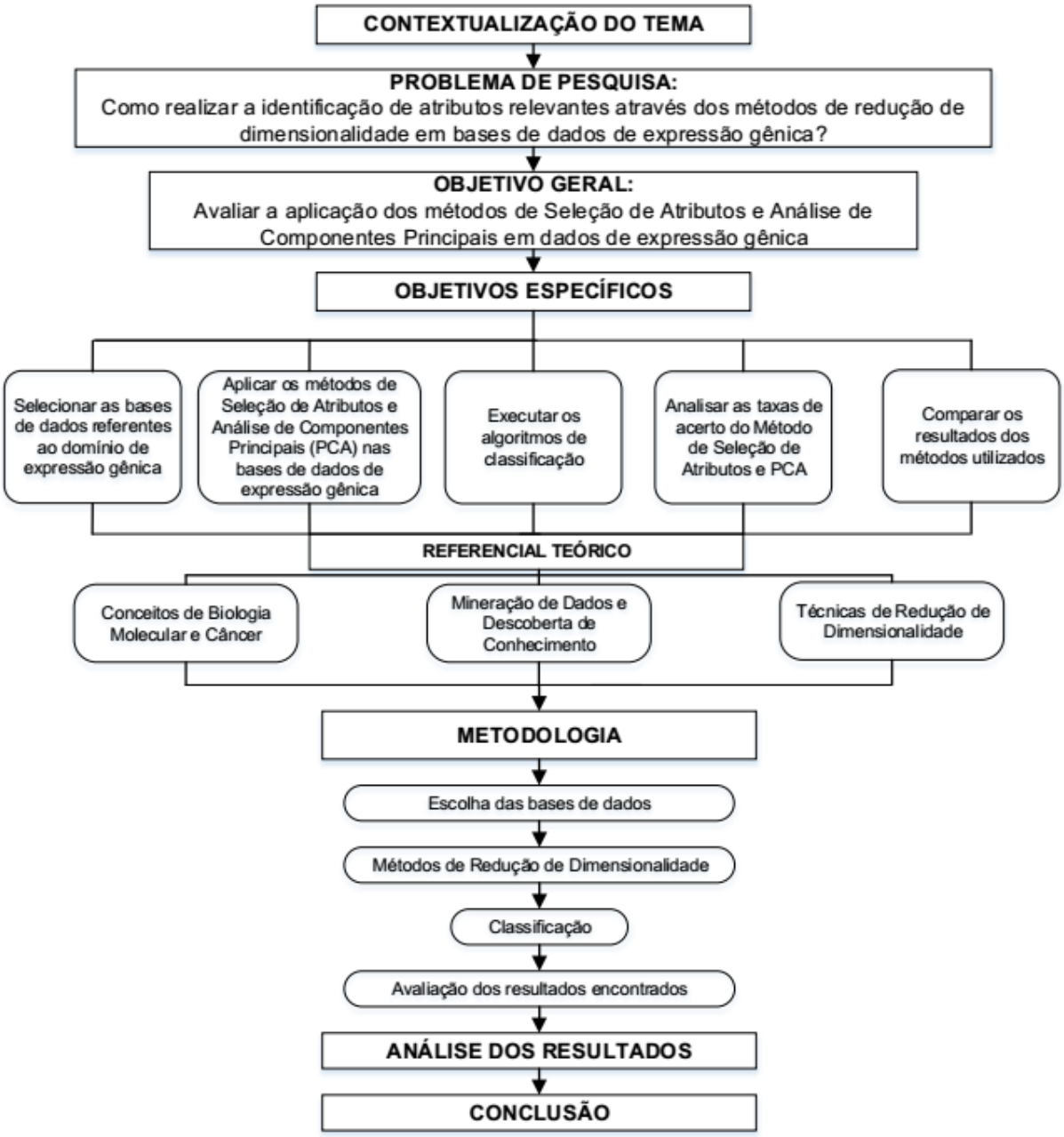
Tendo em vista os conhecimentos adquiridos, permitem-se obter algumas vantagens pelos laboratórios clínicos que usufruem dessa técnica (microarranjos), que poderão prever potenciais surgimentos de neoplasia em indivíduos nos quais os tumores ainda não se manifestaram, bem como analisar e indicar o tratamento mais adequado, além de desenvolver novos produtos, como medicações específicas para as doenças analisadas.

Por fim, estabelecendo uma analogia entre os conceitos referidos com o campo da Engenharia de Produção, podemos dizer que o estudo abarca a área de Gestão do Conhecimento, visto que aborda aspectos referentes à informação, na qual ajuda analisar e trabalhar com dados diversos, ajudando a extrair aquilo que é útil.

1.3 ESTRUTURA DO TRABALHO

Com o intuito de fornecer uma visão geral do desenvolvimento deste estudo, a Figura 1 ilustra os passos delineados para a condução da dissertação.

Figura 1– Passos delineados para a condução da dissertação



Fonte: Autoria própria

2 REFERENCIAL TEÓRICO

Este capítulo retrata a fundamentação ao redor dos estudos envolvendo expressão de genes pela técnica de microarranjos, além de apresentar as definições abrangendo genes, estudados pela Biologia Molecular, câncer e os principais conceitos sobre a Descoberta do Conhecimento em Banco de Dados (*Knowledge Discovery in Databases* – KDD), por meio da Mineração de Dados (*Data Mining*), com suas características, definições e utilidade, juntamente com as Técnicas de Redução de Dimensionalidade, que visam à extração do conhecimento a partir de grandes volumes de dados nas mais diversas aplicações.

2.1 PRINCIPAIS CONCEITOS DE BIOLOGIA MOLECULAR

O estudo em torno da área de Biologia Molecular tem sido de grande interesse para pesquisadores e cientistas, por envolver genes, e ser de alta complexidade em virtude de sua dificuldade de interpretação. Além disso, tem contribuído significativamente para a compreensão de como os genes funcionam quando normais e porque causam doenças quando alterados (ZATZ, 2002). Entender esse processo ajuda consideravelmente no tratamento de doenças posteriores.

Pode-se afirmar que a Biologia Molecular é um ramo que envolve estudos relacionados com as células, moléculas, genes, DNA e proteínas, a partir de suas interações. De acordo com Casley (1992), a Biologia Molecular estuda células e moléculas, blocos básicos na construção de todas as formas de vida. Em particular, estudam-se os genomas dos organismos, definidos como o conjunto de suas informações genéticas.

Segundo Silva (2001), Gregor Mendel foi o primeiro a identificar fatores responsáveis pela hereditariedade nos organismos vivos. Consequentemente, esses fatores foram denominados de genes, os quais correspondem a um segmento de DNA que codifica a informação genética. No intuito de encontrar esses genes, foram identificados os cromossomos, estruturas presentes em todas as células, responsáveis por carregar toda a informação que as células necessitam para seu crescimento, desenvolvimento e produção. Com esses estudos, descobriu-se que os

cromossomos são agrupamentos de genes, constituídos de DNA, e os genes são sequências específicas de DNA.

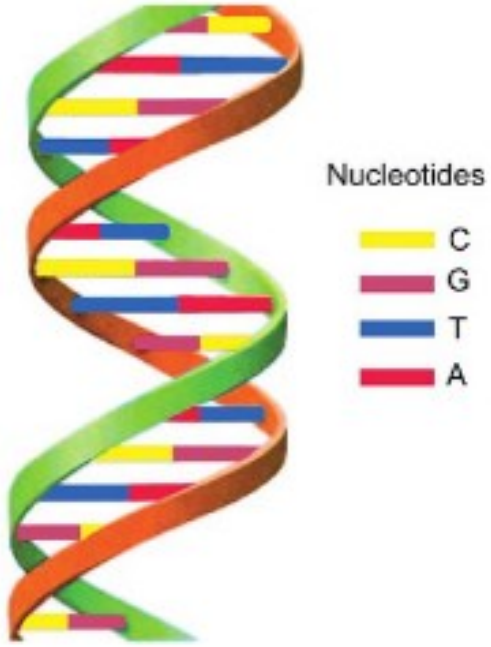
A próxima seção irá abordar os conceitos referentes aos estudos envolvendo genes.

2.1.1 DNA, Expressão Gênica e Proteínas

Em relação aos estudos envolvendo genes, podem ser citados os três aspectos mais importantes para compreensão: DNA, Expressão Gênica e Proteínas. Esses três elementos são a base para o estudo envolvendo genes.

No que se refere à molécula de DNA, esta é formada por duas fitas de nucleotídeos, composta por quatro tipos de bases nitrogenadas essenciais, denominadas de Adenina (A), Citosina (C), Guanina (G) e Timina (T), que se agrupam entre si. Ressalta-se que Adenina só se liga com Timina, e a Citosina só se une à Guanina, formando uma dupla hélice, conforme pode ser visualizado na Figura 2.

Figura 2 – Estrutura do DNA



Fonte: Cole (2016, p.57)

Como referido anteriormente, um segmento de DNA pode conter diversos genes, com a função de codificação de informação genética. Portanto, os genes são as regiões do DNA que codificam uma proteína (LODISH et al., 2014). De maneira

geral, os genes terão como função codificar as proteínas, para que, assim, as proteínas desempenhem sua função no organismo.

No que diz a respeito às proteínas, é importante destacar sua importância na participação de praticamente todos os processos celulares, constituídas de conjuntos de aminoácidos. O processo no qual ocorre a interpretação das sequências de nucleotídeos dos genes na produção de proteínas, é chamado de expressão gênica (BORGES, 2006).

A expressão gênica decorre de dois processos: o primeiro denominado de transcrição, no qual a RNA Polimerase (molécula que transcreve DNA em RNA) se liga a uma região do DNA, através de uma sequência de bases (promotor), gerando uma molécula de RNA mensageiro (mRNA). A segunda etapa é denominada de tradução, que envolve o processo de sintetizar a proteína, utilizando o mRNA. Uma cadeia tripla de nucleotídeos consecutivos (códon) codifica um aminoácido e através desses códons é feito o mapeamento desses aminoácidos, através dos nucleotídeos, processo chamado de código genético (BORGES, 2006; MACEDO, 2015).

Os estudos que envolvem expressão gênica são importantes, pois é possível obter informações primordiais das funções da célula, visto que quando há mudança na fisiologia de um organismo, as alterações referidas são acompanhadas por mudanças nos padrões de expressão de genes (ALBERTS et al., 2010).

A seguir será relatada a técnica que visa à análise de expressão gênica, utilizada nesse trabalho.

2.1.2 Técnica de Microarranjos

Segundo Appelbaum (2015), as técnicas mais comumente utilizadas para análise de expressão gênica são: por meio da Técnica de Microarranjos ou pela Reação em Cadeia da Polimerase via Transcriptase Reversa (*Reverse Transcriptase Polymerase Chain Reaction* – RT-PCR). Ambas podem ser utilizadas com o mesmo propósito, porém a técnica utilizada neste estudo será a de microarranjos de DNA, mais exequível ao que se propõe (base de estudo).

Microarranjo tem sido amplamente utilizada na pesquisa de câncer por mais de uma década (LIU; SO; FAN, 2015), preferencialmente escolhidas para projetos que

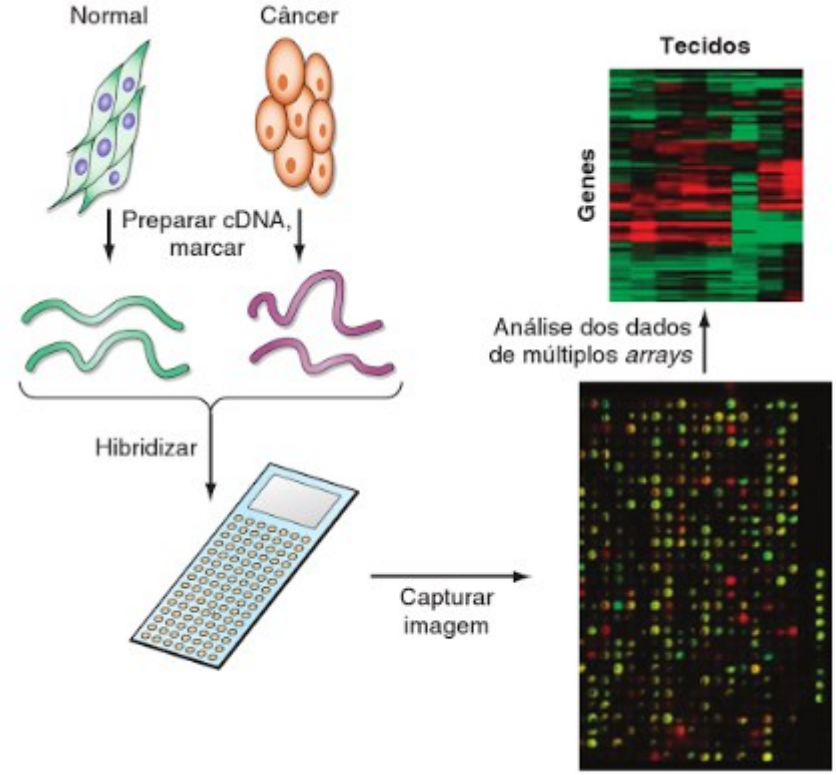
envolvem um grande número de amostras e para a análise de transcriptomas em organismos modelos com genomas bem caracterizados (BAGINSKY et al., 2010).

O estudo da expressão gênica com tais técnicas permite obter um perfil molecular e favorece a identificação de alterações significativas que ocorrem no nível de RNA (BARBOSA et al., 2012).

Utilizando-se da técnica de microarranjo, é possível mapear, em escala genômica, as complexas aberrações moleculares associadas com o desenvolvimento do câncer, base deste estudo, e podem ser correlacionadas com os dados clínicos do paciente, com o objetivo de identificar moléculas alteradas, associadas a algum estágio de câncer ou alguma alteração importante (SORENSEN; ORNTOFT, 2010). A técnica permite também mensurar os níveis de expressão simultaneamente, possibilitando comparar diferentes amostras pelos perfis de expressão.

A Figura 3 apresenta o processo da técnica de microarranjo.

Figura 3 – Processo da Técnica de Microarranjo



Fonte: Morin et al. (2016)

O RNA é organizado a partir das células, transcrito em moléculas de DNA complementar (cDNA) e indicado com corantes fluorescentes, em que geralmente o verde representa as células designadas como normais e o vermelho para as células cancerosas. As sondas fluorescentes são misturadas e hibridizadas, de uma maneira

ordenada (*arrays*) de cDNA onde cada ponto é um fragmento de cDNA (oligonucleotídeo) que representa um gene diferente. A partir disso, a imagem é captada por uma câmera, particularmente sensível à luz fluorescente, e as imagens mostram que os pontos verdes representam expressão menor nas células tumorais, enquanto os pontos vermelhos representam expressão maior. Já os pontos amarelos simbolizam que os níveis de expressão são iguais tanto nas amostras normais quanto nas tumorais (MORIN et al.,2016).

Após o processo acima, os sinais são quantificados, integrados e normalizados a partir de ferramentas computacionais específicas que obtêm o perfil de expressão gênica (COLOMBO; RAHAL, 2010; MORIN et al., 2016).

Os dados de microarranjos são representados geralmente no formato de uma matriz. Segue o formato abaixo:

Figura 4 – Matriz de dados de expressão gênica

Gene 1	Gene 2	. . .	Gene N	
f_{11}	f_{12}	. . .	f_{1N}	c_1
f_{21}	f_{22}	. . .	f_{1N}	c_2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
f_{M1}	f_{M2}	. . .	f_{MN}	c_M

Fonte: Borges (2006)

Cada coluna representa um gene e cada linha representa uma amostra (*C1*), para cada qual a medição do nível de expressão de todos os genes é efetuada. O formato apresentado é semelhante ao que se refere à mineração de dados, pois um gene pode ser representado como atributo ou característica e as amostras como as classes. Geralmente, as bases de dados são formadas por grande quantidade de atributos (genes) e poucas informações referentes às classes (amostras) (BORGES, 2006).

A análise de dados por técnica de microarranjo é um desafio para pesquisadores que visam encontrar os genes mais relevantes de uma amostra e assim construir modelos preditivos, tornando-se necessário encontrar métodos para

validar os modelos, focando principalmente em sua acurácia (SAEYS; INZA; LARRANAGA, 2007).

As próximas seções retratam os conceitos referentes às temáticas da respectiva pesquisa.

2.2 CÂNCER

Os estudos sobre a neoplasia têm aumentado significativamente nos últimos anos, tanto nos países em desenvolvimento como nos países desenvolvidos. O estigma pertinente ao câncer encontra-se desde a Antiguidade. De acordo com a *American Cancer Society* (ACS), aproximadamente metade dos homens e um terço das mulheres do mundo vão desenvolver algum tipo de neoplasia em algum período de sua vida (SOUZA, 2011).

Segundo Voet e Voet (2013), o câncer pode ser considerado como um grupo de doenças caracterizadas por defeitos na transdução de sinal (alterações intra ou extracelular) que causam crescimento desordenado e descontrolado de células. Essas células, segundo os autores, perdem seus controles de crescimento e iniciam uma proliferação excessiva, ocasionando a formação de tumores, designados de dois tipos: tumores benignos e tumores malignos.

Os tumores benignos ou neoplasias benignas apresentam crescimento organizado, em geral, de forma lenta, com limites nítidos. Alguns exemplos desses tumores são o lipoma, o mioma e o adenoma (INCA, 2011). Ocasionalmente, esses tipos de tumores apresentam ameaça à vida, contudo, se esse tipo de tumor ocorre no cérebro, por exemplo, pode ser fatal (VOET; VOET, 2013).

Já os tumores malignos crescem diferentemente dos tumores benignos, de maneira invasiva, invadindo tecidos vizinhos e liberando células, processo denominado de metástase. Esses tipos de tumores representam, quase que constantemente, uma ameaça ao ser humano (INCA, 2011; VOET; VOET, 2013).

2.2.1 Câncer de Pulmão

Neste âmbito, um dos cânceres com maior probabilidade de mortes no mundo é o câncer de pulmão ou carcinoma pulmonar, como também é conhecido, é um dos

cânceres também com maior taxa de incidência. Cerca de 1,6 milhões de novos casos são diagnosticados por ano, equivalendo a aproximadamente 12,7% de todos os casos novos de câncer (SOUZA, 2016). Segundo Cruz, Tanoue e Matthay (2011), o câncer de pulmão é o mais frequentemente diagnosticado.

De acordo com Pereira (2009), os cânceres pulmonares são divididos em dois tipos principais: carcinoma de pequenas células (*Small Cell Lung Cancer* – SCLC) e o carcinoma de células não pequenas (*Non Small Cell Lung Cancer* – NSCLC), com seus três tipos histológicos, adenocarcinoma, carcinoma epidermoide ou escamoso e carcinoma de grandes células.

O primeiro tipo, caracteriza-se por um crescimento tumoral rápido, atuando de forma agressiva, manifestando na forma de tumor central e hilar sendo considerado o mais maligno dentre os cânceres pulmonares (PEREIRA, 2009; SOUZA, 2011). O segundo tipo de câncer, ocorre de maneira mais lenta e manifesta-se, geralmente, por nódulos ou massas periféricas pulmonares (SOUZA, 2012).

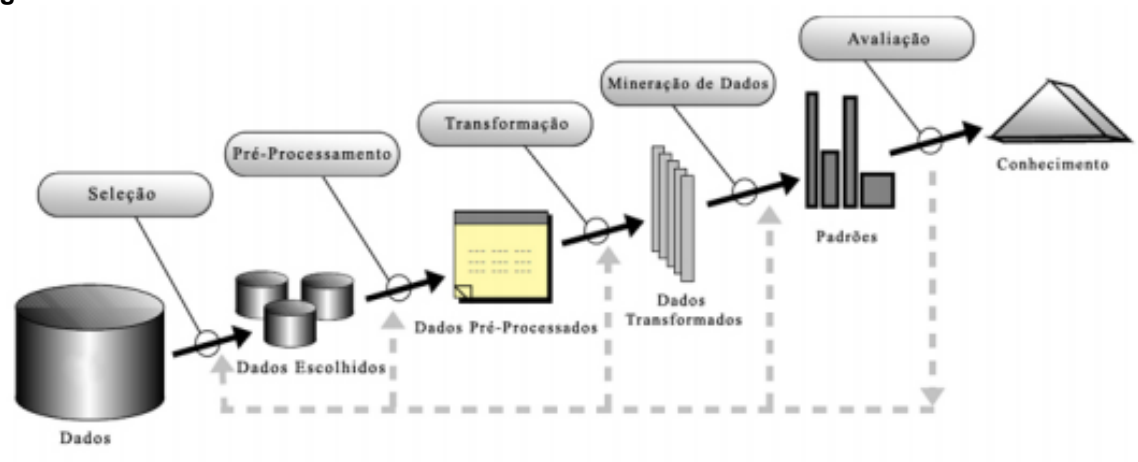
A próxima seção retrata os conceitos fundamentais de Mineração de Dados

2.3 MINERAÇÃO DE DADOS E DESCOBERTA DE CONHECIMENTO

Com a necessidade da busca por soluções viáveis em torno de processos de otimização de dados, a área de pesquisa denominada Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases* – KDD) buscou solucionar problemas envolvendo grande quantidade de informações, tendo em vista transformá-la em conhecimentos úteis. Visando desenvolver uma técnica para extrair esse conhecimento a partir de grandes volumes de dados, foi criada, no final da década de 80, a Mineração de Dados (*Data Mining*) (KAMBER; HAN; PEI, 2012).

A Mineração de Dados é a principal etapa do KDD, cujo papel fundamental é incluir as tarefas de seleção, preparação e exploração das informações, e a análise e interpretação dos resultados, buscando assimilar o conhecimento extraído no processo (GALVÃO; MARIN, 2009). Para Cios et al. (2007) e Fayyad et al. (1996), o KDD é todo o processo de descoberta de conhecimento, e a Mineração de Dados é uma das etapas do processo. A Figura 5 apresenta o processo KDD.

Figura 5 – Processo KDD



Fonte: Fayyad et al (1996)

Primeiramente, é escolhida uma base de dados. Uma porção dos registros é selecionada, pré-processada e submetida ao método de mineração, com algoritmos específicos e ferramentas que encontram padrões para representar o conhecimento aprendido. Após a extração, os padrões passam pelo pós-processamento, e o conhecimento aprendido é avaliado de acordo com sua qualidade para determinar a exequibilidade de sua utilização (YAMAGUCHI et al., 2010).

As etapas do processo KDD são descritas a seguir.

2.3.1 Seleção dos Dados

Esta etapa visa conhecer os dados pertencentes à base de dados, tendo como fundamento a resolução do problema em questão. Nesta fase, é analisada a problemática, juntamente com um profissional da área, para que, assim, sejam especificadas as variáveis a serem resolvidas.

Logo, é o ponto crucial para a pesquisa, pois a partir do problema é possível aprender e desenvolver o conhecimento por meio das ferramentas capazes de extrair informações úteis em uma determinada base de dados.

2.3.2 Pré-processamento

Nesta etapa é contemplada a seleção de dados, isto é, a identificação de um conjunto de dados, em que será possível realizar e aplicar as técnicas para extração do conhecimento. A etapa compreende as correções e formatações de dados para os

algoritmos pertencentes à mineração de dados. Portanto, a qualidade dos dados influencia diretamente na exatidão dos algoritmos.

De acordo com Boente, Goldschimidt e Estrela (2008), as principais funções do pré-processamento são:

- Seleção de dados: visa à identificação dos dados realmente utilizados em uma determinada base.
- Limpeza de dados: é imprescindível os dados estarem totalmente corretos para assegurar a qualidade. Portanto, são verificados e ajustados possíveis erros como falta de dados, caracteres a mais ou inconsistências nas bases.
- Codificação dos dados: existem alguns algoritmos que precisam de códigos especiais para serem executados, podendo a variável ser numérica ou categórica. Por exemplo: para diferenciar masculino de feminino, é preciso diferenciar ambas as variáveis para o atributo sexo, com masculino sendo representando por (1) e feminino por (2), ou seja, o atributo sexo é designado por uma variável categórica e os representativos de masculino e feminino por numérico.
- Enriquecimento dos dados: é o processo no qual há agregação de informações importantes referentes aos conjuntos de dados existentes, fazendo com que haja uma maior acurácia na qualidade do processo.

Definidas as principais funções do pré-processamento, é importante ressaltar que esta etapa é extremamente fundamental, em função de se prepararem os dados, representados futuramente como entrada para os algoritmos.

Outro fator importante nessa fase é o de reduzir dados sem importância na base, pois além de atrapalharem o processo como um todo, dificultam a extração de informações realmente úteis. Para isso, é preciso aplicar métodos de redução com finalidade de excluir dados redundantes antes de iniciar a busca pelos padrões (REZENDE, 2003). Esses métodos são denominados de métodos de redução de dimensionalidade.

Uma das funções do pré-processamento é a busca por padrões ou modelos para tratar o conhecimento obtido, mas para isso é exigida a validação dos mesmos, através da avaliação de características e dos processos envolvidos no modelo. Para

fazer essa validação, é preciso dividir os dados em dois conjuntos, um para treinamento e outro para teste. O conjunto treinamento deve conter os registros a serem utilizados na construção do modelo e o conjunto teste deve conter os registros a serem utilizados na avaliação do modelo (GOLDSCHMIDT; PASSOS, 2005).

O Quadro 1 apresenta os métodos de particionamento para validar os modelos, segundo Goldschmidt e Passos (2005).

Quadro 1 - Métodos de Particionamento

Métodos de Particionamento	Conceito
Validação Cruzada com *K* Conjuntos (*K-fold CrossValidation*)	Este método aplica a divisão aleatória de um conjunto de dados com N elementos em K subconjuntos disjuntos (*folds*), com o mesmo número de elementos (N/K). Logo, cada um dos K subconjuntos é utilizado como conjunto teste e os (K-1) demais conjuntos são reunidos em um conjunto de treinamento. Portanto, o processo é repetido K vezes, sendo gerados e avaliados K modelos de conhecimento. Essa validação é mais utilizada quando se quer avaliar a tecnologia presente na formulação do algoritmo.
Validação Cruzada com *K* Estratificada (*Stratifield K-Fold CrossValidation)*	Essa validação é mais utilizada para problemas envolvendo classificação, tendo similaridade à validação anterior, a qual a geração dos subconjuntos é mutualmente exclusiva. A proporção é considerada durante a amostragem. Por exemplo, se o conjunto original dos dados possui duas classes com distribuição de 20% e 80%, cada subconjunto também deverá conter aproximadamente esta mesma proporção de classes.
Leave-One-Out:	Utilizado apenas para pequenas amostras e computacionalmente dispendioso, este método é um caso particular da Validação Cruzada com *K*, em que cada um dos *K* subconjuntos possui um único registro.
Holdout:	Este procedimento divide de forma aleatória os registros em uma percentagem p fixa para treinamento e (1-p) para teste. É considerada p>1/2. Não existem fundamentos teóricos que dizem que é esta percentagem, porém podem ser utilizados p: 2/3 e (1-p) = 1/3 (REZENDE, 2003). É muito utilizada essa validação quando se deseja produzir apenas um único modelo de conhecimento a ser aplicado posteriormente em algum sistema.
Bootstrap	O conjunto de treinamento é criado a partir de *N* sorteios aleatoriamente e com reposição em torno do conjunto de dados original (contendo N instâncias). Já o conjunto de teste é formado pelas instâncias originais não sorteadas para o conjunto de treinamento. É utilizado para medir um desempenho do algoritmo.

Fonte: Autoria própria

Os métodos de particionamento são de grande importância para o pré-processamento, cuja função está diretamente ligada à garantia da qualidade dos dados.

A próxima etapa refere-se à mineração de dados, na qual será detalhada a seguir.

2.3.3 Mineração de Dados

Esta etapa é considerada o centro da descoberta do conhecimento, na qual se preocupa em extrair padrões a partir de dados observados. São designados os métodos e os algoritmos que irão realizar a busca pelo conhecimento útil nas bases de dados.

As informações contidas nas bases são utilizadas para aprender um determinado conceito alvo ou padrão. Dependendo do conceito, diferentes algoritmos indutivos de aprendizagem são aplicados (MACEDO, 2012). Os conceitos aplicados são denominados de tarefas de mineração. A seguir, serão descritas algumas dessas principais tarefas:

- Associação: o objetivo principal desta tarefa é encontrar, a partir de um conjunto de dados, atributos correlacionados que aparecem com uma maior frequência. Portanto, o propósito das regras de associação é encontrar padrões e tendências em conjunto de informações (CARDOSO; MACHADO, 2008; PERNOMIAN, 2008).
- Classificação: baseia-se na construção de um modelo preditivo, nos quais os atributos são divididos em duas categorias: previsores e meta. Primeiramente, o propósito desta tarefa é encontrar uma correlação entre os atributos previsores e o atributo meta, a partir de instâncias, cuja classe é conhecida, e assim, a partir dos atributos previsores, encontrar instâncias, cuja classe é desconhecida.
- Agrupamento ou Clusterização: esta etapa visa identificar as semelhanças entre os dados existentes e assim dispor estes em grupos, de acordo com suas características similares. Os conjuntos são particionados em subconjuntos, para melhor visualização dos dados.
- Seleção de Atributos: o principal objetivo desta tarefa é a de encontrar um subconjunto de atributos relevantes a partir de atributos originais, reduzindo consideravelmente a otimização do processo de aprendizagem.

Neste estudo, a tarefa de classificação foi escolhida, em razão de sua importância na análise de expressão gênica, além de contemplar com um dos fundamentos do objetivo da pesquisa. A partir dessa tarefa, é possível obter um entendimento maior dos processos envolvidos (CHUANG et al., 2011). Segundo Macedo (2015), é a tarefa com maior destaque nesses tipos de bases.

A seguir é detalhada a última etapa da mineração de dados, que visa à avaliação final do modelo criado.

2.3.4 Pós-processamento

A última etapa do KDD compreende o tratamento do conhecimento adquirido a partir das fases previamente descritas.

Nesta etapa, os padrões descobertos são avaliados no sentido de verificar se satisfazem o critério necessário para constituir um elemento importante para o apoio à tomada de decisão (MALUCELLI et al., 2010), de forma que possam ser facilmente interpretadas pelos usuários. Os padrões são descobertos depois que os dados redundantes e irrelevantes são removidos (SUH, 2012). Também podem ser corrigidos erros que não foram vistos anteriormente.

A seguir, serão descritos os algoritmos de classificação utilizados no estudo.

2.3.5 Algoritmos de Classificação

Existem alguns algoritmos utilizados para tarefas de classificação envolvendo Mineração de Dados, cada qual com um objetivo em específico. Os algoritmos selecionados foram escolhidos porque eles pertencem a diferentes paradigmas de aprendizagem (BORGES; NIEVOLA, 2012), e também por serem considerados como os melhores algoritmos classificadores (WU et al., 2008).

São apresentados no quadro abaixo, os algoritmos empregados no estudo e seus principais conceitos.

Quadro 2 – Algoritmos utilizados na respectiva pesquisa

Algoritmo	Conceito
Naive Bayes (JOHN, 1995)	Os algoritmos Naive Bayes são classificadores estatísticos baseados no Teorema de Bayes, que predizem a probabilidade de um determinado dado pertencer a uma classe em particular. De acordo com Mitchell (2010), o algoritmo Naive Bayes é muito utilizado, tanto para variáveis discretas ou contínuas, pois é de fácil aplicação em um conjunto de amostras. As probabilidades são estimadas de acordo com a frequência de cada valor para os registros de treino. Assim, dada uma nova instância, o classificador faz a estimativa de probabilidade de o registro feito pertencer a uma nova classe específica, considerando que os atributos são condicionalmente independentes (BERTON, 2011).
J48 (QUINLAN, 1993)	O algoritmo J48 permite a criação de modelos de decisão em árvore. O modelo de árvore de decisão é feito a partir da análise dos dados de treino e pelo modelo utilizado para classificar dados ainda não classificados. O algoritmo gera árvores de decisão, a qual cada nó da árvore avalia individualmente a existência ou significância de cada atributo de maneira individual (FRUTUOSO, 2014). As árvores são geradas através da escolha do atributo mais adequado para cada situação e são construídas do topo para a base. Para Tavares, Bozza e Kono (2007), o algoritmo constrói uma árvore de decisão a partir do atributo mais significativo, por meio da abordagem *top-down*. Neste caso, o atributo mais global é escolhido para ser a raiz da árvore, comparando-o com todos os atributos do conjunto. Com isso, para prosseguimento da construção, é considerado o segundo atributo como sendo o próximo nó da árvore, e assim até que se gere o nó folha, que representa o atributo alvo da instância.
SVM (HASTIE, 1998)	É baseado em modelos lineares, abordando aspectos referentes ao aprendizado para problemas de reconhecimento de padrão. Tem como objetivo a determinação de limites de decisão que produzem uma separação ótima entre classes, por meio da minimização de erros, além de realizar a separação entre duas classes distintas, por meio de um hiperplano de separação (VAPNIK, 1995). O algoritmo SVM mapeia cada dado analisado, utilizando um mapeamento fixo, usando os dados de treino, construindo dessa forma um hiperplano com margem de separação máxima, utilizado para classificar exemplos desconhecidos. O algoritmo trabalha com dados linearmente separáveis, no entanto, existe a possibilidade de adaptação para conjuntos não lineares através das funções kernel não lineares (ALVES; FRAGAL, 2011). Mediante essa função, é possível trabalhar com problemas não separáveis linearmente.
KNN (AHA; KIBLER; ALBERT,1991)	De acordo com Frutuoso (2014), o classificador K-NN gera um modelo a partir de uma entrada, e a partir dessa entrada, são caracterizados os k vizinhos mais próximos. A vantagem de utilizar esse algoritmo, é que ele não considera todos os atributos da base de dados, apenas são considerados os atributos de entrada para a geração do modelo, desconsiderando desse modo, os atributos sem relevância para a classificação (YANG; LIU, 1999). Outra vantagem da classificação pelo algoritmo K-NN é a flexibilidade, porém, quando o conjunto de dados é muito grande, podem ocorrer dispersões em relação à precisão da classificação, devido os vizinhos estarem longe da vizinhança do conjunto de dados (BERTON, 2011).

Fonte: Autoria própria

Após a descrição dos algoritmos utilizados neste estudo, a próxima seção visa apresentar as principais ferramentas para mineração de dados.

2.3.6 Ferramentas para Mineração de Dados

Esta seção objetiva apresentar as principais ferramentas computacionais para mineração de dados. O Quadro 3 retrata as ferramentas mais utilizadas neste contexto, com base no relatório da 14° edição anual da KDnuggets Poll (2013), site líder em análise de negócios, mineração de dados e ciência dos dados.

Quadro 3 – Principais ferramentas para mineração de dados

Nome do *Software*	Desenvolvedor	Site para Acesso
Excel	Microsoft	www.microsoft.office.com/excel
R	Bell *Laboratories*	http://www.r-project.org
RapidMiner	RapidMiner	http://www.rapidminer.com
WEKA	*University of Waikato*	www.cs.waikato.ac.nz

Fonte: Adaptado de KDnuggets Poll (2013)

Os *softwares* apresentados no Quadro 3 oferecem interessantes recursos, como a possibilidade de poder trabalhar com uma grande gama de algoritmos de mineração, em diferentes bases de dados, além da utilização de diferentes métodos de avaliação.

O RapidMiner é um *software* de código aberto que trabalha com vários tipos de tarefas de mineração de dados. Os processos de mineração de dados podem ser implementados e executados de forma rápida, intuitiva e sem necessidade de conhecimentos de programação (LAUSCH; SCHMIDT; TISCHENDORF, 2015). O R, por sua vez, segundo os autores, requer aprendizagem em programação R, antes de se utilizar os recursos de mineração de dados provenientes do *software*.

Já, o WEKA (*Waikato Environment for Knowledge Analysis)* é um *software* gratuito, desenvolvido na linguagem Java, que se estabeleceu como a ferramenta mais utilizada na literatura para tarefas envolvendo mineração de dados (MACEDO, 2012). Através de sua interface gráfica, denominada como WEKA *Explorer*, é possível realizar todos os processos envolvendo mineração de dados, avaliando os resultados e comparando os algoritmos escolhidos.

WEKA contém ferramentas para ambas as tarefas da mineração de dados, tais como: classificação, agrupamento (*cluster*), associação e seleção de atributos. Pode ser considerada a mais antiga e bem-sucedida biblioteca de dados de código aberto neste âmbito (LAUSCH; SCHMIDT; TISCHENDORF, 2015).

A próxima seção retrata as técnicas de redução de dimensionalidade que serão utilizadas no trabalho, o qual aborda a Seleção de Atributos e Análise de Componentes Principais.

2.4 TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE

Esta seção aborda as técnicas utilizadas para reduzir a dimensionalidade dos dados nesta pesquisa.

2.4.1 Seleção de Atributos

Uma das técnicas de redução de dimensionalidade mais importante é a de Seleção de Atributos, pois através dessa técnica é possível reduzir atributos redundantes e incoerências, facilitando a melhora do processo como um todo. O método mantém os atributos considerados relevantes para a base a ser estudada (MACEDO, 2015).

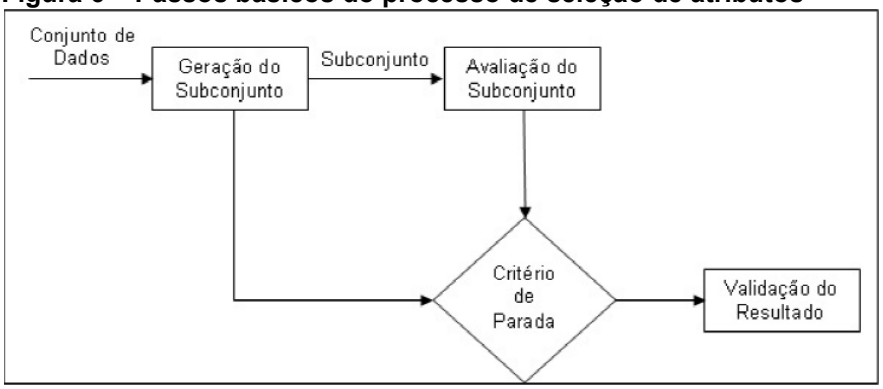
Essa técnica tem sido centro de atenção por seu grande potencial em diversas aplicações, tais como: bioinformática, medicina, processamento de dados e reconhecimento de objetos (WANG et al., 2013a).

Quando se utiliza a tarefa de classificação juntamente com Seleção de Atributos, é minimizada a taxa de erro do classificador, e também o número de atributos selecionados para formar uma base diferente (BORGES, 2006).

No estudo proposto, que envolve a identificação de atributos relevantes em bases de expressão gênica, com estudos envolvendo câncer, presume-se que os dados pertencentes aos bancos de dados estão redundantes e incoerentes, e, por isso, precisam de métodos para organização dos dados, tendo em vista descobrir possíveis genes causadores do câncer.

Segundo Dash e Liu (1997), o método de Seleção de Atributos é composto por quatro passos principais, como apresentado na Figura 6.

Figura 6 – Passos básicos do processo de seleção de atributos



Fonte: Dash e Liu (1997)

A geração do subconjunto é feita a partir do conjunto de dados, no qual é realizado um procedimento de busca que produz subconjuntos de atributos candidatos para avaliação, baseado em uma estratégia de busca (BORGES; NIEVOLA, 2012). As mais utilizadas segundo Liu e Yu (2005) são: busca sequencial, exponencial e aleatória.

Os subconjuntos selecionados são avaliados e posteriormente comparados com um melhor anterior, melhor que ele; se o subconjunto for melhor do que o comparado, ele se torna a referência, baseando-se em um critério de avaliação.

Esse processo é feito até satisfazer um critério de parada, no qual o melhor subconjunto escolhido é avaliado através de diferentes testes em conjuntos de dados reais ou sintéticos (LIU; YU, 2005). O critério de parada é estabelecido quando um processo está suficientemente bom ou quando a busca termina; logo, caracteriza-se como quando a razão do erro da classificação é menor que a razão de erro permitida para uma determinada tarefa, ou quando é limitado o número máximo de atributos ou iterações (BORGES; NIEVOLA, 2012).

Os algoritmos pertencentes ao método de Seleção de Atributos foram desenvolvidos com critérios de avaliação diferentes, sendo necessário utilizar duas abordagens para as diferentes situações. As abordagens avaliam os subconjuntos de atributos de acordo com sua categorização, dependente da participação do algoritmo de aprendizado na avaliação (GUYON; ELISSEEFF, 2003).

A seleção de atributos está dividida em três abordagens: Filtro, *Wrapper* e *Embedded* (SAEYS; INZA; LARRAÑAGA, 2007; LAZAR et al., 2012; GUYON; ELISSEEFF, 2003).

A abordagem Filtro resulta das características dos dados para efetuar a avaliação e selecionar os subconjuntos de características sem envolvimento de um

algoritmo de mineração (WANG et al., 2013b). O método seleciona os dados previamente e subsequentemente realiza o processo de classificação, contudo, não pondera as interações entre os atributos (CHUANG et al., 2011).

Dentre as técnicas mais utilizadas para avaliar um subconjunto de atributos, podem-se destacar as medidas de dependência e consistência, correspondendo aos algoritmos *Correlation-based Feature Selection* (CFS) e *Consistency Subset Eval* (CSE) (LIU; YU, 2005).

O algoritmo CFS classifica os subconjuntos gerados de acordo com uma função de correlação com base em uma recompensa heurística de avaliação (HALL, 1999). Esse algoritmo avalia a importância de um subconjunto de atributos em função da predição individual de cada atributo e o grau de correlação entre eles.

O CFS seleciona um subconjunto de recursos de forma explícita. O algoritmo avalia cada recurso individualmente e atribui uma pontuação com base em sua importância (KOPRINSKA; RANA; AGELIDIS, 2015).

Já o algoritmo CSE reduz o conjunto de atributos preservando a consistência original. O algoritmo avalia o valor de um subconjunto de atributos pelo nível de consistência dos valores verificados para cada classe quando as instâncias de treinamento são projetadas sobre o subconjunto de atributos (WITTEN; FRANK; HALL, 2005).

A abordagem *Wrapper* utiliza um algoritmo de mineração preestabelecido e tem seu desempenho como critério de avaliação. Há a necessidade de terminação de um algoritmo de mineração, sendo seu critério usado para avaliação de subconjunto (WANG et al., 2013b).

Além de utilizar seu desempenho como critério de classificação, o modelo *Wrapper* procura pelo melhor conjunto de características, buscando melhorar o desempenho dos algoritmos, porém tem um valor computacional maior do que o modelo Filtro.

Na abordagem *Embedded*, os conjuntos dos atributos são diretamente incorporados no algoritmo responsável pela indução do modelo de classificação, enquanto na estratégia *Wrapper* e Filtro, são executadas em uma fase de pré-processamento dos dados. Nessa abordagem, o subconjunto de atributos é avaliado por uma medida independente (LINS; MERSCHMANN, 2010).

Resumidamente, a diferença entre os métodos Filtro e *Wrapper*, é que na primeira, os subconjuntos dos atributos são avaliados de acordo com uma medida

independente, à medida que a segunda, o subconjunto, utiliza o próprio algoritmo de classificação para a avaliação.

A validação dos resultados pode ser aplicada para medir o resultado através de um conhecimento *a priori* dos dados. Portanto, se os atributos relevantes já forem conhecidos anteriormente, existe a possibilidade de se comparar o conjunto de atributos conhecidos com os atributos selecionados (MACEDO, 2015).

Outra alternativa é utilizar a razão do erro classificador como um indicador de desempenho, para um subconjunto de atributos selecionado, utilizando a comparação entre a razão do erro classificador treinado no conjunto completo de atributos e no subconjunto de atributos selecionado (BORGES; NIEVOLA, 2012).

2.4.2 Análise de Componentes Principais (PCA)

A Análise de Componentes Principais (PCA) é uma técnica estatística que visa estudar diversas aplicações cuja dimensionalidade é alta. A principal ideia é reduzir a dimensionalidade dos dados do conjunto, transformando subsequentemente em um novo conjunto de variáveis denominado de componentes principais, preservando ao máximo as informações originais. Em outras palavras, o método transforma ortogonalmente um conjunto de variáveis correlacionadas para um conjunto de valores de variáveis linearmente não correlacionadas (componentes principais).

Para Dunteman (1999), o PCA elimina informações redundantes, destacando os recursos ocultos, provenientes das informações contidas nas bases, e visualiza as principais relações existentes entre as observações vistas. Uma das ferramentas mais clássicas e populares para a análise de dados e redução de dimensionalidade, com ampla gama de aplicações bem-sucedidas em toda a ciência e engenharia (JOLLIFFE, 2002).

O método PCA é um método de aprendizagem não supervisionada, que visa encontrar a combinação de condições que explicam a maior variação nos dados (YANG et al., 2008), utilizado em muitos tipos de análises incluindo neurociência e computação gráfica (SHLENS, 2005), além de análises de dados de microarranjos (HOLMES et al., 2011; YANG et al., 2008).

As definições a seguir abrangendo análise de componentes principais surgiram a partir dos estudos de Song et al. (2013).

PCA é uma decomposição de valores próprios da matriz de covariância dos dados, utilizado para aproximação de baixo *rank*, que compara os dados através de uma função linear de variáveis (MARKOS; VOZALIS; MARGARITIS, 2010).

Matematicamente, os componentes principais são obtidos calculando os autovalores da matriz de covariância *C*, como apresentada na equação (1):

$$\mathrm{Cvi} = \lambda_{\mathrm{i}} \mathrm{v}_{\mathrm{i}} \qquad (1)$$

A matriz de covariância dos vetores dos dados originais *X* é representada por *C*, λ_i refere-se aos autovalores da matriz *C* e v_i corresponde aos autovetores correspondentes. Consecutivamente, a fim de reduzir a dimensionalidade dos dados, os autovetores *k*, que correspondem aos maiores autovalores *k*, precisam ser computadorizados (XU et al., 2005).

Considerando $E_k = [v_1, v_2, v_3, \ldots\ldots, v_k]$ e $\Lambda = [\lambda_1, \lambda_2, \lambda_3 \ldots\ldots \lambda_k]$, logo tem-se $CE_k = E_K \Lambda$. Portanto, obtemos a seguinte equação:

$$\mathrm{X}^{\mathrm{PCA}} = \mathrm{E}_{\mathrm{K}}^{\mathrm{T}} \mathrm{X} \qquad (2)$$

Em relação à Equação (2), o número das características da matriz de dados original *X* é reduzido pela multiplicação com a matriz *d* x *k* E_k que tem autovetores *k* correspondentes aos maiores autovalores *k*. O resultado da matriz é X^{PCA} (BINGHAM; MANNILA, 2001).

O próximo capítulo diz respeito à metodologia empregada nesta pesquisa, que aborda as etapas do desenvolvimento do trabalho.

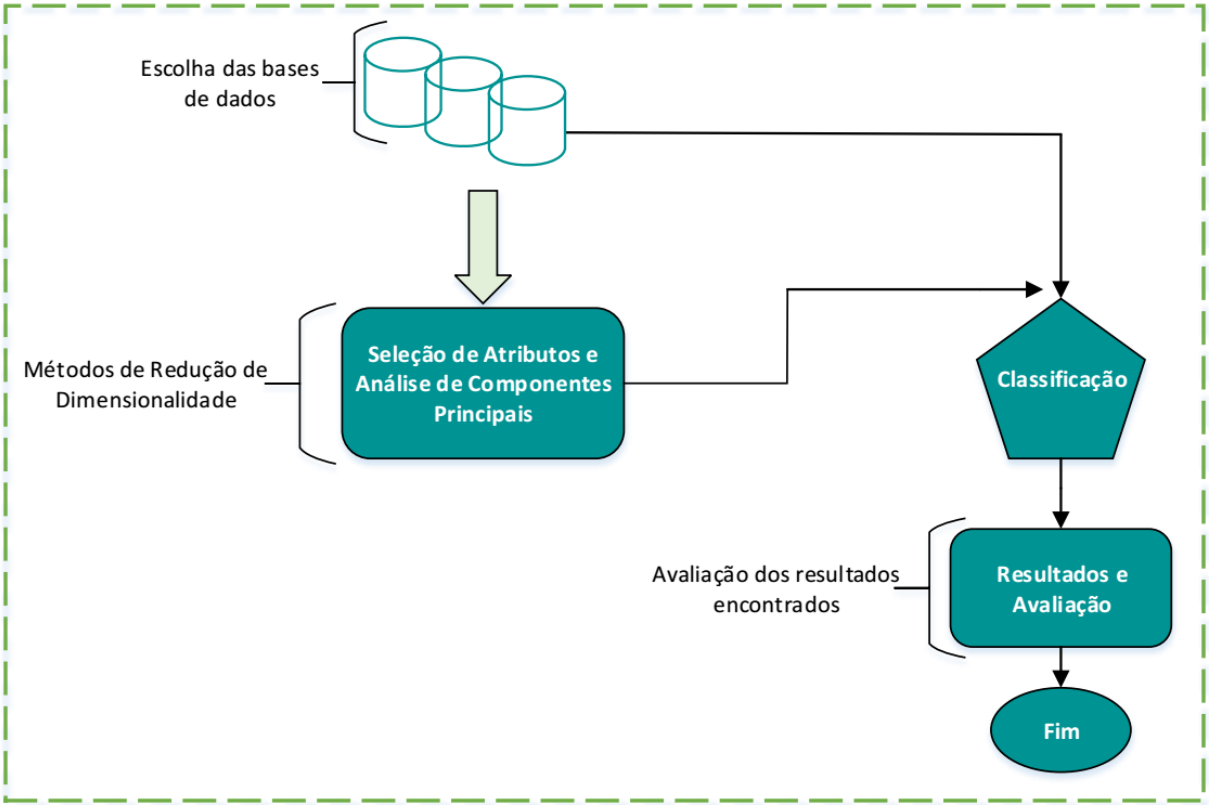
3 PROCEDIMENTOS METODOLÓGICOS

Este capítulo descreve os procedimentos metodológicos utilizados para atingir os objetivos propostos neste trabalho. São descritas as etapas do experimento do estudo, com informações referentes ao processo adotado para o desenvolvimento e aplicação da pesquisa.

3.1 ETAPAS DO EXPERIMENTO

Com o intuito de realizar o experimento foram definidas as seguintes etapas: Escolha das Bases de Dados, Aplicação do Método de Seleção de Atributos e PCA, Execução dos Algoritmos de Classificação e, por fim, avaliação dos resultados encontrados. Os passos são identificados na Figura 7.

Figura 7 – Etapas para realização do experimento



Fonte: Autoria própria

A primeira etapa do processo refere-se à escolha das bases de dados que serão estudadas. Posteriormente, foram aplicados os métodos de Seleção de Atributos e Análise de Componentes Principais. Na terceira etapa, referente à tarefa

de classificação, foram executados os Algoritmos de Classificação, tanto nos dados originais quanto nos dados resultantes gerados pelos dois métodos utilizados. Finalmente, foram avaliados os resultados encontrados juntamente com a comparação de ambas os métodos.

A ferramenta de mineração de dados escolhida para ser utilizado neste estudo foi o *Waikato Environment for Knowledge Analysis* (WEKA) versão 3.9 (WEKA, 2016).

Para evitar viés e ajuste excessivo, foi utilizada a configuração padrão de parâmetros do WEKA. Para utilização do mesmo, os dados devem ser preparados no formato do WEKA (.arff). Nesse estudo, as bases analisadas já estavam adequadas para os experimentos.

Detalha-se a seguir as etapas para realização do experimento.

3.1.1 Descrição das Bases de Dados

As bases de expressão gênica escolhidas são oriundas do repositório de dados biomédicos *Kent Ridge Bio-medical Dataset* (KENT RIDGE, 2016), contemplada como sendo uma das nove melhores, nos últimos anos, referentes a dados de microarranjos, segundo Bolón-Canedo et al. (2014). Para o estudo, serão empregadas três bases de dados. Não se tem uma quantidade mínima ou máxima de bases a serem estudadas, dependentes do propósito do estudo.

Visto ser o trabalho uma comparação entre dois métodos de redução de dimensionalidade, foram levados em considerações três aspectos: quantidade de atributos, informações contempladas nas bases, que nesse caso foram três bases com informações pertinentes a pacientes com câncer de pulmão, e reconhecimento do repositório que fornecem os dados.

As bases são denominadas de LungCancer-Michigan (BEER et al., 2002), LungCancer-Ontario (WIGLE et al., 2002) e LungCancer-Harvard (BHATTACHARJEE et al., 2001). A seguir, serão representadas as descrições das bases.

- Base de dados LungCancer-Michigan

Esse conjunto é formado por 96 amostras, sendo 86 amostras de adenocarcinomas primários de pulmão e 10 amostras pulmonares não neoplásicas. O conjunto possui 7129 atributos (gene).

- Base de dados LungCancer-Ontario

Esse conjunto é formado por 39 amostras de NSCLC (*Non Small Cell Lung Cancer*), que condiz a carcinomas de células não pequenas. Dentre as 39 amostras, 24 correspondem a pacientes que já tiveram recidiva do tumor ou metástase (rotulado na base como *Relapse* (recaída)), e 15 amostras pulmonares não neoplásicas (denominado de *Non-Relapse* (sem recaída)). O conjunto possui 2880 atributos (gene).

- Base de dados LungCancer-Harvard

Esse conjunto é formado por 203 amostras, sendo 139 amostras de adenocarcinomas de pulmão (denominado de ADEN), 21 amostras de carcinomas pulmonares de células escamosas (denominado de SQUA), 20 amostras de tumores carcinoides pulmonares (denominado como COID), 6 amostras de carcinomas de pulmão de células pequenas (rotulado como SCLC) e 17 amostras de pulmões normais (classificado como NORMAL). O conjunto possui 12600 atributos (genes).

A próxima seção irá tratar da segunda etapa do experimento, que retrata a aplicações dos métodos.

3.1.2 Aplicação do Método de Seleção de Atributos e Análise de Componentes Principais

Nesta seção, é apresentada a segunda etapa do estudo, que contempla a aplicação dos métodos de Seleção de Atributos e Análise de Componentes Principais. A subseção 3.1.2.1 retrata o método de Seleção de Atributos e a subseção 3.1.2.2 o método de Análise de Componentes Principais.

3.1.2.1Aplicação do método de seleção de atributos

Para o processo de Seleção de Atributos, foram utilizadas duas abordagens: Filtro e *Wrapper*. Para a abordagem Filtro, empregaram-se as medidas de dependência e consistência, correspondentes ao algoritmo CFS e algoritmo CSE. Já

para a abordagem *Wrapper*, foram utilizados os algoritmos classificadores Naive Bayes, J48, SVM, 1-NN, 3-NN, 5-NN e 7-NN. As abordagens foram especificadas e descritas na seção 2.4.1 (p.36).

Após a utilização do método, para cada abordagem, foram gerados novos subconjuntos de dados, entretanto, apenas com os atributos (genes) considerados relevantes.

Para a primeira abordagem, foram formados 2 subconjuntos e para a segunda abordagem 7 subconjuntos, totalizando 9 subconjuntos para cada uma das bases. No Quadro 4, está a ilustração dos critérios de busca e medidas de avaliação para a geração dos subconjuntos.

Quadro 4 – Algoritmos de Seleção de Atributos

	Algoritmo	Critério de Busca	Medida de Avaliação	Subconjuntos
Filtro	CFS	Sequencial	Dependência	Filtro Dependência (CFS)
	CSE	Sequencial	Consistência	Filtro Consistência (CSE)
Wrapper	Naive Bayes	Sequencial	*Wrapper* (usando Naive Bayes)	*Wrapper* (usando Naive Bayes)
	J48	Sequencial	*Wrapper* (usando J48)	*Wrapper* (usando J48)
	SVM	Sequencial	*Wrapper* (usando SVM)	*Wrapper* (usando SVM)
	1-NN	Sequencial	*Wrappe*r (usando 1-NN)	*Wrapper* (usando 1-NN)
	3-NN	Sequencial	*Wrapper* (usando 3-NN)	*Wrapper* (usando 3-NN)
	5-NN	Sequencial	*Wrapper* (usando 5-NN)	*Wrapper* (usando 5-NN)
	7-NN	Sequencial	*Wrapper* (usando 7-NN)	*Wrapper* (usando 7-NN)

Fonte: Autoria própria

Após a geração dos subconjuntos, estes foram submetidos aos algoritmos de classificação.

A seguir, é descrito o processo de aplicação do método PCA.

3.1.2.2Aplicação do método de análise de componentes principais

Para este método, as bases de dados foram submetidas ao processo de Análise de Componentes Principais (PCA). É importante ressaltar que o método ignora o atributo classe, também denominado de atributo meta, quando os componentes principais são computados. Logo, os valores correspondentes as classes são recolocadas para os dados transformados (componentes principais).

Os critérios de utilização para o método PCA foram definidos de acordo com a porcentagem de variância nos dados originais. As porcentagens escolhidas foram

90%, 95% e 99%. Dessa forma, foi possível comparar os resultados entre as bases através do percentual de utilização de dados da base original.

A próxima seção apresenta a terceira etapa do experimento, que aborda a execução dos algoritmos de classificação.

3.1.3 Classificação

Nesta etapa, os subconjuntos gerados foram submetidos à classificação. Os algoritmos utilizados são o Naive Bayes, J48, SVM, 1-NN, 3-NN, 5-NN e 7-NN. Os algoritmos escolhidos foram utilizados através do *software* WEKA. Definiu-se como algoritmo base o algoritmo Naive Bayes (em destaque nas tabelas abaixo) para as devidas comparações.

Juntamente à aplicação dos algoritmos, é preciso empregar métodos para validar os modelos. O método utilizado é a Validação Cruzada Estratificada com 10 partições, que corresponde em dividir aleatoriamente as bases de dados originais em 10 partições iguais. O resultado final é a média das avaliações e a validação cruzada evita resultados tendenciosos.

Quando se trata de modelos preditivos, é preciso caracterizar a taxa de acerto determinada pelos algoritmos. A taxa de acerto refere-se ao número de instâncias classificados corretamente dividido pelo número total de instâncias. Quanto maior a taxa de acerto, maior a acurácia do modelo preditivo.

A última etapa do experimento, é descrita a seguir.

3.1.4 Resultados e Avaliação

Como essa etapa objetiva apresentar os resultados e avaliação dos dados obtidos, o capítulo 4 irá discutir detalhadamente os resultados através dos procedimentos adotados.

4 ANÁLISE DOS RESULTADOS

Este capítulo apresenta os principais resultados encontrados. A seção 4.1 refere-se aos resultados referentes a utilização do método de Seleção de Atributos e a seção 4.2 refere-se à utilização do método de Análise de Componentes Principais. Posteriormente, a seção 4.3 contemplará a comparação entre os métodos de redução de dimensionalidade escolhidos.

4.1 SELEÇÃO DE ATRIBUTOS

Para medir o desempenho dos modelos preditivos gerados, foi verificada a eficiência dos classificadores para ambos os subconjuntos por meio da taxa de acerto (acurácia) e desvio padrão. Porém, para avaliação dos mesmos, foram executados primeiramente sobre as bases com todos os atributos para as devidas comparações.

A seguir, serão mostrados os resultados referentes às bases de dados escolhidas, tanto para todos os atributos (genes) da base, quanto para os atributos selecionados pelo método, esse último, utilizando as abordagens Filtro e *Wrapper*.

4.1.1 Todos os Atributos

A Tabela 1 apresenta os resultados obtidos ao se aplicar os classificadores usando todos os atributos da base.

Tabela 1 – Resultados da classificação com todos os atributos

	Algoritmos						
Bases de dados	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
LungCancer-Michigan	100,0 ± 0	99,0 ± 3,16	100,0 ± 0	100,0 ± 0	98,89 ± 3,51	98,89 ± 3,51	100,0 ± 0
LungCancer-Ontario	67,50 ± 35,45	84,17 ± 13,86	78,33 ± 22,97	56,67 ± 19,95	55,83 ± 34,26	57,50 ± 39,18	58,33 ± 32,63
LungCancer-Harvard	80,38 ± 6,16	93,12 ± 7,50	95,07 ± 3,34	89,71 ± 5,27	91,21 ± 4,89	89,74 ± 5,70	89,24 ± 5,40

Fonte: Autoria própria

De acordo com a tabela acima, evidencia-se para a base LungCancer-Michigan que as taxas de acerto do algoritmo SVM, 1-NN e 7-NN, apresentaram valores iguais, comparando-se com o algoritmo base (Naive Bayes). Para os demais algoritmos, não houve diferença significativa com o algoritmo Naive Bayes.

Com relação à base LungCancer-Ontario, observou-se que os algoritmos J48 e SVM, apresentaram maiores valores quando comparados com o algoritmo base. Também, nota-se a baixa classificação para os algoritmos KNN. Esta baixa classificação pode ocorrer devido às informações contidas na base, tais como: características das doenças, fatores relacionados à idade ou sexo, resultados referentes às interações entre os genes, entre outras variáveis.

Para a base LungCancer-Harvard, todos os algoritmos classificadores foram melhores estatisticamente, comparando-se com o algoritmo base. O pior algoritmo foi o Naive Bayes.

Dentre as bases estudadas, a base que obteve melhor média foi a LungCancer-Michigan com 99,54% de taxa de acerto.

4.1.2 Abordagem Filtro

Na abordagem Filtro, foram utilizadas as medidas de dependência (algoritmo CFS) e consistência (algoritmo CSE). Para os subconjuntos gerados, foram aplicados os sete classificadores. A Tabela 2 mostra os resultados obtidos, utilizando a medida de dependência.

Tabela 2 – Resultados dos classificadores - Filtro (dependência)

	Algoritmos						
Bases de dados	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
LungCancer-Michigan	100,0 ± 0	99,0 ± 3,16	99,0 ± 3,16	100,0 ± 0	100,0 ± 0	100,0 ± 0	100,0 ± 0
LungCancer-Ontario	74,17± 20,58	71,67 ± 21,94	69,17 ± 15,74	44,17 ± 27,51	59,17 ± 31,29	61,67 ± 33,61	64,17± 33,58
LungCancer-Harvard	95,60 ± 5,45	93,14 ± 7,36	97,05 ± 3,47	95,07 ± 4,72	97,07 ± 2,52	97,55 ± 2,59	97,05 ± 3,47

Fonte: Autoria própria

Os dados referentes à Tabela 2 mostram que para a base LungCancer-Michigan, os algoritmos pertencentes à família K-NN, apresentaram valores equivalentes ao do algoritmo base. Para os algoritmos J48 e SVM, não há diferença significativa com relação ao algoritmo Naive Bayes.

Para a base LungCancer-Ontario, ambos os algoritmos classificadores apresentaram resultados significativamente piores do que o algoritmo base. O algoritmo que apresentou a pior taxa de acerto foi o 1-NN.

Em relação à base LungCancer-Harvard, pode-se afirmar que para todos os algoritmos, a taxa de acerto foi superior a 90%. Evidenciou-se que apenas os algoritmos J48 e 1-NN, apresentaram valores abaixo da taxa de acerto correspondente ao Naive Bayes.

O melhor desempenho da abordagem Filtro, usando a medida de dependência, foi na base LungCancer-Michigan, com média de desempenho de 99,71%. Já a base com o pior desempenho foi na base LungCancer-Ontario, com o valor de 63,46% de taxa de acerto.

A Tabela 3 apresenta os resultados dos classificadores utilizando a medida consistência.

Tabela 3 – Resultados dos classificadores - Filtro (consistência)

	Algoritmos						
Bases de dados	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
LungCancer-Michigan	96,78 ± 5,20	99,0 ± 3,16	95,89 ± 5,32	98,00 ± 4,22	99,00 ± 3,16	99,00 ± 3,16	99,00 ± 3,16
LungCancer-Ontario	74,17 ± 23,12	71,67 ± 14,27	61,67 ± 12,55	79,17 ± 16,32	74,17 ± 20,58	74,17 ± 20,58	74,17 ± 20,58
LungCancer-Harvard	90,19 ± 3,16	89,67 ± 7,45	82,33 ± 7,13	93,07 ± 4,19	90,67 ± 5,82	89,19 ± 5,54	87,74 ± 5,61

Fonte: Autoria própria

Os resultados dos classificadores para a medida consistência mostraram que para a base LungCancer-Michigan, apenas o algoritmo SVM, apresentou desempenho médio abaixo do algoritmo base, já os demais apresentaram valores estatisticamente maiores.

Para a base LungCancer-Ontario, apenas o algoritmo 1-NN apresentou taxa de acerto superior ao algoritmo Naive Bayes. Entretanto, os algoritmos 3-NN, 5-NN e 7-NN apresentaram valores equivalentes ao algoritmo base.

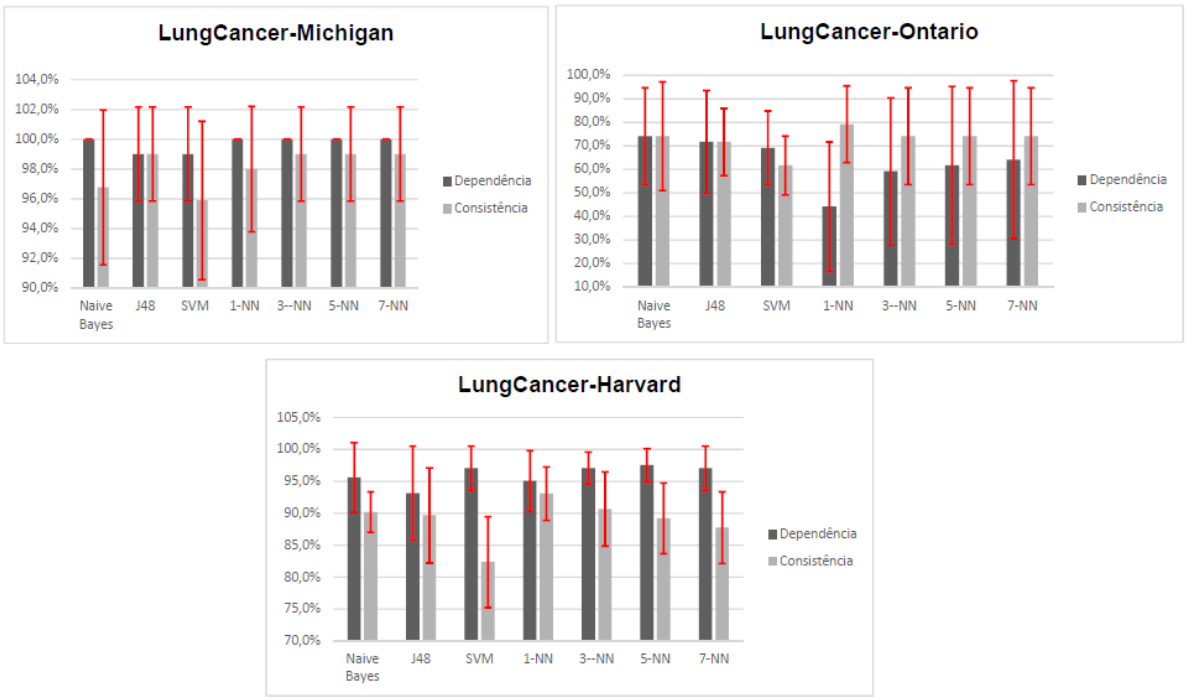
Na base LungCancer-Harvard, ressalta-se os algoritmos 1-NN e 3-NN como os algoritmos que apresentam valores significativamente maiores, se comparados ao algoritmo base.

O melhor desempenho da abordagem Filtro, usando a medida de consistência, foi na base LungCancer-Michigan, com média de desempenho de 98,10%. A base que apresentou menor média foi a base LungCancer-Ontario com 72,74%.

O Gráfico 1 apresenta um comparativo entre as médias das taxas de acerto obtidas pela abordagem Filtro, tanto para a medida de dependência quanto para a medida de consistência, nas três bases de dados estudadas. A barra indica a média

dos algoritmos e as linhas representam o limite inferior e o limite superior por meio do desvio padrão.

Gráfico 1 – Média das taxas de acerto usando abordagem Filtro, nas três bases de dados analisadas



Fonte: Autoria própria

Logo a partir dos dados relatados, observou-se que a medida consistência obteve média de 86,61% de taxa de acerto. Já a medida dependência apresentou na média o valor de 86,42%. Portanto, conclui-se que, a medida consistência apresentou valores de desempenho maiores do que a medida dependência, com diferenças mínimas significativamente.

4.1.3 Abordagem *Wrapper*

Para a abordagem *Wrapper* foram utilizados os próprios algoritmos classificadores como critério de avaliação. A Tabela 4 retrata os resultados referentes a essa abordagem.

Tabela 4 – Resultados dos classificadores para a abordagem *Wrapper*

	Algoritmos						
Bases de dados	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
LungCancer-Michigan	100,0 ± 0	98,89 ± 3,51	100,0 ± 0	100,0 ± 0	99,00 ± 3,16	100,0 ± 0	100,0 ± 0
LungCancer-Ontario	84,17 ± 18,19	85,00 ± 17,48	82,50 ± 16,87	95,00 ± 10,54	95,00 ± 10,54	92,50 ± 12,08	84,17 ± 18,19
LungCancer-Harvard	98,05 ± 4,12	94,60 ± 3,56	97,02 ± 2,58	99,50 ± 1,58	99,02 ± 2,06	97,07 ± 3,45	98,05 ± 2,52

Fonte: Autoria própria

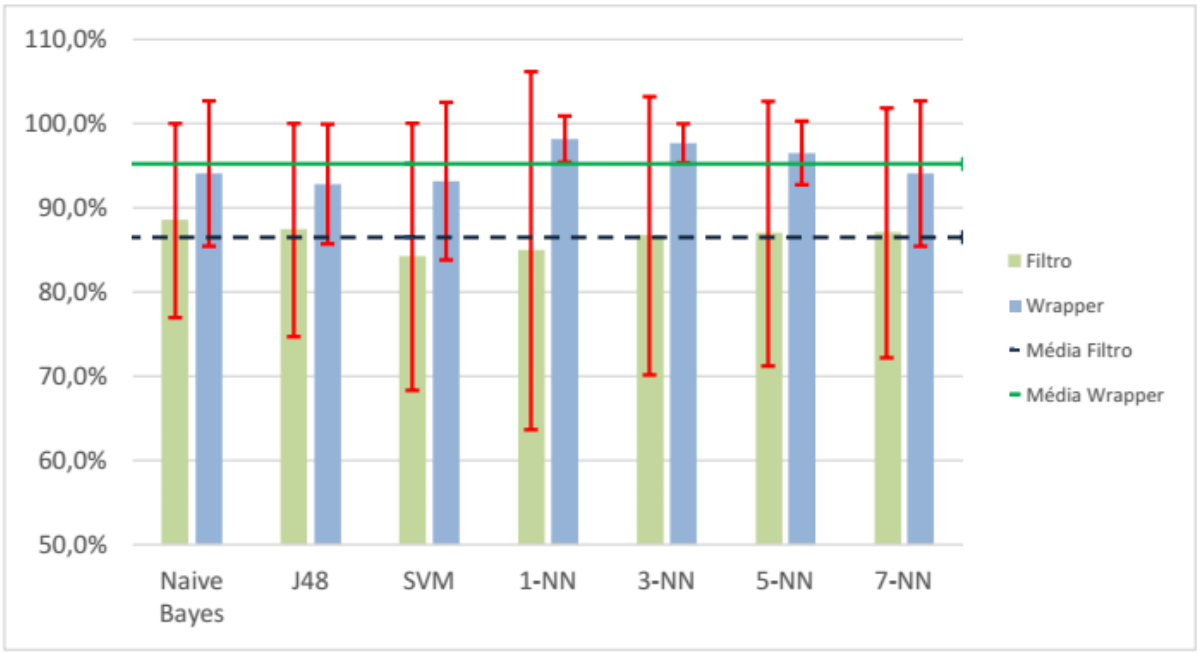
Em relação à base LungCancer-Michigan, observa-se que obteve o melhor desempenho dentre as bases analisadas, com média de 99,7% de taxa de acerto. Apenas os classificadores J48 e 3-NN, ficaram abaixo da média, comparados ao algoritmo Naive Bayes, entretanto com diferenças mínimas, significativamente.

Para a base LungCancer-Ontario, apenas o algoritmo SVM, apresentou resultado estatisticamente pior que o algoritmo base. Obteve média de 88,33% de taxa de acerto, o que corresponde à base com menor desempenho.

A base LungCancer-Harvard teve como melhores classificadores os algoritmos 1-NN, 3-NN e 7-NN, por apresentarem valores superiores ou equivalentes ao algoritmo Naive Bayes. Para os demais algoritmos, não há diferenças significativas ao se comparar com o algoritmo base. A média de taxa de acerto para essa base é de 97,62%.

No Gráfico 2, a partir dos dados descritos em ambas as abordagens, uma comparação das médias é apresentada entre as abordagens Filtro e *Wrapper* para as três bases de dados analisadas.

Gráfico 2 – Média das taxas de acerto para as abordagens Filtro e *Wrapper*, nas três bases de dados analisadas



Fonte: Autoria própria

Analisando as duas abordagens, é possível concluir que, a abordagem *Wrapper* apresentou média de 95,22% de taxa de acerto, enquanto que a abordagem

Filtro apresentou 86,51%. Logo, têm-se que a abordagem *Wrapper* apresentou desempenho superior à abordagem Filtro.

Alguns dos genes que foram selecionados com mais frequência são apresentados no Apêndice A. Os genes escolhidos poderão auxiliar na descoberta de possíveis alterações nos padrões de expressão.

4.2 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

Neste tópico, serão vistos os resultados condizentes ao método PCA, apresentando os resultados das taxas de acerto para cada algoritmo, além de retratar o subconjunto que apresentou melhor desempenho preditivo. O critério para sua utilização foi elaborado de acordo com a porcentagem de variância nos dados originais, que, no estudo em questão, foi estabelecido 90%, 95% e 99%.

As Tabelas 5, 6 e 7 abaixo, correspondem à utilização do método, nas bases LungCancer-Michigan, LungCancer-Ontario e LungCancer-Harvard, respectivamente.

Tabela 5 – Resultados dos classificadores utilizando PCA para a base LungCancer-Michigan

	Algoritmos						
Variância	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
90% dos atributos	99,00 ± 3,16	98,89 ± 3,51	100,0 ± 0	95,78 ± 7,22	96,78 ± 5,20	92,67 ± 8,69	93,78 ± 7,15
95% dos atributos	99,00 ± 3,16	98,89 ± 3,51	98,00 ± 4,22	96,89 ± 5,02	96,89 ± 5,02	92,89 ± 4,92	92,89 ± 4,92
99% dos atributos	99,00 ± 3,16	98,89 ± 3,51	91,67 ± 4,42	91,67 ± 4,42	89,56 ± 0,57	89,56 ± 0,57	89,56 ± 0,57

Fonte: Autoria própria

Tabela 6 – Resultados dos classificadores utilizando PCA para a base LungCancer-Ontario

	Algoritmos						
Variância	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
90% dos atributos	69,17 ± 10,43	67,50 ± 28,99	63,33 ± 19,72	68,33 ±25,09	73,33 ± 23,17	63,33 ± 19,72	58,33 ± 25,46
95% dos atributos	75,83 ± 24,67	57,50 ± 28,99	55,83 ± 21,89	54,17 ± 27,85	60,83 ± 22,92	60,83 ± 25,78	53,33 ± 20,86
99% dos atributos	58,33 ± 28,05	62,50 ± 29,46	62,50 ± 27,00	55,83 ± 21,89	74,17 ± 16,87	64,17 ± 12,45	58,33 ± 15,21

Fonte: Autoria própria

Tabela 7 – Resultados dos classificadores utilizando PCA para a base LungCancer-Harvard

	Algoritmos						
Variância	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
90% dos atributos	79,45 ± 10,50	84,21 ± 8,10	87,69 ± 2,55	75,86 ± 3,68	77,90 ± 5,40	75,40 ± 3,66	75,40 ± 3,73
95% dos atributos	74,98 ± 10,53	84,71 ± 7,97	79,31 ± 3,10	73,38 ± 4,27	73,93 ± 3,68	72,95 ± 4,39	72,98 ± 4,23
99% dos atributos	69,55 ± 7,57	84,71 ± 8,95	72,95 ± 4,45	69,95 ± 3,58	68,50 ± 3,88	68,50 ± 2,00	68,50 ± 2,00

Fonte: Autoria própria

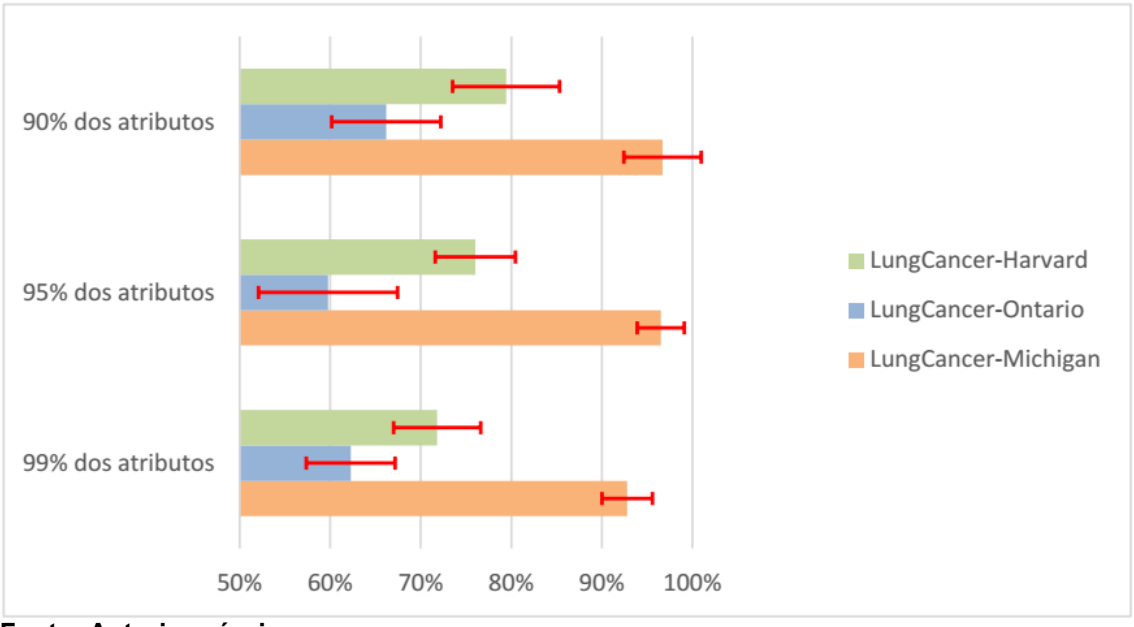
Analisando a Tabela 5, observa-se que apenas o algoritmo SVM com porcentagem de 90% dos atributos, apresentou taxa de acerto superior ao algoritmo base. Constatou-se que os classificadores 5-NN e 7-NN apresentaram os piores resultados nas três experiências, as feitas com 90%, 95% e 99% dos atributos originais. Nessa última experiência, o algoritmo 3-NN também apresentou valor proporcional aos piores resultados.

Na Tabela 6, evidencia-se que apenas o algoritmo 3-NN apresentou, para as experiências feitas com 90% dos atributos originais, desempenho estatisticamente melhor do que o algoritmo Naive Bayes. Todos os algoritmos classificadores obtiveram resultados estatisticamente piores para as experiências feitas com 95% dos atributos, ao se comparar com o algoritmo Naive Bayes. Os algoritmos J48, SVM, 3-NN, 5-NN e 7-NN apresentaram valores significativamente maiores e equivalentes ao algoritmo base, para experiências com 99% dos atributos originais.

Em relação à Tabela 7, nota-se que os algoritmos J48 e SVM apresentaram desempenho superior, comparando-se com o algoritmo Naive Bayes, nas três experiências. Para o subconjunto com 99% dos atributos, o algoritmo 1-NN, apresentou aumento mínimo significativo em relação ao algoritmo base. Os demais algoritmos apresentaram resultados estatisticamente piores.

O Gráfico 3 apresenta uma comparação das médias das taxas de acerto, utilizando os três experimentos (90%, 95% e 99%) nas três bases estudadas.

Gráfico 3 – Média das taxas de acerto para as experiências com 90%, 95% e 99%, nas três bases de dados analisadas



Fonte: Autoria própria

Por meio da análise do Gráfico 3, a experiência que apresentou melhor desempenho foi aquela realizada com 90% dos atributos, tendo como média 80,77%, seguida das experiências com 95% dos atributos, com média de 77,43% e, posteriormente, 99% dos atributos com 75,62%.

A próxima seção irá comparar os métodos empregados nessa pesquisa.

4.3 COMPARAÇÃO ENTRE OS MÉTODOS DE REDUÇÃO DE DIMENSIONALIDADE

Esta seção tem o intuito de comparar o método que obteve melhor desempenho nas bases de dados escolhidas. Primeiramente, serão evidenciadas a quantidade de subconjuntos selecionados/transformados após a aplicação dos dois métodos empregados no estudo nas três bases de dados e, em seguida, suas respectivas comparações.

A Tabela 8 apresenta a quantidade de atributos selecionados nas bases LungCancer-Michigan, LungCancer-Ontario e LungCancer-Harvard para o método de Seleção de Atributos nas três bases estudadas.

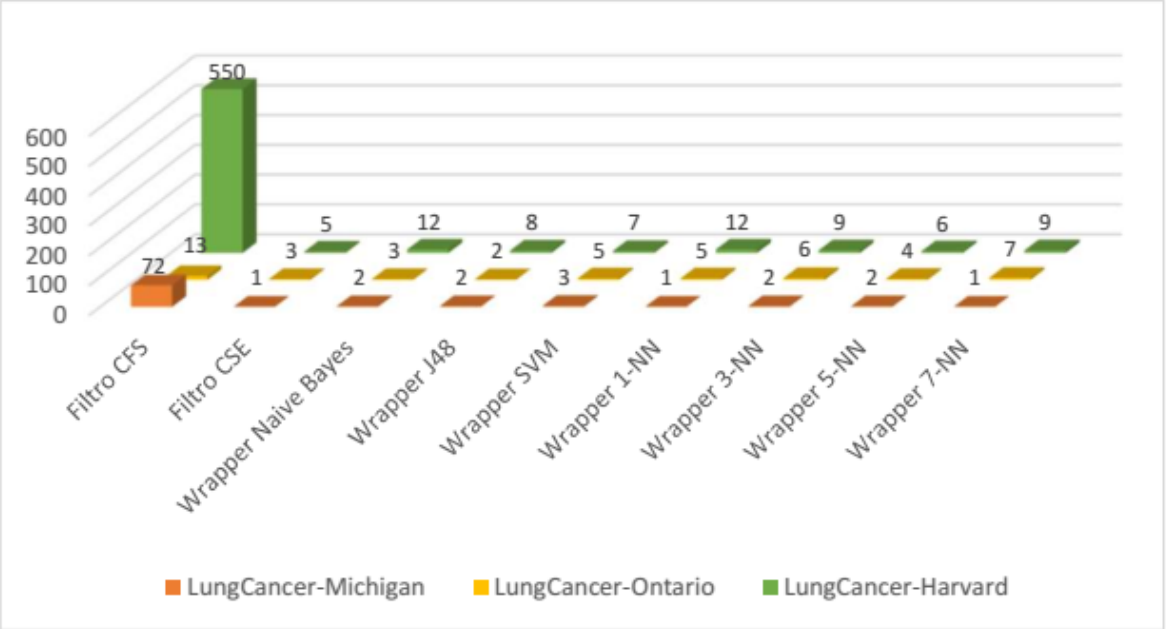
Tabela 8 – Quantidade de atributos selecionados pelo método de Seleção de Atributos nas bases de dados analisadas

			Base de dados		
		Algoritmo	**LungCancer-Michigan**	**LungCancer-Ontario**	**LungCancer-Harvard**
Seleção de Atributos	**Total**	–	7129	2880	12600
	Filtro	CFS	72	13	550
		CSE	1	3	5
	Wrapper	Naive Bayes	2	3	12
		J48	2	2	8
		SVM	3	5	7
		1-NN	1	5	12
		3-NN	2	6	9
		5-NN	2	4	6
		7-NN	1	7	9

Fonte: Autoria própria

Analisando os subconjuntos gerados, destaca-se que o método reduziu significativamente o número de atributos pelo método de Seleção de Atributos. O Gráfico 4 apresenta a quantidade de atributos selecionados graficamente.

Gráfico 4 – Quantidade de atributos selecionados pelo método de Seleção de Atributos nas três bases de dados analisadas



Fonte: Autoria própria

Já, em relação ao método PCA, a quantidade de novos atributos derivados dos atributos originais, são apresentados na Tabela 9.

Tabela 9 – Quantidade de atributos transformados a partir do método PCA nas três bases de dados analisadas

		Base de dados		
	Variância	**LungCancer-Michigan**	**LungCancer-Ontario**	**LungCancer-Harvard**
PCA	–	7129	2880	12600
	90%	64	9	108
	95%	77	15	146
	99%	91	28	188

Fonte: Autoria própria

A fim de comparar os métodos utilizados, se faz necessário apresentar os resultados encontrados, para assim verificar a eficiência dos mesmos. Os resultados serão discutidos separadamente para cada base de dados.

A Tabela 10 apresenta os resultados referentes à base LungCancer-Michigan, seguida da Tabela 11, que se refere à base LungCancer-Ontario e, por fim, a Tabela 12 que apresenta os dados pertencentes à base LungCancer-Harvard.

Tabela 10 – Resultados dos métodos de Seleção de Atributos e Análise de Componentes Principais para a base LungCancer-Michigan

	Algoritmos						
Subconjuntos	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
Todos os atributos	100,0 ± 0	99,0 ± 3,16	100,0 ± 0	100,0 ± 0	98,89 ± 3,51	98,89 ± 3,51	100,0 ± 0
Filtro-Dependência	100,0 ± 0	99,0 ± 3,16	99,0 ± 3,16	100,0 ± 0	100,0 ± 0	100,0 ± 0	100,0 ± 0
Filtro-Consistência	96,78 ± 5,20	99,0 ± 3,16	95,89 ± 5,32	98,00 ± 4,22	99,00 ± 3,16	99,00 ± 3,16	99,00 ± 3,16
Wrapper	100,0 ± 0	98,89 ± 3,51	100,0 ± 0	100,0 ± 0	99,00 ± 3,16	100,0 ± 0	100,0 ± 0
90% dos atributos	99,00 ± 3,16	98,89 ± 3,51	100,0 ± 0	95,78 ± 7,22	96,78 ± 5,20	92,67 ± 8,69	93,78 ± 7,15
95% dos atributos	99,00 ± 3,16	98,89 ± 3,51	98,00 ± 4,22	96,89 ± 5,02	96,89 ± 5,02	92,89 ± 4,92	92,89 ± 4,92
99% dos atributos	99,00 ± 3,16	98,89 ± 3,51	91,67 ± 4,42	91,67 ± 4,42	89,56 ± 0,57	89,56 ± 0,57	89,56 ± 0,57

Fonte: Autoria própria

Observa-se, pelos dados da Tabela 10, que para o método de seleção de atributos, a abordagem que apresentou média de desempenho maior foi a abordagem *Wrapper* com 99,70% de taxa de acerto, contra 98,91% da abordagem Filtro. Evidencia-se, também, que os algoritmos que tiveram média estatisticamente significativas foram os algoritmos Naive Bayes,1-NN, 3-NN, 5-NN e 7-NN para a medida dependência e, J48, 1-NN, 3-NN, 5-NN e 7-NN, para a medida consistência, comparando-se com o algoritmo base. Para a abordagem *Wrapper*, apenas os algoritmos J48 e 3-NN apresentaram taxas de acertos inferiores ao do algoritmo Naive Bayes.

Em relação ao método PCA, apenas o algoritmo SVM com 90% dos atributos da base original apresentou taxa de acerto superior ao algoritmo base. O subconjunto que obteve maior média foi o subconjunto com 90% de atributos, com 96,70% de taxa de acerto.

Comparando-se os métodos empregados, obtivemos como média para utilização com todos os atributos da base de 99,54% de taxa de acerto. Para o método de seleção de atributos o valor equivale a 99,17%, e para o método de PCA, a média é de 95,35% de taxa de acerto.

Portanto, a utilização com todos os atributos da base apresentou a melhor taxa de acerto, porém pelo método de seleção de atributos, o valor se aproximou ao da base original e, também, não se pode excluir o método PCA, pois apresentou excelentes resultados, pelo fato da taxa de acerto ser superior a 90%, ou seja, ambos os métodos possuem condições de aplicação para essa base de dados.

A Tabela 11 apresenta as informações pertinentes à base LungCancer-Ontario.

Tabela 11 – Resultados dos métodos de Seleção de Atributos e Análise de Componentes Principais para a base LungCancer-Ontario

	Algoritmos						
Subconjuntos	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
Todos os atributos	67,50 ± 35,45	84,17 ± 13,86	78,33 ± 22,97	56,67 ± 19,95	55,83 ± 34,26	57,50 ± 39,18	58,33 ± 32,63
Filtro-Dependência	74,17± 20,58	71,67 ± 21,94	69,17 ± 15,74	44,17 ± 27,51	59,17 ± 31,29	61,67 ± 33,61	64,17± 33,58
Filtro-Consistência	74,17 ± 23,12	71,67 ± 14,27	61,67 ± 12,55	79,17 ± 16,32	74,17 ± 20,58	74,17 ± 20,58	74,17 ± 20,58
Wrapper	84,17 ± 18,19	85,00 ± 17,48	82,50 ± 16,87	95,00 ± 10,54	95,00 ± 10,54	92,50 ± 12,08	84,17 ± 18,19
90% dos atributos	69,17 ± 10,43	67,50 ± 28,99	63,33 ± 19,72	68,33 ±25,09	73,33 ± 23,17	63,33 ± 19,72	58,33 ± 25,46
95% dos atributos	75,83 ± 24,67	57,50 ± 28,99	55,83 ± 21,89	54,17 ± 27,85	60,83 ± 22,92	60,83 ± 25,78	53,33 ± 20,86
99% dos atributos	58,33 ± 28,05	62,50 ± 29,46	62,50 ± 27,00	55,83 ± 21,89	74,17 ± 16,87	64,17 ± 12,45	58,33 ± 15,21

Fonte: Autoria própria

A partir dos dados da Tabela 11, nota-se que a abordagem que apresentou maior taxa de acerto foi a abordagem *Wrapper* para o método de seleção de atributos, com 88,33% de taxa de acerto, contra 67,80% da abordagem Filtro. O algoritmo com melhor desempenho foi o algoritmo 1-NN, para a medida consistência, além dos algoritmos 3-NN, 5-NN e 7-NN, que apresentaram resultados equivalentes ao algoritmo Naive Bayes, utilizando a mesma medida. Para a abordagem *Wrapper*, apenas o algoritmo SVM apresentou taxa de acerto inferior ao do algoritmo Naive Bayes.

Em relação ao método PCA, o algoritmo que apresentou melhor taxa de acerto foi o 3-NN, para subconjuntos com 90% e 99% de atributos da base original, comparando-se com o algoritmo Naive Bayes. Além disso, os algoritmos J48, SVM, 5-NN e 7-NN apresentaram resultados estatisticamente significativos para os subconjuntos com 99% dos atributos. O subconjunto com desempenho melhor é a experiência feita com 90% dos atributos da base, obtendo uma média de 66,19% de taxa de acerto.

Por meio da comparação entre ambos os métodos, observou-se que, o método com melhor desempenho é o de seleção de atributos com 74,84% de taxa de acerto. O método PCA obteve valor de 62,74% e a média para todos os atributos da base foi de 65,48% de taxa de acerto. É importante ressaltar que, embora o método PCA apresentou menor média comparado com todos os atributos, esse obteve média de desempenho superior em alguns dos casos.

Logo, por meio dos valores relatados anteriormente, tem-se que o método de seleção de atributos obteve desempenho estatisticamente superior ao método PCA.

Já na Tabela 12 são analisadas as informações referentes à base LungCancer-Harvard.

Tabela 12 – Resultados dos métodos de Seleção de Atributos e Análise de Componentes Principais para a base LungCancer-Harvard

	Algoritmos						
Subconjuntos	**Naive Bayes**	**J48**	**SVM**	**1-NN**	**3-NN**	**5-NN**	**7-NN**
Todos os atributos	80,38 ± 6,16	93,12 ± 7,50	95,07 ± 3,34	89,71 ± 5,27	91,21 ± 4,89	89,74 ± 5,70	89,24 ± 5,40
Filtro-Dependência	95,60 ± 5,45	93,14 ± 7,36	97,05 ± 3,47	95,07 ± 4,72	97,07 ± 2,52	97,55 ± 2,59	97,05 ± 3,47
Filtro-Consistência	90,19 ± 3,16	89,67 ± 7,45	82,33 ± 7,13	93,07 ± 4,19	90,67 ± 5,82	89,19 ± 5,54	87,74 ± 5,61
Wrapper	98,05 ± 4,12	94,60 ± 3,56	97,02 ± 2,58	99,50 ± 1,58	99,02 ± 2,06	97,07 ± 3,45	98,05 ± 2,52
90% dos atributos	79,45 ± 10,50	84,21 ± 8,10	87,69 ± 2,55	75,86 ± 3,68	77,90 ± 5,40	75,40 ± 3,66	75,40 ± 3,73
95% dos atributos	74,98 ± 10,53	84,71 ± 7,97	79,31 ± 3,10	73,38 ± 4,27	73,93 ± 3,68	72,95 ± 4,39	72,98 ± 4,23
99% dos atributos	69,55 ± 7,57	84,71 ± 8,95	72,95 ± 4,45	69,95 ± 3,58	68,50 ± 3,88	68,50 ± 2,00	68,50 ± 2,00

Fonte: Autoria própria

Para a Tabela 12, é visto que os dados através da seleção de atributos apresentaram valores superiores aos descritos pelo método PCA. Em relação às duas abordagens pertencentes à seleção de atributos, destaca-se a abordagem *Wrapper* como sendo a que obteve melhor desempenho, com 97,62% de taxa de acerto, enquanto que a abordagem Filtro obteve 92,53%. Os algoritmos com melhor desempenho foram o SVM, 3-NN, 5-NN e 7-NN para a medida dependência e 1-NN e 3-NN para a medida consistência.

No que se refere ao método PCA, conclui-se que os algoritmos J48 e SVM apresentaram desempenho superior nos três subconjuntos (90%, 95% e 99%), comparando-se com o Naive Bayes. A média com maior desempenho foi para o subconjunto com 90% dos atributos, tendo 79,42% de taxa de acerto.

Portanto, a partir da análise dos dados da Tabela 12, constata-se que o a base com todos os atributos obteve média de 89,78%. Já os métodos Seleção de Atributos e PCA, obtiveram 94,23% e 75,8% de taxa de acerto, respectivamente. Com isso, apresenta-se como o melhor método, dentre as bases analisadas, o método de Seleção de Atributos. A utilização do método PCA, apresentou resultados inferiores, porém na maioria dos casos, obteve uma boa taxa de acerto.

Assim, a partir dos resultados encontrados, conclui-se que o método de Seleção de Atributos, apresentou os melhores resultados para ambas as bases de dados, seguidamente o método de Análise de Componentes Principais. O PCA, mesmo apresentando resultados inferiores, possui valores dentro da normalidade.

Segundo Borges (2006), um valor ideal para taxa de acerto deve estar entre 70% a 100%. Diante disso, não se deve descartar a aplicação do mesmo.

Com base em levantamentos encontrados na literatura, observou-se que o método de Seleção de Atributos através das abordagens Filtro e *Wrapper*, apresentaram melhor resultado, comparando-se com outros métodos: Projeção Aleatória e DRM-F, tanto no estudo de Borges e Nievola (2012), quanto no estudo de Macedo (2015), para bases envolvendo expressão gênica. O método apresentou médias de 90% de taxa de acerto nos estudos mencionados. Esses resultados são similares aos encontrados no presente estudo, demonstrando sua aplicabilidade e efetividade nesses tipos de dados.

Por fim, através das comparações realizadas, foi possível avaliar os melhores resultados para cada base de dados, utilizando os dois métodos de redução de dimensionalidade.

Após as informações apresentadas, serão relatadas as principais conclusões do trabalho realizado.

5 CONCLUSÃO

Neste trabalho, o objetivo geral foi avaliar a aplicação dos métodos de redução de dimensionalidade, denominados de Seleção de Atributos e Análise de Componentes Principais (PCA), em dados de expressão gênica. Para isso, foi necessário responder aos objetivos específicos traçados.

Todos os objetivos específicos foram trabalhados a partir do processo KDD. O primeiro objetivo específico foi a seleção das bases de dados para a respectiva pesquisa, nesse caso, a partir de dados de expressão gênica. As bases foram: LungCancer-Michigan, LungCancer-Ontario e LungCancer-Harvard, todas referentes ao câncer de pulmão. Levou-se em consideração, principalmente, a credibilidade do repositório de dados biomédicos, sendo o repositório designado foi o *Kent Ridge Biomedical*.

O segundo objetivo específico, tratou-se de aplicar os métodos de redução de dimensionalidade: Seleção de Atributos e Análise de Componentes Principais (PCA) nas respectivas bases. Para o primeiro método, utilizaram-se duas abordagens – a Filtro e a *Wrapper*, enquanto que, para o segundo, as experiências foram de acordo com a porcentagem de variância nos dados originais, que neste trabalho foram 90%, 95% e 99%.

Em seguida, no terceiro objetivo específico tratou-se de aplicar os algoritmos de classificação. Os algoritmos utilizados foram Naive Bayes, J48, SVM, 1-NN, 3-NN, 5-NN e 7-NN, por meio do *software* WEKA, a fim de analisar as taxas de acerto (acurácia), para ambos os algoritmos.

Por fim, no quarto objetivo específico, referente aos resultados e avaliação, notou-se, primeiramente, certa dificuldade quanto ao processamento dos dados, em virtude do alto número de genes em alguma das bases, e poucas informações referentes às amostras.

Em relação à avaliação dos métodos, foram constatadas melhorias com a utilização destes. Em referência ao primeiro método, podemos afirmar que o desempenho, ao utilizar as duas abordagens, Filtro e *Wrapper*, aumentou significativamente a taxa de acerto em relação à obtida com todos os atributos da base, para as três bases de dados estudadas. Além disso, houve uma alta redução na quantidade de atributos ao se utilizar esse método.

A abordagem que apresentou melhores resultados foi a abordagem *Wrapper*, seguida da medida consistência a partir da abordagem Filtro. Destaca-se também que a abordagem *Wrapper* demandou alto tempo computacional, o mesmo não ocorrendo com a abordagem Filtro.

Em relação ao método de Análise de Componentes Principais, verificou-se que os melhores resultados foram encontrados em experiências com 90% dos atributos da base original, ou seja, com uma porcentagem de variância menor do que com as outras experiências analisadas. Notou-se também que, nas bases de dados estudadas, o método PCA apresentou resultados superiores a 80% nas taxas de acerto em alguns casos, tornando-se, portanto, um método viável para redução.

No estudo realizado, os algoritmos que apresentaram os melhores desempenhos nas bases foram: Naive Bayes, SVM e 1-NN.

Assim, tem-se como recomendação, a partir dos estudos feitos previamente, a utilização do método de seleção de atributos para bases de expressão gênica, devido à apresentação de bons resultados para esse tipo de domínio. O método PCA é um método alternativo que, em alguns casos, produz bons resultados.

Deste modo, apresentam-se algumas sugestões para trabalhos futuros.

5.1 SUGESTÃO DE TRABALHOS FUTUROS

No decorrer do desenvolvimento deste estudo foram identificadas algumas oportunidades para o desenvolvimento de trabalhos futuros relacionados ao tema deste trabalho. São elas:

- Aplicar os métodos utilizados nesta pesquisa em outras bases de dados, a fim de comparar o desempenho preditivo entre as bases;
- Comparar outros métodos de redução de dimensionalidade para os mesmos algoritmos classificadores, além das mesmas bases utilizadas neste estudo;
- Utilizar outros algoritmos de classificação, bem como mudança de parâmetros de entrada dos algoritmos;
- Facilitar futuros estudos comparativos quando um pesquisador propõe um novo método, devido à grande quantidade de experimentos realizados.

REFERÊNCIAS

AHA, D.W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. **Machine learning**, v. 6, n. 1, p. 37-66, 1991.

ALBERTS, B.; JOHNSON, A.; LEWIS, J.; RAFF, M.; ROBERTS, K.; WALTER, P. **Biologia Molecular da Célula.** Artes Médicas. 5ª ed. Artmed, 2010.

ALVES, F. C.; FRAGAL, E. H. Avaliação dos Algoritmos MAXVER e SVM na Classificação da Cobertura Vegetacional da Planície de Inundação do Alto Rio Paraná. In: ENCONTRO ESTADUAL DE GEOGRAFIA E ENSINO, II, SEMANA DE GEOGRAFIA, XX, 2011, Maringá. **Anais.**.. Maringá: 2011. p. 0621-0632.

APOLLONI, J.; LEGUIZAMÓN, G; ALBA, E. Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. **Applied Soft Computing**, v. 38, p. 922-932, 2016.

APPELBAUM F.R. Transplante de Células Hematopoiéticas. In: LONGO, D.L. **Hematologia e Oncologia de Harrison**. AMGH Editora, cap.30, p.316, 2015.

BAGINSKY, S.; et al. Update on gene expression analysis, proteomics, and network discovery. **Plant Physiol**., v. 152, 2010.

BARBOSA, E. B.; et al. Proteômica: metodologias e aplicações no estudo de doenças humanas. **Revista da Associação Médica Brasileira.** São Paulo, v. 58, n. 3, p. 366-375, 2012.

BARSHAN, E.; GHODSJ, A.; AZIMIFAR, Z.; JAHROMI, M. Z. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. **Pattern Recognition**, 2011.

BARTL, E.; REZANKOVA,H.; SOBISEK, L. Comparison of classical dimensionality reduction methods with Novel approach based on formal concept analysis. In: YAO, J.;RAMANNA, S.;WANG, G.; SURAJ, Z.(Eds.). Rough sets and knowledge technology (RSKT 2011), October 9–12 2011, Banff, Canada. **Lecture notes in computer science**, v. 6954, p. 26-35, Springer, 2011.

BEER, D. G.; et al. Gene-expression Profiles Predict Survival of Patients with Lung Adenocarcinoma. **Nature Medicine**, v. 8, n. 8, p. 816-824, ago. 2002.

BERTON, L. **Caracterização de classes e detecção de outliers em redes complexas**. 2011. 89f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Universidade de São Paulo. São Carlos, 2011.

BHATTACHARJEE, A.; et al. Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses. **PNAS**, v. 98, n. 24, p. 13790-13795, nov. 2001.

BINGHAM, E.; MANNILA, H. Random projection in dimensionality reduction: applications to image and text data. In: **Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining** (KDD 2001), August 26–29 2001, ACM: San Francisco, CA, USA. pp. 245–250, 2001.

BOENTE, A. N. P.; GOLDSCHMIDT, R. R.; ESTRELA, V. V. Uma metodologia de suporte ao processo de descoberta de conhecimento em bases de dados. In: SIMPÓSIO DE EXCELÊNCIA EM GESTÃO E TECNOLOGIA, 5., 2008. Resende (RJ). **Anais...**2008. v. 1. p. 4-5. Disponível em: <http://www.boente.eti.br/publica/seget2008kdd.pdf>. Acesso em: 10 jun. 2015.

BOLÓN-CANEDO, V.; et al. A review of microarray datasets and applied feature selection methods. **Information Sciences**, v. 282, p. 111-135, 2014.

BORGES, H. B. **Redução de dimensionalidade em bases de dados de expressão gênica**. 2006. 123 f. Dissertação (Mestrado em Informática) - Pontifícia Universidade Católica do Paraná. Curitiba, 2006.

BORGES, H. B.; NIEVOLA, J. C. Comparing the dimensionality reduction methods in gene expression databases. **Expert Systems with Applications**, vol. 39, p. 10780–10795, 2012.

CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudos de caso na Universidade Federal de Lavras. **Revista de Administração Pública,** Fundação Getúlio Vargas, Rio de Janeiro, v. 42, n. 3, p. 495-528, maio/jun. 2008.

CASLEY, D. **Primer on molecular biology**. Technical report, U. S. Department of Energy, Office of Health and Environmental Research, 1992.

CHUANG, L.T.; YANG, C.H.; WU, K.C.; YANG, C.H. A hybrid feature selection method for DNA microarray data. **Computers in Biology and Medicine**, v. 41, p. 228-237, 2011.

CIOS, K. J.; PEDRYCZ, W.; SWINIARSKI, R. W.; KURGAN, L. A. **Data Mining** – A Knowledge Discovery Approach. Springer, 2007.

COLE, Laurence A. Biology of Life: Biochemistry, Physiology and Philosophy. Academic Press, p.57, 2016.

COLOMBO J.; RAHAL, P. A tecnologia de microarray no estudo do câncer de cabeça e pescoço, **Revista Brasileira de Biociências**, v. 8, n. 1, p. 64-72, 2010.

CORTES, C.; VAPNIK, V. Support vector networks. **Machine Learning**, v. 20, n 3, p. 273-297, 1995.

CRUZ, C. S. D.; TANOUE, L. T.; MATTHAY, R.A. Lung cancer: epidemiology, etiology, and prevention. **Clinics in chest medicine**, v. 32, n. 4, p. 605-644, 2011.

DASH, M.; LIU, H. **Feature selection for classification.** Intelligent Data Analysis, 1 (1-4), p. 131–15, 1997.

DUGGAN, D. J.; BITTNER, M.; CHEN, Y.; MELTZER, P.; TRENT, J.M. Expression profiling using cDNA microarrays. **Nature Genetics**, v. 21, p. 10-14, 1999.

DUNTEMAN, G. H. **Principal components analysis**. Sage University paper series on quantitative applications in the social sciences, Newbury Park, CA, USA, 1999.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases.** American Association for Artificial Intelligence, 1996.

FRUTUOSO, D. G. **Recuperação de informação e classificação de entidades organizacionais em textos não estruturados.** 2014. 86 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Pernambuco. Recife, 2014.

GALVÃO, N. D.; MARIN, H.F. Técnica de mineração de dados: uma revisão da literatura. **Acta Paulista de Enfermagem (online),** São Paulo, v. 22, n. 5, p. 686-690. set/out, 2009.

GIL, A. C. Como elaborar projetos de pesquisa. São Paulo: Atlas, 2002. **Métodos e Técnicas de Pesquisa Social**, v. 5, p. 64-73, 2001.

GHOSH, Antara; BARMAN, Soma. Application of Euclidean distance measurement and principal component analysis for gene identification. **Gene**, v. 583, n. 2, p. 112-120, 2016.

GOLDSCHIMIDT, R.; PASSOS, E. **Data mining:** um guia prático. Rio de Janeiro: Campus, 2005.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research,** v. 3, p. 1157-1182, 2003.

HALL, M. A. **Correlation-based feature selection for machine learning**. 198 f. 1999. Thesis (Doctor of Philosophy) - Department of Computer Science. University of Waikato. Hamilton (New Zealand), 1999.

HASTIE, T.; et al. Classification by pairwise coupling. **Annals of statistics**, v. 26, n. 2, p. 451-471, 1998.

HOLMES, S.; et al. Visualization and statistical comparisons of microbial communities using R packages on phylochip data. Bioscomputing 2011: Proceedings of the Pacific Symposium. Hawaii, USA, p. 142-153, 2011.

INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA. **Ministério da Saúde**. Instituto Nacional de Câncer José Alencar Gomes da Silva. Coordenação de Prevenção e Vigilância. Estimativa 2016: incidência de câncer no Brasil / Instituto Nacional de Câncer José Alencar Gomes da Silva. Rio de Janeiro: INCA, 122p, 2015.

INSTITUTO NACIONAL DE CÂNCER JOSÉ ALENCAR GOMES DA SILVA. COORDENAÇÃO GERAL DE AÇÕES ESTRATÉGICAS. COORDENAÇÃO DE EDUCAÇÃO. **ABC do câncer**: abordagens básicas para o controle do câncer, 2011. Disponível em: < http://bvsms.saude.gov.br/bvs/publicacoes/inca/abc_do_cancer_2ed.pdf>. Acesso em: 10 fev. 2017.

JOHN, G. H.; LANGLEY, Pat. Estimating continuous distributions in Bayesian classifiers. In: **Proceedings of the Eleventh conference on Uncertainty in artificial intelligence**. Morgan Kaufmann Publishers Inc., 1995. p. 338-345.

JOLLIFFE, I.T. **Principal Component Analysis** (2. ed.) Springer, New York, 2002.

KAMBER, M.; HAN, J.; PEI, J. **Data mining:** Concepts and techniques. Elsevier, 2012.

KDNUGGETS POLL. **KDnuggets Annual Software Poll**: Using Data Science software in 2013. Disponível em: < http://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html>. Acesso em: 01 abr. 2017.

KENT RIDGE Bio-medical Dataset. Disponível em: <http://datam.i2r.a-star.edu.sg/datasets/krbd//>. Acesso em: 30 abr. 2015.

KHALILABAD, N. D.; HASSANPOUR, H. Employing image processing techniques for cancer detection using microarray images. **Computers in Biology and Medicine**, v. 81, p. 139-147, 2017.

KIRA, K.; RENDELL, L. A. The feature selection problem: traditional methods and a new algorithm. In: CONFERENCE ON ARTIFICIAL INTELLIGENCE, 10., **Proceedings...** Menlo Park (CA), p. 129-136, 1992.

KOHAVI R.; JOHN, G. H. The Wrapper Approach. In: LIU, H.; MOTODA (Eds.). **Proceedings of the feature extraction, construction and selection**: A data mining perspective, 1998, p. 33–49.

KOPRINSKA, I.; RANA, M.; AGELIDIS, V. G. Correlation and instance based feature selection for electricity load forecasting. **Knowledge-Based Systems**, v.82, p.29-40, 2015.

KUMAR, Vinay; ASTER, Jon C.; ABBAS, Abbas. **Robbins & Cotran Patologia-Bases Patológicas das Doenças**. Elsevier Brasil, 2015.

LATKOWSKI, T.; OSOWSKI, S. Data mining for feature selection in gene expression autism data. **Expert Systems with Applications**, v. 42, n. 2, p. 864-872, 2015.

LAUSCH, A.; SCHMIDT, A.; TISCHENDORF, L. Data mining and linked open data– New perspectives for data analysis in environmental research. **Ecological Modelling**, v. 295, p. 5-17, 2015.

LAZAR, C.; TAMINAU, J.; MEGANCK, S.; STEENHOFF, D.; COLETTA, A.; MOLTER, C.; SCHAETZEN, V.; DUQUE, R.; BERSINI, H.; NOWE, A. A survey on Filtro techniques for feature selection in gene expression microarray analysis. **IEEE/ACM Transactions on Computacional Biology and Bioinformatics**, v. 9, p.1106-1119, 2012.

LINS, T. S.; MERSCHMANN, L. H. C. **Técnicas de seleção de atributos utilizando paradigmas de algoritmos Disciplina de Projeto e Análise de Algoritmos**. Universidade Federal de Ouro Preto. Ouro Preto, Minas Gerais, 2010.

LIU, H.; YU, L. Toward Integrating Feature Selection Algorithms for Classification Vand Clustering. **IEEE Transactions on Knowledge and Data Engineering**, v.17, n. 4, p. 491-502, 2005.

LIU, L.; SO, A. Y.L.; FAN, J.B. Analysis of cancer genomes through microarrays and next-generation sequencing. **Translational Cancer Research**, v. 4, n. 3, p. 212-218, 2015.

LODISH, H.; et al. **Biologia celular e molecular**. Artmed Editora, 7° ed. 2014.

MACEDO, D.C. **Comparação da redução de dimensionalidade de dados usando seleção de atributos e conceitos de framework:** um experimento no domínio de clientes. Dissertação (Mestrado em Engenharia de Produção) – Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2012.

MACEDO, D.C.; ISHIKAWA, E.C.M.; SANTOS, C.B.; MATOS, S.N.; BORGES, H.B.; FRANCISCO, A.C. Proposed method for dimensionality reduction based on framework in gene expression domain. **Genetics and Molecular Research**, v.13, p.10597-10606, 2014.

MACEDO, D.C. **Redução de dimensionalidade de dados derivados do domínio de expressão gênica:** uma proposta de método. 2015. 138 f. Tese (Doutorado em Engenharia de Produção) – Universidade Tecnológica Federal do Paraná. Ponta Grossa, 2015.

MALUCELLI, A.; et al. Classificação de microáreas de risco com uso de mineração de dados. **Revista de Saúde Pública**, São Paulo, v. 44, n. 2, abr. 2010.

MARKOS, A. I.; VOZALIS, M. G.; MARGARITIS, K. G. **An optimal scaling approach to collaborative filtering using categorical principal component analysis and neighborhood formation**. In H. Papadopoulos, A. S. Andreou, & M. Bramer (Eds.), Artificial intelligence applications and innovations (AIAI 2010), October 6–7 2010. Larnaca, Cyprus: Proceedings. IFIP Advances in information and communication technology (v. 339, p. 22-29). Springer, 2010.

MITCHELL, T. M. **Machine learning.** Boston: WCB/McGraw-Hill, 1997.

MORIN, P.J.; et al. Genética do Câncer. In: KASPER, D. L.; HAUSER, S.L.; JAMESON, J.L.; FAUCI, A.S.; LONGO, D.L.; LOSCALZO, J. **Medicina Interna de Harrison**. McGraw, cap.7, p101e., 2016.

_____. **Generative and Discriminative classifiers**: Naive Bayes and Logistic Regression. Machine Learning, Projeto da 2.ed, jan. 2010.

NASCIMENTO, M. C.; TOLEDO, F. M.; DE CARVALHO, A. C. Investigation of a new GRASP-based clustering algorithm applied to biological data. **Computers & Operations Research**, v. 37, n. 8, p. 1381-1388, 2010.

PEREIRA, M. A.L. S. C. **Manifestações neurológicas como forma de apresentação do cancro do pulmão**. 2009. 44f. Dissertação (Mestrado Integrado em Medicina) – Faculdade de Medicina da Universidade de Coimbra, Coimbra, 2009.

PERMONIAM, V. A. **Visualização Exploratória de Dados do Desempenho na Aprendizagem em um Ambiente Adaptável**. 2008. 124f. Tese (Doutorado em Engenharia Elétrica) – Universidade de São Paulo, Escola de Engenharia de São Carlos, São Carlos, 2008.

PRIETO-MORENO, A.; LLANES-SANTIAGO, O.; GARCÍA-MORENO, E. Principal components selection for dimensionality reduction using discriminant information applied to fault diagnosis. **Journal of Process Control**, v. 33, p. 14-24, 2015.

QUINLAN, J. R. **C4.5**: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

REZENDE, S. O. **Sistemas inteligentes:** fundamentos e aplicações. Barueri (SP): Manole, 2003.

SAEYS, Y.; INZA, I.; LARRANAGA, P. A review of feature selection techniques in bioinformatics, **Bioinformatics**, v. 23, p. 2507-2517, 2007.

SAKTHIVEL, N. R.; et al. Comparison of dimensionality reduction techniques for the fault diagnosis of mono block centrifugal pump using vibration signals. **Engineering Science and Technology, an International Journal**, v. 17, n. 1, p. 30-38, 2014.

SCHENA, M.; SHALON, D.; DAVIS, R. W.; BROWN, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. **Science**, v. 270, p.467–470, 1995.

SCHUCH, R.; DILL, S. L.; SUASEN, P. S.; PADOIN, E. L.; CAMPOS, M. Mineração de dados em uma subestação de energia elétrica. In: PROCEEDINGS OF THE 9TH BRAZILIAN CONFERENCE ON DYNAMICS, CONTROLAND THEIR APPLICATIONS – DINCON'10. **Anais...** Serra Negra, p. 804-810, jun. 2010.

SETÚBAL, J. C.; MEIDANIS, J. **Introduction to Computational Molecular Biology**. PWS Publishing Company, 1997.

SHLENS, J. A **Tutorial on Principal Component Analysis**. La Jolla, California, USA: Systems Neurobiology Laboratory, Salk Institute for Biological Studies, 2005.

SILVA, F. H. **Apostila - curso de biologia molecular**. INBIO - I Escola Brasileira de Inteligência Artificial e Bioinformática, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. 2001. Disponível em:<http:// genfis40.esalq.usp.br/downloads/biologia_molecular.pdf>. Acesso em: 04 jun. 2015.

SILVA, L. A. F.; PASSOS, N. S. **DNA Forense** - coleta de amostras biológicas em locais de crime para estudo do DNA. Maceió: Edufal - Editora Gráfica da Universidade Federal de Alagoas, 2002.

SONG, M.; YANG, H.; SIADAT, S.; PECHENIZKIY, M. A comparative study of dimensionality reduction techniques to enhance trace clustering performances. **Expert Systems with Applications**, v. 40, pp. 3722–3737, 2013.

SORENSEN, K.D.; ORNTOFT, T.F. Discovery of prostate cancer biomarkers by microarray gene expression profiling. **Expert Review of Molecular Diagnostics**, v.10, n.1, p.49-64, 2010.

SOUZA, C. P. **Perfil de expressão de microRNAs e seus alvos moleculares em carcinoma pulmonar**. 2016. 81f. Tese (Doutorado em Bases Gerais da Cirurgia). Universidade Estadual Paulista. Botucatu, 2016.

SOUZA, M.C. **Câncer de pulmão:** tendências de mortalidade e fatores associados à sobrevida dos pacientes do Instituto Nacional de Câncer José Alencar Gomes da Silva. 2012. 172 f. Tese (Doutorado em Ciências) – Escola Nacional de Saúde Pública Sergio Arouca. Rio de Janeiro, 2012.

SOUZA, P. C. **Estudo da participação do colágeno V no câncer de pulmão, especificamente no carcinoma não de pequenas células**. 2011. 219f.Tese (Doutorado em Ciências). Faculdade de Medicina da Universidade de São Paulo. São Paulo, 2011.

SUH, S. C. **Practical applications of data mining**. Jones & Bartlett Publishers, 2012.

TAVARES, C.; BOZZA, D.; KONO, F. Descoberta de conhecimento aplicado a dados eleitorais. **Revista Gestão & Conhecimento,** v. 5, n. 1, p. 54-9, jan./jun. 2007: Disponível em: <http://gc.facet.br/v5n1/pdf/descoberta_de_conhecimento_aplicado_a_dados_eleitorais.pdf>. Acesso em: 08 jun. 2015.

TORRE, L. A.; BRAY, F.; SIEGEL, R.L.; FERLAY, J.; LORTET-TIEULENT, J.; JERNAL, A. **Global cancer statistics**, 2012. *CA Cancer J Clin.*65: p.87-108, 2015.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. Springer-Verlag, 1995.

VOET, D.; VOET, J.G.; **Bioquímica**, 4ª edição. Artmed Editora, 2013.

WANG, G.; SONG, Q.; XU, B.; ZHOU, Y. Selecting feature subset for high dimensional data via the propositional foil rules. **Pattern Recognition**, v. 46, p.199-214, 2013a. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320312003469>. Acesso em: 30 abr. 2016.

WANG, J.; WU, L.; KONG, J.; YUXIN, LI.; ZHANG, B. Maximum weight and minimum redundancy: a novel *Framework* for feature subset selection. **Pattern Recognit**, v. 46, p. 1616-1627, jun. 2013b.

WEKA. The University of Waikato. Software. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 26 abr. 2016.

WORLD HEALTH ORGANIZATION. **Cancer**. Disponível em: <http://www.who.int/cancer/en/>. Acesso em: 30 abr.2017.

WIGLE, D. A.; et al. Molecular Profiling of Non-Small Cell Lung Cancer and Correlation with Disease-free Surviva. **Cancer Research**, v. 62 p. 3005-3008, jun. 2002.

WITTEN, I. H.; FRANK, E.; HALL, M, A. **Data mining**: Practical machine learning tools and techniques (2nd ed.). San Francisco: Morgan Kaufmann, 2005.

WU, X.; et al. Top 10 algorithms in data mining. **Knowledge and information systems**, v. 14, n. 1, p. 1-37, 2008.

XU, X.; WANG, X**. An adaptive network intrusion detection method based on PCA and support vector machines**. In X. LI, S.; WANG, Z. Y. Dong (Eds.), Advanced data mining and applications, first international conference (ADMA 2005), July 22–24, 2005. Wuhan, China: Proceedings. Lecture notes in computer science (v. 3584, p. 696–703). Springer, 2005.

YAMAGUCHI, J. K. **Diretrizes para a escolha de técnicas de visualização aplicadas no processo de extração do conhecimento**. 2010. 182f. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual de Maringá, Maringá, 2010.

YANG, H. HARRINGTON, C. A.; VARTANIAN, K.; COLDREN, C. D.; HALL, R.; CHURCHILL, G. A. Randomization in laboratory procedure is key to obtaining reproducible microarray results. **PLoS One**, 3:e3724, 2008.

YANG, Y.; LIU, X. **A re-examination of text categorization methods**. In: SIGIR'99 Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley. ACM Press, p. 42-49, 1999.

YIN, H.; HUANG, W. Adaptive nonlinear manifolds and their applications to pattern recognition**. Information Sciences**, 180(14):2649-62, 2010.

ZATZ, M. A biologia molecular contribuindo para a compreensão e a prevenção das doenças hereditárias. **Ciência e Saúde Coletiva**, v.7, n.1, p.85-99, 2002.

ZHANG, Z.; JIANG, M.; YE, N. Effective multiplicative updates for non-negative discriminative learning in multimodal dimensionality reduction. **Artificial Intelligence Review**, v.34, n.3, p. 235–260, 2010.

ZOU, D.; MA, L.; YU, J.; ZHANG, Z. Biological databases for human research. **Genomics, proteomics & bioinformatics**, v. 13, n. 1, p. 55-63, 2015.

APÊNDICE A - Genes Selecionados

Genes selecionados

A fim de apresentar os conhecimentos gerados, serão descritos os genes que foram selecionados com mais frequência e o percentual de vezes que apareceram nos experimentos nas bases de dados estudadas para o método de Seleção de Atributos. A Tabela 13 refere-se aos genes com relação a base LungCancer-Michigan, seguidamente da Tabela 14, que se refere à base LungCancer-Ontario e a Tabela 15 para a base LungCancer-Harvard.

Outra questão importante é que os genes identificados são fundamentados nos estudos feitos pelos autores de cada base de dados. As informações pertinentes aos genes podem ser vistas em instituições de pesquisa como *National Center For Biotechnology Informations* – NCBI, *Affymetric*, *Ensembl*, entre outros.

Tabela 13 - Genes selecionados com mais frequência na base LungCancer-Michigan pelo Método de Seleção de Atributos

Atributo	Percentual
J02874_at	33,3%
L34657_at	44,4%
U60115_at	44,4%
X64559_at	22,2%

Fonte: Autoria própria

Tabela 14 – Genes selecionados com mais frequência na base LungCancer-Ontario pelo método de Seleção de Atributos

Atributo	Percentual
281898	44,4%
302402	33,3%
259592	22,2%
50416	22,2%
470571	22,2%

Fonte: Autoria própria

Tabela 15 – Genes selecionados com mais frequência na base LungCancer-Harvard pelo método de Seleção de Atributos

Atributo	Percentual
33529_at	22,2%
34981_at	22,2%
36743_at	33,3%
38138_at	33,3%

(continua)

(continuação)

41453_at	33,3%
31791_at	33,3%
32650_at	22,2%
33322_i_at	22,2%
34708_at	22,2%
35669_at	44,4%
37639_at	22,2%
38689_at	33,3%
38700_at	22,2%
33904_at	22,2%
36105_at	22,2%
37406_at	22,2%
38032_at	22,2%
41325_at	22,2%
1814_at	44,4%
598_at	22,2%
39990_at	22,2%

Fonte: Autoria própria