UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

FERNANDA LORENÇON

TÉCNICAS DE MINERAÇÃO DE DADOS PARA ANÁLISE DE UMA BASE DE PACIENTES COM CÂNCER DE ESÔFAGO NO PERÍODO DE 1998 A 2017

TRABALHO DE CONCLUSÃO DE CURSO

MEDIANEIRA

2019

FERNANDA LORENÇON

TÉCNICAS DE MINERAÇÃO DE DADOS PARA ANÁLISE DE UMA BASE DE PACIENTES COM CÂNCER DE ESÔFAGO NO PERÍODO DE 1998 A 2017

Trabalho de Conclusão de Curso apresentado ao Departamento Acadêmico de Computação da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do título de "Bacharel em Computação".

Orientador: Prof. Dr. Alan Gavioli

MEDIANEIRA

2019

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Diretoria de Graduação e Educação Profissional
Coordenação do Curso de Ciência da Computação

TERMO DE APROVAÇÃO

TÉCNICAS DE MINERAÇÃO DE DADOS PARA ANÁLISE DE UMA BASE DE PACIENTES COM CÂNCER DE ESÔFAGO NO PERÍODO DE 1998 A 2017

Por

FERNANDA LORENÇON

Este Trabalho de Conclusão de Curso foi apresentado às 10:20h do dia 21 de novembro de 2019 como requisito parcial para a obtenção do título de Bacharel no Curso de Ciência da Computação, da Universidade Tecnológica Federal do Paraná, Câmpus Medianeira. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. Alan Gavioli
UTFPR - Câmpus Medianeira

Prof. Everton Coimbra de Araújo
UTFPR - Câmpus Medianeira

Prof. Pedro Luiz de Paula Filho
UTFPR - Câmpus Medianeira

A folha de aprovação assinada encontra-se na Coordenação do Curso.

RESUMO

LORENÇON, Fernanda. TÉCNICAS DE MINERAÇÃO DE DADOS PARA ANÁLISE DE UMA BASE DE PACIENTES COM CÂNCER DE ESÔFAGO NO PERÍODO DE 1998 A 2017. 70 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2019.

O câncer é uma doença responsável por milhares de novos casos por ano, no qual são coletados e armazenados dados para acompanhamento de casos e de pacientes da doença. No Brasil esses dados são armazenados em uma base de dados nacional, chamada registros hospitalares de câncer. Dados na área da saúde têm ganhado interesse para utilização na descoberta de conhecimento em bases de dados, portanto este trabalho visou aplicar técnicas de mineração de dados para identificação de padrões relevantes e perfis de pacientes diagnosticados com câncer de esôfago, já que o diagnóstico precoce é essencial. Realizando mineração de regras de associação, classificação e agrupamento dos dados e aplicando métodos de seleção de variáveis em variáveis relacionadas a fatores anteriores ao diagnóstico, com auxílio da ferramenta WEKA e a base de dados disponibilizada no integrador de registros hospitalares de câncer. Como resultados, o algoritmo *apriori* minerou regras de associação, mas acabou influenciado pelos valores predominantes de algumas variáveis. Os algoritmos de classificação, J48 e *reptree* foram testados com diversas configurações e conseguiram um percentual de acerto satisfatório, cerca de 71%, ao classificar corretamente as instâncias para a classe referente a localização primária do tumor. Os algoritmos de agrupamento de dados, *k-means* e sobretudo o EM, mostraram bom desempenho ao agrupar avaliando *clusters* com relação a classe quando utilizados principalmente sem o filtro de instâncias *resample*. Os métodos aplicados mostraram ser capazes de classificar e agrupar os casos de câncer de esôfago, assim como revelaram padrões interessantes na relação das variáveis selecionadas por métodos de seleção de variáveis.

Palavras-chave: Análise de componentes principais, regras de associação, agrupamento, classificação

ABSTRACT

LORENÇON, Fernanda. TÉCNICAS DE MINERAÇÃO DE DADOS PARA ANÁLISE DE UMA BASE DE PACIENTES COM CÂNCER DE ESÔFAGO NO PERÍODO DE 1998 A 2017. 70 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2019.

Cancer is a disease responsible for thousands of new cases per year, so data are collected and stored to track cases and patients of the disease. In Brazil this data is stored in a national database called hospital cancer records. Health data has gained interest for use in knowledge discovery in databases, so this work aimed to apply data mining techniques to identify relevant patterns and profiles of patients diagnosed with esophageal cancer, since early diagnosis is essential. Mining rules of association, classification and grouping of data and applying methods of variable selection in variables related to factors prior to diagnosis, with the help of WEKA tool and the database available in the hospital cancer registry integrator. As a result, the apriori algorithm mined association rules, but was eventually influenced by the predominant values of some variables. The classification algorithms, J48 and reptree were tested with several configurations and achieved a satisfactory hit percentage, about 71%, by correctly classifying instances for the class referring to the primary tumor location. Data grouping algorithms, k-means and especially EM, showed good performance when grouping by evaluating clusters against class when used mainly without the resample instance filter. The applied methods were able to classify and group the esophageal cancer cases, as well as revealing interesting patterns in the relation of the variables selected by variable selection methods.

Keywords: Principal components analysis, association rules, clustering, classification

LISTA DE FIGURAS

FIGURA 1 – Etapas do processo de KDD. 12
FIGURA 2 – Hierarquia do aprendizado. 14
FIGURA 3 – Conjuntos obtidos na primeira etapa do *apriori*. 18
FIGURA 4 – Exemplos de árvore de decisão. 19
FIGURA 5 – Algoritmos hierárquicos aglomerativos e divisivos. 21
FIGURA 6 – Exemplo de agrupamento hierárquico. 21
FIGURA 7 – Exemplo de agrupamento particional. 22
FIGURA 8 – Exemplo de agrupamento com o algoritmo *k-means*. 23
FIGURA 9 – Exemplo de agrupamento baseado em modelos. 23
FIGURA 10 – O que é o câncer 27
FIGURA 11 – Fluxo de informações do IRHC. 31
FIGURA 12 – Número total de registros de casos de câncer de esôfago no Brasil. 31
FIGURA 13 – Interface da página de download da base de dados do IRHC. 32
FIGURA 14 – Interface inicial WEKA. 34
FIGURA 15 – Fluxograma de atividades para realização do trabalho. 35
FIGURA 16 – Diagrama de atividades da aplicação do método de associação para a Base de dados 1. 37
FIGURA 17 – Diagrama de atividades de aplicação dos métodos de classificação para a Base de dados 2. 38
FIGURA 18 – Diagrama de atividades de aplicação dos métodos de agrupamento para a Base 2. 39
FIGURA 19 – Diagrama de atividades de aplicação dos métodos de agrupamento para a Base 1. 39
FIGURA 20 – Resultado do agrupamento na Base de Dados 1 com o algoritmo EM. 53
FIGURA 21 – Visualização dos agrupamentos considerando idade X sexo 53
FIGURA 22 – Visualização dos agrupamentos considerando tabagismo X alcoolismo . . . 54

LISTA DE TABELAS

TABELA 1 – Carrinhos de compras 16
TABELA 2 – Resultados algoritmo Apriori utilizando 18 variáveis 42
TABELA 3 – Resultados algoritmo Apriori utilizando 17 variáveis 43
TABELA 4 – Resultados algoritmo Apriori utilizando 15 variáveis 43
TABELA 5 – Resultados algoritmo Apriori utilizando 14 variáveis 44
TABELA 6 – Resultados da aplicação do algoritmo J48 utilizando variáveis anteriores ao diagnóstico 45
TABELA 7 – Resultados algoritmo J48 com variáveis selecionadas por *Info gain* e *Gain ratio* 45
TABELA 8 – Resultados algoritmo J48 utilizando variáveis selecionadas por PCA 46
TABELA 9 – Resultados do algoritmo *reptree* utilizando variáveis anteriores ao diagnóstico 47
TABELA 10 – Resultados algoritmo *reptree* utilizando variáveis selecionadas por *info gain* e *gain ratio* 48
TABELA 11 – Resultados do algoritmo *reptree* utilizando variáveis selecionadas por PCA 49
TABELA 12 – Resultados algoritmo *k-means* utilizando filtro *resample* e variáveis anteriores ao diagnóstico 50
TABELA 13 – Resultados algoritmo *k-means* utilizando filtro *resample* e variáveis selecionadas por *Info Gain* e *Gain Ratio* 50
TABELA 14 – Resultados do algoritmo *k-means* utilizando filtro *resample* e variáveis selecionadas por PCA 50
TABELA 15 – Resultados do algoritmo EM com a utilização do filtro *resample* 51
TABELA 16 – Resultados do algoritmo *k-means* utilizando variáveis anteriores ao diagnóstico 51
TABELA 17 – Resultados do algoritmo *k-means* utilizando variáveis selecionadas por *Info Gain* e *Gain Ratio* 51
TABELA 18 – Resultados do algoritmo *k-means* utilizando variáveis selecionadas por PCA 52
TABELA 19 – Resultados do algoritmo EM 52
TABELA 20 – Resultados do algoritmo *k-means* na Base de Dados 1 com variáveis selecionadas por PCA 52
TABELA 21 – Dicionário das variáveis da base de dados do SisRHC disponível para download no IRHC. 63

LISTA DE SIGLAS

BD	Base de Dados
CID-O	Classificação Internacional de Doenças para Oncologia
CP	Componente Principal
csv	*Comma Separated Values*
dbf	*Data Base File*
DM	Mineração de Dados
DNA	Ácido Desoxirribonucleico
DATASUS	Departamento de Informática do Sistema Único de Saúde
EM	*Expectation Maximization*
GCO	*Global Cancer Observatory*
INCA	Instituto Nacional do Câncer José Alencar Gomes da Silva
IRHC	Integrador de Registros Hospitalares do Câncer
KDD	Descoberta de Conhecimento em Base de Dados
KNN	K-vizinhos mais próximos
PCA	Análise de Componentes Principais
RHC	Registros Hospitalares de Câncer
sql	*Structured Query Language*
UCI	Universidade da Califórnia em Irvine
WEKA	*Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO 8
1.1 OBJETIVOS GERAL E ESPECÍFICOS 9
1.2 JUSTIFICATIVA 9
1.3 ORGANIZAÇÃO DO DOCUMENTO 10
2 LEVANTAMENTO BIBLIOGRÁFICO 11
2.1 KDD E MINERAÇÃO DE DADOS 11
2.1.1 Regras de Associação 15
2.1.2 Métodos de Classificação 18
2.1.3 Métodos de Agrupamentos de Dados 19
2.1.3.1 Métodos Hierárquicos 20
2.1.3.2 Métodos Particionais 21
2.1.3.3 Métodos baseados em modelos 23
2.1.4 Métodos de Seleção de Variáveis 24
2.2 MINERAÇÃO DE DADOS REFERENTES AO CÂNCER 26
2.2.1 O Câncer no Brasil 26
2.2.2 Trabalhos Correlatos 28
3 MATERIAIS E MÉTODOS 30
3.1 BASE DE DADOS 30
3.2 SOFTWARES 32
3.2.1 TabWin 32
3.2.2 PostgreSQL 33
3.2.3 WEKA 33
3.3 FLUXO DE ATIVIDADES 34
3.3.1 Obtenção da base de dados e pré-processamento 35
3.3.2 Seleção de variáveis 36
3.3.3 Mineração de dados 37
4 RESULTADOS E DISCUSSÃO 41
4.1 RESULTADOS 41
4.1.1 Regras de associação 41
4.1.2 Classificação 44
4.1.3 Agrupamento 49
4.2 DISCUSSÃO 54
5 CONCLUSÕES 57
REFERÊNCIAS 59
Anexo A – DICIONÁRIO DE DADOS 63

1 INTRODUÇÃO

Anualmente são diagnosticados milhares de novos casos de câncer, uma doença genética que corresponde a um conjunto de mais de 100 doenças que têm em comum o crescimento celular anormal, podendo se desenvolver em qualquer tecido ou órgão (INCA, 2019c). As maiores taxas da doença acontecem em países desenvolvidos onde os tipos de cânceres estão mais relacionados ao desenvolvimento e a urbanização. Já em países menos desenvolvidos são diagnosticados mais casos de cânceres relacionados a infecções (INCA, 2017).

O surgimento do câncer acontece em três estágios: primeiramente as células sofrem alterações, porém o tumor ainda não é identificável; no segundo estágio é quando a célula alterada é transformada para célula maligna; e o último é quando as células se multiplicam de forma anormal e podem surgir os sintomas da doença (INCA, 2019a).

De acordo com o Ministério da Saúde (2017), há fatores internos e externos que colaboram no desenvolvimento do câncer. Os fatores internos referem-se a questão genética e ao organismo do pacientes, e os fatores externos estão relacionados ao estilo de vida, hábitos do paciente e a aspectos ambientais. Estes fatores são considerados de risco, pois podem aumentar as chances de se desenvolver a doença.

Com o avanço da tecnologia, cada vez estão sendo coletados mais dados das mais diversas áreas, inclusive na saúde, porém esses dados são desperdiçados quando não explorados (BRAMER, 2016). A mineração de dados visa detectar padrões, relacionamentos entre variáveis e descobrir conhecimento em uma grande quantidade de dados por meio de técnicas como regras de associação, classificação, agrupamento, e regressão. A mineração de dados é uma das etapas da Descoberta de Conhecimento em Base de Dados, assim como o pré-processamento dos dados, que visa o tratamento e preparação dos dados para a mineração e o pós processamento, que envolve a interpretação e a análise das informações obtidas (GOLDSCHMIDT; PASSOS, 2015).

As bases de dados de saúde podem conter informações para identificação de causas e tratamento de doenças (BRAMER, 2016). A mineração de dados tem sido aplicada na área da saúde, de modo a gerar conhecimento para auxílio na tomada de decisão em diagnósticos,

tratamento e prevenção a doenças.

1.1 OBJETIVOS GERAL E ESPECÍFICOS

Esse trabalho tem como objetivo aplicar técnicas de mineração de dados para análise e obtenção de conhecimento em uma base de pacientes diagnosticados com câncer de esôfago no Brasil no período de 1998 a 2017. Esse objetivo principal pode ser dividido nos seguintes objetivos específicos:

- Aplicar algoritmo *apriori* para mineração de regras de associação;
- Aplicar métodos de agrupamento de dados para segmentação da base de dados segundo características relevantes;
- Aplicar métodos de classificação de dados para classificar as instâncias da base de dados de acordo com características relevantes;
- Avaliar a relevância dos resultados obtidos com a mineração de dados;
- Apresentar algoritmos e parâmetros de configuração que proporcionaram os melhores resultados para as regras de associação, classificação e para a segmentação do conjunto de dados.

1.2 JUSTIFICATIVA

O Globocan (GCO, *Global Cancer Observatory*) é uma plataforma de informações sobre incidência e mortalidade de câncer no mundo. Segundo este, a estimativa para 2018 era de 18.078.957 novos casos de câncer em todo o mundo e 9.555.027 de mortes causadas pela doença, sendo o câncer de esôfago o 7º tipo com maior número de novos casos e o 6º maior responsável por mortes (GCO, 2018).

Com a enorme quantidade de dados e informações coletadas na área da saúde, o estudo dessas bases é capaz de gerar conhecimento para prevenção de doenças, melhora no diagnóstico, tratamento e conscientização, consequentemente melhorando qualidade de vida da população e proporcionando diminuição de custos.

Considerando a importância dos estudos nessa área e sendo o câncer de esôfago uma doença agressiva, com crescente número de casos e levando em conta que o diagnóstico precoce é essencial para a eficácia do tratamento, com a aplicação de técnicas de mineração de dados sobre bases de dados correspondentes a esta doença, pode ser possível a descoberta de conhecimento e a apresentação desse conhecimento de forma relevante para a população. Embora existam trabalhos correlatos Preissler (2016), Minelli (2013) e Viana (2018), nenhum destes trabalhos executou mineração de dados especificamente para o câncer de esôfago num período de tempo tão longo quanto o proposto neste trabalho.

1.3 ORGANIZAÇÃO DO DOCUMENTO

Esse documento será organizado da seguinte forma. O Capítulo 2 apresentará as etapas do processo de descoberta de conhecimento em bases de dados e uma breve abordagem sobre o câncer. Em seguida, são apresentados os trabalhos correlatos recentes, com o intuito de situar este trabalho no estágio atual do conhecimento. A metodologia utilizada se encontra no Capítulo 3, nele são descritas todas as etapas para o desenvolvimento do projeto, primeiramente esta uma descrição da base de dados, em seguida em softwares utilizados e por fim um fluxograma de atividades para realização do trabalho. No Capítulo 4 são mostrados os resultados obtidos com as técnicas aplicadas e a discussão sobre os mesmos e, por fim, no Capítulo 5 encontra-se a conclusão do trabalho.

2 LEVANTAMENTO BIBLIOGRÁFICO

Nessa seção é descrito o estado da arte do tema escolhido. Primeiramente são abordadas as etapas do processo de descoberta de conhecimento em bases de dados e, em seguida, é apresentada uma breve abordagem sobre câncer e trabalhos correlatos.

2.1 KDD E MINERAÇÃO DE DADOS

O avanço da tecnologia possibilitou que diversas áreas de conhecimento coletassem diariamente dados de diferentes tipos, formando volumosos bancos de dados que, quando examinados, podem ajudar na preparação de campanhas de saúde e marketing mais eficazes e no aumento de lucros e vendas de uma organização a partir de informações de compras e perfis de clientes, por exemplo. Essa grande quantidade de dados gerados tem tornado inviável a análise e a interpretação manual dos mesmos. Portanto, tem-se utilizado técnicas e ferramentas computacionais que fazem parte de uma área chamada descoberta de conhecimento em bases de dados (KDD, do inglês *Knowledge Discovery in Databases*) para automatizar tais análises (FAYYAD et al., 1996; BRAMER, 2016).

A descoberta de conhecimento em bases de dados trata de extrair de fontes de dados informações até então desconhecidas, que sejam possivelmente úteis. Este processo pode acontecer de maneira iterativa e interativa (FAYYAD et al., 1996). O KDD compreende todo o processo desde a obtenção dos dados até o conhecimento gerado a partir dos mesmos, descritos pelas etapas de seleção, pré-processamento, transformação, mineração de dados (DM) e interpretação, como mostrado na Figura 1. Quando iterativo significa que podem acontecer repetições em partes do processo e, caso interativo, carece de intervenção humana (FAYYAD et al., 1996; GOLDSCHMIDT; PASSOS, 2015).

A facilidade de armazenamento e o crescimento da coleta de dados combinados com a competitividade das empresas, são fatores que contribuíram para a popularização da descoberta

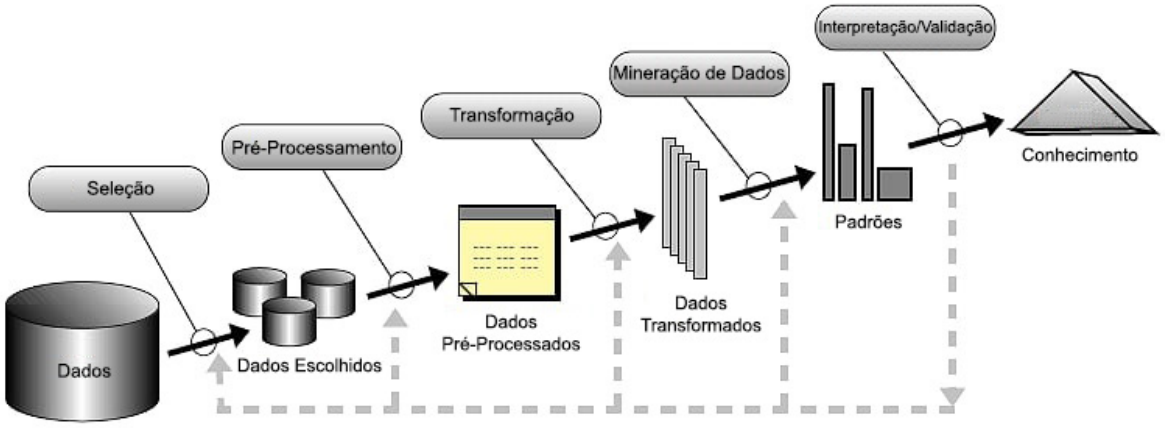


Figura 1 – Etapas do processo de KDD.

Fonte: Adaptado de Fayyad et al. (1996)

de conhecimento e mineração de dados (LAROSE; LAROSE, 2014). Porém, ainda que estes dados tenham sido colhidos de maneira organizada e elaborada, para que sejam realmente úteis, é preciso convertê-los em informação e posteriormente em conhecimento (VERCELLIS, 2009). Portanto, dados são os elementos na forma como estão armazenados na fonte de dados; quando os dados são tratados eles se tornam informações; e conhecimento é interpretar e tomar decisões tendo como base as informações (ROSINI; PALMISANO, 2011; GOLDSCHMIDT; PASSOS, 2015).

Ter um conhecimento prévio sobre a área ou sistema de onde os dados serão analisados, conhecer quais informações são mais relevantes para o problema de estudo (WU et al., 2014), assim como saber e definir o propósito para realização da descoberta de conhecimento é uma parte fundamental para aplicação do processo (FAYYAD et al., 1996).

Na etapa de seleção é feita a integração dos dados, visto que estes podem vir de diversas fontes de dados, como planilhas, bancos de dados e *data warehouses* (HAN et al., 2012). Também acontece a apuração dos possíveis dados que serão utilizados, uma vez que dependendo do domínio do problema apenas uma parte dos dados armazenados serão necessários e relevantes para a mineração (BRAMER, 2016; CARVALHO, 2002).

O pré-processamento envolve tarefas como a limpeza e tratamento dos dados. Levando em conta que estes podem vir de diversas fontes (AGGARWAL, 2015) e também conter dados de um longo período de tempo (LAROSE; LAROSE, 2014), é comum estes não estarem seguindo um mesmo padrão ou apresentarem falhas. Ruídos, inconsistências e incompletude precisam ser tratados para que os resultados dos algoritmos não sejam afetados (HAN

et al., 2012; CAMILO; SILVA, 2009). Os ruídos são dados divergentes ou discrepantes, também conhecidos como *outliers* (GOLDSCHMIDT; PASSOS, 2015); dados inconsistentes apresentam valores incompatíveis nos atributos ou anormais; os incompletos não contêm informações relevantes (HAN et al., 2012).

Uma das formas de tratar valores incorretos é a eliminação da instância, porém esta solução pode afetar negativamente os resultados (BRAMER, 2016), além de correr o risco de excluir uma quantidade significativa de dados. Então, outra saída é a substituição por um valor que seja frequente para aquele tipo de dado, estimar o valor levando em conta outras instâncias (LAROSE; LAROSE, 2014; AGGARWAL, 2015).

Sendo a etapa de pré-processamento de grande importância para bons resultados, esta juntamente com as anteriores costumam utilizar de 10 a 60% do tempo de todo o processo de descoberta de conhecimento (LAROSE; LAROSE, 2014).

Apesar das etapas anteriores já realizarem alterações na base, os dados ainda não se encontram na forma adequada para a mineração; logo, é necessário realizar alterações para melhor representação dos dados (HAN et al., 2012; AGGARWAL, 2015). Cabe a etapa de transformação modificar os valores, que podem assumir forma numérica ou categórica, para a forma aceita pelos algoritmos (CAMILO; SILVA, 2009) e para os formatos aceitos pelas ferramentas de mineração a serem utilizadas. Valores numéricos ou quantitativos podem ser discretos, que assumem valores contáveis (idade, valores binários) ou contínuos, podem assumir valores infinitos (medidas) (FACELLI et al., 2011). Valores categóricos são rótulos qualitativos, que podem ser nominais, em que não há uma ordem lógica (feminino, masculino) e os ordinais, que seguem algum tipo de ordem (pequeno, médio, grande) (LOUZADA-NETO; DINIZ, 2002).

Ao transformar dados, é possível diminuir a quantidade de variáveis e valores que estas podem assumir quando se altera a dimensionalidade e representação dos dados (FAYYAD et al., 1996; AGGARWAL, 2015). Como exemplos de técnicas de transformação, tem-se:

- Agregação: são aplicadas operações para reduzir atributos de dados (HAN et al., 2012), por exemplo, nem sempre é interessante trabalhar com os 365 dias do ano, então pode-se utilizar os meses para representá-los (TAN et al., 2006);
- Discretização: valores contínuos são transformados em categóricos, dividindo esses valores em intervalos e então deve-se rotular cada intervalo (VERCELLIS, 2009). Por exemplo, o atributo idade pode ser segmentado em intervalos de 0 a 12, 13 a 18 e 19 a 59, e esses rotulados como “criança”, “jovem” e “adulto”;
- Normalização: há casos em que a unidade de medida de uma variável quantitativa pode levá-la a ter um intervalo de valores muito grande, fazendo com que esse atributo se

sobressaia dos demais (TAN et al., 2006); neste caso, os dados são normalizados para que todos os atributos tenham um peso semelhante, ou seja, são organizados na forma de intervalos menores como, por exemplo, de 0,1 a 0,5. É interessante o uso dessa técnica quando se utiliza funções de distância para medir similaridade. (HAN et al., 2012).

Comumente, a mineração de dados é tratada como sinônimo de KDD; porém se trata apenas de uma das etapas de todo o processo de KDD (HAN et al., 2012; CIOS et al., 2007). Esta fase é responsável pela descoberta e extração de padrões em um grande conjunto de dados pré-processados (LAROSE; LAROSE, 2014), que visam descrever padrões até então desconhecidos e representá-los de uma forma clara, bem como a previsão de valores futuros que são usados para tomada de decisão (FAYYAD et al., 1996).

Os métodos clássicos de estatística, baseados em hipótese e teste, quando trabalhados com grandes volumes de dados se tornam penosos (TAN et al., 2006). Porém, a mineração de dados emprega técnicas de estatística, como a regressão e classificador Bayesiano (VERCELLIS, 2009), além de reconhecimento de padrões e aprendizado de máquina para atingir seu objetivo (FAYYAD et al., 1996).

O aprendizado indutivo procura gerar um modelo ou tirar conclusões, a partir dos dados ou fatos fornecidos, e pode ser dividido em aprendizado supervisionado e aprendizado não supervisionado. Quando supervisionado, significa que já se sabe o que deverá estar na saída, pois a amostra fornecida para treinamento é conhecida. Os métodos de classificação e regressão pertencem a este tipo de aprendizado (KANTARDZIC, 2003). E para o aprendizado não supervisionado, não são determinados atributos alvos, apenas se quer que seja feita uma varredura para encontrar associações e agrupamentos que demonstrem o comportamento dos dados (CATTRAL et al., 2001). A Figura 2 demostra a hierarquia do aprendizado indutivo.

Figura 2 – Hierarquia do aprendizado.

Fonte: Facelli et al. (2011)

Portanto, a mineração de dados aplica técnicas como regras de associação, agrupamento, classificação e regressão. Assim, a partir de um conjunto de dados são descobertos padrões que podem ser aplicados a volumes de dados muito maiores (VERCELLIS, 2009). Alguns exemplos onde a mineração de dados é usada: marketing dirigido apenas para clientes em potencial, detecção de fraudes em cartões de crédito, análise de imagens, previsão climática, determinar o hábito de consumo de clientes e a relação entre sintomas e perfil de pacientes (BRAMER, 2016; CARVALHO, 2002).

Por fim, na etapa de interpretação é efetuada a análise dos resultados obtidos, verificação da utilidade, validação e como poderão ajudar na tomada de decisão. É válido ressaltar que todas as etapas anteriores associadas a um conhecimento prévio sobre o domínio do problema são de grande importância para a interpretação dos resultados obtidos na mineração de dados e para a extração de conhecimento útil (FAYYAD et al., 1996).

2.1.1 Regras de Associação

As regras de associação servem para identificar padrões frequentes na relação entre atributos ou conjuntos de atributos, criando regras para descrever esses padrões. Essas regras costumam ser escritas no formato "se-então", ou seja, se antecedente então consequente. Isto significa que a ocorrência do antecedente leva a ocorrência do consequente de acordo com medidas de interesse (LAROSE; LAROSE, 2014).

O grau de interesse e qualidade das regras é medido pelo suporte e pela confiança. Sendo o suporte o indicador de quantas vezes os itens do conjunto aparecem juntos, ou seja, Equação 1 (GOLDSCHMIDT; PASSOS, 2015):

$$suporte(X \rightarrow Y) = \frac{nr(X,Y)}{nr} \tag{1}$$

O $suporte$ é a porcentagem de ocorrência do conjunto, X representa o conjunto de itens antecedente, Y o conjunto consequente, logo $nr(X,Y)$ se refere ao número de registros que contêm X e Y e nr o números total de registros da base de dados. Para uma regra ser considerada frequente é necessário atender ao valor de suporte mínimo estabelecido (GOLDSCHMIDT; PASSOS, 2015).

Já a $confiança$ representa a probabilidade da regra levando em conta a ocorrência de X. Sendo $nr(X,Y)$ se refere ao número de registros que contêm X e Y e $nr(X)$ o número total

de registros que contem X, Equação 2. Portanto, para a regra ser válida é preciso que atenda a uma medida de confiança mínima (GOLDSCHMIDT; PASSOS, 2015).

$$\text{confiança}(X \rightarrow Y) = \frac{nr(X,Y)}{nr(X)} \tag{2}$$

A mineração de regras de associação é frequentemente utilizada para tomada de decisão, aplicada em bases de dados de carrinhos de compras para identificar hábitos de consumo, orientando a organização de produtos em um supermercado para estimular o cliente à compra de produtos ou à realização de promoções de determinados produtos (BRAMER, 2016; CIOS et al., 2007; CARVALHO, 2002).

Por exemplo, a Tabela 1 mostra compras de quatro clientes e itens comprados:

Tabela 1 – Carrinhos de compras

	Itens					
Transação	leite	pão	biscoito	ovos	queijo	geleia
1	•	•	•			
2	•	•		•	•	
3	•	•				•
4				•	•	

Fonte: Autoria própria

Considerando que um especialista da área definiu que deseja suporte $\geq 25\%$ e confiança $\geq 30\%$, alguns exemplos de regras válidas que podem ser extraídas do exemplo são:

• leite $\rightarrow$ *pão*

$$suporte(leite \rightarrow \text{pão}) = \frac{nr(leite, \text{pão})}{nr} = \frac{3}{4} = 75\% \tag{3}$$

A Equação 3 mostra que leite e pão foram vendidos juntos em 75% das vendas.

$$\text{confiança}(leite \rightarrow \text{pão}) = \frac{nr(leite, \text{pão})}{nr(leite)} = \frac{3}{3} = 100\% \tag{4}$$

Pela Equação 4 vê-se que, 100% dos clientes que compraram leite, também compraram pão.

• leite, pão $\rightarrow$ *queijo*

$$suporte(leite, \text{pão} \rightarrow queijo) = \frac{nr(leite, \text{pão}, queijo)}{nr} = \frac{1}{4} = 25\% \tag{5}$$

A Equação 5 mostra que leite, pão e queijo foram vendidos juntos em 25% das vendas.

$$\text{confiança}(leite, \text{pão} \rightarrow queijo) = \frac{nr(leite, \text{pão}, queijo)}{nr(leite, \text{pão})} = \frac{1}{3} = 33\% \quad (6)$$

Pela Equação 6 vê-se que, 33% dos clientes que compraram leite e pão, também compraram queijo.

O *apriori* é um algoritmo para mineração de regras de associação proposto por Agrawal et al. (1993), muito conhecido e utilizado na mineração de dados. Segundo Tan et al. (2006) o princípio do *apriori* diz que se um conjunto de itens é frequente, então os subconjuntos desse conjunto serão frequentes também .

Utiliza dados categóricos e pode ser dividido em duas etapas, na primeira o objetivo é encontrar os conjuntos de itens que satisfaçam o valor de suporte mínimo estabelecido e na segunda é calcular a confiança dos conjuntos e encontrar as regras válidas, que satisfaçam o valor de confiança mínimo (GOLDSCHMIDT; PASSOS, 2015).

Ou seja, primeiramente são informados os valores mínimos de suporte e confiança desejados. Na primeira iteração são encontrados os conjuntos, compostos por um item, que satisfaçam o valor de suporte mínimo. A partir desses conjuntos é feita a combinação deles para formar novos conjuntos com dois itens, onde são selecionados apenas os que satisfaçam o valor mínimo de suporte. Depois, são usados os conjuntos com dois itens para formar conjuntos com três itens, e então são calculados seus suportes e assim sucessivamente (AGRAWAL et al., 1993).

Após serem encontrados todos os conjuntos que apresentem suporte válidos, são então geradas as regras de associação e calculadas suas confianças, onde as regras válidas serão somente aquelas que atendam ao valor de confiança miníma (AGRAWAL et al., 1993).

Considerando o exemplo da Tabela 1, e adotando os valores de mínimos para suporte e confiança como 30% e 50%, tem-se que C representa todos com conjuntos de itens possíveis, com seus respectivos valores de suporte e L representa os conjuntos de itens que satisfazem o valor de suporte mínimo na Figura 3. Onde os itens de $C3$ não obtiveram valor de suporte válido, portanto $L3$ não está apresentado.

Assim, serão geradas regras de acordo com $L2$ e calculadas suas confianças:

•leite $\rightarrow pão = 100\%$

•pão $\rightarrow leite = 100\%$

• ovos $\rightarrow queijo = 100\%$

•queijo $\rightarrow ovos = 100\%$

Percebe-se que todas as regras de associação geradas a partir de $L2$, satisfazem o valor mínimo de confiança, portanto são regras válidas.

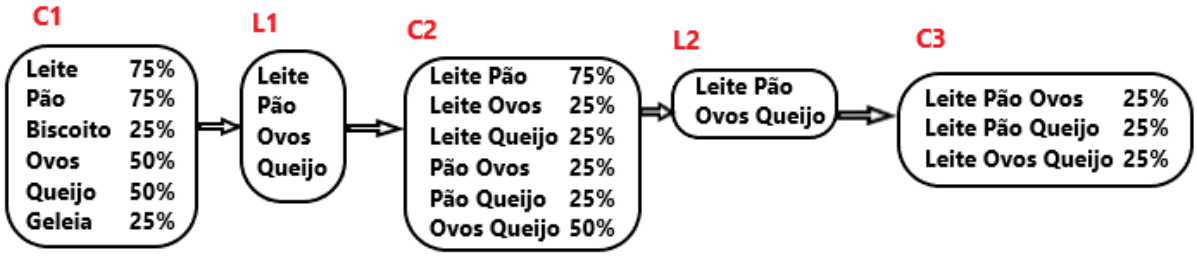


Figura 3 – Conjuntos obtidos na primeira etapa do *apriori*.

Fonte: Autoria própria

2.1.2 Métodos de Classificação

Os métodos de classificação têm como objetivo relacionar cada instância de uma base de dados a uma classe (GOLDSCHMIDT; PASSOS, 2015) por meio do aprendizado de uma função, também chamada de modelo, que realize esta tarefa (FAYYAD et al., 1996).

Sendo um método de aprendizado supervisionado, este relacionamento é efetuado utilizando um conjunto de treinamento, onde entre os atributos, há um atributo especial que possuirá os rótulos (categóricos) das classes, para então construir um modelo capaz de realizar a predição da classe para novas instâncias (AGGARWAL, 2015; TAN et al., 2006). Portanto a classificação consiste em duas etapas: o aprendizado do modelo e a classificação dos dados. E a representação pode ser feita, por exemplo, por árvores de decisão e regras de classificação (HAN et al., 2012).

As árvores de decisão utilizam a técnica de dividir para conquistar (WITTEN et al., 2011), nesta representação cada nó interno representa um teste para uma variável que define como será a divisão dos dados nos nós filhos, para essa separação é escolhido o teste que melhor divide os dados. As arestas representam os resultados dos testes e os nós folhas representam as classes (HAN et al., 2012; FACELLI et al., 2011), como pode ser observado na Figura 4 que demonstra a classe ”comprar Computador” que define se o cliente comprará ou não um computador, a partir do conhecimento sobre os valores dos atributos idade, se ele é ou não estudante e a classificação de crédito.

O C4.5 é um exemplo tradicional de algoritmo de classificação que utiliza árvore de decisão. Durante a construção da árvore as principais atividades são a escolha e quais serão os nós internos e a criação dos nós filhos (GOLDSCHMIDT; PASSOS, 2015).

O J48 é a implementação da WEKA do algoritmo C4.5. Assim como a maioria dos algoritmos de árvore de decisão, a árvore é construída da raiz para as folhas de forma recursiva,

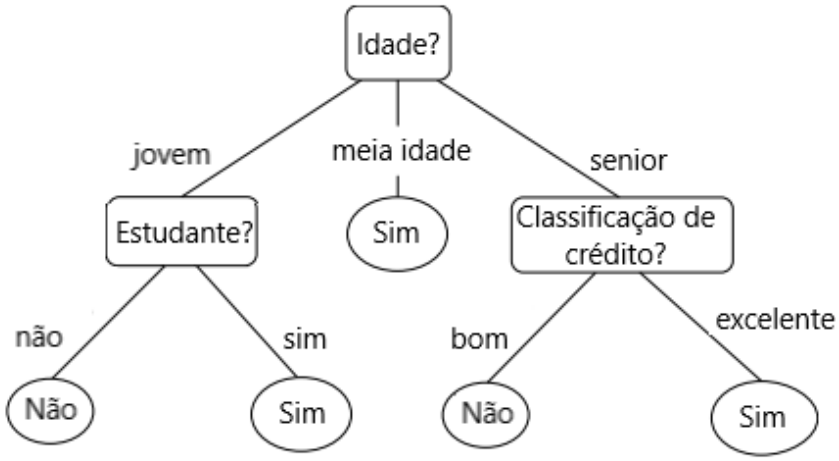


Figura 4 – Exemplo de árvore de decisão.
Fonte: Adaptado de Han et al. (2012)

onde o conjunto de treinamento vai sendo dividido em conjuntos menores conforme a árvore é formada (HAN et al., 2012). Para escolha dos nós internos o C4.5 utiliza o conceito de ganho de informação (LAROSE; LAROSE, 2014). Para mais detalhes veja Han et al. (2012) e Larose e Larose (2014).

O *reptree* é um algoritmo baseado no C4.5 (SNOUSY et al., 2011). Constrói a árvore de decisão usando ganho de informação e realiza a poda utilizando poda de erro reduzido (WITTEN et al., 2011). A amostra é dividida em conjunto de treinamento e validação, a árvore é construída utilizando o conjunto de treinamento e o conjunto de validação é usado para validar a classificação (PAULO et al., 2012). Para mais detalhes veja Witten et al. (2011).

2.1.3 Métodos de Agrupamentos de Dados

Os métodos de agrupamentos, também conhecidos como métodos de clusterização, buscam segmentar os dados em grupos/*clusters* de acordo com as semelhanças das características dos dados. Como é um método de aprendizado não supervisionado, não se sabe quais classes serão geradas e como serão compostas (TAN et al., 2006; CARVALHO, 2002). Os dados podem ser apresentados em forma de pontos em um espaço multidimensional, onde para medir a dissimilaridade entre elementos algumas abordagens utilizam funções de distâncias. Quando a dimensão, que equivale a quantidade de variáveis, é muito grande,

os dados podem ficar muito espalhados, pois a distância entre as instâncias aumenta e isso pode dificultar a definição dos grupos e prejudicar os resultados dos métodos de agrupamento (GOLDSCHMIDT; PASSOS, 2015; FACELLI et al., 2011).

As funções de distância ajudam a delimitar a abrangência dos grupos. Para dados quantitativos as medidas mais utilizadas são a distância Euclidiana (Equação 7) e a distância Manhattan (Equação 8) (FACELLI et al., 2011).

$$d(\hat{X},\hat{Y}) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{7}$$

$$d(\hat{X},\hat{Y}) = \sum_{i=1}^{k}|x_i - y_i| \tag{8}$$

Onde, tendo dois pontos, $X = (x_1, x_2, ..., x_k)$ e $Y = (y_1, y_2, ..., y_k)$, com k dimensões e sendo que cada x representa o valor do elemento para uma das variáveis consideradas e y sendo semelhante a x, portanto $d(\hat{X},\hat{Y})$ é a distância entre esses dois pontos (FACELLI et al., 2011).

Portanto, dentro de cada *cluster* se busca que os elementos sejam o mais similares possível e entre os *clusters* os elementos sejam mais diferentes. Os algoritmos de agrupamento são divididos principalmente como hierárquicos e particionais, mas há outros tipos, como agrupamento baseado em modelo (HAN et al., 2012).

2.1.3.1 Métodos Hierárquicos

Os algoritmos hierárquicos organizam os dados de forma hierárquica, isto é, um grupo irá possuir subgrupos aninhados, se baseando, principalmente na distância dos grupos ou na distância (Euclidiana ou Manhattan) dos centróides, pontos que representam os centros dos *clusters*, para realizar esta tarefa (FACELLI et al., 2011; TAN et al., 2006).

Esses algoritmos são distribuídos em abordagens aglomerativa e divisiva. A Figura 5 demonstra a diferença entre as duas abordagens (FACELLI et al., 2011).

A abordagem aglomerativa, também chamada *bottom-up*, inicia com cada dado sendo um *cluster* e vai agrupando *clusters* de acordo com sua similaridade até que todos os elementos estejam em um único *cluster*, ou atinja uma condição de parada definida (GOLDSCHMIDT; PASSOS, 2015).

Já a abordagem divisiva, ou *top-down*, começa com todos os dados em um único *cluster* e então vai realizando a divisão para novos *clusters* até cada dado formar um *cluster*

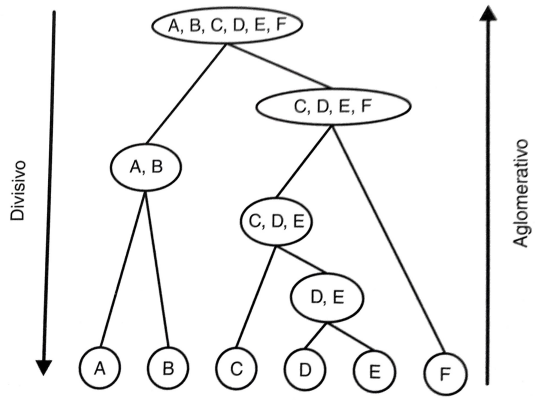


Figura 5 – Algoritmos hierárquicos aglomerativos e divisivos.
Fonte: Facelli et al. (2011)

ou atingir outra condição de parada (GOLDSCHMIDT; PASSOS, 2015).

A Figura 6 demonstra um exemplo de agrupamento hierárquico no formato de árvore, onde é possível formar o grupo localização primária de um tumor, que contém os grupos localização detalhada e estadiamento.

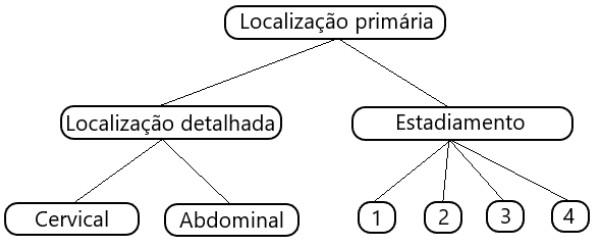


Figura 6 – Exemplo de agrupamento hierárquico.
Fonte: Autoria própria

2.1.3.2 Métodos Particionais

O método de agrupamento por particionamento é a forma mais simples de clusterização. O usuário informa a quantidade de grupos e o algoritmo separa os dados de acordo com a similaridade por meio de função de distância. As principais técnicas realizam a separação exclusiva, ou seja, um elemento só pode estar em um grupo e são eficazes para bases

de dados de pequeno a médio porte (HAN et al., 2012).

Após a separação inicial, os dados vão sendo realocados até se obter boa divisão. Para escolha do representante de cada *cluster*, duas entre as possíveis formas de determinar o representante é calcular a média dos elementos do grupo ou selecionar o elemento mais próximo do centro (GOLDSCHMIDT; PASSOS, 2015).

A Figura 7 demonstra um exemplo simples de agrupamento particional, onde tendo dados de peso e altura, pode-se formar grupos que representam crianças e adultos.

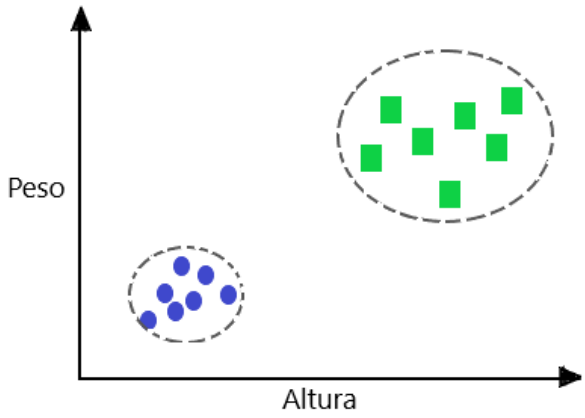


Figura 7 – Exemplo de agrupamento particional.

Fonte: Autoria própria

O *k-means* é um famoso método de agrupamento por particionamento, simples e rápido de ser executado. Utiliza funções de distância para medir a similaridade, tenta formar agrupamentos de forma a minimizar a função de erro quadrado (aprofundado em (FACELLI et al., 2011)) e necessita que seja informado o número de *clusters* a serem formados (GOLDSCHMIDT; PASSOS, 2015).

Considerando a Figura 8, inicialmente é informado o número de k *clusters* que se deseja formar, três neste caso, então são escolhidos k pontos aleatórios para serem os centróides (representados pelo símbolo $+$ na figura) de cada *cluster*. Com os centróides determinados, é realizado o cálculo das distâncias de cada instância com os centróides, de modo a definir a qual *cluster* (o mais próximo) estas serão associadas. Após serem feitos os agrupamentos iniciais, os centróides são novamente calculados de acordo com a média das instâncias de cada *cluster*. Então as instâncias são reagrupadas levando em consideração a sua distância com o novo centróide. Este processo ocorre até que não sejam mais modificados os centros e os *clusters*, resultando num agrupamento final (FACELLI et al., 2011; HAN et al., 2012; GOLDSCHMIDT; PASSOS, 2015).

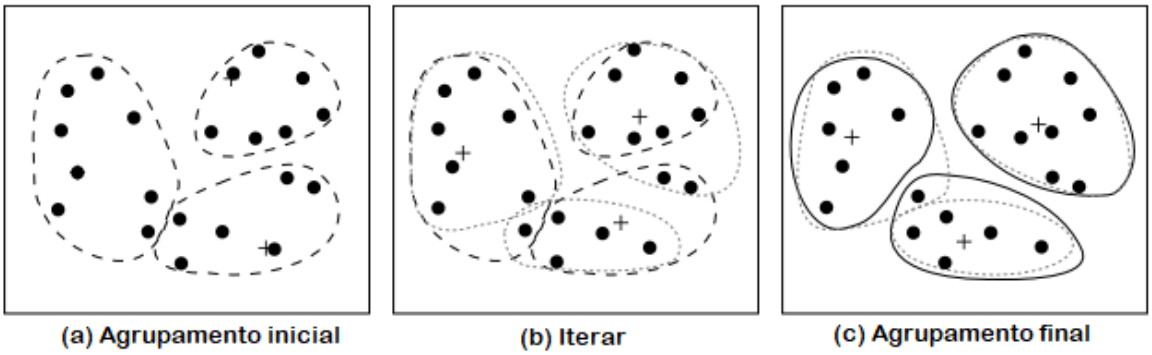


Figura 8 – Exemplo de agrupamento com o algoritmo *k-means*.

Fonte: Adaptado de Han et al. (2012)

2.1.3.3 Métodos baseados em modelos

Os métodos de agrupamento baseados em modelos utilizam modelos de mistura finita para realizar os agrupamentos. Os modelos de mistura finita são formados por k distribuições de probabilidades que correspondem a k *clusters*. Por exemplo, tendo um conjunto de dados descritos por uma variável x, e sabendo que tem-se $k = 2$, para cada *cluster* C, o modelo deste *cluster* será composto pela média e pelo desvio padrão dos valores da variável para este, como mostra a Figura 9, e também pela probabilidade de C em relação à variável (MARKOV; LAROSE, 2007).

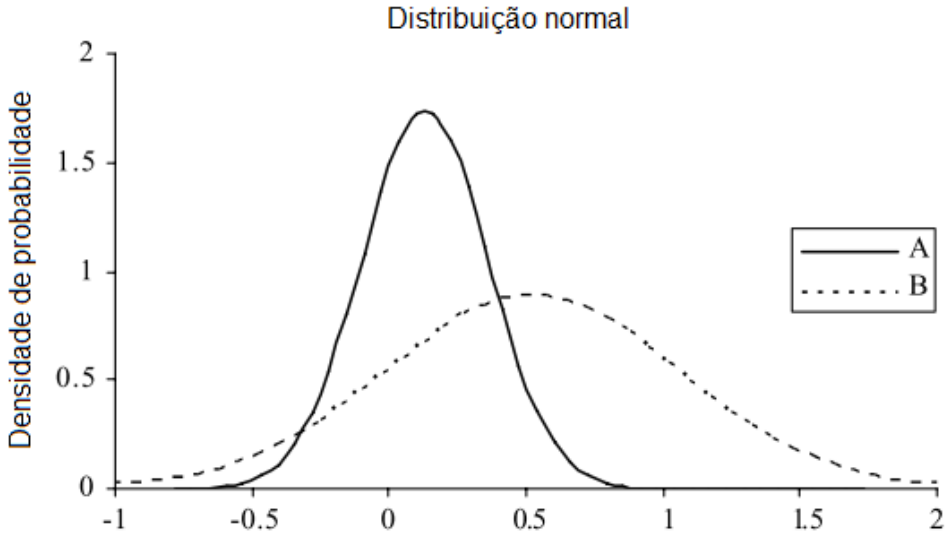


Figura 9 – Exemplo de agrupamento baseado em modelos.

Fonte: Markov e Larose (2007)

Os modelos de misturas finitas fornecem as características que descrevem quais valores de instâncias são esperados para que estas pertençam a um *cluster*, as instâncias têm probabilidade de constituírem um *cluster*, ao invés de serem inclusas ao mesmo. Sendo assim, são criadas hipóteses para relacionar as instâncias aos *clusters*. (WITTEN et al., 2011).

O EM (*Expectation Maximization*) é um algoritmo de agrupamento baseado em modelos. Utiliza probabilidade para definição dos *clusters* e assim como o *k-means* este também necessita informar previamente o número de *clusters* (MARKOV; LAROSE, 2007). A escolha desse algoritmo se deve ao fato de ter sido utilizado no trabalho correlato de Viana (2018), onde proporcionou bons resultados, por isso há potencial de apresentar bons resultados neste trabalho também.

Seu funcionamento é iterativo e envolve duas etapas, a primeira é o cálculo das probabilidades, que correspondem as expectativas e a segunda é o cálculo dos parâmetros de distribuição, que busca maximizar as probabilidades (expectativas) de distribuição. Como não se conhece os parâmetros, inicialmente são definidos de forma aleatória os parâmetros de distribuição esperados para cada *cluster*, ou seja, a média, desvio padrão e probabilidade. Esses parâmetros são usados para calcular a probabilidade de cada instância pertencer a cada *cluster*. Por meio dessas probabilidades é realizada uma nova estimativa para os parâmetros e assim novamente são calculadas probabilidades das instâncias para os *clusters* (WITTEN et al., 2011). Essas iterações ocorrem até que os parâmetros não estejam mais alterados ou que essa alteração seja mínima (TAN et al., 2006).

2.1.4 Métodos de Seleção de Variáveis

Para aplicação de algoritmos de classificação e agrupamento de dados, é importante realizar previamente uma seleção de variáveis para o processo. Isto porque, grandes quantidades de variáveis e valores que estas são capazes de assumir, podem gerar ruídos e prejudicar resultados e desempenho dos algoritmos, diante disso deve-se utilizar apenas variáveis relevantes (FACELLI et al., 2011).

As técnicas de seleção de variáveis buscam eliminar campos desnecessários que tendem a atrapalhar os algoritmos de mineração de dados. Uma das técnicas mais aplicadas é a análise de componentes principais (VERCELLIS, 2009).

A Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*)

é utilizada como um mecanismo para extração e para análise de dados. Tendo um conjunto ou vetor de dados X com n variáveis $X_1, X_2 ... X_n$ a ideia é criar um novo conjunto de n variáveis $Y_1, Y_2 ... Y_n$ a partir de uma combinação linear (Equação 9) (JOLLIFFE, 2002):

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{in}X_n \tag{9}$$

para $i = 1...n$, onde a_{ij}, $j = 1...n$ representa o coeficiente, ou seja, o nível de influência de cada variável para Y_i. Tendo o vetor de dados, são efetuadas operações e então é calculado a matriz de covariância e seus autovalores e autovetores, onde os autovetores são os valores dos coeficientes a_{ij}, logo os valores a_{ij} estão associados ao vetor X. As novas variáveis desse novo conjunto Y são chamadas de componentes principais (CPs). Os CPs são organizados de acordo com o componente que mais influencia na variância do conjunto de dados original para o componente que menos influencia (JOLLIFFE, 2002) e carregam quase tanta informação quanto as variáveis originais (LAROSE, 2006).

Ao organizar os componentes desta forma, faz-se a soma da porcentagem de contribuição de cada componente para variância do conjunto de dados original, até que uma taxa desejada seja atingida, mantendo a representação confiável dos dados. Então o restante dos componentes podem ser descartados (JOLLIFFE, 2002). Desta forma reduz-se a dimensão dos dados mantendo apenas atributos úteis e que preservam as características que definem sua variância (VERCELLIS, 2009; GOLDSCHMIDT; PASSOS, 2015)

Para execução da PCA, os dados podem ser normalizados para que os atributos estejam em um mesmo intervalo; então é feita a aplicação da técnica de PCA que dá origem aos CPs (HAN et al., 2012).

Entre outros métodos de seleção de variáveis tem-se o *info gain* e o *gain ratio*. O *info gain* avalia as variáveis medindo o ganho de informações em relação à classe (WITTEN et al., 2011), de modo a reduzir a entropia, quanto maior for o valor do ganho de informação maior é a contribuição da variável. Já o *gain ratio* é um aprimoramento do *info gain* que tenta normalizar o valor de contribuição da variável (SHARMA; DEY, 2012) e avalia o valor de uma variável considerando a taxa de ganho em relação a classe (WITTEN et al., 2011). Para mais detalhes consulte Han et al. (2012) e Sharma e Dey (2012).

2.2 MINERAÇÃO DE DADOS REFERENTES AO CÂNCER

A Informática na Saúde, que engloba a tecnologia da informação e a área da saúde, compreende atividades como captura, armazenamento e processamento de informações (HERLAND et al., 2014; FICHMAN et al., 2011).

Os dados armazenados pelos órgãos de saúde têm aumentado e, com isso, cresceu o interesse de se utilizar a mineração de dados para auxiliar na tomada de decisão para melhora no setor da saúde, no diagnóstico e tratamento de pacientes de diversas enfermidades, assim como redução de gastos tanto para governos quanto para a população (HERLAND et al., 2014).

Diversos trabalhos têm aplicado técnicas de mineração de regras de associação, agrupamento e classificação na área da saúde, como, Bharati et al. (2019) que utilizou mineração de dados para prever e analisar dados de câncer de mama. Por meio da ferramenta WEKA, foram aplicados e comparados cinco algoritmos de classificação em dados obtidos do repositório de aprendizado de máquina da Universidade da Califórnia em Irvine (UCI), onde o algoritmo K-vizinhos mais próximo (KNN, do inglês *K-Nearest Neighbor*) obteve melhores resultados. E Ksiazek et al. (2019) que se basearam em aprendizado de máquina e mineração de dados para projetar uma abordagem de diagnóstico do câncer de fígado, utilizando um conjunto de dados de um centro hospitalar de Portugal.

2.2.1 O Câncer no Brasil

No Brasil a estimativa de casos de câncer para 2018 era de aproximadamente 580 mil novos casos, sendo aproximadamente 300 mil em homens e 280 mil em mulheres. Deste total, quase 10.800 casos previstos eram de câncer de esôfago (INCA, 2018b). Em 2018 os cânceres de maior mortalidade foram pulmão e próstata para os homens, e mama e pulmão para as mulheres, sendo o câncer de esôfago o 5º maior responsável por mortes em homens e o 14º em mulheres, totalizando aproximadamente 9.800 óbitos (GCO, 2018).

É normal a multiplicação da maioria das células, porém quando ocorrem alterações no Ácido Desoxirribonucleico (DNA) da célula, essa multiplicação acontece de forma desenfreada e surge a neoplasia (Figura 10). O câncer é a neoplasia maligna, e uma característica é a possibilidade de ocorrer metástase, ou seja, as células cancerígenas invadirem outros tecidos

ou órgãos. Os cânceres que têm seu início em tecidos como pele ou mucosa são chamados carcinoma, e os cânceres que começam, por exemplo, em ossos, músculos ou cartilagem são denominados sarcoma (INCA, 2019c).

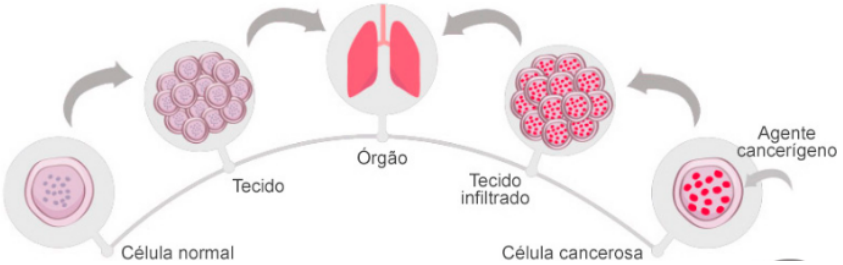


Figura 10 – O que é o câncer

Fonte: INCA (2019c)

O câncer pode ser consequência de diversas causas, entre elas fatores hereditários. Porém há alguns fatores de risco responsáveis por boa parte dos casos de câncer, como tabagismo, consumo de bebidas alcoólicas, sedentarismo, má alimentação, sexo sem proteção e poluição. Por exemplo, sabe-se que o tabagismo está altamente relacionado aos cânceres de pulmão, esôfago, estômago e boca (KASPER et al., 2017).

A detecção precoce busca a descoberta da doença nos estágios inicias. Um modo de detecção é chamando rastreamento, em que sem que existam sintomas, são feitos exames periódicos. E outro modo é o diagnóstico precoce, onde são realizados exames em pessoas que já apresentam sintomas da doença (INCA, 2018a).

O tipo do câncer é definido conforme a localização primária do tumor. Um importante fator para o tratamento adequado do câncer é o estadiamento que define a extensão da doença. O sistema de estadiamento mais comum é o TNM. O T refere-se a características do tumor primário, como tamanho e a abrangência do câncer; o N representa a presença do tumor em linfonodos vizinhos; e o M está relacionado a ocorrência de metástase em locais distantes. Os componentes TNM são combinados e organizados em estágios (I, II, III, IV), onde o estágio IV representa o mais elevado da doença (KASPER et al., 2017).

O câncer de esôfago é uma neoplasia agressiva e que geralmente tem seu diagnóstico tardio. O esôfago é um importante órgão do aparelho digestivo, localizado entre a faringe e o estômago. Os principais tipos de câncer de esôfago são o carcinoma epidermoide escamoso e o adenocarcinoma, onde o primeiro é responsável por 96% dos casos da doença (INCA, 2018a).

Embora não exista um sintoma específico que determine câncer, a ocorrência persistente de sintomas deve ser investigada por meio de exames físicos, clínicos, entre outros (KASPER et al., 2017). Sintomas comuns para o câncer de esôfago são dificuldade de engolir,

refluxo, dor abdominal e perda de peso (INCA, 2018a).

Os objetivos do tratamento são a cura, manter a qualidade vida, amenizar sintomas ou prolongar o tempo vida. Para escolha do tratamento a ser realizado deve-se levar em conta o estágio da doença e as condições de saúde do paciente. Para o câncer de esôfago, pode ser realizado por meio da combinação ou não de tratamentos locais, como cirurgia e radioterapia e tratamento sistêmico, a quimioterapia (INCA, 2018a; KASPER et al., 2017).

O Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA) é um órgão do Ministério da Saúde encarregado pelo controle e prevenção ao câncer e também por pesquisas e informações na área (INCA, 2019e). Os Registros Hospitalares de Câncer (RHC) são bases de dados que contém informações dos casos de câncer diagnosticados em um hospital. O armazenamento dessas informações auxilia no acompanhamento de diagnósticos e no planejamento de ações relacionadas a câncer. Esses RHCs são integrados e formam o Integrador de Registros Hospitalares de Câncer (IRHC), que é mantido pelo INCA (INCA, 2019d).

Os RHCs utilizam a classificação internacional de doenças para oncologia (CID-O) para codificação de neoplasias, sendo as neoplasias malignas agrupadas de C00 a C97, onde C15 corresponde ao câncer de esôfago (DATASUS, 2007).

2.2.2 Trabalhos Correlatos

O trabalho de Minelli (2013) teve como objetivo avaliar RHCs, aplicar técnicas de mineração de dados e representação de conhecimento para descoberta de regras de associação e prognóstico de pacientes com câncer de pulmão. A partir da avaliação de quatro RHCs, a base de dados selecionada foi da plataforma IRHC. Para aplicação das técnicas de mineração de dados foi utilizada a ferramenta WEKA. Nela, para descoberta de regras de associação foi escolhido o algoritmo *apriori* e para predição foi adotado o algoritmo de árvore de decisão C4.5. As regras de associação obtiveram melhores resultados que a árvore de decisão que se mostrou inviável devido a baixa quantidade de registros que restaram após reduções necessárias em razão do desbalanceamento dos dados.

O trabalho de Preissler (2016) teve como meta utilizar mineração de dados para traçar o perfil de pacientes do Rio Grande do Sul com câncer de estômago. A base de dados utilizada envolveu dados do período de 2000 a 2013, obtidos pela plataforma IRHC.

Utilizando a ferramenta WEKA, foram aplicados o algoritmo *apriori* para extração de regras de associação e o algoritmo *k-means* de agrupamento para análise do banco de dados, com base nos seguintes atributos: local de nascimento, estadiamento, idade, gênero, profissão, diagnóstico, tratamento, base de diagnóstico, morfologia e ocorrência de mais de um tumor. Foi possível concluir que a aplicação de ambas as técnicas foi viável, pois utilizando os mesmos atributos conseguiram gerar conhecimento. Porém o algoritmo *k-means* foi o que obteve melhores resultados, pois conseguiu definir o perfil dos pacientes de forma mais clara e possibilitou identificar características dos pacientes.

O trabalho de Viana (2018) tinha como objetivo analisar a base do RHC, de modo a detectar por meio de técnicas de agrupamento padrões no perfil de pacientes dos 10 tipos de cânceres mais incidentes no Brasil. Utilizando a ferramenta WEKA foram aplicados os algoritmos *k-means* e EM, em dados do período de 2005 a 2015 utilizando os mesmos atributos para os 10 tipos de cânceres: idade, sexo, raça, histórico familiar, alcoolismo, tabagismo, TNM e localização do câncer. O EM foi o algoritmo que melhor conseguiu diferenciar os *clusters*.

O trabalho de Bonini (2016) tinha como objetivo obter conhecimento relacionado a características do câncer de mama por meio da aplicação de algoritmos de árvores de decisão. O estudo foi realizado na ferramenta WEKA, onde foram aplicados os algoritmos de classificação J48 e *reptree* em uma base de dados da *Wisconsin Breast Cancer* colhida pelo Dr. William H. Wolberg, da Universidade de Wisconsin Hospitals e utilizando as variáveis espessura, tamanho, forma, adesão, células epiteliais, núcleos nus, cromatina, núcleos normais e mitoses. Ambos os algoritmos geraram resultados relevantes, pois conseguiram classificar corretamente aproximadamente 96% das instâncias.

3 MATERIAIS E MÉTODOS

Nesse capítulo será descrita a metodologia utilizada para o desenvolvimento deste projeto. Serão descritas as etapas do projeto e os principais fundamentos e tecnologias empregados. Primeiramente é descrita a base de dados, em seguida os softwares utilizados e por fim o fluxograma com a sequência das atividades.

3.1 BASE DE DADOS

A base de dados foi coletada por meio do integrador de registros hospitalares de câncer (IRHC). O IRHC é um sistema *web* desenvolvido e implantado em 2007 pelo INCA para divulgação de dados dos registros hospitalares de câncer (RHC), de modo a agilizar a transmissão e padronizar os dados, e formar uma base de dados nacional. Os RHCs são ferramentas descritas como centros de coleta, armazenamento, análise e divulgação de informações de pacientes diagnosticados com neoplasias malignas. Os RHCs utilizam o SisRHC, um sistema elaborado pelo INCA, que coleta, consolida e popula a base de dados do IRHC. Antes dos dados serem disponibilizados no integrador, os dados de cada estado passam pela coordenação estadual e depois pela coordenação nacional (INCA, 2011). Este processo pode ser visualizado na Figura 11 .

A base de dados possui 46 variáveis entre demográficas e clínicas de pacientes diagnosticados com diversos tipos de câncer. Essas informações são disponibilizadas de forma pública e não permitem a identificação de pacientes, também é válido ressaltar que os registros não representam e não devem ser utilizados para cálculo de incidência de casos de câncer (INCA, 2011).

O IRHC disponibiliza dados de 1985 a 2017 (exceto 1987), onde foram coletados cerca de 3.269.715 instâncias. Porém para o período de 1985 a 1997 é baixa a quantidade de casos consolidados no integrador e não constam casos para o câncer de esôfago. A Figura 12 mostra

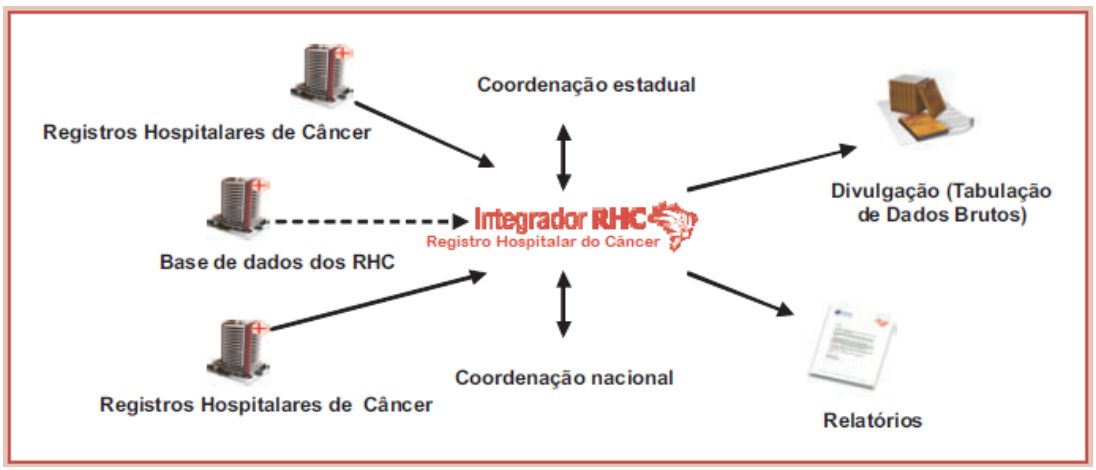


Figura 11 – Fluxo de informações do IRHC.

Fonte: INCA (2011)

o número total de registros de casos de câncer de esôfago no Brasil por ano, no período de 1998 a 2017, totalizando 78.420 instâncias. O fato dos anos 2016 e 2017 terem decaído o número de instâncias na base se deve ao fato da demora de aproximadamente dois anos para consolidação dos dados na plataforma, o que também explica a ausência de dados do ano de 2018.

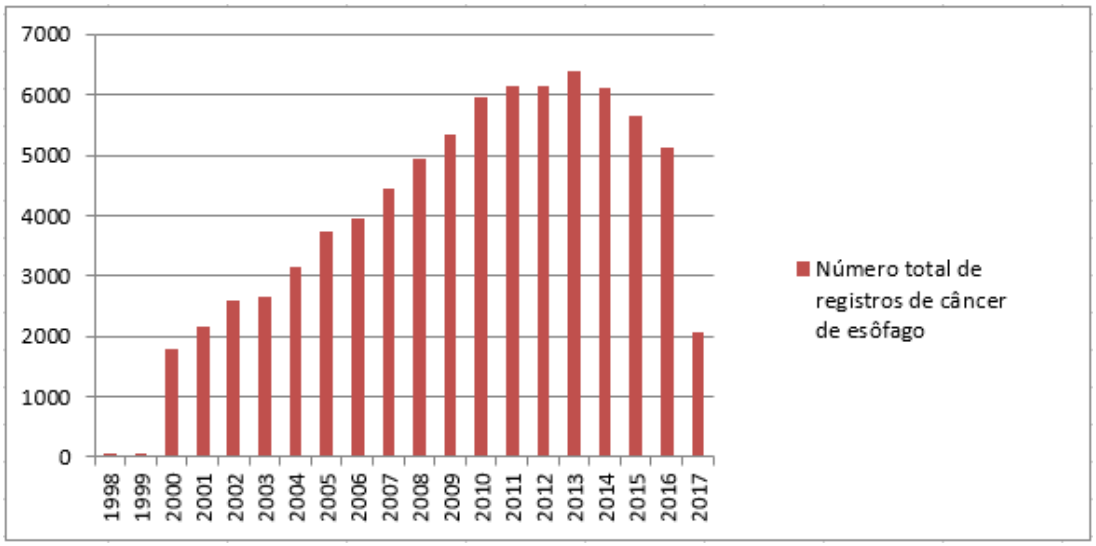


Figura 12 – Número total de registros de casos de câncer de esôfago no Brasil.

Fonte: Autoria própria

Além de poder selecionar apenas dados dos anos de interesse, a plataforma também permite baixar dados de todos os estados brasileiros ou de cada estado separadamente. Os arquivos com os dados são fornecidos pelo IRHC no formato *dbf* (*Data Base File*). Também são concedidos o dicionário de dados, a tabela de códigos das clínicas e arquivos com as descrições

do que cada código representa para o valor das variáveis (Figura 13).

Figura 13 – Interface da página de download da base de dados do IRHC.
Fonte: INCA (2019b)

3.2 SOFTWARES

Nesta seção serão descritos os softwares utilizados para realização deste trabalho.

3.2.1 TabWin

Arquivos em formato *dbf* podem ser visualizados utilizando o *software* TabWin[1]. Esta ferramenta gratuita é um tabulador desenvolvido pelo departamento de informática do sistema único de saúde (DATASUS) (MARCELINO, 2011). Nele, é possível, por exemplo realizar

[1] http://www2.datasus.gov.br/DATASUS/index.php?area=060805&item=3

operações aritméticas e gerar gráficos, além de realizar a criação de arquivos para formatos como *csv* (*Comma Separated Values*) e *sql* (*Structured Query Language*). Foi utilizado o TabWin na versão 3.5 para criação de arquivos *sql* a partir dos arquivos *dbf*.

3.2.2 PostgreSQL

O PostgreSQL[2] é um sistema de banco de dados objeto-relacional livre, cujo desenvolvimento iniciou em 1986 em um projeto chamado Postgres da Universidade da Califórnia em Berkeley. O PostgreSQL é compatível com os principais sistemas operacionais, oferece diversos recursos (POSTGRESQL, 2018) e pode ser gerenciado por meio de uma ferramenta chamada pgAdmin.

Foi utilizado PostgreSQL na versão 9.6, com pgAdmin 4 na versão 4.6. Este banco de dados foi escolhido devido a familiaridade com a ferramenta e esta ser gratuita. Os arquivos gerados em formato *sql* foram utilizados para criação e população das tabelas com os registros. Em seguida, foi criada uma tabela contendo apenas os registros de casos de câncer de esôfago, ou seja, onde a variável loctudet, que representa a localização primária do tumor, fosse igual a C15.

3.2.3 WEKA

A WEKA[3] (*Waikato Environment for Knowledge Analysis*) é uma ferramenta gratuita e de código aberto, desenvolvida na Universidade de Waikato na Nova Zelândia. A ferramenta incorpora diversos algoritmos de aprendizado de máquina e pré-processamento de dados (WITTEN et al., 2011).

A ferramenta pode ser utilizada por meio de quatro diferentes interfaces: *Explorer*, *Experimenter*, *Simple CLI* e *KnowledgeFlow*, como mostra a Figura 14, onde a interface de uso mais comum é a *Explorer*. Os formatos suportados pela ferramenta são *arff* (*Attribute Relation*

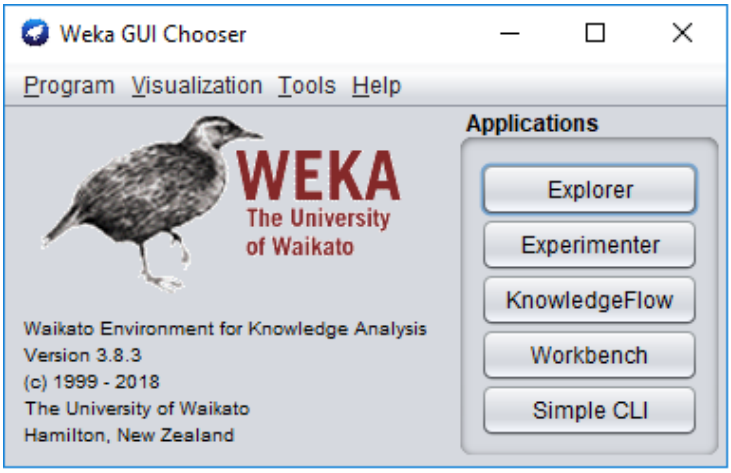


Figura 14 – Interface inicial WEKA.

Fonte: Autoria própria

File Format) para manipulação dos dados, *csv* e *c45* (GOLDSCHMIDT; PASSOS, 2015).

A WEKA foi escolhida devido a sua interface amigável e familiaridade com a ferramenta, sendo utilizada a versão 3.8.3. Onde foi realizada a aplicação dos métodos de seleção de atributos, *info gain*, *gain ratio* e PCA, aplicação do algoritmo de extração de regras de associação, *apriori*, algoritmos de classificação, J48 e *reptree* e algoritmos de agrupamento de dados, *K-means* e EM.

3.3 FLUXO DE ATIVIDADES

A Figura 15 demonstra o fluxo das atividades para a realização do trabalho. O fluxo do processo foi iniciado com a obtenção da base de dados por meio do IRHC e do TabWin, em seguida a seleção dos dados referentes ao câncer de esôfago e o pré-processamento dos dados utilizando o PostgreSQL e a WEKA. Então, a aplicação do *info gain*, *gain ratio* e análise de componentes principais, para seleção de variáveis, assim como a mineração de regras de associação com o algoritmo *apriori* e a aplicação de algoritmos de classificação e de agrupamento de dados, também realizados com a ferramenta WEKA. Por fim a análise e interpretação dos resultados obtidos.

[2]https://www.postgresql.org/about/
[3]https://www.cs.waikato.ac.nz/ml/weka/index.html

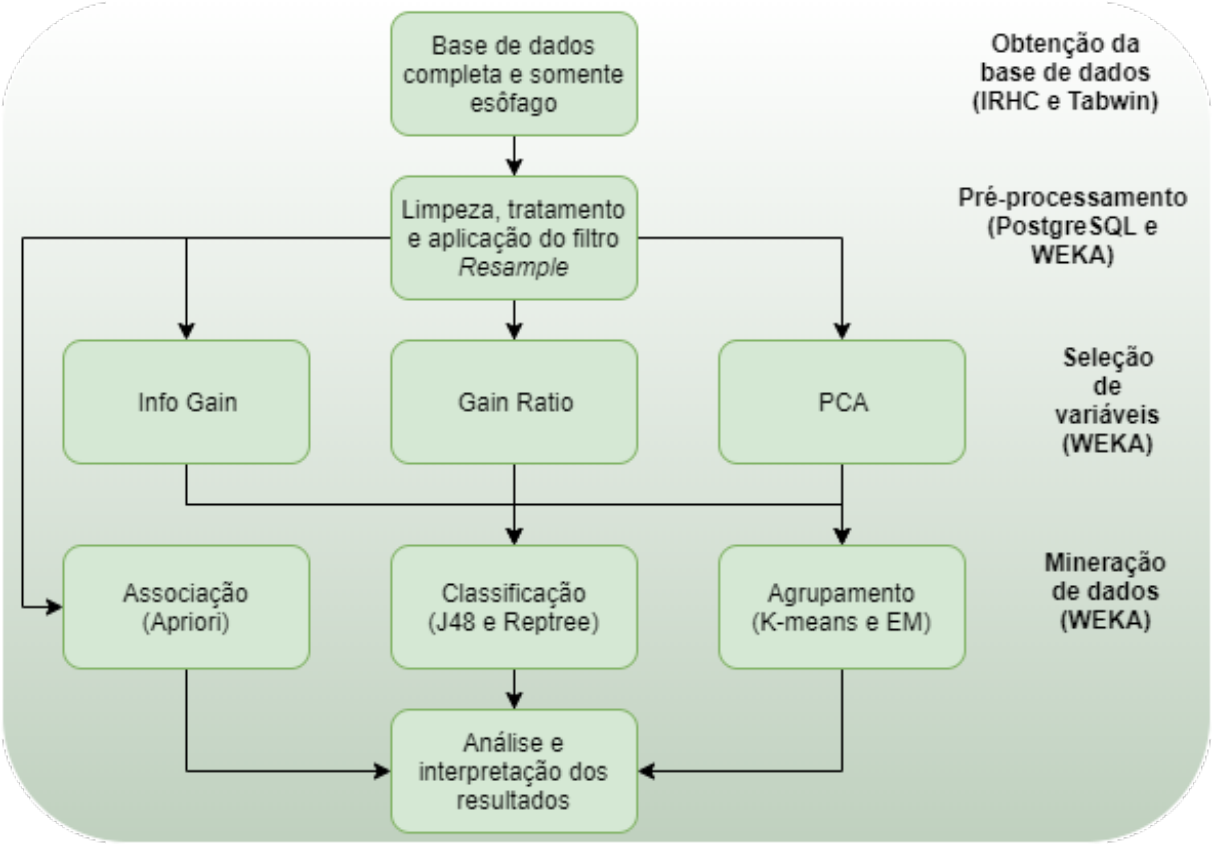


Figura 15 – Fluxograma de atividades para realização do trabalho.

Fonte: Autoria própria

3.3.1 Obtenção da base de dados e pré-processamento

Após obtenção da base de dados na plataforma do IRHC, foram criadas duas BDs, a BD 1 contendo apenas casos de câncer de esôfago (variável LOCTUDET com valor igual a C15) e a BD 2 contendo todos os casos, onde foi atribuído ”sim” a variável LOCTUDET (localização primária do tumor) para casos de câncer de esôfago e ”não” para o restante dos casos de câncer.

O pré-processamento iniciou no PostgreSQL onde inicialmente os valores dos atributos foram mapeados de acordo com o dicionário de dados, transformando as variáveis numéricas para categóricas, melhorando assim sua visualização e entendimento.

Apesar dos dados já passarem por um tratamento antes de serem consolidados no IRHC, foi necessário realizar uma limpeza onde foram removidos valores como ”sem informação”, pois não seriam úteis para realização do trabalho e dados inconsistentes, que não estavam mapeados no dicionário de dados.

A variável referente a ocupação foi modificada criando grupos com ocupações semelhantes: ”agrícola” para trabalhos relacionados a agricultura e pecuária, ”serviços” para trabalhos referentes a prestação de serviços sem esforço físico, ”braçal” para prestação de serviços com trabalho braçal, ”construção” para ocupações relacionadas a construção civil e ”outros” para o restante das ocupações que não se encaixavam a estes grupos ou que não estavam especificadas na base de dados original.

A variável idade, foi alterada criando três grupos de idades, ”criançaEjovem” para idade de 0 a 17 anos, ”adulto” de 18 a 59 e ”idoso” para idades de 60 anos ou mais. Após todas as atividades de pré-processamento, na Base 1 restaram 11.986 instâncias e 444.072 instâncias na Base 2.

As bases passaram por uma pré seleção onde foram descartadas variáveis relacionadas a data, localização geográfica, entre outros atributos que não fossem relevantes, por exemplo, a lateralidade do tumor que considerando o tipo de tumor tratado neste trabalho, é um atributo que não se aplica por não se tratar de um órgão par.

Para utilização da Base 2, que continha todos os casos de câncer, devido ao desbalanceamento dos valores da variável LOCTUDET foi necessário a utilização do filtro *Resample* para criação de uma amostra com o mesmo número de instâncias para as duas classes, já que a classe onde LOCTUDET igual a ”não” possuía um número muito elevado de instâncias.

O filtro *Resample* é um filtro de instância supervisionado que pode ser usado para gerar uma amostra dos dados com ou sem substituição das instâncias e preservando a distribuição dos dados originais na amostra (WITTEN et al., 2011). O *Resample* foi utilizado com o percentual de tamanho da amostra de 2,7%, bias igual a 1 e o restante de suas configurações padrão, resultando 5.994 instâncias para cada classe.

3.3.2 Seleção de variáveis

Para seleção de variáveis foram aplicados os métodos *Info Gain*, *Gain Ratio* e PCA, onde a seleção foi feita apenas sobre as variáveis que correspondiam a fatores anteriores ao diagnóstico: sexo, idade, raça, instrução, histórico familiar, alcoolismo, tabagismo e ocupação, e utilizando apenas a BD 2 devido a esta ser a base de dados onde foram aplicados os algoritmos de classificação e agrupamento .

O *Info Gain* e o *Gain Ratio* foram executados utilizando o método de pesquisa *ranker*.

Os procedimentos geraram resultados semelhantes, onde as variáveis selecionadas foram as mesmas: sexo, tabagismo e alcoolismo.

Para o PCA também foi utilizado o método de pesquisa *ranker*, foram selecionadas as variáveis com maior influência sobre a variável alvo, LOCTUDET, portanto foram escolhidas as variáveis com maior influência nas três primeiras componentes principais: idade, sexo, tabagismo e alcoolismo.

3.3.3 Mineração de dados

Para a mineração de regras de associação foi utilizada a base que continha apenas casos de câncer de esôfago (BD 1), já para classificação foi utilizada a base com todos os casos (BD 2) e para agrupamento foram aplicados os algoritmos às duas bases.

Em relação a mineração de regras de associação inicialmente foi acessado a BD 1, em seguida foi feito um pré-processamento onde foram deixadas apenas as variáveis a serem utilizadas: ALCOOLISM, BASMAIMP, DIAGANT, ESTADIAG, ESTDFIMT, HISTFAMC, IDADE, INSTRUC, LOCTUPRI, MAISUMTU, OCUPACAO, PRITRATH, RACACOR, SEXO, TABAGISM, TIPOHIST e TPCASO. Os valores de suporte e confiança mínimos foram definidos de modo que fossem geradas no máximo 20 regras. Em seguida foi feita a mineração de regras de associação com o algoritmo *apriori*. A escolha deste algoritmo se deve a sua popularidade. Devido a base de dados conter variáveis onde um grande número de instâncias contém um determinado valor, ao gerar regras de associação essas variáveis predominavam criando regras de associação repetitivas, por isso foram geradas regras e logo após foi feito a remoção da variável predominante para geração de novas regras e assim sucessivamente, como pode ser observado no diagrama da Figura 16.

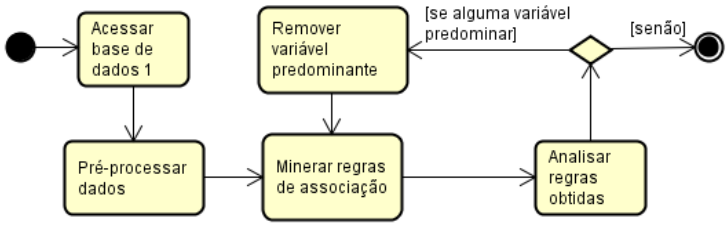


Figura 16 – Diagrama de atividades da aplicação de mineração de regras de associação para a Base de dados 1.

Fonte: Autoria própria

Conforme a Figura 17, os algoritmos de classificação foram aplicados a BD 2. Utilização o filtro *Resample* e os três conjuntos de variáveis: as variáveis anteriores ao diagnóstico, as variáveis selecionadas por *Info Gain* e *Gain Ratio* e as variáveis selecionadas pelo PCA. A cada conjunto de variáveis foram executados os algoritmos J48 e *reptree*,

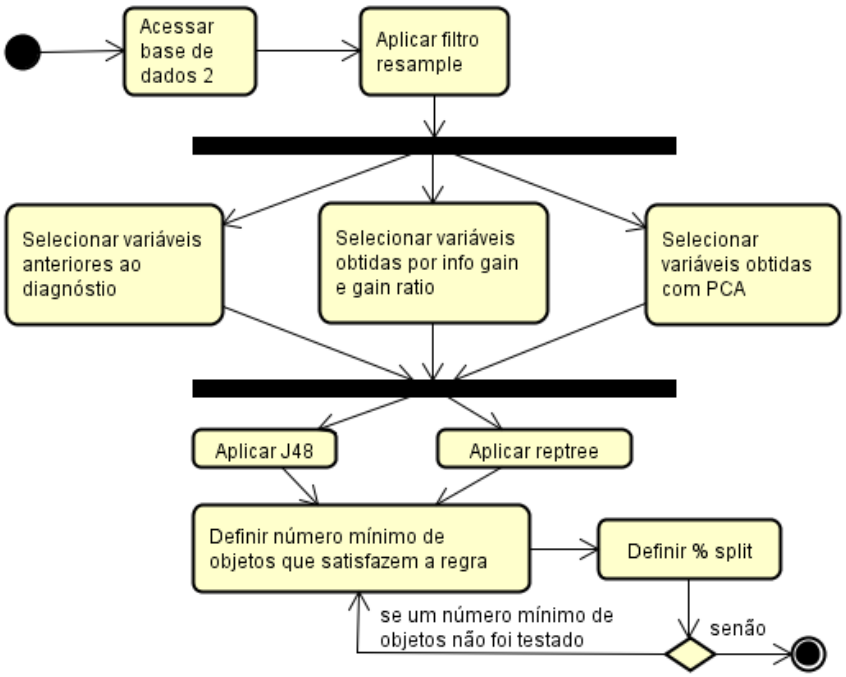


Figura 17 – Diagrama de atividades de aplicação dos métodos de classificação para a Base de dados 2.

Fonte: Autoria própria

utilizando os seguintes números mínimos de elementos que devem satisfazer cada regra de classificação gerada, 2, 10 e 100, no qual para cada número mínimo de objetos foram aplicados percentuais de *split* de 25%, 50%, 66% e 75%, que representam o percentual do conjunto de dados que são usados como conjunto de treinamento do classificador, sendo assim, o percentual restante é usado como conjunto de teste do classificador. Esses valores de *split* foram definidos de forma empírica, de modo a testar a estabilidade dos modelos de classificação com diferentes configurações desses parâmetros. Foram escolhidos estes algoritmos para classificação devido a sua utilização em trabalhos correlatos de Minelli (2013) e Bonini (2016), além do C4.5 ser um tradicional algoritmo de classificação, que possibilita a fácil interpretação da árvore e que corresponde ao J48 na implementação da WEKA.

Os algoritmos de agrupamento de dados foram utilizados nas duas bases de dados, BD 1 e BD 2. A escolha dos algoritmos *k-means* e EM foi devido a sua aplicação nos trabalhos correlatos de Preissler (2016) e Viana (2018), onde foram utilizados sem a realização de métodos de seleção de variáveis.

Para a BD 2, como mostra a Figura 18, foram aplicados com e sem o filtro *Resample* e para cada conjunto de variáveis, o algoritmo *k-means* utilizando as distâncias Euclidiana e Manhattan e o algoritmo EM, ambos realizando a avaliação dos *clusters* em relação a classe LOCTUDET, com número de *cluster* igual a 2, devido a classe LOCTUDET poder receber dois valores (”sim” e ”não”) e com o restante das configurações padrão.

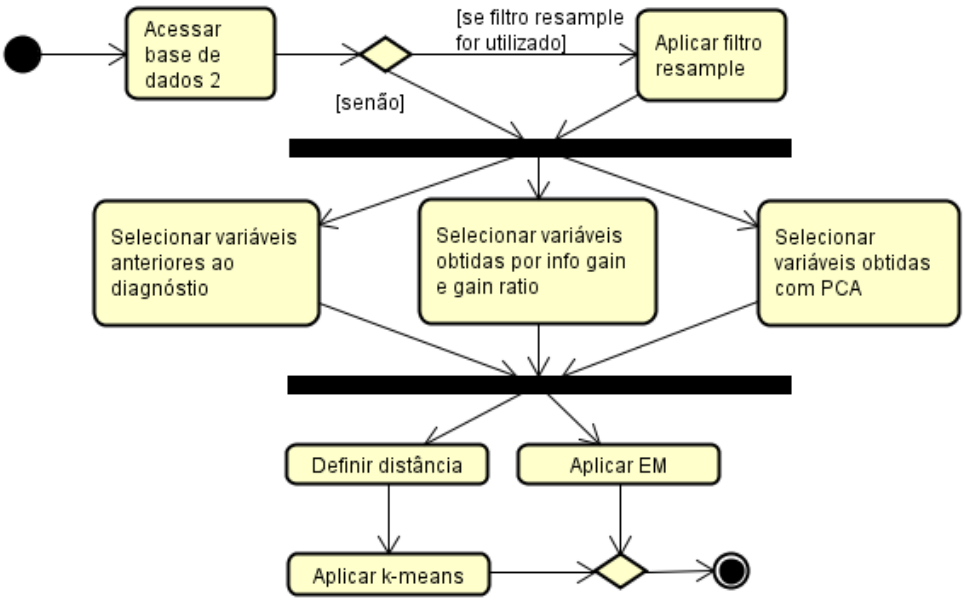


Figura 18 – Diagrama de atividades de aplicação dos métodos de agrupamento para a Base 2.

Fonte: Autoria própria

Para o agrupamento na BD 1, Figura 19, foi escolhido utilizar apenas as variáveis selecionadas pelo PCA devido aos trabalhos correlatos de (VIANA, 2018) e (PREISSLER, 2016) já terem realizado o agrupamento de dados utilizando variáveis escolhidas de forma empírica. Foram aplicados os algoritmos *k-means* e EM, ambos com número de *clusters* igual a 2 e o restante de suas configurações padrões.

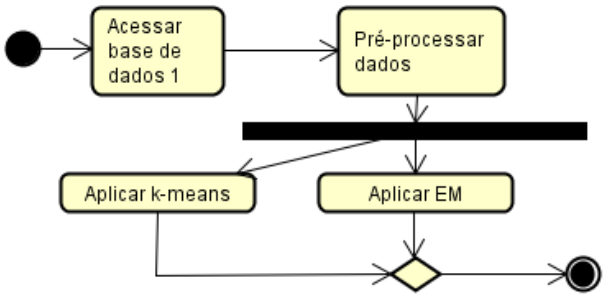


Figura 19 – Diagrama de atividades de aplicação dos métodos de agrupamento para a Base 1.

Fonte: Autoria própria

Os resultados, assim como a avaliação e interpretação dos mesmos, são apresentados no próximo capítulo.

4 RESULTADOS E DISCUSSÃO

Neste capítulo serão apresentados os resultados obtidos com a mineração de regras de associação, classificação e agrupamento de dados, assim como uma análise e discussão destes resultados.

4.1 RESULTADOS

Primeiramente serão apresentados os resultados das aplicações dos algoritmos *apriori*, J48, *reptree*, *k-means* e EM de acordo com as bases de dados utilizadas. Os valores dos parâmetros foram definidos após vários testes empíricos, que incluíram várias tentativas de valores e que foram selecionados por serem considerados satisfatórios.

4.1.1 Regras de associação

Com uma análise inicial percebeu-se que aproximadamente 96% dos casos de câncer de esôfago tinham como base mais importante para o diagnóstico (BASMAIMP) a histologia do tumor. Aproximadamente 96% das instâncias tinham como ”sim” a ocorrência de mais de um tumor primário (MAISUMTU). Por volta de 74% das instâncias como casos (TPCASO) analíticos, classificação que se dá de acordo com a abordagem realizada no hospital e ao estado do paciente. Aproximadamente 82% dos casos tinham o tipo histológico do tumor primário (TIPOHIST) como 8070/3 que se refere ao carcinoma epidermoide escamoso [1].

[1] https://www.hospitalsiriolibanes.org.br/hospital/especialidades/centro-oncologia/esofago/Paginas/diagnosticos.aspx

Devido a predominância dessas variáveis e ao fato de possuírem uma relação muito grande, ao minerar regras de associação foram geradas uma grande quantidade de regras, portanto após uma mineração inicial de regras, essas variáveis foram sendo removidas.

Utilizando todas as 18 variáveis, com suporte igual a 70% e confiança igual a 90% foram mineradas 18 regras que podem ser observadas na Tabela 2.

Tabela 2 – Resultados algoritmo Apriori utilizando 18 variáveis

Regras de associação	Suporte	Confiança
$tipohist = 8070/3 \Longrightarrow basmaimp = histTumorPrim$	78%	97%
$tipohist = 8070/3, maisumtu = sim \Longrightarrow basmaimp = histTumorPrim$	75%	97%
$diagant = comDiagSemTrat \Longrightarrow maisumtu = sim$	73%	97%
$diagant = comDiagSemTrat, basmaimp = histTumorPrim \Longrightarrow maisumtu = sim$	70%	97%
$basmaimp = histTumorPrim \Longrightarrow maisumtu = sim$	91%	96%
$maisumtu = sim \Longrightarrow basmaimp = histTumorPrim$	91%	96%
$tipohist = 8070/3 \Longrightarrow maisumtu = sim$	78%	96%
$diagant = comDiagSemTrat \Longrightarrow basmaimp = histTumorPrim$	72%	96%
$basmaimp = histTumorPrim, tipohist = 8070/3 \Longrightarrow maisumtu = sim$	72%	96%
$sexo = masculino \Longrightarrow maisumtu = sim$	72%	96%
$sexo = masculino \Longrightarrow basmaimp = histTumorPrim$	72%	96%
$tpcaso = analitico \Longrightarrow basmaimp = histTumorPrim$	71%	96%
$rzntr = naoSeAplica \Longrightarrow maisumtu = sim$	71%	96%
$rzntr = naoSeAplica \Longrightarrow basmaimp = histTumorPrim$	71%	96%
$diagant = comDiagSemTrat, maisumtu = sim \Longrightarrow basmaimp = histTumorPrim$	70%	96%
$tpcaso = analitico \Longrightarrow maisumtu = sim$	70%	96%
$tipohist = 8070/3 \Longrightarrow basmaimp = histTumorPrim, maisumtu = sim$	75%	93%
$diagant = comDiagSemTrat \Longrightarrow basmaimp = histTumorPrim, maisumtu = sim$	70%	93%

Fonte: Autoria própria

Com a remoção da variável BASMAIMP e executando o algoritmo *apriori* nas variáveis restantes, utilizando suporte e confiança iguais a 60% e 85% respectivamente, foram mineradas 13 regras mostradas na Tabela 3.

Tabela 3 – Resultados algoritmo Apriori utilizando 17 variáveis

Regras de associação	Suporte	Confiança
$diagant = comDiagSemTrat \Longrightarrow maisumtu = sim$	73%	97%
$diagant = comDiagSemTrat, tipohist = 8070/3 \Longrightarrow maisumtu = sim$	60%	97%
$tipohist = 8070/3 \Longrightarrow maisumtu = sim$	78%	96%
$sexo = masculino \Longrightarrow maisumtu = sim$	72%	96%
$rzntr = naoSeAplica \Longrightarrow maisumtu = sim$	71%	96%
$tpcaso = analitico \Longrightarrow maisumtu = sim$	70%	96%
$tpcaso = analitico, rzntr = naoSeAplica \Longrightarrow maisumtu = sim$	63%	96%
$tpcaso = analitico \Longrightarrow rzntr = naoSeAplica$	66%	90%
$tpcaso = analitico, maisumtu = sim \Longrightarrow rzntr = naoSeAplica$	63%	90%
$rzntr = naoSeAplica \Longrightarrow tpcaso = analitico$	66%	89%
$maisumtu = sim, rzntr = naoSeAplica \Longrightarrow tpcaso = analitico$	63%	89%
$tpcaso = analitico \Longrightarrow maisumtu = sim, rzntr = naoSeAplica$	63%	86%
$rzntr = naoSeAplica \Longrightarrow tpcaso = analitico, maisumtu = sim$	63%	85%

Fonte: Autoria própria

Removendo a variável MAISUMTU e TPCASO e realizando a mineração de regras de associação com as variáveis restantes, utilizando suporte e confiança iguais a 35% e 85%, foram mineradas 9 regras, apresentadas na Tabela 4.

Tabela 4 – Resultados algoritmo Apriori utilizando 15 variáveis

Regras de associação	Suporte	Confiança
$alcoolis = sim \Longrightarrow sexo = masculino$	38%	87%
$tabagism = sim, rzntr = naoSeAplica \Longrightarrow tipohist = 8070/3$	37%	87%
$sexo = masculino, tabagism = sim \Longrightarrow tipohist = 8070/3$	40%	86%
$tabagism = sim, diagant = comDiagSemTrat \Longrightarrow tipohist = 8070/3$	37%	86%
$alcoolis = sim \Longrightarrow tabagism = sim$	37%	86%
$alcoolis = sim \Longrightarrow tipohist = 8070/3$	37%	86%

Continua na página seguinte.

Tabela 4–Resultados algoritmo Apriori utilizando 15 variáveis

Continuação da página anterior.

Regras de associação	Suporte	Confiança
$tabagism = sim \Longrightarrow tipohist = 8070/3$	49%	85%
$sexo = masculino, alcoolis = sim \Longrightarrow tipohist = 8070/3$	38%	85%
$sexo = masculino, idade = adulto \Longrightarrow tipohist = 8070/3$	37%	85%

Fonte: Autoria própria

Removendo a variável TIPOHIST e executando o algoritmo *apriori* nas variáveis restantes, utilizando suporte e confiança iguais a 25% e 85% respectivamente, foram mineradas 9 regras conforme mostra a Tabela 5.

Tabela 5 – Resultados algoritmo Apriori utilizando 14 variáveis

Regras de associação	Suporte	Confiança
$pritrath = radioQuimio \Longrightarrow rzntr = naoSeAplica$	29%	90%
$alcoolis = sim, rzntr = naoSeAplica \Longrightarrow sexo = masculino$	28%	88%
$alcoolis = sim \Longrightarrow sexo = masculino$	38%	87%
$alcoolis = sim, tabagism = sim \Longrightarrow sexo = masculino$	33%	87%
$alcoolis = sim, diagant = comDiagSemTrat \Longrightarrow sexo = masculino$	28%	87%
$alcoolis = sim \Longrightarrow tabagism = sim$	37%	86%
$sexo = masculino, alcoolis = sim \Longrightarrow tabagism = sim$	33%	86%
$alcoolis = sim, diagant = comDiagSemTrat \Longrightarrow tabagism = sim$	28%	86%
$alcoolis = sim, rzntr = naoSeAplica \Longrightarrow tabagism = sim$	27%	85%

Fonte: Autoria própria

4.1.2 Classificação

Iniciando a execução com as variáveis referentes a informações anteriores ao diagnóstico (variáveis citadas na seção 3.3.2) para aplicação do algoritmo de classificação J48, a Tabela 6 apresenta os resultados obtidos. Conforme aumentou-se o número mínimo de instâncias na base de dados necessários para uma regra de classificação ser considerada válida,

Tabela 6 – Resultados da aplicação do algoritmo J48 utilizando variáveis anteriores ao diagnóstico

Número mínimo de instâncias	*% split*	Número de regras	% de acerto
2	25%	295	71,66%
2	50%	295	71,8%
2	66%	295	72,47%
2	75%	295	72,6%
10	25%	159	71,64%
10	50%	159	72,12%
10	66%	159	72,64%
10	75%	159	72,13%
100	25%	39	71,38%
100	50%	39	71,63%
100	66%	39	71,44%
100	75%	39	72%

Fonte: Autoria própria

percebe-se que o número de regras mineradas diminuiu e o percentual de acerto manteve-se, conforme explicado no Capítulo 3 estes valores foram definidos de forma empírica.

Para as variáveis geradas com o *Info gain* e o *Gain ratio* os resultados do algoritmo J48 são apresentados na Tabela 7. Observa-se que com apenas três variáveis, com o aumento do número mínimo de instâncias na base de dados necessários para uma regra de classificação ser considerada válida, o número de regras mineradas diminuiu e o percentual de acerto foi mantido.

Tabela 7 – Resultados algoritmo J48 com variáveis selecionadas por *Info gain* e *Gain ratio*

Número mínimo de instâncias	*% split*	Número de regras	% de acerto
2	25%	20	71,36%
2	50%	20	71,07%
2	66%	20	72,15%

Continua na página seguinte.

Tabela 7–Resultados algoritmo J48 com variáveis selecionadas por *Info gain* e *Gain ratio*

Continuação da página anterior.

Número mínimo de instâncias	*% split*	Número de regras	% de acerto
2	75%	20	73%
10	25%	16	71,36%
10	50%	16	71,07%
10	66%	16	72,08%
10	75%	16	72,93%
100	25%	16	71,36%
100	50%	16	71,07%
100	66%	16	72,08%
100	75%	16	72,93%

Fonte: Autoria própria

Para as variáveis selecionadas com a utilização do PCA, os resultados da aplicação do algoritmo J48, são apresentados na Tabela 8. Tem-se que elevando o número mínimo de instâncias na base de dados necessários para uma regra de classificação ser considerada válida, percebe-se que o número de regras mineradas foi semelhante e mesmo utilizando apenas quatro variáveis o percentual de acerto se manteve.

Tabela 8 – Resultados algoritmo J48 utilizando variáveis selecionadas por PCA

Número mínimo de instâncias	*% split*	Número de regras	% de acerto
2	25%	22	71,3%
2	50%	22	71,1%
2	66%	22	72%
2	75%	22	72,7%
10	25%	22	71,29%
10	50%	22	71,08%
10	66%	22	71,9%

Continua na página seguinte.

Tabela 8–Resultados algoritmo J48 utilizando variáveis selecionadas por PCA

Continuação da página anterior.

Número mínimo de instâncias	% *split*	Número de regras	% de acerto
10	75%	22	72,6%
100	25%	20	71,3%
100	50%	20	71,08%
100	66%	20	71,9%
100	75%	20	72,6%

Fonte: Autoria própria

Foram testados outros algoritmos de classificação (*RandomTree*[2], *RandomForest*[2] e LMT[3]), porém estes geraram resultados do percentual de acerto semelhantes, portanto devido ao trabalho correlato de Bonini (2016) ter utilizado *reptree* e conseguido bons resultados, optou-se por apresentar neste trabalho os resultados obtidos com este algoritmo. Assim, para o algoritmo *reptree*, a Tabela 9 mostra os resultados alcançados utilizando as variáveis relacionadas a fatores anteriores ao diagnóstico com o algoritmo *reptree*. Percebe-se que à medida que o número mínimo de objetos necessários para uma regra de classificação ser considerada válida foi aumentado, o número de regras obtidas decresceu enquanto o percentual de acerto permaneceu estável.

Tabela 9 – Resultados do algoritmo *reptree* utilizando variáveis anteriores ao diagnóstico

Número mínimo de instâncias	% *split*	Número de regras	% de acerto
2	25%	428	70,74%
2	50%	428	71,38%
2	66%	428	71,49%
2	75%	428	71,37%
10	25%	193	71,01%
10	50%	193	71,45%
10	66%	193	71,54%

Continua na página seguinte.

[2]http://www.irdindia.in/journal_ijaece/pdf/vol3_iss4/2.pdf
[3]http://mtc-m21b.sid.inpe.br/col/sid.inpe.br/mtc-m21b/2017/03.23.16.13/doc/publicacao.pdf

Tabela 9–Resultados do algoritmo *reptree* utilizando variáveis anteriores ao diagnóstico

Continuação da página anterior.

Número mínimo de instâncias	*% split*	Número de regras	% de acerto
10	75%	193	72,7%
100	25%	32	68,52%
100	50%	32	71,2%
100	66%	32	71,73%
100	75%	32	72,3%

Fonte: Autoria própria

Com as variáveis obtidas por *info gain* e *gain ratio*, conforme é mostrado na Tabela 10 o número de regras permaneceu inalterado independente do número mínimo de instâncias na base de dados necessários para uma regra ser considerada válida, enquanto o percentual de acerto se manteve.

Tabela 10 – Resultados algoritmo *reptree* utilizando variáveis selecionadas por *info gain* e *gain ratio*

Número mínimo de instâncias	*% split*	Número de regras	% de acerto
2	25%	25	71,69%
2	50%	25	71,43%
2	66%	25	72,08%
2	75%	25	72,97%
10	25%	25	71,69%
10	50%	25	71,43%
10	66%	25	72,08%
10	75%	25	72,97%
100	25%	25	68,52%
100	50%	25	71,43%
100	66%	25	72,08%

Continua na página seguinte.

Tabela 10–Resultados algoritmo *reptree* utilizando variáveis selecionadas por *info gain* e *gain ratio*

Continuação da página anterior.

Número mínimo de instâncias	*% split*	Número de regras	% de acerto
100	75%	25	72,93%

Fonte: Autoria própria

Utilizando as variáveis selecionadas pelo PCA, como mostrado na Tabela 11, o número de regras permaneceu o mesmo, apesar da alteração no número mínimo de instâncias na base de dados necessários para uma regra ser considerada válida, assim como o percentual de acerto se manteve sem grandes alterações.

Tabela 11 – Resultados do algoritmo *reptree* utilizando variáveis selecionadas por PCA

Número mínimo de instâncias	*% split*	Número de regras	% de acerto
2	25%	29	70,68%
2	50%	29	71,27%
2	66%	29	72,08%
2	75%	29	72,93%
10	25%	29	70,68%
10	50%	29	71,27%
10	66%	29	72,08%
10	75%	29	72,93%
100	25%	29	69,99%
100	50%	29	71,07%
100	66%	29	71,68%
100	75%	29	72,93%

Fonte: Autoria própria

4.1.3 Agrupamento

Aplicando agrupamento de dados na BD 2 e avaliando os *clusters* com respeito a classe, primeiramente foram executados os algoritmos utilizando o filtro *resample*. Pela Tabela

12 pode-se observar os resultados obtidos com o algoritmo *k-means* utilizando as variáveis anteriores ao diagnóstico, o algoritmo obteve aproximadamente 50% de acerto, onde entre os casos de câncer de esôfago foi 52% de acerto.

Tabela 12 – Resultados algoritmo *k-means* utilizando filtro *resample* e variáveis anteriores ao diagnóstico

Distância	% de acerto	% de erro
Euclidiana	50,50%	49,50%
Manhattan	50,50%	49,50%

Fonte: Autoria própria

Conforme é mostrado na Tabela 13, com as variáveis selecionadas por *Info Gain* e *Gain Ratio* e filtro *resample*, o algoritmo *k-means* obteve aproximadamente 52% de acerto, onde apenas 28% do casos de câncer de esôfago foram agrupados corretamente.

Tabela 13 – Resultados algoritmo *k-means* utilizando filtro *resample* e variáveis selecionadas por *Info Gain* e *Gain Ratio*

Distância	% de acerto	% de erro
Euclidiana	51,18%	48,82%
Manhattan	51,18%	48,82%

Fonte: Autoria própria

A Tabela 14 apresenta os resultados obtidos por meio das variáveis obtidas pelo PCA, onde com a utilização do filtro *resample* o algoritmo *k-means* conseguiu 56% de acerto de forma geral e 30% de acerto apenas entre os casos de câncer de esôfago.

Tabela 14 – Resultados do algoritmo *k-means* utilizando filtro *resample* e variáveis selecionadas por PCA

Distância	% de acerto	% de erro
Euclidiana	56,08%	43,92%
Manhattan	56,08%	43,92%

Fonte: Autoria própria

Pela Tabela 15 são demonstrados os resultados obtidos com o algoritmo EM com cada grupo de variáveis utilizando filtro *resample*, com todas as variáveis relacionadas a fatores anteriores ao diagnóstico obteve-se 60,31% de acerto no geral e 70% de acerto entre as instâncias com câncer de esôfago, com as variáveis selecionadas com *Info Gain* e *Gain Ratio* no geral 68,74% de acerto e 63% entre as instâncias com câncer de esôfago e com as variáveis obtidas com o PCA 88,62% de acerto no geral e 64% entre as instâncias com câncer de esôfago.

Tabela 15 – Resultados do algoritmo EM com a utilização do filtro *resample*

Variáveis	% de acerto	% de erro
Anteriores ao diagnóstico	60,31%	39,69%
Info Gain e *Gain Ratio*	68,74%	31,26%
PCA	68,82%	31,18%

Fonte: Autoria própria

Em seguida, foram aplicados os algoritmos de agrupamento de dados na BD 2 sem utilizar o filtro *resample*. Sendo assim, a Tabela 16 demonstra o percentual de acerto do agrupamento de modo geral, em relação a todos os casos de câncer, onde obteve-se 61,75% de acerto ao agrupar as variáveis considerando LOCTUDET como Sim e Não. De forma mais específica, entre os casos de câncer de esôfago, o algoritmo conseguiu agrupar corretamente 67,3% das instâncias, utilizando as variáveis anteriores ao diagnóstico.

Tabela 16 – Resultados do algoritmo *k-means* utilizando variáveis anteriores ao diagnóstico

Distância	% de acerto	% de erro
Euclidiana	61,75%	38,25%
Manhattan	61,75%	38,25%

Fonte: Autoria própria

Conforme pode ser observado na Tabela 17, utilizando as variáveis obtidas por *Info Gain* e *Gain Ratio* o algoritmo *k-means* sem o filtro *resample* conseguiu 71,57% de acerto, onde agrupou corretamente 67,51% das instâncias de casos de câncer de esôfago.

Tabela 17 – Resultados do algoritmo *k-means* utilizando variáveis selecionadas por *Info Gain* e *Gain Ratio*

Distância	% de acerto	% de erro
Euclidiana	71,57%	28,43%
Manhattan	71,57%	28,43%

Fonte: Autoria própria

Sem o filtro *resample* e utilizando as variáveis selecionadas por PCA, como pode ser observado na Tabela 18, o algoritmo *k-means* alcançou 76,15% de acerto no agrupamento de modo geral e agrupou corretamente 58,94% dos casos de câncer de esôfago.

Por fim, pode-se observar pela Tabela 19 que o algoritmo EM sem a utilização do filtro *resample*, obteve 67,03% de acerto com as variáveis anteriores ao diagnóstico onde acertou 74% entre as instâncias com câncer de esôfago, 70,4% com as variáveis obtidas por *Info Gain* e *Gain Ratio* onde obteve 64% de acerto entre as instâncias com câncer de esôfago e 73,52% de

Tabela 18 – Resultados do algoritmo *k-means* utilizando variáveis selecionadas por PCA

Distância	% de acerto	% de erro
Euclidiana	76,15%	23,85%
Manhattan	76,15%	23,85%

Fonte: Autoria própria

Tabela 19 – Resultados do algoritmo EM

Variáveis	% de acerto	% de erro
9	67,03%	32,97%
Info Gain e *Gain Ratio*	70,4%	29,6%
PCA	73,52%	26,48%

Fonte: Autoria própria

acerto de modo geral com as variáveis obtidas por PCA onde acertou 62% das instâncias com câncer de esôfago.

No segundo tipo de experimento com agrupamento de dados, foi utilizado conjunto de treinamento na Base de Dados 1 ao invés de avaliar os *clusters* em relação a classe como foi feito para Base 2.

Na Tabela 20 que demonstra o resultado do algoritmo *k-means*, tem-se que o *cluster* 0 ficou composto por adultos do sexo masculino que possuem histórico de consumo de bebida alcoólica e de tabaco, contendo 9.077 instâncias. Já o *cluster* 1 foi constituído por idosos do

Tabela 20 – Resultados do algoritmo *k-means* na Base de Dados 1 com variáveis selecionadas por PCA

	Cluster	
Variável	0	1
sexo	masculino	feminino
idade	adulto	idoso
tabagismo	sim	nunca
alcoolismo	sim	nunca
Instâncias	9077	2909

Fonte: Autoria própria

sexo feminino que não possuem histórico de consumo de álcool e de tabaco, 2.909 instâncias. Resultados estes obtidos com duas iterações e soma dos erros quadrados igual a 13907,0.

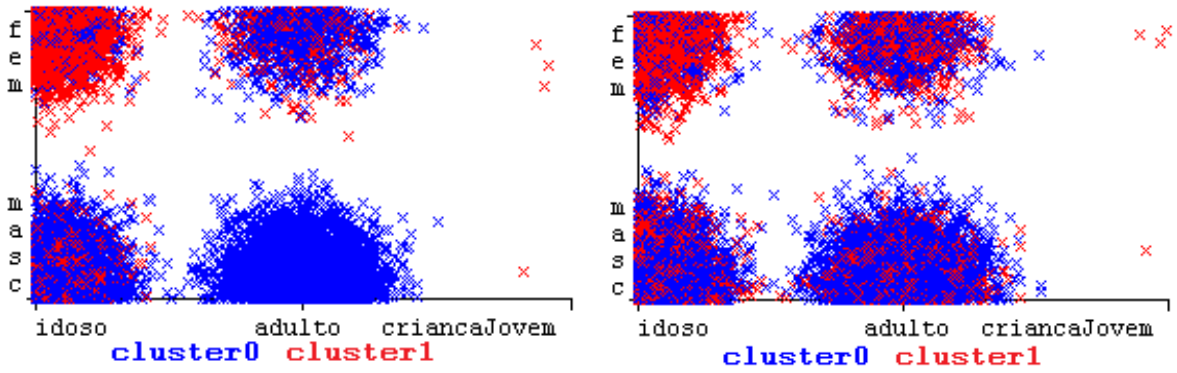
O EM realizou 39 iterações e obteve *log* de verossimilhança igual a -3.09564. O *cluster* 0 ficou composto por 7.716 instâncias e o *cluster* 1 por 4.270 instâncias. A Figura 20 apresenta o resultado do agrupamento, onde entre parênteses tem-se a probabilidade dos *clusters* e para cada variável a contagem de frequência dos valores.

Attribute	Cluster 0 (0.67)	1 (0.33)
sexo		
masculino	7058.2589	2069.7411
feminino	1012.4602	1849.5398
[total]	8070.7191	3919.2809
idade		
idoso	3815.6183	2701.3817
adulto	4255.0876	1213.9124
criancaJovem	1.0132	4.9868
[total]	8071.7191	3920.2809
alcoolis		
sim	5168.4438	139.5562
exConsumidor	2408.2575	50.7425
nunca	477.3116	3686.6884
naoAvaliado	13.3042	24.6958
naoSeAplica	6.4019	20.5981
[total]	8073.7191	3922.2809
tabagism		
nunca	470.6362	2399.3638
sim	5776.1365	1155.8635
exConsumidor	1814.3283	339.6717
naoAvaliado	3.6265	14.3735
naoSeAplica	8.9916	13.0084
[total]	8073.7191	3922.2809

Figura 20 – Resultado do agrupamento na base de dados 1 com o algoritmo EM.

Fonte: Autoria própria

Na Figura 21 podem ser visualizados a separação dos *clusters* em relação as variáveis idade e sexo, onde em (a) tem-se o agrupamento realizado pelo algoritmo *k-means* e em (b) pelo algoritmo EM.

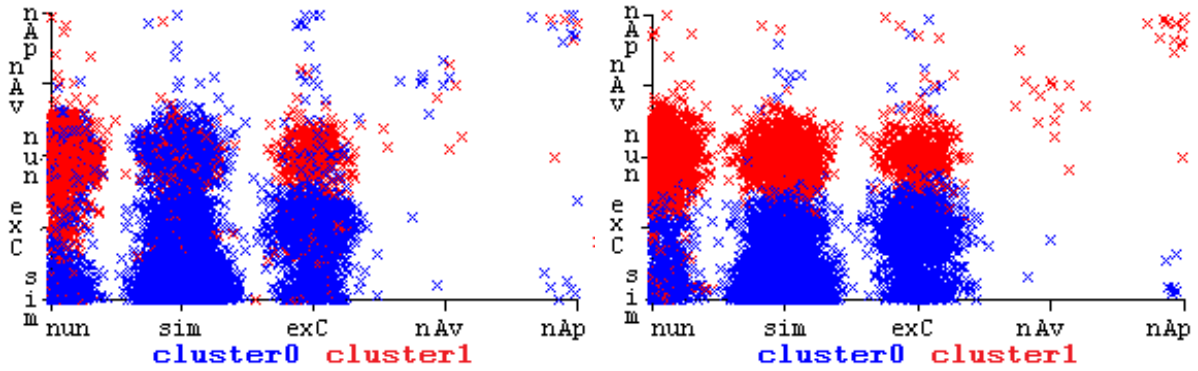


(a) (b)

Figura 21 – Visualização dos agrupamentos considerando idade X sexo: (a) *k-means*, (b) EM.

Fonte: Autoria própria

Na Figura 22 podem ser visualizados a separação dos *clusters* em relação as variáveis tabagismo e alcoolismo, onde em (a) tem-se o agrupamento realizado pelo algoritmo *k-means* e em (b) pelo algoritmo EM.



(a) (b)

Figura 22 – Visualização dos agrupamentos considerando tabagismo X alcoolismo, onde nAp = não se aplica, nAv = não avaliado, nun = nunca, exC = ex comsumidor: (a) *k-means*, (b) EM.

Fonte: Autoria própria

4.2 DISCUSSÃO

Na etapa de pré-processamento dos dados ocorreu a eliminação de muitas instâncias, fato que pode ter tornado algumas variáveis predominantes. Apesar dos RHCs possuírem uma documentação bem completa com orientações para preenchimento das fichas dos pacientes e dos dados passarem por um tratamento antes de sua consolidação na plataforma do IRHC, ainda percebe-se uma falta de padronização e deficiência no preenchimento das fichas do pacientes, que resultou em uma quantidade elevada de variáveis com valores como ”sem informação”.

A mineração de regras de associação demonstrou-se influenciada pelos valores predominantes da maioria das variáveis, o que pode atrapalhar na mineração de regras úteis. Porém é um método capaz de proporcionar conhecimento relacionado ao perfil dos pacientes quando avaliado por um profissional.

Foi possível perceber, por meio das regras mineradas, que grande parte dos casos de câncer de esôfago são diagnosticados em outras unidades hospitalares, mas todo o tratamento é realizado nos RHCs. A alta relação entre as instâncias de sexo masculino com o histórico

de consumo de álcool e tabaco. A ocorrência de mais de um tumor primário quando o sexo é masculino. O tipo histológico do tumor primário com o consumo de álcool e tabaco. O trabalho correlato de Minelli (2013), realizou a mineração de regras de associação em uma base de câncer de pulmão utilizando valores mínimos de suporte e confiança de 10% o que resultou em um elevado número de regras. O trabalho de Preissler (2016) utilizou suporte e confiança mínimos de 20% e 80% para minerar regras relacionadas ao câncer de estômago. Neste trabalho foram utilizados valores de suporte e confiança mínimos de forma que fossem geradas no máximo 20 regras, de modo a analisar somente as regras com o maior grau de confiança a partir do valor mínimo estabelecido.

Nos métodos de classificação J48 e *reptree* apresentaram bons resultados com o filtro *resample*, mantendo aproximadamente 71% de acerto mesmo quando foram utilizadas apenas as variáveis obtidas com os métodos de seleção de variáveis. Ao aumentar o número mínimo de instâncias necessários para tornar uma regra válida, o número de regras obtidas diminuiu, mas o percentual de acerto se manteve mesmo quando alterou-se o percentual *split*.

No trabalho correlato de Minelli (2013) foi utilizada árvore de decisão para prognóstico de pacientes baseado no tempo de sobrevida do paciente, onde se mostrou uma técnica inviável devido desbalanceamento das variáveis. Neste trabalho foi utilizada árvore de decisão para classificar as instâncias de acordo com a localização primária do tumor, foi aplicado o filtro de instâncias *resample* para balanceamento das classes e realizado um comparativo dos resultados obtidos com variáveis selecionadas de forma empírica e entre variáveis selecionadas com métodos de seleção de variáveis.

Para algoritmos de agrupamento em sua utilização na Base de Dados 2, o algoritmo *k-means* não obteve bons resultados quando utilizado com o filtro *resample* e quando avaliados os resultados apenas para a classe referente ao câncer de esôfago o desempenho foi pior ao se utilizar variáveis selecionados pelos métodos de seleção de variáveis.

Quando aplicados os algoritmos de agrupamento na Base de Dados 2, sem a utilização do filtro *resample*, com ambos obteve-se um bom percentual de acerto ao agrupar as instâncias em relação a classe tanto com as variáveis selecionadas de forma empírica quanto pelos métodos de seleção de variáveis.

O algoritmo EM foi o que se mostrou mais estável com e sem a utilização do filtro *resample*, pois também conseguiu manter um percentual de acerto semelhante entre as instâncias referentes a câncer de esôfago. Os trabalhos correlatos de Preissler (2016) e Viana (2018) que utilizaram métodos de agrupamento de dados, não realizaram agrupamento no modo de avaliar *clusters* em relação a classe como realizado neste trabalho.

No agrupamento com a base de dados 1, percebe-se que utilizando variáveis obtidas

com método de seleção de variáveis PCA, os *cluters* ficaram bem definidos assim como nos trabalhos correlatos de Preissler (2016) e Viana (2018) que utilizaram variáveis selecionadas de forma empírica.

É possível observar que não é completamente correto afirmar que o histórico de consumo de álcool e tabaco possui total ligação com a ocorrência de câncer de esôfago, pois percebe-se que em mulheres essa relação não é tão forte quanto é para homens. Em mulheres observa-se que a ocorrência do câncer de esôfago está mais relacionada a idade do que para homens, dado que em mulheres a doença predomina em idosas e em homens na faixa etária adulto. Grande parte dos homens que relataram consumir ou serem ex-consumidores de tabaco, também relataram consumir ou serem ex-consumidores de álcool. Ambos os algoritmos, na visualização dos *clusters*, conseguiram diferenciá-los bem, mas o EM demostrou uma melhor separação.

5 CONCLUSÕES

Neste documento foram mostradas aplicações de técnicas de mineração de dados para descoberta de conhecimento relacionados ao câncer de esôfago em uma base de dados do IRHC.

A plataforma do IRHC é de fácil utilização e disponibiliza toda a documentação necessária para entendimento dos valores das variáveis. Apesar da utilização de sistemas para preenchimento de fichas de pacientes, a falta de padronização e erros no preenchimento são problemas que persistem e que refletem na quantidade de instâncias eliminadas da base de dados durante todo o pré-processamento e possivelmente na qualidade dos dados.

A utilização de métodos de seleção de variáveis, para escolha de variáveis a serem utilizadas pelos algoritmos de classificação e agrupamento nesta base de dados de pacientes com câncer de esôfago, se mostrou viável. Obteve resultados satisfatórios, em média aproximadamente 72% de acerto, empregando quantidades de variáveis bem menores do que a quantidade empregada de forma empírica, o que possibilita somente a utilização de variáveis que sejam realmente relevantes.

A aplicação do algoritmo *apriori* para mineração de regras de associação, foi possível e se mostrou viável. Encontrou alguns padrões já conhecidos, como a relação do consumo de tabaco com o tipo histológico carcinoma epidermoide escamoso. Um padrão interessante encontrado foi que pacientes que realizaram o tratamento foram considerados como casos analíticos, a casos deste tipo são destinados maiores recursos devido ao estadiamento e estado de saúde do paciente. Para identificação de outros padrões úteis seria interessante uma análise minuciosa de um profissional da saúde.

Os métodos de classificação aplicados na base de dados com todos os casos de câncer, conseguiram bons percentuais de acerto, aproximadamente 72%, ao tentar classificar corretamente as instâncias para as classes LOCTUDET igual a ”sim” e ”não”, mantendo o percentual de acerto quase inalterado mesmo quando alterado o número de instâncias mínimas para a regra ser válida e alterado o percentual de treinamento.

Os métodos de agrupamento de dados, se mostram relevantes nos dois modos de análises realizados. No primeiro foi realizada a avaliação dos *clusters* em relação a classe, com a base de dados contendo todos os casos de câncer onde a variável LOCTUDET igual ”sim” para câncer de esôfago e ”não” para o restante dos casos, em que o algoritmo EM conseguiu melhor acerto ao agrupar as instâncias de cada classe do que o algoritmo *k-means*.

No segundo modo, usando conjunto de treinamento com a base de dados contendo apenas registros de câncer de esôfago, para definição de grupos mais propensos a desenvolverem a doença, o algoritmo EM também demonstrou conseguir definir melhor os *clusters* principalmente na visualização com as variáveis tabagismo e alcoolismo. Uma importante contribuição foi a descoberta da diferença da incidência do câncer de esôfago entre homens e mulheres, no que diz respeito às variáveis selecionadas pelo método de seleção de variáveis PCA.

Portanto as técnicas de mineração de dados aplicadas a base de dados de câncer de esôfago se mostrou viável e capaz de gerar conhecimento relacionado a doença em estudo.

Como sugestões para trabalhos futuros tem-se que na mineração de regras de associação seria interessante uma análise com valores de suporte e confiança mínimos mais baixos para tentar descobrir possíveis padrões com valores de variáveis menos frequentes.

Também podem ser aplicados métodos de seleção de variáveis para variáveis posteriores ao diagnóstico, e a aplicação das técnicas de mineração de dados para descoberta de conhecimento em relação ao andamento e tratamento da doença.

REFERÊNCIAS

AGGARWAL, C. C. **Data mining: the textbook**. Londres: Springer, 2015.

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases, sigmod conference. In: ____. [S.l.: s.n.], 1993. v. 22, p. 207–216.

BHARATI, S.; RAHMAN, M.; PODDER, P. Breast cancer prediction applying different classification algorithm with comparative analysis using weka. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2019. p. 581–584. ISBN 9781538682791.

BONINI, J. A. Aplicação de algoritmos de Árvore de decisão sobre uma base de dados de câncer de mama. **Communications and Innovations Gazzete**, v. 1, p. 57–67, 2016.

BRAMER, M. **Principles of data mining**. Londres: Springer-Verlag London, 2016.

CAMILO, C. O.; SILVA, J. C. da. Mineração de dados: conceitos, tarefas, métodos e ferramentas. **Relatório Técnico**, p. 29, 2009. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.>

CARVALHO, L. A. V. de. **Data mining: a mineração de dados no marketing, medicina, economia, engenharia e administração**. São Paulo: Editora Érica, 2002.

CATTRAL, R.; OPPACHER, F.; DEUGO, D. Supervised and unsupervised data mining with an evolutionary algorithm. In: . [S.l.: s.n.], 2001. p. 767 – 774 vol. 2. ISBN 0-7803-6657-3.

CIOS, K. J.; PEDRYCZ, W.; SWINIARSKI, R. W.; KURGAN, L. A. **Data mining a knowledge discovery approach**. USA: Springer Science+Business Media, 2007.

DATASUS. **Morfologia de neoplasias**. 2007. Disponível em: <http://www.datasus.gov.br/cid10/V2008/WebHelp/morfoneo.htm>. Acesso em: 27/05/2019.

FACELLI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. de Leon Ferreira de. **Inteligência artificial: uma abordagem de apredizado de máquina**. Rio de Janeiro: LTC - Livros Técnicos e Científicos, 2011.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. In: **AI Magazine Volume 17 Number 3**. [S.l.: s.n.], 1996. p. 37–54.

FICHMAN, R.; KOHLI, R.; KRISHNAN, R. The role of information systems in healthcare: Current research and future trends. **Information Systems Research**, INFORMS Inst.for Operations Res.and the Management Sciences, v. 22, n. 3, p. 419–428, 2011. ISSN 10477047.

GCO, Global Cancer Observatory. **Cancer today**. 2018. Disponível em: <https://gco.iarc.fr/today/home>. Acesso em: 01/05/2019.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining um guia prático**. Rio de Janeiro: Elsevier Editora Ltda, 2015.

HAN, J.; KAMBER, M.; PEI, J. **Data mining concepts and techniques**. 225 Wyman Street, Waltham, MA 02451, USA: Morgan Kaufmann Publishers, 2012.

HERLAND, M.; KHOSHGOFTAAR, T.; WALD, R. A review of data mining using big data in health informatics. **Journal of Big Data**, SpringerOpen, v. 1, n. 1, 2014.

INCA, Instituto Nacional de Câncer José Alencar Gomes da Silva. **Integrador RHC**. 2011. Disponível em: <https://www.inca.gov.br/sites/ufu.sti.inca.local/files//media/document//folder-integrador-rhc-2011.pdf>. Acesso em: 17/05/2019.

INCA, Instituto Nacional de Câncer José Alencar Gomes da Silva. **Estimativa 2018: incidência de câncer no Brasil**. Rio de Janeiro: INCA, 2017.

INCA, Instituto Nacional de Câncer José Alencar Gomes da Silva. **Câncer de esôfago**. 2018. Disponível em: <https://www.inca.gov.br/tipos-de-cancer/cancer-de-esofago>. Acesso em: 29/04/2019.

INCA, Instituto Nacional de Câncer José Alencar Gomes da Silva. **Estatísticas de câncer**. 2018. Disponível em: <https://www.inca.gov.br/numeros-de-cancer>. Acesso em: 23/04/2019.

INCA, Instituto Nacional de Câncer José Alencar Gomes da Silva. **Como surge o câncer?** 2019. Disponível em: <https://www.inca.gov.br/como-surge-o-cancer>. Acesso em: 01/05/2019.

INCA, Instituto Nacional de Câncer José Alencar Gomes da Silva. **Integrador RHC**. 2019. Disponível em: <https://irhc.inca.gov.br/RHCNet/visualizaTabNetExterno.action>. Acesso em: 17/05/2019.

INCA, Instituto Nacional de Câncer José Alencar Gomes da Silva. **O que é câncer?** 2019. Disponível em: <https://www.inca.gov.br/o-que-e-cancer>. Acesso em: 22/04/2019.

INCA, Instituto Nacional de Câncer José Alencar Gomes da Silva. **Registros Hospitalares de Câncer**. 2019. Disponível em: <https://www.inca.gov.br/numeros-de-cancer/registros-hospitalares-de-cancer-rhc>. Acesso em: 23/04/2019.

INCA, Instituto Nacional de Câncer José Alencar Gomes da Silva. **Sobre o INCA**. 2019. Disponível em: <https://www.inca.gov.br/institucional>. Acesso em: 23/04/2019.

JOLLIFFE, I. **Principal Component Analysis**. New York: Springer, 2002.

KANTARDZIC, M. **Data Mining: Concepts, Models, Methods, and Algorithms**. New Jersey: John Wiley & Sons, 2003.

KASPER, D. L.; HAUSER, S. L.; JAMESON, J. L.; FAUCI, A. S.; LONGO, D. L.; LOSCALZO, J. **Medicina interna de Harrison**. Porto Alegre: AMGH Editora Ltda, 2017.

KSIAZEK, W.; ABDAR, M.; ACHARYA, U.; PLAWIAK, P. A novel machine learning approach for early detection of hepatocellular carcinoma patients. **Cognitive Systems Research**, Elsevier B.V., v. 54, p. 116–127, 2019. ISSN 13890417.

LAROSE, D. T. **Data mining methods and models**. New Jersey: John Wiley & Sons, 2006.

LAROSE, D. T.; LAROSE, C. D. **Discovering knowledge in Data: an introduction to data mining**. USA: John Wiley & Sons, 2014.

LOUZADA-NETO, F.; DINIZ, C. A. R. **Técnicas Estatísticas Em Data Mining**. [S.l.]: IMCA, 2002.

MARCELINO, V. L. R. F. **Apresentando o TabNet e o TabWin**. 2011. Disponível em: <http://www.saude.sp.gov.br/resources/ses/perfil/gestor/homepage/auditoria/reunioes/tabwin_funcionalidades_acesso_e_uso_da_ferramenta.pdf>. Acesso em: 20/05/2019.

MARKOV, Z.; LAROSE, D. T. **Uncovering Patterns in Web Content, Structure and Usage**. [S.l.]: Wiley, 2007.

MINELLI, L. **Representação de conhecimento relacionado ao prognóstico de pacientes com câncer de pulmão**. Dissertação (Mestrado) — Universidade Federal de Santa Maria, 2013.

Ministério da Saúde. **Câncer: o que é, causas, tipos, sintomas, tratamentos, diagnóstico e prevenção**. 2017. Disponível em: <http://portalms.saude.gov.br/saude-de-a-z/cancer>. Acesso em: 23/04/2019.

PAULO, M. C. M. de; MATHIAS, S. B. B. R. P.; LACERDA, M.; KORTING, T. S.; FONSECA, L. M. G. Comparação dos atributos ecolhidos pelo treinamento de classificadores de árvores de decisão com seleção de atributos por filtro. **Revista Brasileira de Cartografia**, 2012. Disponível em: <http://wiki.dpi.inpe.br/lib/exe/fetch.php?media=wiki:mauriciodepaulo:podas_geodma.pdf>.

POSTGRESQL. **O que é o PostgreSQL?** 2018. Disponível em: <https://www.postgresql.org/about/>. Acesso em: 26/05/2019.

PREISSLER, A. **Data Mining para definição dos perfis de pacientes com câncer de estômago**. Dissertação (Trabalho de Conclusão de Curso) — Universidade Regional do Noroeste do Estado do Rio Grande do Sul – UNIJUI, 2016.

ROSINI, A. M.; PALMISANO, A. **Administração de Sistemas de Informação e a Gestão do Conhecimento**. São Paulo: Cengage Learning, 2011.

SHARMA, A.; DEY, S. Article: Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. **IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications**, ACCTHPCA, n. 3, p. 15–20, 2012.

SNOUSY, M. B. A.; EL-DEEB, H. M.; BADRAN, K.; KHLIL, I. A. A. Suite of decision tree-based classification algorithms on cancer gene expression data. **Egyptian Informatics Journal**, v. 12, n. 2, p. 73 – 82, 2011. ISSN 1110-8665. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1110866511000223>.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. Boston: Pearson Education, 2006.

VERCELLIS, C. **Business Intelligence: data mining and optimization for decision making**. West Sussex: Wiley, 2009.

VIANA, A. **Reconhecimento de Padrões em base de dados de neoplasia maligna utilizando algoritmo de clusterização**. Dissertação (Trabalho de Conclusão de Curso) — Universidade Federal do Sul e do Sudeste do Pará, 2018.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining Practical Machine Learning Tools and Techniques**. Massachusetts: Morgan Kaufmann Publishers, 2011.

WU, X.; ZHU, X.; WU, G.; DING, W. Data mining with big data. **IEEE Transactions on Knowledge and Data Engineering**, v. 26, n. 1, p. 97–107, Jan 2014. ISSN 1041-4347.

ANEXO A – DICIONÁRIO DE DADOS

Tabela 21 – Dicionário das variáveis da base de dados do SisRHC disponível para download no IRHC.

Nº do campo na ficha de cadastro	Variável	Descrição	Domínio
44	ALCOOLIS	Histórico de consumo de bebida alcoólica	1.Nunca; 2.Ex-consumidor; 3.Sim; 4.Não avaliado; 8.Não se aplica; 9.Sem informação
22	ANOPRIDI	Ano do diagnóstico	aaaa
42	ANTRI	Ano da triagem	dd/mm/aaaa
base de dados SP	BASDIAGSP	Base mais importante para o diagnóstico do tumor	1.Exame clínico 2.Recursos auxiliares não microscópicos 3.Confirmação microscópica 4.Sem informação
24	BASMAIMP	Base mais importante para o diagnóstico do tumor	1.Clínica; 2.Pesquisa clínica; 3.Exame por imagem; 4.Marcadores tumorais; 5.Citologia; 6.Histologia da metástase; 7.Histologia do tumor primário; 9.Sem informação

Continua na página seguinte.

Tabela 21–Dicionário das variáveis da base de dados do SisRHC disponível para download no IRHC.

Continuação da página anterior.

Nº do campo na ficha de cadastro	Variável	Descrição	Domínio
47	CLIATEN	Clínicas do primeiro atendimento - entrada do paciente	Codificação segundo Tabela de Clínicas do SisRHC
31	CLITRAT	Clínica de início do tratamento	Codificação segundo Tabela de Clínicas do SisRHC
SIS	CNES	Número do CNES do Hospital	Codificação segundo tabela do Cadastro Nacional de Estab. de Saúde
32	DATAINITRT	Data do início do primeiro tratamento específico para o tumor, no hospital	dd/mm/aaaa
36	DATAOBITO	Data do óbito	dd/mm/aaaa
21	DATAPRICON	Data da 1ª consulta	dd/mm/aaaa
23	DIAGANT	Diagnóstico e tratamento anteriores	1.Sem diag./Sem trat.; 2.Com diag./Sem trat.; 3.Com diag./Com trat.; 4.Outros; 9. Sem informação
22	DTDIAGNO	Data do primeiro diagnóstico	dd/mm/aaaa
32	DTINITRT	Ano do início do primeiro tratamento específico para o tumor, no hospital	aaaa
21	DTPRICON	Ano da 1ª consulta	aaaa
42	DTTRIAGE	Data da triagem	dd/mm/aaaa
28a	ESTADIAG	Estadiamento clínico do tumor (TNM) - Grupo	Codificação do grupamento do estádio clínico segundo classificação TNM

Continua na página seguinte.

Tabela 21–Dicionário das variáveis da base de dados do SisRHC disponível para download no IRHC.

Continuação da página anterior.

Nº do campo na ficha de cadastro	Variável	Descrição	Domínio
28a	ESTADIAM	Estadiamento clínico do tumor (TNM)	Codificação do grupamento do estádio clínico segundo classificação TNM
17	ESTADRES	UF de procedência (residência)	Sigla da UF de procêdencia
41	ESTCONJ	Estado conjugal atual	1.Solteiro; 2.Casado; 3.Viúvo; 4.Separado judicialmente; 5.União consensual; 9.Sem informação
35	ESTDFIMT	Estado da doença ao final do primeiro tratamento no hospital	1.Sem evidência da doença (remissão completa); 2.Remissão parcial; 3.Doença estável; 4.Doença em progressão; 5.Suporte terapêutico oncológico; 6. Óbito; 8. Não se aplica; 9. Sem informação
48	EXDIAG	Exames relevantes para o diagnóstico e planejamento da terapêutica do tumor	1.Exame clínico e patologia clínica; 2.Exames por imagem; 3.Endoscopia e cirurgia exploradora; 4.Anatomia patológica; 5.Marcadores tumorais; 8.Não se aplica; 9. Sem informação

Continua na página seguinte.

Tabela 21–Dicionário das variáveis da base de dados do SisRHC disponível para download no IRHC.

Continuação da página anterior.

Nº do campo na ficha de cadastro	Variável	Descrição	Domínio
43	HISTFAMC	Histórico familiar de câncer	1.Sim; 2.Não; 9.Sem informação
SIS	IDADE	Idade na 1ª consulta (calculada pela diferença entre a data da 1ª consulta e a data do nascimento)	Idade, em anos; valor igual a zero para crianças menores de 1 ano
11	INSTRUC	Escolaridade	1.Nenhuma; 2.Fundamental incompleto; 3.Fundamental completo; 4.Nível médio; 5.Nível superior incompleto; 6.Nível superior completo; 9.Sem informação
50	LATERALI	Lateralidade do tumor	1.Direita; 2. Esquerda; 3.Bilateral; 8.Não se aplica; 9.Sem informação
9	LOCALNAS	Local de nascimento	Sigla da UF de nascimento
25	LOCTUDET	Localização primária (Categoria 3d)	Código da CID-O, 3 dígitos
25	LOCTUPRI	Localização primária detalhada (Subcategoria 4d)	Código da CID-O, 4 dígitos
49	LOCTUPRO	Localização provável do tumor primário (somente para os casos em que a localização primária do tumor é desconhecida)	CID-O, 4 dígitos

Continua na página seguinte.

Tabela 21–Dicionário das variáveis da base de dados do SisRHC disponível para download no IRHC.

Continuação da página anterior.

Nº do campo na ficha de cadastro	Variável	Descrição	Domínio
51	MAISUMTU	Ocorrência de mais um tumor primário	1.Sim; 2.Não; 3.Duvidoso
	MUUH	Município da unidade hospitalar	Tabela de municípios do IBGE
12	OCUPACAO	Ocupação principal	Codificação pela Tabela da Código Brasileiro de Ocupações; mais de três 9 representa Ocupação ignorada
46	ORIENC	Origem do encaminhamento	1.SUS; 2.Não SUS; 3.Veio por conta própria;8.Não se aplica; 9. Sem informação
28b	OUTROESTA	Outros estadiamentos clínicos do tumor	Codificação do grupamento do estádio clínico segundo outras classificações que não a TNM
34	PRITRATH	Primeiro tratamento recebido no hospital	1.Nenhum; 2. Cirurgia; 3.Radioterapia; 4.Quimioterapia; 5.Hormonioterapia; 6.Transplante de medula óssea; 7.Imunoterapia; 8.Outras; 9.Sem informação

Continua na página seguinte.

Tabela 21–Dicionário das variáveis da base de dados do SisRHC disponível para download no IRHC.

Continuação da página anterior.

Nº do campo na ficha de cadastro	Variável	Descrição	Domínio
13	PROCEDEN	Código do Município de procedência (residência)	Tabela de municípios do IBGE
10	RACACOR	Raça/cor	1.Branca; 2.Preta; 3.Amarela; 4.Parda; 5.Indígena; 9.Sem informação
33	RZNTR	Principal razão para a não realização do tratamento antineoplásico no hospital	1.Recusa do tratamento; 2.Tratamento realizado fora; 3.Doença avançada, falta de condições clínicas ou outras doenças associadas; 4.Abandono do tratamento; 5.Complicações de tratamento; 6.Óbito; 7.Outras razões; 8.Não se aplica; 9. Sem informação
6	SEXO	Sexo	1. Masculino; 2. Feminino
45	TABAGISM	Histórico de consumo de tabaco	1.Nunca; 2.Ex-consumidor; 3.Sim; 4.Não avaliado; 8.Não se aplica; 9.Sem informação
26	TIPOHIST	Tipo histológico do tumor primário	Codificação da morfologia do tumor pela CID-O
27	TNM	TNM	Codificação do estádio clínico segundo classificação TNM
38	TPCASO	Tipo de caso	1. Sim (Analítico); 2. Não (Não analítico)

Continua na página seguinte.

Tabela 21–Dicionário das variáveis da base de dados do SisRHC disponível para download no IRHC.

Continuação da página anterior.

Nº do campo na ficha de cadastro	**Variável**	**Descrição**	**Domínio**
SIS	UFUH	UF da unidade hospitalar	Sigla da Unidade da Federação da unidade hospitalar (IBGE)
SIS	VALOR_TOT	Text	-

Fonte: INCA (2019b)