UNIVERSIDADE FEDERAL DO PARANÁ

RUAN CARLOS TORRES

MINERAÇÃO DE DADOS EM BASE DE AVALIAÇÃO NUTRICIONAL

CURITIBA
2017

RUAN CARLOS TORRES

MINERAÇÃO DE DADOS EM BASE DE AVALIAÇÃO NUTRICIONAL

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção de grau de Bacharel no Curso de Gestão da Informação, Departamento de Ciência e Gestão da Informação, do Setor de Ciências Sociais Aplicadas, da Universidade Federal do Paraná.

Orientadora: Prof.ª Dr.ª Denise Fukumi Tsunoda

CURITIBA
2017

TERMO DE APROVAÇÃO

RUAN CARLOS TORRES

MINERAÇÃO DE DADOS EM BASE DE AVALIAÇÃO NUTRICIONAL

Trabalho apresentado como requisito parcial à obtenção do grau de bacharel em Gestão da Informação no curso de graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, pela seguinte banca examinadora:

_____

Prof.ª Dr.ª Denise Fukumi Tsunoda

Orientadora - Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

_____

Prof. Dr. Cícero Aparecido Bezerra

Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

_____

Prof. André José Ribeiro Guimarães

Setor de Ciências Sociais Aplicadas da Universidade Federal, UFPR

Curitiba, 05 de dezembro de 2017

AGRADECIMENTOS

Primeiramente, à minha família, que sempre prestou apoio e suporte em todos os momentos. À minha mãe Janete, meus irmãos Rafael e Roberta, e demais familiares.

Dedico em especial ao meu pai, Eudes Ferme Torres, que infelizmente nos deixou no fim dessa jornada.

À minha orientadora Prof.ª Drª. Denise Tsunoda, pela paciência e total suporte necessário para a conclusão desse trabalho.

Aos demais professores, que nessa jornada de 4 anos repassaram da melhor forma o conhecimento.

À minha namorada, Amanda, que esteve todo momento ao meu lado, incentivando, apoiando e dando o suporte necessário.

E a todos os colegas e amigos que conheci ao longo desse período, que sem dúvida contribuíram para que esse dia chegasse.

“Não tenha medo de desistir do bom para ir para o ótimo”

John Davison Rockefeller

RESUMO

Estudo de caso sobre técnicas de mineração de dados aplicadas à uma base de avaliação nutricional contendo informações coletadas por um profissional de nutrição, o qual utilizou o Procedimento Operacional Padronizado (POP) de triagem e avaliação nutricional do Hospital de Clínicas da Universidade Federal do Paraná como guia. Objetiva a aplicação de técnicas de mineração de dados em uma base de dados de nutrição, a fim de identificar padrões e apresentá-los de forma visual. Constitui-se de um estudo de caso com pesquisa exploratória, finalidade aplicada e abordagem quantitativa. Realiza as tarefas de classificação e associação por meio de métodos Apriori, J48, Decision Table e PART. Demonstra que é possível descobrir padrões dentro de bases de avaliação nutricional, bem como os principais fatores de influência no estado de saúde dos pacientes. Propõe como continuidade da pesquisa a aplicação de técnicas de mineração de dados em outras bases de nutrição, a fim de comparar e validar esta proposta.

Palavras-chave: Nutrição. Gestão da Informação. Descoberta de Conhecimento em Base de Dados. Mineração de Dados. Tomada de Decisão. Business Intelligence.

ABSTRACT

Case study on data mining techniques applied to a nutritional assessment base containing information collected by a nutrition professional, who used the Standardized Operational Procedure (SOP) for screening and nutritional evaluation of the Hospital de Clínicas of Universidade Federal do Paraná as a guide. It is aimed to apply data mining techniques in a nutrition database in order to identify patterns and present them visually. This is a case study with exploratory research, applied purpose and quantitative approach. Classification and association tasks are performed by using Apriori, J48, Decision Table and PART methods. It's shows that it is possible to predict patterns within nutritional assessment bases, as well as the main influential factors on the health status of patients. It is suggested the applicattion data mining techniques in other nutrition bases, in order to compare and validate the proposal through the results obtained.

Keywords: Nutrition. Information management. Knowledge Discovery in Database. Data Mining. Decision Making. Business Intelligence.

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

A era da informação prometeu mudar a forma como se trabalha, compete e se pensa no mercado. De acordo com Marr (2015), a cada segundo nós criamos novas informações. Ainda segundo o autor, 73% das organizações já investiram ou planejaram investir em big data no ano de 2016, e para a surpresa de todos, no momento menos de 0,5% de todos os dados já foram analisados ou utilizados.

Dessa forma, a descoberta de conhecimento (KDD) e a etapa de mineração de dados têm atraído a mídia, pesquisadores e indústria. Segundo Fayaad (1996), existe uma demanda urgente por novas teorias computacionais e ferramentas para auxiliar os usuários a extrair informações dos grandes volumes de dados gerados. Dentro do processo de KDD, a mineração de dados por ser uma etapa que consiste na aplicação de análise de dados e algoritmos de descoberta, os quais dentro das limitações da eficiência computacional, gera um número de padrões sobre os dados.

No século XXI, já existem empresas que preocupadas em empoderar a capacidade de análise de dados dos usuários, são responsáveis por entregar ferramentas que permitam esse tipo de atividade. É o caso da Microsoft, Qlik e Tableau, as quais compõem o quadrante mágico de 2017, disponibilizado pela empresa de consultoria em tecnologia, Gartner.

Portanto, é possível inferir que a utilização de técnicas de mineração de dados em conjunto com as ferramentas de business intelligence possibilita a descoberta de novos conhecimentos e facilita a cognição dos usuários. Logo, a presente pesquisa visa a aplicação de técnicas de mineração de dados em uma base de avaliação nutricional buscando a identificação de padrões para posterior aplicação na área da saúde.

1.1 PROBLEMATIZAÇÃO

Na era do Big Data, o desafio não é coletar dados mas selecionar aqueles corretos e utilizar computadores para aumentar o domínio do conhecimento e identificar padrões que não foram percebidos ou não encontrados previamente (DEAN, 2014, p. 5).

Segundo Marr (2015), o volume de dados está crescendo mais rápido do que nunca, e no ano de 2020, a previsão é de que entre 1,7 megabytes de novas informações sejam criadas a cada segundo por cada habitante do planeta. Além disso, pelo menos um terço de todos estes dados estarão armazenados em nuvem. Conforme Mcafee & Brynjolfsson (2012), esta medida em bytes, é usada para especificar o tamanho da capacidade de armazenamento de um dispositivo, e que demonstra a grandiosidade da escalada da geração transmissão de dados pelas redes informatizadas, já que 2,8 exabytes correspondem a uma quantidade mais de um bilhão de vezes maior daquilo que era a referência na transação de dados de uma década atrás.

Nesse sentido, os protocolos eletrônicos da área de saúde contribuem para disponibilizar um grande volume de dados organizados e estruturados. Na área de Nutrição não é diferente, porém verifica-se um cenário de disponibilidade de um volume considerável de dados e poucas técnicas de mineração de dados aplicadas para o processamento dessas informações em busca de novos conhecimentos.

Esta pesquisa objetiva aplicar técnicas de mineração de dados sobre uma base de dados de acompanhamento nutricional, criada por um nutricionista profissional por meio do monitoramento de pacientes, identificando por meio da aplicação de uma metodologia possíveis padrões de diagnóstico e dieta dos pacientes. A base de dados possui informações que identificam cada um dos pacientes, bem como informações levantadas a partir de avaliação e diagnóstico, tais como: total de dias internado, peso, altura, IMC, sintomas, exame físico, diagnóstico médico e prescrição dentre outros. Entretanto, a identificação dos pacientes não será necessária em nenhuma etapa da análise, garantindo sigilo a esse tipo de informação.

A pesquisa busca aplicar uma metodologia para descoberta de padrões em formulários de avaliação nutricional a partir da aplicação de técnicas de mineração de dados. Dessa forma, a pesquisa tem como foco a seguinte pergunta: como identificar padrões em uma base de dados de avaliação nutricional e apresentá-los de forma visual por meio da aplicação de técnicas de mineração de dados?

1.2 OBJETIVOS

Para responder à questão levantada na pesquisa foram definidos os objetivos a serem alcançados, sendo estes desmembrados em objetivo geral e específicos.

1.2.1 Objetivo Geral

O objetivo geral consiste em aplicar técnicas de mineração de dados em uma base de dados de nutrição, a fim de identificar padrões e apresentá-los de forma visual.

1.2.2 Objetivos Específicos

São definidos os objetivos específicos para que se possa alcançar o objetivo geral previamente estabelecido, sendo eles:

- pesquisar e definir o(s) método(s) de mineração de dados que será(ão) utilizado(s) na base de dados de nutrição;
- classificar os pacientes com base nos padrões de decisões identificados;
- selecionar uma ferramenta de Business Intelligence para apresentação dos resultados encontrados.

1.3 JUSTIFICATIVA

Foi realizado um levantamento em 05 de maio de 2017 na base principal da Web of Science, com o objetivo de verificar as pesquisas existentes na área considerando o acervo bibliométrico. Como parâmetro para a pesquisa, foi o utilizado o termo “Data mining”, pesquisando pelo título e considerando todos os anos e índices. Os resultados trouxeram 11.476 registros. Para cruzar com o foco do trabalho, foi realizado outro levantamento combinando os termos “Data mining” e “Nutrition”, o qual obteve apenas 6 registros.

Os 6 artigos retornados não se assemelham ao trabalho de conclusão proposto. Destes 6, um deles se repete. Logo, ao total são 5 artigos diferentes. No primeiro artigo retornado, os autores buscaram demonstrar e comparar a utilidade dos métodos

de mineração de dados na classificação de um resultado categórico derivado de uma intervenção relacionada à nutrição. No segundo artigo, foram aplicadas as técnicas de mineração de dados RSM (*Response Surface Methodology*) e CHAID (*Chi-Squared Automatic Interaction Detection*) para definição de concentrações ótimas de sal mineral para o desenvolvimento de planta in vitro. Já o terceiro artigo utilizou técnicas de mineração de dados para explorar o número de alimentos para ingestão era necessário para prever com precisão a realização ou não de recomendações dietéticas fundamentais. Para a visualização do resultado, foram construídas árvores de decisão para a realização de recomendações para frutas e legumes, sódio, gordura, gordura saturada e açúcares livres, usando dados de um conjunto nacional de dados de vigilância dietética. O quarto artigo propõe exibir os resultados do método de mineração de dados Apriori, implantado no SINPE (Sistema Integrado de Protocolos Eletrônicos) em uma pesquisa de Nutrição Enteral. Por fim, o quinto e último artigo propõe uma análise de alimentos para a nutrição esportiva.

Logo, é possível inferir que existe um pequeno volume de pesquisa desenvolvida na área, e os que possuem, não existe uma relação exata com o trabalho proposto, o que ressalta a relevância da contribuição do trabalho para a comunidade científica.

A Figura 1 apresenta as dez principais áreas de desenvolvimento de pesquisas em mineração de dados (data mining), em que a área de Ciência da Computação (Computer Science) representa 61% do volume de registros.

Figura 1 - Áreas de pesquisa Data mining

Campo: Áreas de pesquisa	Contagem do registro	% de 11476	Gráfico de barras
COMPUTER SCIENCE	7096	61.833 %	
ENGINEERING	3527	30.734 %	
OPERATIONS RESEARCH MANAGEMENT SCIENCE	677	5.899 %	
TELECOMMUNICATIONS	562	4.897 %	
AUTOMATION CONTROL SYSTEMS	553	4.819 %	
BUSINESS ECONOMICS	521	4.540 %	
MATHEMATICS	457	3.982 %	
BIOCHEMISTRY MOLECULAR BIOLOGY	291	2.536 %	
MEDICAL INFORMATICS	270	2.353 %	
PHARMACOLOGY PHARMACY	258	2.248 %	

FONTE: Web of Science (2017)

A Figura 2 apresenta a área correspondente aos 6 resultados da combinação dos termos “Data mining” e “Nutrition”, em que temos o domínio da Ciência da Computação (Computer Science).

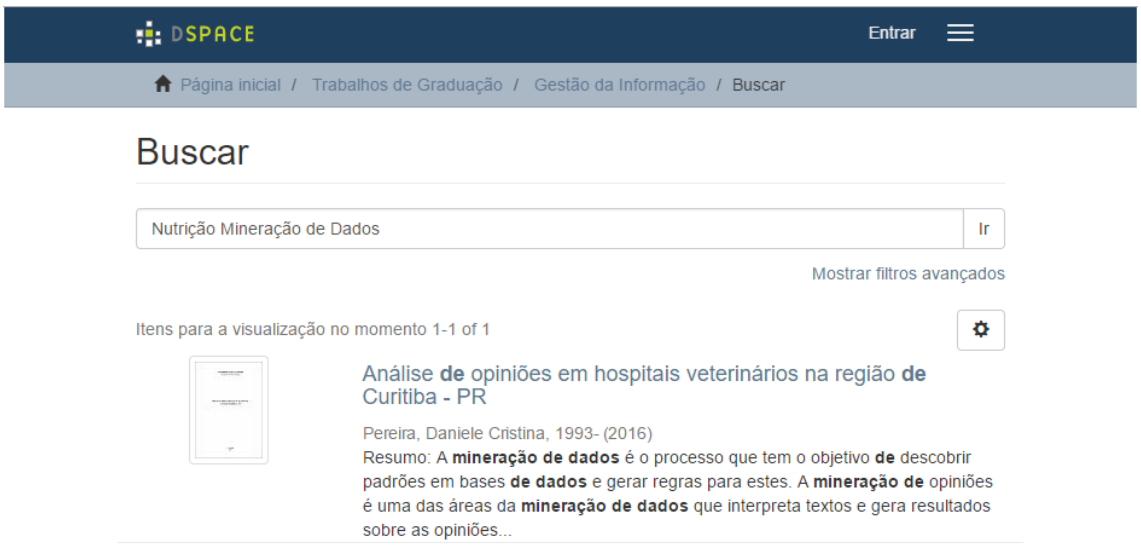
Figura 2 - Áreas de pesquisa - Data mining and Nutrition

Campo: Áreas de pesquisa	Contagem do registro	% de 6	Gráfico de barras
BIOTECHNOLOGY APPLIED MICROBIOLOGY	2	33.333 %	
NUTRITION DIETETICS	2	33.333 %	
PLANT SCIENCES	2	33.333 %	

FONTE: Web of Science (2017).

Outro levantamento realizado foi no repositório de monografias do curso de Gestão da Informação da Universidade Federal do Paraná, com objetivo de validar a contribuição da pesquisa para o curso. Conforme apresentado na Figura 3, após a pesquisa dos termos “nutrição” e “mineração de dados”, a qual retornou apenas um registro, foi possível validar a carência de pesquisas nesta área para o curso de Gestão da Informação.

Figura 3 - Pesquisa em repositório de monografias



FONTE: Dspace (2017).

Além da carência de estudos desenvolvidos na área, outra motivação para o desenvolvimento da pesquisa é o interesse na área de nutrição, que surgiu em decorrência de atividades esportivas e o ânimo em poder contribuir com a área da saúde, e consequentemente com a sociedade. A identificação de padrões possibilita ao profissional de nutrição tomar decisões mais assertivas, o que contribui com mais agilidade na saúde dos pacientes.

2 REVISÃO DA LITERATURA

A seguir é apresentada a fundamentação teórica proposta para nortear a pesquisa de acordo com a problemática investigada e os objetivos traçados. Para isso, foram abordados como temas: gestão da informação, nutrição, recuperação da informação, KDD (*Knowledge Discovery in Databases*), mineração de dados (*data mining*) e visualização da informação.

2.1 GESTÃO DA INFORMAÇÃO

Segundo Davenport (1998), é indiscutível a crescente no uso da tecnologia nas empresas para a potencialização no modo de realizar o trabalho, porém a tecnologia é apenas uma ferramenta para a administração da informação. Ressalta-se a importância do tratamento e gestão das informações, visto que a tecnologia por si só não irá fornecer as informações que os gerentes necessitam para administrar os negócios. Em seu livro “Ecologia da Informação”, o autor argumenta que os problemas informacionais não fazem parte apenas do cotidiano de empresas pequenas, e cita como exemplo a pobreza informacional que a IBM vivia em meados de 1993. O problema persiste até mesmo em empresas pioneiras em sistemas de informações.

Para realizar a gestão da informação, é preciso entender os conceitos que a envolvem. A definição de dado, informação e conhecimento auxilia a no momento de construção de informações relevantes e de valor para o negócio. Os Dados possuem características de serem facilmente estruturados, facilmente obtidos por máquinas, frequentemente quantificados e facilmente transferível. Já a Informação são os dados dotados de relevância e propósito, que requerem unidade de análise, exige o consenso em relação ao significado e exige necessariamente a mediação humana. Ao passo que o Conhecimento é a informação valiosa da mente humana, que inclui reflexão, síntese e contexto, possui difícil estruturação, difícil captura em máquinas, frequentemente tácito (comprova mas não explica) e difícil transferência.

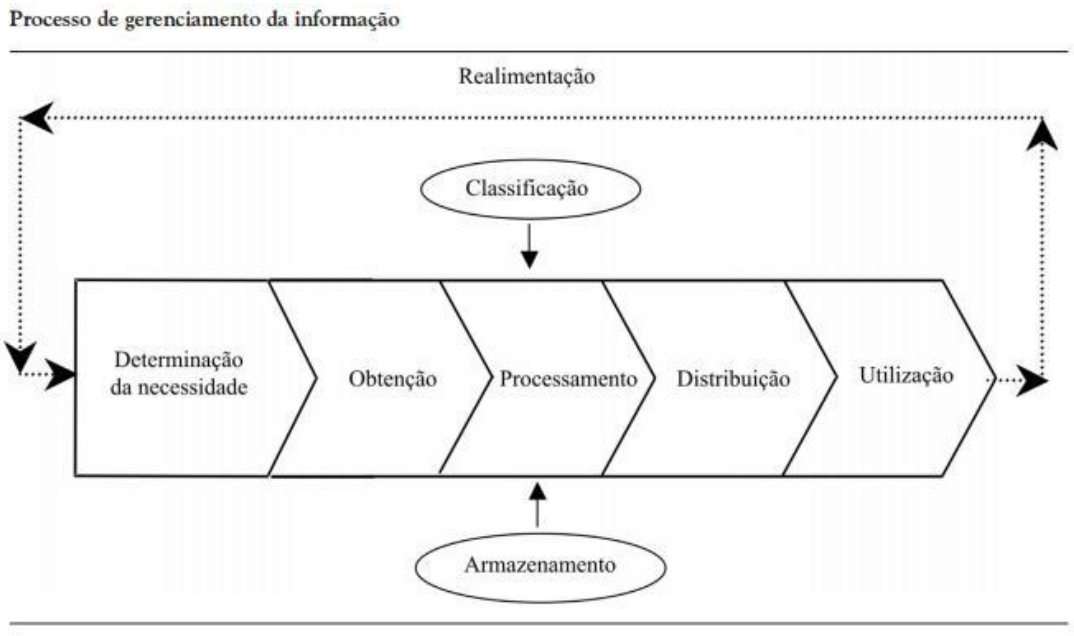
Davenport (1998) ainda cita o processo de gerenciamento da informação, utilizando como base a definição da IBM, que é constituída dos 7 passos seguintes:

1. Administração de exigências. Combinar as necessidades de informação dos usuários e as dos canais fornecedores, conhecer as exigências das pessoas

que necessitam da informação e conquistar a confiança de fornecedores e usuários.

2. Plano de ativos informacionais. Classificar a informação, assim que é obtida, de acordo com sua confidencialidade, como tempo pelo qual ela deve ser mantida, com a maneira como deve ser protegida.
3. Plano de sistemas informacionais (opcional). Planejar o armazenamento e a distribuição dos dados (em meios eletrônicos ou em papel).
4. Aquisição. Obter a informação.
5. Análise. Analisar o conteúdo da informação para estabelecer os níveis de confiabilidade, segurança e qualidade.
6. Disseminação. Distribuir a informação aos que necessitam dela.
7. Feedback. Perguntar aos receptores, por meio de entrevistas, se a informação adequada foi recebida e distribuída de maneira correta, e se foi dado treinamento suficiente quanto ao uso dela.

Figura 4 - Processo de gerenciamento da informação



FONTE: MORAES, 2006, p. 3.

Com base em Marchiori (2002), busca-se a prova de que uma ampla conectividade em que um grande volume de informações está disponível para competências profissionais, individuais e/ ou organizacionais heterogêneas que se encontram em um tempo voltado à informação e ao aprendizado. Alguns pressupostos

que possam justificar, de algum modo, a necessidade de profissionais gestores da informação são:

> • O reconhecimento de que a informação, para ser acessível, deve ser organizada e gerenciada;
> • A percepção de que as necessidades de informação se tornam cada vez mais complexas e dependentes de diferentes e múltiplas fontes – cuja correta avaliação e qualidade é fator crucial para os processos de tomada de decisão;
> • A percepção de que as áreas e os setores econômicos se tornarão dependentes de uma força de trabalho que tenha acesso e possa compartilhar informação. (MARCHIORI, 2002, p.72)

A Gestão da Informação deve estar inserida nos níveis operacional ao estratégico da organização e planejar, organizar, dirigir, distribuir e controlar recursos humanos, tecnológicos, financeiros, materiais e físicos e a partir disso gerar valor para o indivíduo entregando-lhe no momento certo informações certas.

Esses pontos relatados pela autora evidenciam que a informação pode ser um recurso estratégico. Reflete-se que pode, não deve. A informação deve possuir uma série de características que remetem sua qualidade como: confidencialidade, disponibilidade, atualidade, raridade, contextualização, precisão, confiabilidade, originalidade, existência, pertinência e audiência. No pensamento estratégico enxerga-se que alguns atributos da informação, definem ela e como consequência uma possível estratégia gerada a partir dela serão. Segundo De Sordi (2008), com o passar do tempo, as informações têm forte propensão a se desvincularem, a se desatualizarem da realidade que representam, ou seja, se uma estratégia organizacional for baseada em uma informação com esse atributo rompido, a estratégia será invalidada. Em relação à vantagem competitiva sustentável podemos vincular ainda o pensamento de que a dimensão do ineditismo reforça o quão raro é determinada informação, considerando-se a sua existência, seja no ambiente informacional da organização, seja no ambiente externo dela (De SORDI, 2008). Uma série de outros atributos podem definir o quão importante é uma informação na formulação da estratégia, seja ela uma estratégia de posicionamento, de truque, perspectiva, padrão ou pretexto, seguindo a visão porteriana, ou um recurso responsável pela geração de valor e vantagem competitiva segundo a VBR.

2.2 NUTRIÇÃO

De acordo com o Conselho Regional de Nutricionistas (CRN), nutrição é a ciência que estuda as diversas etapas que um alimento sofre, desde a sua introdução no organismo (mastigação) até sua eliminação, também relacionando estes fatores à presença ou não de consequências maléficas ou benéficas. É nessa etapa que ocorre os processos de digestão, absorção, metabolismo e eliminação dos nutrientes. É um ato involuntário e ocorre apenas com a introdução do alimento no sistema digestório. Outro ponto é a necessidade de saber que os alimentos são complexos e precisam ser decompostos em elementos simples para depois serem reconstruídos nas formas que o organismo deseja.

Ainda com base no CRN, o mercado de atuação para os nutricionistas possui diferentes áreas de atuações, como: alimentação coletiva, nutrição clínica, saúde coletiva, docência, indústria de alimentos, nutrição em esportes e marketing na área de alimentação e nutrição. Dentro da nutrição clínica, o nutricionista é responsável por prestar assistência dietética e promover educação nutricional a indivíduos, sadios ou enfermos, em nível hospitalar, ambulatorial, domiciliar e em consultórios de nutrição e dietética, visando a promoção, manutenção e recuperação da saúde.

2.3 RECUPERAÇÃO DA INFORMAÇÃO

A Recuperação de Informação é uma área da Ciência da Computação que lida com armazenamento automático e recuperação de documentos, que são de grande importância devido ao uso universal da linguagem para comunicação (CARDOSO, 2004).

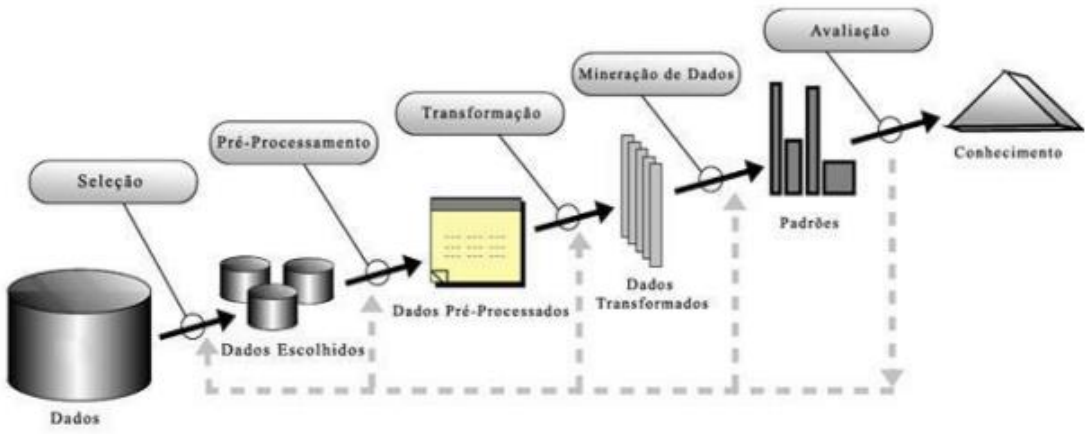
De acordo com Cardoso apud Frakes (2004), a partir do crescimento do volume de publicações, ao longo dos anos, foram desenvolvidas técnicas de recuperação de informação para responder às necessidades dos usuários de bibliotecas, tradicionais ou digitais. A ferramenta mais importante para auxiliar o processo de recuperação é denominada índice, que é uma coleção de termos que indicam o local onde a informação desejada pode ser localizada. Estes termos devem ser organizados de forma a facilitar sua busca.

2.4 KDD (KNOWLEDGE DISCOVERY IN DATABASES)

De acordo com Fayyad (1996), KDD é “o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados”. Este processo é necessário devido ao grande avanço da tecnologia e consequentemente maior volume e utilização de dados, tornando mais complexo o entendimento de tais dados sem que haja um filtro do que é realmente útil.

Fayyad também define o KDD em cinco fases: seleção; pré-processamento; transformação; data mining; e interpretação/avaliação. Estão fases estão representadas na Figura 5.

Figura 5 - Processos de KDD



FONTE: Fayyad, Piatetsky-Shapiro e Smyth (1996, p. 41)

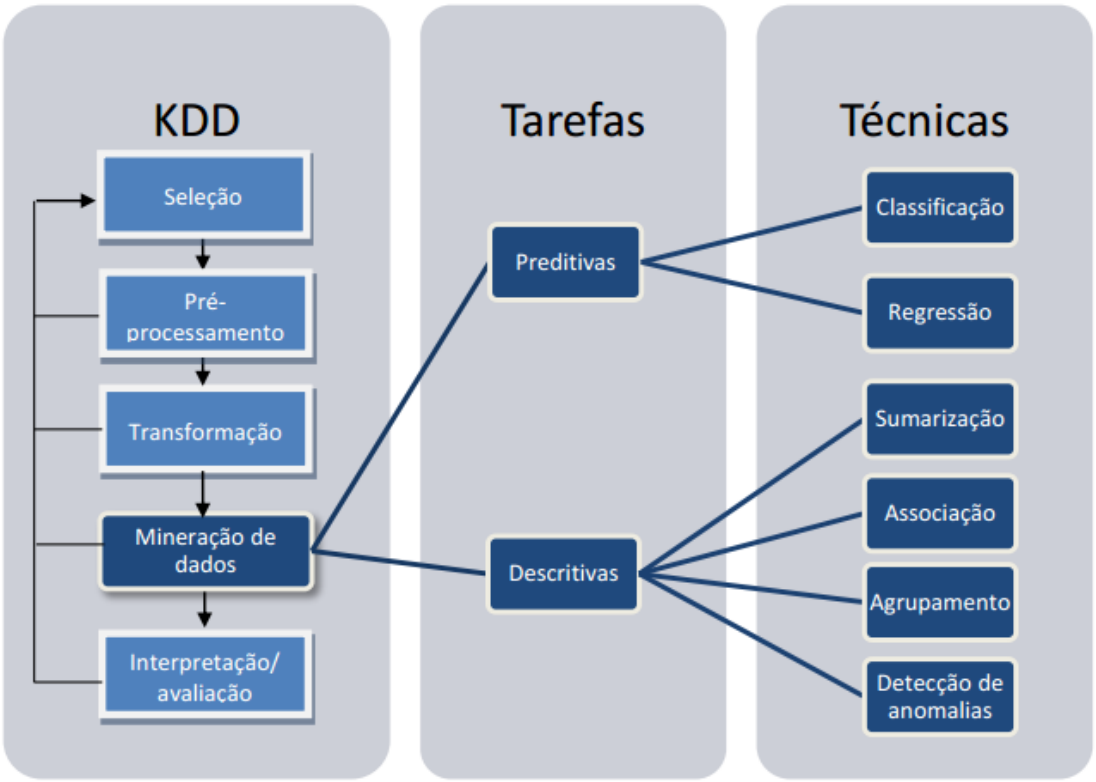
2.5 MINERAÇÃO DE DADOS

A Mineração de dados é definida como o processo de descoberta de padrões em um conjunto de dados. O processo necessita ser automático ou semiautomático. Além disso, os padrões descobertos necessitam possuir um significado que seja interpretado como uma vantagem, geralmente uma vantagem econômica. A prática de minerar dados consiste em solucionar problemas analisando dados já armazenados em uma determinada base de dados (ELSEVIER, 2009).

Lobur et. al. (2008, p. 95) disserta que a mineração de dados é uma etapa dentro do processo KDD, “consistindo na aplicação de análise de dados e algoritmos

de descoberta que, sob limitações aceitáveis de eficiência computacional, produzem um particular número de padrões a partir dos dados", conforme demonstrado na Figura 6.

Figura 6 - A Mineração de Dados como uma importante etapa do processo KDD e principais tarefas e técnicas relacionadas



FONTE: NOGUEIRA, 2015.

2.6 VISUALIZAÇÃO DA INFORMAÇÃO

A Visualização da Informação é imprescindível para a cognição do usuário. Segundo Pereira (2015), visualização da informação consiste no processo de transformar dados em imagens ou representações gráficas com a finalidade de serem interpretadas e/ou apresentadas.

Para Santos (2017), a prática da visualização de dados além de melhorar a tomada de decisão ajuda a aumentar o poder da análise em toda a organização. O poder explicativo e exploratório da visualização de dados permite que os usuários baseiem suas decisões em recursos visuais. Segundo o autor os meios visuais são mais eficazes do que os dados brutos. As ferramentas gráficas evoluíram para

analisar dados, por conseguinte, a detecção de anomalias nos dados, relacionamentos, padrões ou tendências são altamente evidenciadas através de representações gráficas. As ferramentas oferecem aos utilizadores recursos visuais para expor dados e funcionalidades de interação para uma análise rápida e vantajosa. (PEREIRA, 2015).

De acordo com Duarte (2012 p. 7), a visualização da informação permite aos gerentes:

a) explorar o sistema visual para extrair informação dos dados;

b) proporcionar uma visão global do conjunto de dados;

c) identificar estrutura, padrões, tendências, anomalias e relações entre os dados;

d) auxiliar na identificação das áreas de interesse.

Algumas formas de visualização da informação disponíveis para utilização em aplicações são os *scorecards* e *dashboards*. Os *scoreboards* são formas de visualizações focadas em monitorar o progresso dos objetivos planejados de forma estratégica pela alta gerência e são representados em diversos indicadores-chave, que representam o desempenho. Já os *dashboards* são ferramentas de diagnósticos que possuem como objetivo fornecer aos usuários um painel com informações relevantes, que auxiliam na tomada de decisão mais acertiva por parte dos gestores

Após apresentação dos conceitos relacionados ao tema, a próxima seção aborda sobre os encaminhamentos metodológicos para o desenvolvimento do estudo e alcance dos resultados.

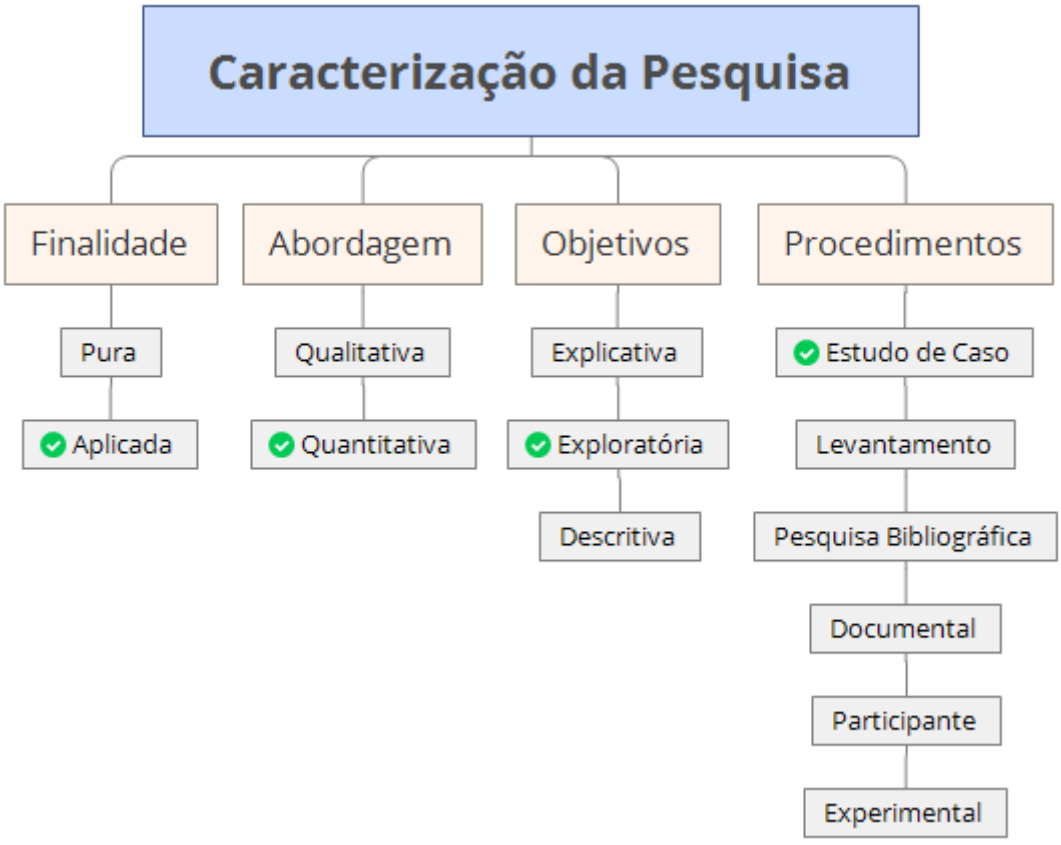
3 ENCAMINHAMENTOS METODOLÓGICOS

Esta seção apresenta a caracterização da pesquisa, descrição da base de dados, ferramentas que serão utilizadas e validação dos resultados.

3.1 CARACTERIZAÇÃO DA PESQUISA

De acordo com Gil (2008), a caracterização da pesquisa é realizada com base em métodos e técnicas sob o ponto de vista de sua finalidade, abordagem, objetivos e procedimentos, conforme demonstra a Figura 7.

Figura 7 - Caracterização da pesquisa



FONTE: O autor (2017).

Quanto ao propósito, a pesquisa se caracteriza exploratória por ter como objetivo a descrição das características de determinada população e o relacionamento de relações entre variáveis.

Em relação à abordagem, a pesquisa se mostra quantitativa, na medida que todos atributos podem ser quantificáveis e busca transformar em números, opiniões e informações com o objetivo de classificá-las e quantificá-las.

Quanto à finalidade, a pesquisa pode ser caracterizada como aplicada, visto que segundo Gil (2008), esse tipo de pesquisa objetiva gerar conhecimentos para aplicação prática em solução de problemas específicos e sua preocupação está menos voltada para o desenvolvimento de teorias de valor universal do que para a aplicação imediata numa realidade circunstancial.

Quanto aos procedimentos e técnicas, a presente pesquisa constitui-se em um estudo de caso considerando que investiga um fenômeno dentro do seu contexto de realidade. Segundo Yin (2005), o estudo de caso pode ser utilizado tanto em pesquisas exploratórias quanto descritivas e explicativas e chama atenção de pesquisadores devido sua capacidade de servir a pesquisas com diferentes propósitos. Ainda de acordo com Yin (2005), para a realização de um estudo de caso são destacadas cinco componentes importantes: as questões de um estudo; suas proposições, caso haja; sua unidade de análise; a lógica que une os dados às proposições; e os critérios para realizar a interpretação das descobertas.

3.2 BASE DE DADOS

A base de dados de avaliação nutricional que será utilizada para desenvolvimento do trabalho contém trezentos e sessenta e seis registros. O arquivo encontra-se em formato xls, sendo possível encontrar alguns atributos, como: número de registro, sexo, idade, data de internação na clínica, data de alta, peso atual, altura, IMC, prescrição kcal, prescrição ptn, sintomas, exame físico entre outras.

A coleta foi realizada pelo profissional de nutrição, o qual utilizou o Procedimento Operacional Padronizado (POP) de triagem e avaliação nutricional do Hospital de Clínicas da Universidade Federal do Paraná como guia. O profissional de nutrição preenchia a ficha física, e posteriormente, realizava a transferência para uma planilha eletrônica.

3.3 PROCEDIMENTOS E FERRAMENTAS

A partir da disponibilização da base de dados, se faz necessária a utilização de ferramentas para cumprir os objetivos destacados no trabalho. Para realização das etapas de pré-processamento e transformação dos dados, as quais compõem o processo de descoberta de conhecimento (KDD), utilizou-se a ferramenta da Microsoft, o Excel, devido a facilidade e familiaridade com o mesmo.

Antes da execução das técnicas de mineração, é importante descrever a base de dados a fim de obter o máximo de informações sobre os dados que compõem a base. Para isso, a ferramenta SPSS, da IBM, será utilizada para aplicação de estatística descritiva. O motivo da escolha do SPSS passa pela afinidade com a ferramenta e pelo seu reconhecimento no mercado em relação à análise estatística.

Com a base preparada, transformada e devidamente descrita de forma detalhada, é o momento de aplicar as técnicas de mineração de dados através do sistema Weka, desenvolvido pela Universidade de Waikato na Nova Zelândia, o qual incorpora os principais algoritmos utilizados na mineração de dados. O Weka foi o sistema utilizado pelo docente para ministrar a matéria de Mineração de Dados, a qual agregou conhecimento na ferramenta, permitindo aplicação dos principais métodos.

Por fim, após o processo de mineração de dados e posterior avaliação, é necessário apresentar as descobertas. Para tanto, será utilizado um software de Business Intelligence. Dentre as opções de ferramenta de BI para desenvolvimento do trabalho proposto são o Qlikview, da empresa Qlik, e o PowerBI da Microsoft. As opções foram expostas com base no quadrante mágico, apresentado no Gráfico 1, disponibilizado pela empresa de consultoria em tecnologia, Gartner. Dessa forma, optou-se pela utilização do Power BI, por ser uma ferramenta que apresenta maior usabilidade em relação às demais.

Gráfico 1 - Quadrante Mágico

FONTE: Gartner, 2017.

Quadro 1 - Ferramentas e procedimentos

(continua)

Ferramenta	Características	Procedimentos
Microsoft Excel	É um editor de planilhas que possui recursos que incluem uma interface intuitiva e capacitadas ferramentas de cálculo e de construção de gráficos.	Pré-processamento e transformação da base de dados

(continuação)

IBM SPSS Statistics	É um software estatístico que fornece uma variedade de técnicas, incluindo análise ad hoc, testes de hipóteses e relatórios, para facilitar o acesso e gerenciamento de dados, selecionar e realizar análises e compartilhar seus resultados.	Estatística descritiva da base de dados
Weka	É uma coleção de algoritmos de aprendizagem de máquinas para tarefas de mineração de dados.	Mineração de dados
QlikView	É uma plataforma de BI orientada ao usuário, que auxilia na tomada de decisões a partir de fontes diversas de conhecimento, dados, pessoas e ambiente.	Visualização da informação
Microsoft Power BI	É um pacote de ferramentas de análise de negócios que oferece insights nas organizações. É possível conectar-se a uma diversidade de fontes de dados, simplifica a preparação dos dados e conduz a análise ad-hoc.	Visualização da informação

FONTE: O autor (2017).

3.4 VALIDAÇÃO DE RESULTADOS

A validação dos resultados é a etapa em que os resultados serão analisados, primeiramente pelo sistema utilizado para a realização da mineração de dados e posteriormente por um especialista da área em questão com o objetivo de identificar se houve alguma descoberta de conhecimento. A validação por meio do sistema é verificada através da taxa de acerto, que indica a porcentagem de acerto das previsões geradas por meio do software. Em um segundo momento, os resultados serão apresentados para um profissional da área de nutrição para verificar se os padrões identificados a partir da aplicação de técnicas de mineração de dados na base

de dados são relevantes e contribuem no processo de tomada de decisão do profissional ou se os padrões identificados já estavam evidenciados a partir de outros tipos, que não mineração, de análise de dados.

Para validação dos resultados será aplicado um questionário com o profissional nutricionista que disponibilizou a base de dados, a fim de não só quantificar os resultados apresentados, mas também obter sugestões de melhorias.

A partir da definição da forma de validação, foi realizada a análise da base de dados e posterior mineração, seguida de suas respectivas visualizações, apresentadas na próxima seção de resultados.

4 RESULTADOS

A seguir é apresentada a análise e descrição estatística da base de dados, a escolha do algoritmo de mineração e os resultados alcançados.

4.1 ANÁLISE DA BASE DE DADOS

A base de dados trabalhada é resultado da coleta de informações de pacientes atendidos por um profissional de nutrição no Hospital de Clínicas. Setenta e duas variáveis compõem a base de dados com 364 registros no total. Embora seja conhecida e respeitada a diversidade de gêneros, para a variável “Sexo”, foi considerado apenas o gênero masculino e feminino no momento da coleta. As variáveis são listadas a seguir:

Quadro 2 - Lista de atributos que compõem a base de dados

(continua)

VARIÁVEL	DESCRIÇÃO
Nome	Nome do paciente.
Registro	Número do prontuário registrado no sistema do hospital.
Sexo	Gênero do paciente. 1=F 2=M.
Idade	Idade do paciente.
HAS	Identifica presença de hipertensão. NÃO=0 SIM=1.
DM	Identifica presença de diabetes. NÃO=0 SIM=1.
Admissão hospitalar	Data em que o paciente chegou ao hospital (pronto atendimento ou outra clínica).
Data internação na clinica	Data em que o paciente chegou na clínica médica (ala de nutrição).
Data alta	Data em que o paciente recebeu alta da clínica médica.
dias internados na clinica	Quantidade de dias internados na clínica.
total de dias internados	Quantidade total de dias internados.
Unidade Internação	Clínica médica feminina ou masculina. 1=CMM 2=CMF.
Status-Paciente	No momento da alta, se foi pra casa ou transferido ou morreu. 1=alta 2=transf 3=óbito.

(continuação)

Diagnóstico médico	Diagnóstico segundo códigos no prontuário
Redução alimentar	Se paciente apresenta redução da quantidade de alimento que habitualmente ingere (redução de 0% ou 25% ou 50% ou 75%). De 0-25-50-75.
Modificação no tipo de dieta	Modificação no tipo de dieta (pastosa ou liquida ou jejum (está sem comer a alguns dias) ou terapia nutricional enteral (via sonda)). 1 =pastosa 2=liquida 3=jejum 4= TNE (inicial).
SGI	Presença de sintomas gastrointestinais (diarreia, obstipação, vômitos, náusea...). 0=não 1= sim.
Diarréia	Presença ou ausência de Diarréia. 0=não 1=sim.
Obstipac	Presença ou ausência de Obstipação
Nauseas	Presença ou ausência de Náuseas
Vomitos	Presença ou ausência de Vômitos
Disfagia	Presença ou ausência de Disfagia (dificuldade em mastigar)
Odinofa	Presença ou ausência de Odinofa (dor ao engolir)
Inapet	Presença ou ausência de Inapet (falta de apetite)
Perda peso	Identifica se o paciente perdeu de peso. 0=não 1=sim.
% perda peso	Porcentagem de peso que perdeu em relação ao peso usual
quantos meses	Em quanto tempo observou-se a perda de peso
Exame físico	Exame físico que identifica a perda aparente de gordura e massa muscular leve, moderada ou grave. 1=leve 2=moderado 3=grave.
Edema	Identifica a presença de inchaço (1- pés, 2- membro inferior, 3- membros inferiores e superiores, 4 - generalizado). 0=não 1=+ 2= ++ 3=+++ 4=++++.
Ascite	Identifica o acúmulo de líquido na barriga (1- pouco, 2- moderado, 3- grave (muito)). 0=não 1=leve 2=mod 3=grave.
Capacidade funcional 1=leve 2=mod 3=grave	Identifica se o paciente se sente fraco. A nutrição está influenciando na sua capacidade física de desempenhar as tarefas do dia a dia (tomar banho, andar, comer, etc.)
Peso usual	Peso mais frequente
Peso seco	Peso descontando o edema e a ascite
Peso Atual (real)	Peso medido (real - na balança, estimado - através de medidas e fórmulas, 3- estimado somente pela circunferência do braço)
Peso cálculo	Peso utilizado no cálculo de necessidades nutricionais. 1=Real 2=estimado 3=est CB.
Peso ideal	Peso ideal segundo altura
Último peso seco	Peso seco registrado no último internamento

(continuação)

Último peso real	Peso medido (real - na balança, estimado - através de medidas e fórmulas, 3- estimado somente pela circunferência do braço)
Altura referida	Altura segundo o paciente
Altura real	Altura medida
AJ	Altura do joelho
Altura estimada AJ	Altura estimada por fórmula utilizando a medida altura do joelho
CH	Chanfradura (medida de um dedo polegar a outro, paciente em posição de cruz)
Altura estimada CH	Altura estimada por fórmula utilizando a medida da chanfradura
Altura utilizada	Altura utilizada nos cálculos de necessidades nutricionais. 1=real 2=AJ 3=CH 4=refe.
IMC	Índice de massa corporal (relaciona peso e altura)
CA	Circunferência abdominal
CB	Circunferência do braço
%CB	Percentual de adequação para idade e sexo da circunferência do braço
CMB	Circunferência muscular do braço
%CMB	Percentual de adequação para idade e sexo da circunferência muscular do braço
PCT	Prega cutânea tricipital
%PCT	Percentual de adequação para idade e sexo da prega cutânea tricipital
PSE	Prega cutânea subescapular
Cpanturilha	Circunferência da panturrilha
Linfócito	Número de linfócitos segundo exames de sangue
Albumina	Proteína albumina segundo exames de sangue
Glicemia JEJUM	Glicemia em jejum segundo exames de sangue
PCR	Proteína Creatina segundo exames de sangue
Diagnóstico nutricional	Diagnostico nutricional (desnutrido, eutrófico, sobrepeso/obesidade, risco nutricional). 1= desn 2=eutr 3=sobre/obesid 4=risco.
Grau de desnutrição	Grau da desnutrição (risco, aguda ou crônica). 1= risco 2=aguda 3=crônica.
Intensidade da desnutrição	Intensidade da desnutrição (leve, moderada, grave). 1=leve 2=mod 3=grave.
nível atendi	Nível de atendimento (primário, secundário ou terciário)
GET	Gasto energético total (quantidade de calorias gastas por dia)
GET kcal/kg	Recomendação nutricional de calorias por peso
PTN	Recomendação nutricional de proteínas por dia

(conclusão)

PTN g/kg	Recomendação nutricional de proteínas por peso
prescrição kcal inicial	Quantidade de calorias calculadas da dieta ofertada
prescição ptna inicial	Quantidade de proteínas calculadas da dieta ofertada
prescrição kcal final	Quantidade de calorias calculadas da dieta ofertada no último dia de internamento
prescrição ptna final	Quantidade de proteínas calculadas da dieta ofertada no último dia de internamento
NEVO	Uso de suplementação via oral. 1=sim 0=não.
TNE	Uso de terapia nutricional enteral. 1=sim 0=não.

FONTE: Elaborado pelo autor (2017).

Para a realização da análise, foi definido como atributo meta a coluna “Status-Paciente”, com o objetivo de identificar padrões que influenciaram em um tratamento eficaz, resultando em alta, se o paciente precisou ser transferido ou se o paciente foi a óbito. Primeiramente, foram desconsideradas as colunas da base de dados que identificariam o paciente, com o objetivo de manter a devida confidencialidade dos dados e as colunas que não apresentariam influência sobre o atributo meta analisado. São elas: “Nome”, “Registro”, “Admissão hospitalar”, “Data internação na clínica”, “Data alta”, “quantos meses”, “%CB”, “%CMB”, “%PCT” e “dias internados na clínica”. Em seguida, foram tratadas as células não preenchidas e preenchidas com o símbolo asterisco, substituindo-as por “ND”, conforme figura abaixo.

Figura 8 - Base de dados sem tratamento de preenchimento

Sexo	Idade	Idade_DISC	Reduçã	% perd	quanto	Exame	Edema	Ascite	Capaci	Peso us	Peso us	Peso se
F	95	IDOSO	0	*	*	0	2	NÃO	0	85	85	*
F	93	IDOSO	75	*	*	3	0	NÃO	0	*	*	*
M	91	IDOSO	50	0	0	3	0	NÃO	0	57	57	*
M	90	IDOSO	*	19,5	2	2	0	NÃO	0	77	77	*
M	89	IDOSO	0	7,46	1	1	0	NÃO	0	67	67	*
F	89	IDOSO	25	6,4	1	1	0	NÃO	3	62	62	*
F	87	IDOSO	75	*	6	3	0	NÃO	3	*	*	*
M	87	IDOSO	0	0	0	0	0	NÃO	2	75	75	*
F	86	IDOSO	75	5,8	1	0	0	NÃO	3	85	85	*
M	85	IDOSO	50	28,4	12	2	0	NÃO	2	90	90	*
F	85	IDOSO	0	5,4	3	0	1	NÃO	1	62	62	53
F	85	IDOSO	0	22,8	12	2	0	NÃO	0	49	49	*
F	85	IDOSO	75	6,67	2	0	0	NÃO	0	60	60	*
F	83	IDOSO	0	0	0	0	0	NÃO	0	61	61	*
F	83	IDOSO	50	5	4	2	0	NÃO	0	60	60	*
M	83	IDOSO	0	0	0	1	0	NÃO	0	61	61	*
F	82	IDOSO	0	0	0	0	0	NÃO	0	72	72	*
M	82	IDOSO	75	0	0	2	0	NÃO	2	55	55	*
M	82	IDOSO	0	0	0	0	3	NÃO	1	98	98	111,1
F	81	IDOSO	75	*	12	3	0	NÃO	3	38	38	*
F	81	IDOSO	0	0	0	2	0	NÃO	0	51	51	*
F	81	IDOSO	50	*	*	1	2	NÃO	0	69	69	*
F	81	IDOSO	50	*	5	1	0	NÃO	0	62	62	*
M	81	IDOSO	50	0	0	1	0	NÃO	0	75	75	*

FONTE: Elaborado pelo autor (2017).

Figura 9 - Base de dados após tratamento de preenchimento

Sexo	Idade	Idade_DISC	Reduçã	% perd	quanto	Exame	Edema	Ascite	Capaci	Peso us	Peso us	Peso se
F	95	IDOSO	0	ND	ND	0	2	NÃO	0	85	85	ND
F	93	IDOSO	75	ND	ND	3	0	NÃO	0	ND	ND	ND
M	91	IDOSO	50	0	0	3	0	NÃO	0	57	57	ND
M	90	IDOSO	ND	19,5	2	2	0	NÃO	0	77	77	ND
M	89	IDOSO	0	7,46	1	1	0	NÃO	0	67	67	ND
F	89	IDOSO	25	6,4	1	1	0	NÃO	3	62	62	ND
F	87	IDOSO	75	ND	6	3	0	NÃO	3	ND	ND	ND
M	87	IDOSO	0	0	0	0	0	NÃO	2	75	75	ND
F	86	IDOSO	75	5,8	1	0	0	NÃO	3	85	85	ND
M	85	IDOSO	50	28,4	12	2	0	NÃO	2	90	90	ND
F	85	IDOSO	0	5,4	3	0	1	NÃO	1	62	62	53
F	85	IDOSO	0	22,8	12	2	0	NÃO	0	49	49	ND
F	85	IDOSO	75	6,67	2	0	0	NÃO	0	60	60	ND
F	83	IDOSO	0	0	0	0	0	NÃO	0	61	61	ND
F	83	IDOSO	50	5	4	2	0	NÃO	0	60	60	ND
M	83	IDOSO	0	0	0	1	0	NÃO	0	61	61	ND
F	82	IDOSO	0	0	0	0	0	NÃO	0	72	72	ND
M	82	IDOSO	75	0	0	2	0	NÃO	2	55	55	ND
M	82	IDOSO	0	0	0	0	3	NÃO	1	98	98	111,1
F	81	IDOSO	75	ND	12	3	0	NÃO	3	38	38	ND
F	81	IDOSO	0	0	0	2	0	NÃO	0	51	51	ND
F	81	IDOSO	50	ND	ND	1	2	NÃO	0	69	69	ND
F	81	IDOSO	50	ND	5	1	0	NÃO	0	62	62	ND
M	81	IDOSO	50	0	0	1	0	NÃO	0	75	75	ND

FONTE: Elaborado pelo autor (2017).

Outro ponto de tratamento são as informações inconsistentes encontradas na base, como um registro em que o paciente foi cadastrado com idade 0 (zero). Logo, esse registro foi eliminado para a realização da análise, totalizando 364 registros a serem analisados.

Para os campos “total de dias internados na clínica” e “%perda de peso”, foi aplicado um método de corte com o objetivo de discretizar os dados em 4 (quatro) grupos intitulados “A”, “B”, “C” e “D”, os quais compreendem um determinado intervalo de valores.

$$Maior - Menor = x$$

$$NumGrupos: y$$

$$Z = \frac{x}{y}$$

$$Intervalo\ 1\ (A): Menor + Z = A \qquad \rightarrow I1\ \leq A$$

$$Intervalo\ 2\ (B): A + Z = B \qquad \rightarrow A < I2 \leq B$$

$$Intervalo\ 3\ (C): B + Z = C \qquad \rightarrow B < I3 \leq C$$

$$Intervalo\ 4\ (D): C + Z = D \qquad \rightarrow I4 > C$$

Já para os campos “Linfócito”, “Albumina”, “Glicemia JEJUM”, “PCR”, “GET”, “PTN”, “prescrição kcal inicial”, “prescrição ptna inicial”, “prescrição kcal final”, “prescrição ptna final”, “Peso usual”, “Peso seco”, “Peso Atual”, “Peso ideal”, “Último peso seco”, “Último peso real”, “Altura referida”, “Altura real”, “AJ”, “Altura estimada”, “CH”, “Altura estimada CH”, “CA”, “CB”, “CMB”, “PCT”, e “PSE”, foi aplicado o método de corte para classificar os dados em 3 (três) grupos intitulados “A”, “B” e “C”, conforme fórmula abaixo.

$$Maior - Menor = x$$
$$NumGrupos: y$$
$$Z = \frac{x}{y}$$
$$Intervalo\ 1\ (A): Menor + Z = A \qquad \rightarrow I1\ \leq A$$
$$Intervalo\ 2\ (B): A + Z = B \qquad \rightarrow A < I2 \leq B$$
$$Intervalo\ 3\ (C): B + Z = C \qquad \rightarrow I3 > B$$

Para os campos “CPanturrilha”, “GET kcal/kg” e “PTN g/kg”, com o objetivo de evitar que um intervalo apresente uma quantidade muito maior que os demais, tornando tendenciosa a análise, foi aplicado o método de corte para classificação dos dados em 2 (dois) grupos, buscando equilíbrio entre os intervalos, intitulados “A” e “B”, conforme fórmula abaixo.

$$Maior - Menor = x$$
$$NumGrupos: y$$
$$Z = \frac{x}{y}$$
$$Intervalo\ 1\ (A): Menor + Z = A \qquad \rightarrow I1\ \leq A$$
$$Intervalo\ 2\ (B): A + Z = B \qquad \rightarrow I2 > A$$

A base de dados possui algumas colunas que foram preenchidas com números, porém esses números equivalem a um determinado grupo, o que caracteriza atributos nominais. Nesse sentido, as colunas “HAS”, “DM”, “Unidade Internação”, “Status-Paciente”, “Redução alimentar de 0-25-50-75”, “1 =pastosa 2=liquida 3=jejum 4= TNE (inicial)”, “SGI”, “Diarréia NÃO=não SIM=sim”, “Obstipac”, “Nauseas”, “Vomitos”, “Disfagia”, “Odinofa”, “Inapet”, “Perda peso NÃO=não SIM=sim”, “Exame físico 0=

sem perda 1=leve 2=moderado 3=grave", "Edema 0=não 1=+ 2= ++ 3=+++ 4=++++", "Ascite NÃO=não LEVE=leve MODERADO=mod GRAVE=grave", "Capacidade funcional 1=leve 2=mod 3=grave", "1=Real 2=estimado 3=est CB", "Altura utilizada 1=real 2=AJ 3=CH 4=refe", "1= desn 2=eutr 3=sobre/obesid 4=risco", "1= risco 2=aguda 3=crônica", "1=leve 2=mod 3=grave", "NEVO SIM=sim NÃO=não" e "TNE SIM=sim NÃO=não" foram ajustadas, substituindo o valor pelo respectivo nome, conforme exemplo abaixo.

Figura 10 - Tratamento de categorias nominais

D	E	F	G
HAS NÃO=0	DM NÃO=0	HAS	DM
1	0	SIM	NÃO
1	1	SIM	SIM
0	0	NÃO	NÃO
1	1	SIM	SIM
0	0	NÃO	NÃO
1	1	SIM	SIM
1	0	SIM	NÃO
1	0	SIM	NÃO
0	0	NÃO	NÃO
0	0	NÃO	NÃO
1	0	SIM	NÃO
0	0	NÃO	NÃO
1	1	SIM	SIM
1	1	SIM	SIM
1	1	SIM	SIM
1	1	SIM	SIM
1	1	SIM	SIM
0	0	NÃO	NÃO
1	0	SIM	NÃO
1	0	SIM	NÃO
1	0	SIM	NÃO
1	1	SIM	SIM
1	0	SIM	NÃO
1	0	SIM	NÃO

FONTE: Elaborado pelo autor (2017).

Para as colunas "Idade" e "IMC", foram utilizadas referências na literatura para auxílio na classificação dos dados. Para a classificação da coluna "Idade", foi utilizada

a seguinte tabela de faixa etária disponibilizada pelo IBGE (Instituto Brasileiro de Geografia e Estatística):

Quadro 3 - Classificação da Faixa Etária

CLASSIFICAÇÃO	FAIXA ETÁRIA
Jovem	De 0 a 19 anos
Adulto	De 20 a 58 anos
Idoso	Mais de 58 anos

FONTE: IBGE (2017).

Para a classificação do “IMC”, é necessário verificar a faixa etária do indivíduo. De acordo com a faixa etária, os valores de classificação se alteram. A seguir, encontra-se a classificação de IMC para jovens e adultos, e classificação de IMC para idosos, respectivamente:

Quadro 4 - Classificação de IMC para jovens e adultos

MÍNIMO	CATEGORIA	MÁXIMO
0 >=	Baixo Peso	<= 18,5
18,5 >	Peso Normal	<= 24,9
24,9 >	Sobrepeso	<= 29,9
29,9 >	Obeso I	<= 34,9
34,9 >	Obeso II	<= 40
40 >	Obeso III	

FONTE: BLACKBURN; THORNTON, 1979.

Quadro 5 - Classificação de IMC para Idosos

MÍNIMO	CATEGORIA	MÁXIMO
0 >=	Baixo Peso	<= 22
22 >	Risco de Deficit	<= 24
24 >	Eutrofia	<= 27
27 >	Sobrepeso	

FONTE: BLACKBURN; THORNTON, 1979.

4.2 ANÁLISE ESTATÍSTICA DA BASE

Para identificar a distribuição dos pacientes por status, foi realizada a contagem pela coluna “Status-Paciente”. O resultado está disposto na Tabela 1, classificado em ordem decrescente.

Tabela 1 - Distribuição dos pacientes por Status

Status-Paciente	Quantidade	%
ALTA	325	89,29%
OBITO	31	8,52%
TRANSF	7	1,92%
ND	1	0,27%
total	**364**	**100,00%**

FONTE: Elaborado pelo autor (2017).

A partir dessa análise, é possível identificar certa desproporcionalidade na distribuição dos pacientes por status. Um dos fatores a ser considerado é o IMC (Índice de Massa Corporal), o qual está distribuído conforme Tabela 2, respeitando os cálculos com base na faixa etária de cada paciente.

Tabela 2 - Distribuição dos pacientes por IMC

IMC	Quantidade	%
Sobrepeso	87	23,90%
Peso Normal	83	22,80%
Baixo Peso	81	22,25%
Eutrofia	29	7,97%
Obeso I	28	7,69%
Risco de Deficit	25	6,87%
ND	17	4,67%
Obeso II	8	2,20%
Obeso III	6	1,65%
total	**364**	**100,00%**

FONTE: Elaborado pelo autor (2017).

Outro fator a ser considerado é o número total de dias em que o paciente permaneceu internado no hospital. Como demonstrado na análise da base de dados, foi utilizado um método de corte para dividir esse número em 4 (quatro) categorias

distintas, intituladas A, B, C e D, as quais compreendem os intervalos dispostos na Tabela 3 classificados em ordem decrescente.

Tabela 3 - Distribuição dos pacientes pelo total de dias internados

TotalDiasInternados	Quantidade	%
A <= 7	105	25,85%
B <= 15	91	25,00%
C <=36	85	23,35%
D >36	83	22,80%
total	**364**	**100,00%**

FONTE: Elaborado pelo autor (2017).

Após a análise do total de dias internados, que se mostrou equilibrada, foi verificada a distribuição dos pacientes por TNE, ou seja, pacientes que fazem uso de terapia nutricional enteral (via sonda). O resultado é apresentado na Tabela 4, disposta logo abaixo. Percebe-se que existe uma grande diferença de 320 pacientes na distribuição por uso de Terapia Nutricional Enteral. A grande maioria disposta na base de dados não utiliza TNE.

Tabela 4 - Distribuição dos pacientes por TNE

TNE 1=sim 0=não	Quantidade	%
NÃO	342	94,00%
SIM	22	6,00%
Total	**364**	**100,00%**

FONTE: Elaborado pelo autor (2017).

Outro fator de impacto na análise é a distribuição dos pacientes pela presença ou ausência de SGI (sintomas gastrointestinais), o que interfere diretamente na dieta do paciente. A Tabela 5 apresenta o resultado obtido na distribuição dos pacientes, o qual é possível perceber uma diferença de 54 pacientes com incidência de SGI quando comparados aos que não apresentam tais sintomas.

Tabela 5 - Distribuição dos pacientes pela presença ou ausência de SGI

SGI	Quantidade	%
SIM	209	57,42%
NÃO	155	42,58%
Total	**364**	**100,00%**

FONTE: Elaborado pelo autor (2017).

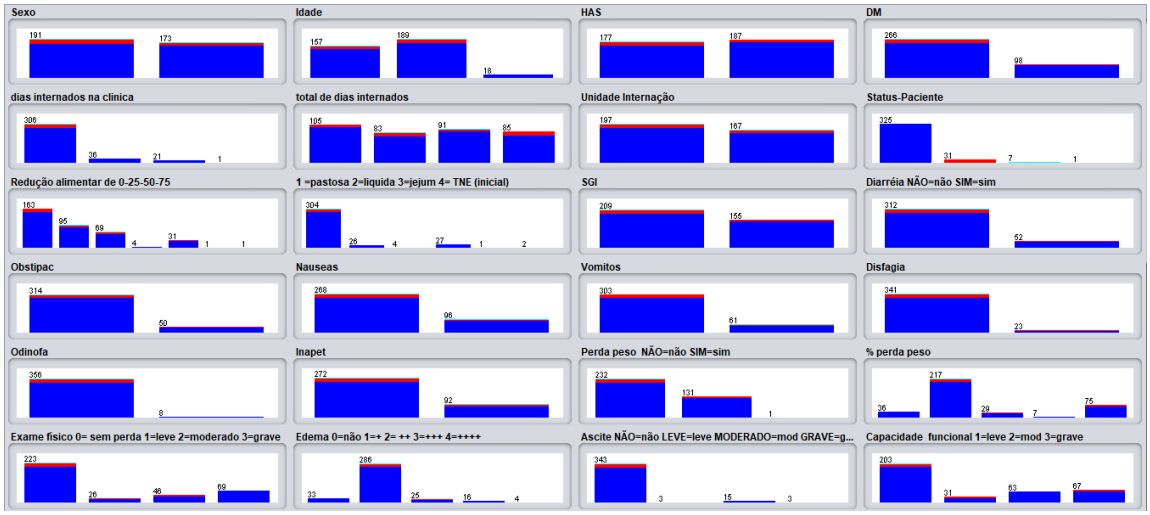
Ao concluir a análise e descrição estatística da base de dados, a próxima subseção aborda o processo de mineração de dados.

4.3 MINERAÇÃO DE DADOS

Após identificar os atributos presentes na base de dados, apresentar a forma de categorização e a distribuição dos principais atributos, é necessário preparar a base de dados para utilização da ferramenta Weka, onde será realizada a etapa de mineração dos dados. Para isso, a planilha é salva no formato .csv. Entretanto, é necessário abrir a planilha no bloco de notas e utilizar a função de localizar e substituir, com o objetivo de substituir todos os “ponto e vírgula” por “vírgula”. Após a realização desse procedimento, é necessário converter o arquivo .csv para .arff (*Attribute-Relation File Format*). Dessa forma, é possível importar a base de dados para o software Weka e iniciar o processo de mineração de dados.

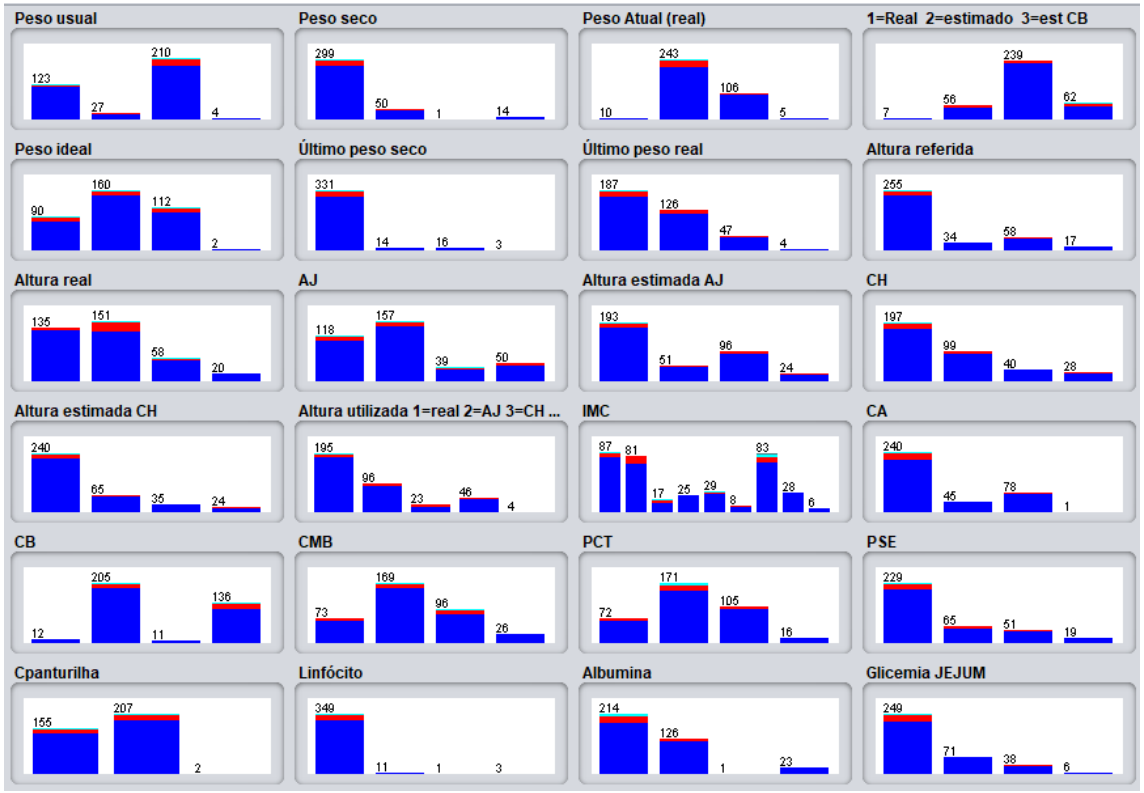
As Figuras 11, 12 e 13 apresentam a visualização gráfica dos atributos importados. As cores são separadas de acordo com a quantidade de atributos correspondentes ao atributo meta “Status-Paciente”. Conforme recorte 1, no atributo “Status-Paciente” podemos identificar as cores e suas respectivas correspondências. O status “Alta” é representado pela cor azul, “Óbito” pela cor vermelha e “Trasnf” pela cor azul claro.

Figura 11 - Histograma valores dos atributos da base de dados (recorte 1)



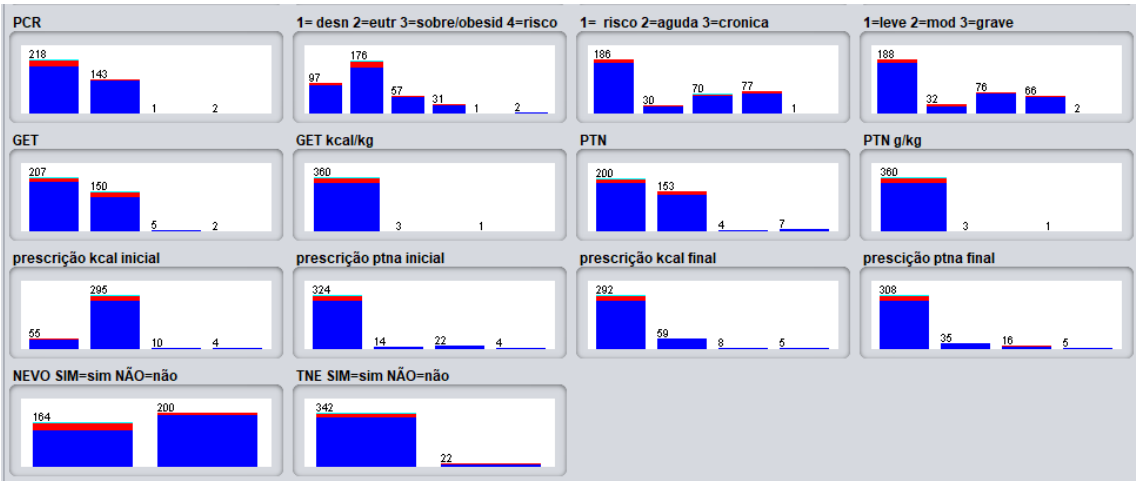
FONTE: Dados da pesquisa utilizando o software Weka (2017)

Figura 12 - Histograma valores dos atributos da base de dados (recorte 2)



FONTE: Dados da pesquisa utilizando o software Weka (2017)

Figura 13 - Histograma valores dos atributos da base de dados (recorte 3)



FONTE: Dados da pesquisa utilizando o software Weka (2017)

Para a realização da análise, é necessário selecionar os métodos que serão utilizados para minerar a base de dados. As heurísticas mais conhecidas são as de classificação e associação. Para as heurísticas de classificação, o software Weka dispõe de: *bayes*, *functions*, *lazy*, *meta*, *misc*, *rules* e *trees*.

Já para as heurísticas de associação, o sistema dispõe de 2 (três) algoritmos ativados: Apriori e FilteredAssociator. Todavia, para o presente trabalho, é utilizado apenas o algoritmo Apriori, por maior afinidade e por ser o mais conhecido na categoria de associação.

Para a interpretação dos resultados, é necessário conhecer alguns conceitos básicos apresentados por Martinez, Casal e Janeiro (2009) logo abaixo.

- Kappa Statistic: índice que compara o valor encontrado nas observações com aquele que se pode esperar do acaso. É o valor calculado dos resultados encontrados nas observações e relatado como um decimal (0 a 1). Quanto menor o valor de Kappa, menor a confiança de observação, o valor 1 implica a correlação perfeita.
- mean absolute error: média da diferença entre os valores atuais e os preditos em todos os casos, é a média do erro da predição.
- true Positives (TP): são os valores classificados verdadeiramente positivos.

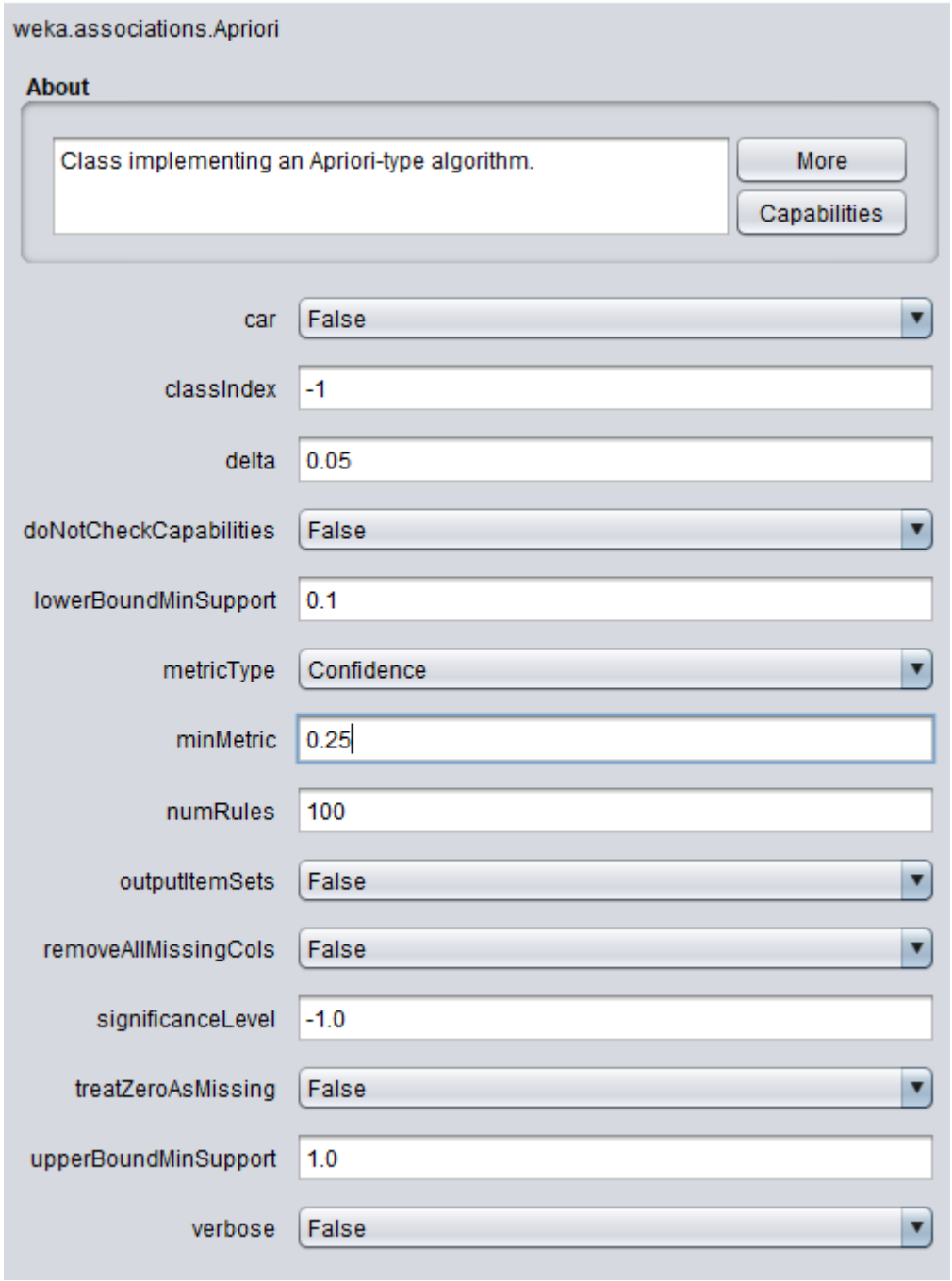
- false Positives (FP): são os falsos positivos, são os dados classificados erroneamente como positivos pelo classificador.
- precision (Precisão): é o valor da predição positiva (número de casos positivos por total de casos cobertos), muito influenciada pela especificidade e pouco pela sensitividade. Sensitividade é o número de casos positivos que são verdadeiramente positivos e especificidade é o número de casos negativos que são verdadeiramente negativos.
- recall (Cobertura): é o valor da cobertura de casos muito influenciada pela sensitividade e pouco pela especificidade. É calculada por número de casos cobertos pelo número total de casos aplicáveis.
- f-measure: usada para medir o desempenho, pois combina valores de cobertura e precisão de uma regra numa única fórmula [2*Prec*Rec / (Prec+Rec)].
- root relative squared error: reduz o quadrado do erro relativo na mesma dimensão da quantidade sendo predita incluindo raiz quadrada. Assim como a raiz quadrada do erro significativo (root mean-squared error), este exagera nos casos em que o erro da predição foi significativamente maior do que o erro significativo.
- relative absolute error: é o erro total absoluto. Em todas as mensurações de erro, valores mais baixos significam maior precisão do modelo, com o valor próximo de zero temos o modelo estatisticamente perfeito.
- root mean-squared error: usado para medir o sucesso de uma predição numérica. Este valor é calculado pela média da raiz quadrada da diferença entre o valor calculado e o valor correto. O root mean-squared error é simplesmente a raiz quadrada do mean-squared-error (dá o valor do erro entre os valores atuais e os valores preditos).

4.3.1 Apriori

De acordo com o sistema Weka, o algoritmo Apriori é definido como “iterativamente reduz o suporte mínimo até encontrar o número necessário de regras com o dado de confiança mínimo. O algoritmo tem uma opção para minerar regras de

associação de classe”. Ainda no software, é apresentado na Figura 14 os parâmetros com seus valores padrão.

Figura 14 - Valores padrão dos parâmetros do algoritmo Apriori



FONTE: Dados da pesquisa utilizando o software Weka (2017)

O software Weka disponibiliza a explicação de cada um dos parâmetros apresentados no sistema:

- verbose - Se ativado o algoritmo será executado no modo detalhado. Parâmetro mantido como default.
- minMetric - pontuação métrica mínima. Considera apenas as regras com pontuações superiores a este valor. Corresponde ao valor mínimo para a métrica selecionada em metricType. Parâmetro mantido como default.
- numRules - número máximo de regras que serão mostradas na tela de resultados. Parâmetro alterado para 100.000.
- lowerBoundMinSupport - um limite inferior para o suporte mínimo. Parâmetro mantido como default.
- classIndex - Índice do atributo de classe. Se for definido como -1, o último atributo é tomado como atributo de classe. Parâmetro mantido como default, pois o último atributo corresponde ao atributo meta motivo_arquivamento.
- outputItemSets - se configurado como “true”, na saída, além de exibir as regras mineradas, exibirá também os itemsets frequentes. Parâmetro mantido como default.
- car - Se as regras de associação de classe são extraídos em vez de regras de associação (geral). Parâmetro mantido como default.
- doNotCheckCapabilities - Se for definido, as capacidades do associador não são verificados antes do associados ser construído. Parâmetro mantido como default.
- removeAllMissingCols - Remover colunas com todos os valores em falta. Parâmetro mantido como default.
- significanceLevel - O nível de significância ou teste de significância (única métrica de confiança). Parâmetro mantido como default.
- treatZeroAsMissing - Se estiver ativado, a zero (isto é, o primeiro valor de um valor nominal) é tratado da mesma forma que um valor em falta. Parâmetro mantido como default.
- delta - reduz o suporte iterativamente por este valor, partindo do limite superior até que o limite inferior seja alcançado. Parâmetro mantido como default.
- metricType – trata-se da especificação da medida de interesse que irá determinar a validade da regra. O conjunto de resultados minerados será

ordenado de acordo com essa medida. É possível escolher a medida: confiança, lift, conviction e leverage. Parâmetro mantido como default.

- upperBoundMinSupport - Limite superior para o apoio mínimo. Comece de forma iterativa diminuindo o apoio mínimo a partir deste valor. Parâmetro mantido como default.

O primeiro experimento foi realizado considerando os valores padrão, exceto o número de regras que foi alterado para “10000000”, de forma que apresente o número máximo de regras encontradas. Todavia, devido ao número excessivo de regras que se encaixam nesses parâmetros, a máquina não foi capaz de processar o algoritmo.

Logo, o segundo experimento foi realizado alterando o padrão de confiança para “0.9” e o número de regras para no máximo “10000”. Foram apresentadas 10000 regras, das quais 2381 relacionadas ao atributo meta “Status-Paciente”.

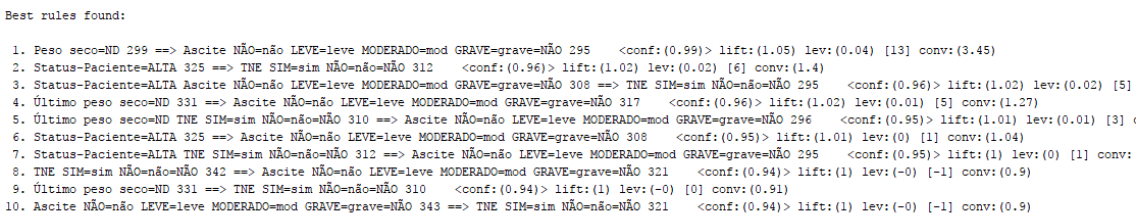
Devido à dificuldade de análise em consequência do alto número de regras encontradas, foi validado com um profissional de nutrição e foi possível excluir alguns atributos da análise, como: “Diarréia 0=não 1=sim”, “Obstipac”, “Nauseas”, “Vomitos”, “Disfagia”, “Odinofa” e “Inapet”. Esses atributos são representados pelo atributo “SGI”, que como mencionado na análise da base de dados, diz respeito a presença ou ausência de sintomas gastrointestinais.

Outro atributo excluído foi o “% perda peso”, já representado pelo atributo “Perda peso”. Foram também excluídos os atributos “Linfócitos”, “Glicemia” e “Albumina”, pela presença de muitos valores divergentes do padrão considerado normal.

Por último, foi excluído o atributo “dias internados na clínica”, visto que o mesmo é representado pelo atributo “total de dias internados”.

O terceiro experimento, foi limitado para as 10 primeiras regras, confiança com limite inferior de “0.9” e suporte com limite inferior de “1”, com o objetivo de facilitar a análise, conforme representado na Figura 15.

Figura 15 - Resultados Apriori Weka



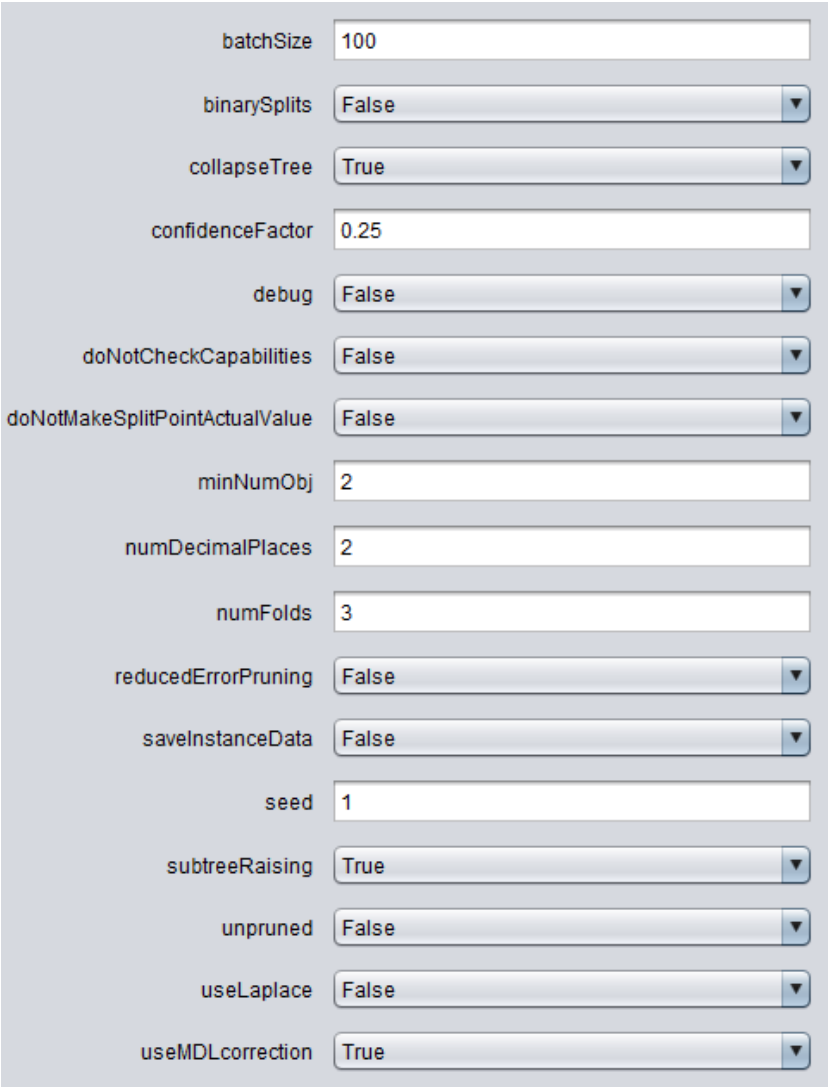
FONTE: Resultados da pesquisa no software Weka

Dos resultados obtidos a partir do algoritmo Apriori, apenas as linhas 2 e 6 possuem relação com o atributo meta. A linha 2 ressalta uma grande relação dos pacientes que obtiveram alta do hospital com a ausência de sonda (Terapia Nutricional Enteral), o que não se configura como uma nova descoberta. Já a linha 6 ressalta que 95% dos pacientes que não foram diagnosticados com Ascite obtiveram alta. Essa não é uma relação obvia para os profissionais da área, todavia representa uma nova descoberta.

4.3.2 J48

O algoritmo J48 foi escolhido na categoria de árvores, o qual é definido pelo sistema Weka como: “Classe para gerar uma árvore de decisão C4.5 podada ou não podada”. Apresenta como os seguintes parâmetros padrão os itens dispostos na Figura 16.

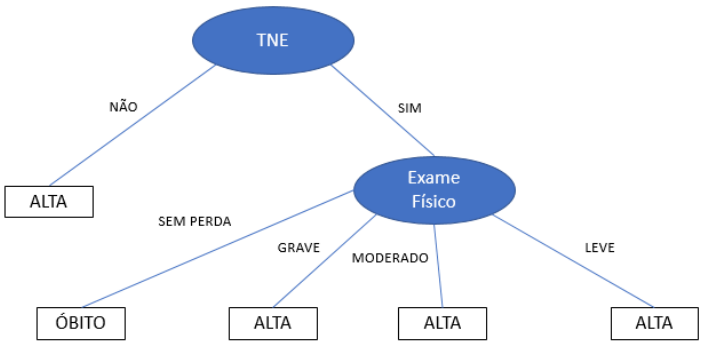
Figura 16 - Parâmetros J48



FONTE: Dados da pesquisa utilizando o software Weka (2017)

Após a execução do experimento 1 nos valores padrões, foi gerada uma árvore com 5 folhas e de tamanho 7, apresentada na Figura 17.

Figura 17 - Árvore adaptada a partir dos resultados do algoritmo J48 (default)



FONTE: Adaptado pelo autor (2017).

Analisando a árvore gerada, é possível identificar que o atributo que corresponde a raiz da árvore, ou seja, possui maior influência, é em relação a Terapia Nutricional Enteral (TNE), que identifica o uso ou não de sonda por parte do paciente. O segundo atributo mais influente é o Exame Físico, o qual identifica a perda aparente de gordura e massa muscular como sem perda, leve, moderada ou grave. Já no experimento, foi alterado o valor mínimo de confiança para 0.90, resultando uma árvore de 57 folhas e de tamanho 77, impossibilitando uma visualização eficiente da árvore.

A Tabela 6 apresenta de forma resumida a comparação entre os experimentos 1 e 2, no quesito de precisão.

Tabela 6 - Comparação dos resultados dos experimentos - J48

Parâmetro	Exp.1	%	Exp.2	%
Correctly Classified Instances	328	90,11%	310	85,16%
Incorrectly Classified Instances	36	9,89%	54	14,84%
Kappa Statistic	0,2219		0,18	
Mean absolute error	0,0877		0,0865	
Root mean squared error	0,2161		0,2690	
Relative absolute error %	86,9959		85,8063	
Root relative squared error %	78,4263		121,6101	
Total Number of Instances	**364**		**364**	

FONTE: Elaborado pelo autor (2017)

A partir dos resultados apresentados na Tabela 6 é possível identificar qual experimento obteve maior sucesso no quesito de precisão. No experimento 1, com o parâmetro de confiança mínimo em 0.25 (padrão), em um total de 364, o algoritmo classificou corretamente 328 e incorretamente 36, o que representa um resultado satisfatório, atingindo 90,11% de taxa acerto. Nesse quesito, o experimento 1 apresentou melhores resultados que o experimento 2, o qual atingiu a taxa de acerto de 85,16%.

O *Kappa Statistic* (Estatística Kappa) é responsável por indicar o grau de concordância. No experimento 1, obteve-se o valor de 0,2219, já no experimento 2, o resultado foi de 0,18. Apesar de aparentar pequena diferença, os experimentos se enquadram em diferentes classificações. O experimento 1 é classificado como grau de concordância razoável, enquanto o experimento 2 indica grau de concordância baixo. Logo, no quesito Estatística Kappa o experimento 1 apresentou melhor resultado.

O *Mean Absolute Error* (Erro Médio Absoluto) foi de 0,0877 no experimento 1 e 0,0865 no experimento 2. Nesse parâmetro, o experimento 2 apresentou leve vantagem quanto ao experimento 1. O valor não representa diferença considerável entre os experimentos.

O *Root Mean Squared Error* (Erro Quadrado Médio) no experimento 1 foi de 0,2161, enquanto o experimento 2 apresentou 0,2690. Considerando esse parâmetro, o experimento 1 apresentou resultados melhores.

O *Relative Absolute Error* (Erro Absoluto Relativo) foi de 86,99% no experimento 1 e 85,80% no experimento 2. Nesse parâmetro, o experimento 2 apresentou uma leve vantagem em relação ao experimento 1. Entretanto, os dois experimentos indicam um erro absoluto relativo elevado.

Por último, o *Root Relative Squared Error* (Raiz do Erro Quadrado Relativo) do experimento 1 foi de 78,42% e no experimento 2 foi de 121,61%. O resultado indica que o experimento 1 apresentou diferença consideravelmente menor em relação ao experimento 2.

As Figuras 18 e 19, respectivamente, apresentam os resultados no quesito de acurácia.

Figura 18 - Acurácia do experimento 1

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,991	0,846	0,907	0,991	0,947	0,288	0,374	0,858	ALTA
0,194	0,009	0,667	0,194	0,300	0,332	0,337	0,186	OBITO
0,000	0,000	0,000	0,000	0,000	0,000	0,130	0,020	TRANSF
0,000	0,000	0,000	0,000	0,000	0,000	0,040	0,003	ND
0,901	0,756	0,867	0,901	0,871	0,285	0,365	0,782	

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

Figura 19 - Acurácia do experimento 2

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,926	0,769	0,909	0,926	0,918	0,169	0,506	0,884	ALTA
0,290	0,063	0,300	0,290	0,295	0,231	0,557	0,161	OBITO
0,000	0,008	0,000	0,000	0,000	-0,013	0,337	0,017	TRANSF
0,000	0,000	0,000	0,000	0,000	0,000	0,475	0,003	ND
0,852	0,692	0,837	0,852	0,844	0,170	0,507	0,804	

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

Novamente, o experimento 1 apresentou melhores resultados no quesito de acurácia. Para o parâmetro *True Positive* (TP Rate), o experimento 1 apresentou valores 6,5% melhores que o experimento 2 no valor mais alto de TP Rate, representado pela classificação de Status Paciente igual a Alta.

As Figuras 20 e 21 apresentam as matrizes confusão obtidas dos experimentos 1 e 2, respectivamente, demonstrando na diagonal principal as instâncias que foram corretamente classificadas.

Figura 20 - Matriz confusão do experimento 1

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 322   3   0   0 |   a = ALTA
  25   6   0   0 |   b = OBITO
   7   0   0   0 |   c = TRANSF
   1   0   0   0 |   d = ND
```

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

Figura 21 - Matriz confusão do experimento 2

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 301  21   3   0 |   a = ALTA
  22   9   0   0 |   b = OBITO
   7   0   0   0 |   c = TRANSF
   1   0   0   0 |   d = ND
```

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

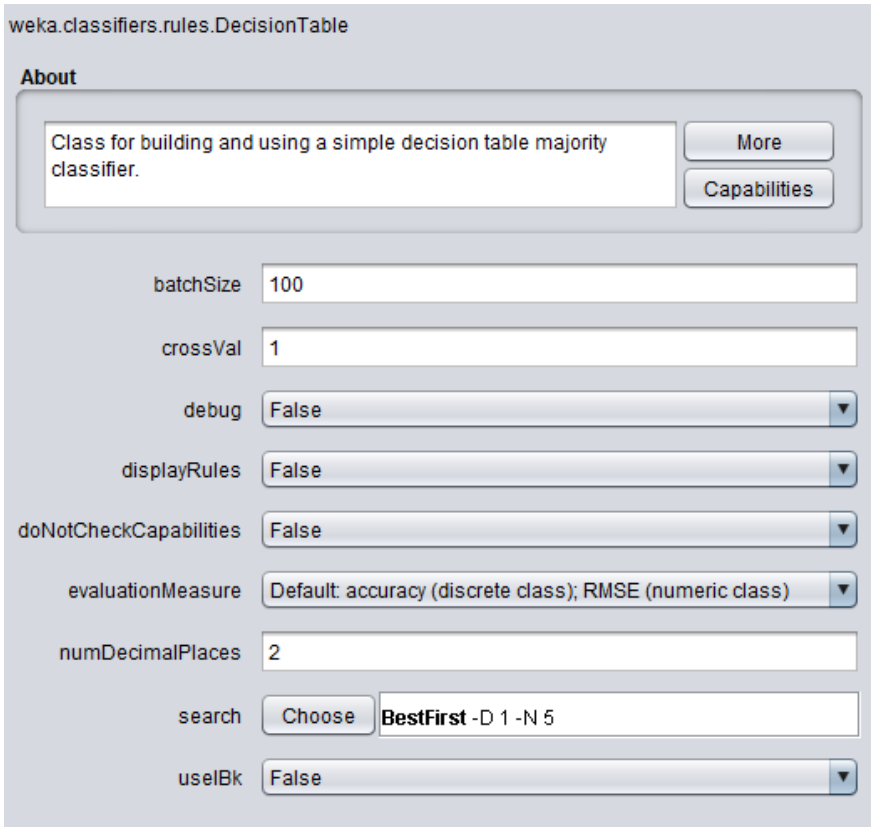
Em ambos os experimentos, a categoria com maior taxa de acerto foi “Alta”. No experimento 1, a taxa de acerto foi de 90,70%, enquanto no experimento 2 foi de 90,93%. Embora na categoria Alta o experimento 2 tenha apresentado melhores resultados, é possível identificar que em relação aos resultados obtidos na matriz confusão, as instâncias foram classificadas corretamente 4,95% a mais no experimento 1 em relação ao experimento 2.

Após concluir a análise dos resultados apresentados na heurística J48, foi realizada a mineração de dados com os demais algoritmos.

4.3.3 *Decision Table*

Para as heurísticas de regras, outro algoritmo escolhido foi o Decision Table. O software Weka define como: “classe para a construção e utilização de uma simples tabela de decisão pela classificação da maioria”. A Figura 22 apresenta os parâmetros padrões disponibilizados pelo Weka.

Figura 22 - Valores padrão dos parâmetros do algoritmo Decision Table

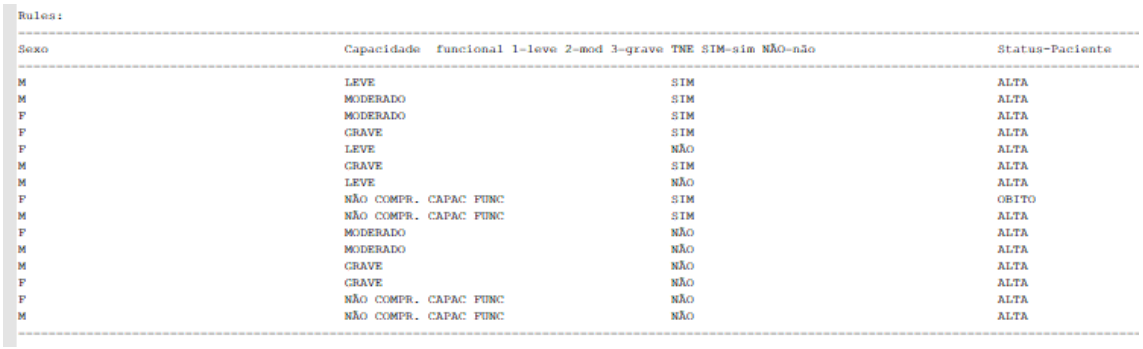


FONTE: Dados da pesquisa utilizando o software Weka (2017)

O experimento 1 foi realizado com os valores de parâmetros padrão, alterando apenas o parâmetro “displayRules” para True, com objetivo de apresentar as regras identificadas pelo algoritmo. O algoritmo demorou 0,15 segundos para execução e identificou os atributos Sexo, Perda peso, Ascite, Altura real e TNE como influenciadores para determinar o status do paciente. Ao total, foram geradas 37 regras dispostas no Apêndice A.

O experimento 2 foi realizado de forma semelhante ao experimento 1, exceto a alteração do parâmetro crossVal para 10. Esse parâmetro define o número de dobras para validação cruzada. O experimento dois demorou 0,13 segundos para execução e identificou 15 regras, das quais os atributos influenciadores identificados foram: Sexo, Capacidade funcional e TNE. Na Figura 23 são apresentadas as regras identificadas.

Figura 23 - Regras identificadas no experimento 2 Decision Table



Rules:

Sexo	Capacidade funcional 1-leve 2-mod 3-grave	TNE SIM-sim NÃO-não	Status-Paciente
M	LEVE	SIM	ALTA
M	MODERADO	SIM	ALTA
F	MODERADO	SIM	ALTA
F	GRAVE	SIM	ALTA
F	LEVE	NÃO	ALTA
M	GRAVE	SIM	ALTA
M	LEVE	NÃO	ALTA
F	NÃO COMPR. CAPAC FUNC	SIM	OBITO
M	NÃO COMPR. CAPAC FUNC	SIM	ALTA
F	MODERADO	NÃO	ALTA
M	MODERADO	NÃO	ALTA
M	GRAVE	NÃO	ALTA
F	GRAVE	NÃO	ALTA
F	NÃO COMPR. CAPAC FUNC	NÃO	ALTA
M	NÃO COMPR. CAPAC FUNC	NÃO	ALTA

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

As regras apresentadas na Figura 23 podem ser entendidas da seguinte forma. A primeira regra, por exemplo, indica que o sistema identificou que pacientes do sexo feminino, que possuem a capacidade funcional leve e que utilizam sonda, tendem a receber alta. E assim respectivamente para as demais regras.

A Tabela 7 apresenta de forma resumida a comparação entre os experimentos 1 e 2, no quesito de precisão.

Tabela 7 - Comparação dos resultados dos experimentos - Decision Table

Parâmetro	Exp. 1	%	Exp. 2	%
Correctly Classified Instances	328	90,11%	328	90,11%
Incorrectly Classified Instances	36	9,89%	36	9,89%
Kappa Statistic	0,1635		0,2219	
Mean absolute error	0,1285		0,1359	
Root mean squared error	0,2299		0,2340	
Relative absolute error %	127,5047		134,9196	
Root relative squared error %	103,9544		105,8072	
Total Number of Instances	**364**		**364**	

FONTE: Elaborado pelo autor (2017)

No total de instâncias (364), ambos experimentos 1 e 2 classificaram corretamente 328 instâncias e incorretamente 36. Logo, obteve-se um resultado satisfatório do ponto de vista de classificação correta, atingindo 90,11% de taxa de acerto.

O Kappa Statistic (Estatística Kappa) obtido no experimento 1 foi de 0,1635, enquanto o experimento 2 resultou em 0,2219. Novamente, os valores são próximos,

porém são classificados em categorias distintas. O experimento 1 se enquadra em baixo grau de concordância, enquanto o experimento 2 indica grau de concordância razoável. Embora o experimento 2 tenha obtido resultado mais satisfatório que o experimento 1, ambos estão longe de indicar alto grau de concordância.

O *Mean Absolute Error* (Erro Médio Absoluto) foi de 0,1285 no experimento 1 e 0,1359 no experimento 2. Levando em consideração esse parâmetro, o experimento 1 apresentou melhores resultados.

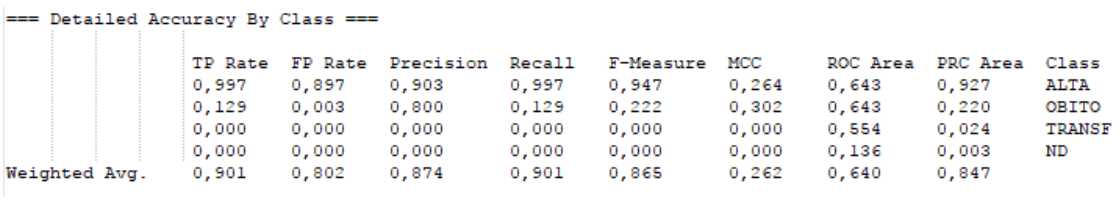
O *Root Mean Squared Error* (Erro Quadrado Médio) do experimento 1 foi de 0,2299 e 0,2340 no experimento 2. Considerando esse parâmetro, o experimento 1 apresentou menor erro entre os valores atuais e previstos, consequentemente, apresentou resultados melhores resultados.

O *Relative Absolute Error* (Erro Absoluto Relativo) do experimento 1 foi de 127,50% e do experimento 2 foi de 134,91%. Ambos apresentaram erro absoluto relativo extremamente elevado. Todavia, o experimento 1 apresentou resultados mais eficazes nesse quesito.

O último quesito, o *Root Relative Squared Error* (Raiz do Erro Quadrado Relativo) apresentou 103,95% para o experimento 1 e 105,80% para o experimento 2. Isso indica que o experimento 1 apresentou resultados melhores em relação ao experimento 2.

As Figuras 24 e 25, respectivamente, apresentam os resultados no quesito de acurácia.

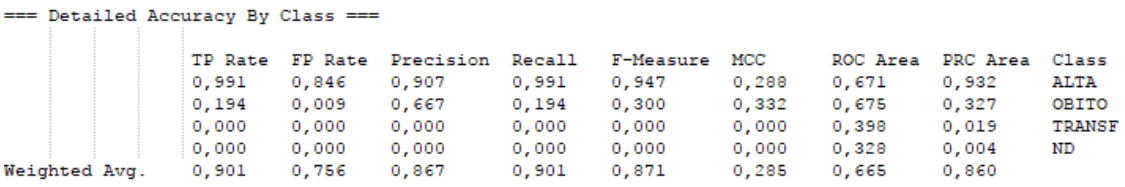
Figura 24 - Acurácia do experimento 1 Decision Table



```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,997    0,897    0,903      0,997    0,947      0,264    0,643     0,927     ALTA
               0,129    0,003    0,800      0,129    0,222      0,302    0,643     0,220     OBITO
               0,000    0,000    0,000      0,000    0,000      0,000    0,554     0,024     TRANSF
               0,000    0,000    0,000      0,000    0,000      0,000    0,136     0,003     ND
Weighted Avg.  0,901    0,802    0,874      0,901    0,865      0,262    0,640     0,847
```

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

Figura 25 - Acurácia do experimento 2 Decision Table



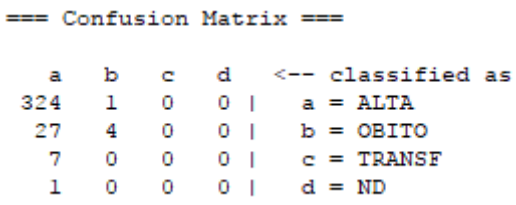
```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,991    0,846    0,907      0,991    0,947      0,288    0,671     0,932     ALTA
               0,194    0,009    0,667      0,194    0,300      0,332    0,675     0,327     OBITO
               0,000    0,000    0,000      0,000    0,000      0,000    0,398     0,019     TRANSF
               0,000    0,000    0,000      0,000    0,000      0,000    0,328     0,004     ND
Weighted Avg.  0,901    0,756    0,867      0,901    0,871      0,285    0,665     0,860
```

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

Com relação a acurácia, o experimento 1 apresenta um resultado 0,6% maior que o experimento 2 para a classificação de status Alta do parâmetro TP Rate.

As Figuras 26 e 27 apresentam as matrizes confusão obtidas dos experimentos 1 e 2, respectivamente, demonstrando na diagonal principal as instâncias que foram corretamente classificadas.

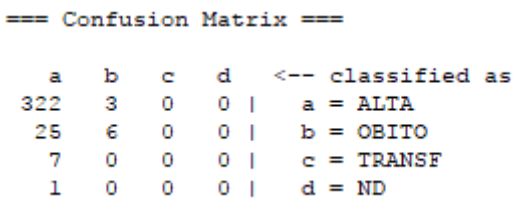
Figura 26 - Matriz confusão do experimento 1 Decision Table



```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 324   1   0   0 |   a = ALTA
  27   4   0   0 |   b = OBITO
   7   0   0   0 |   c = TRANSF
   1   0   0   0 |   d = ND
```

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

Figura 27 - Matriz confusão do experimento 2 Decision Table



```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 322   3   0   0 |   a = ALTA
  25   6   0   0 |   b = OBITO
   7   0   0   0 |   c = TRANSF
   1   0   0   0 |   d = ND
```

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

Em ambos os experimentos, a categoria com maior taxa de acerto foi “Alta”. No experimento 1, a taxa de acerto foi de 90,25%, enquanto no experimento 2 foi de 90,70%. Embora na categoria Alta o experimento 2 tenha apresentado melhores

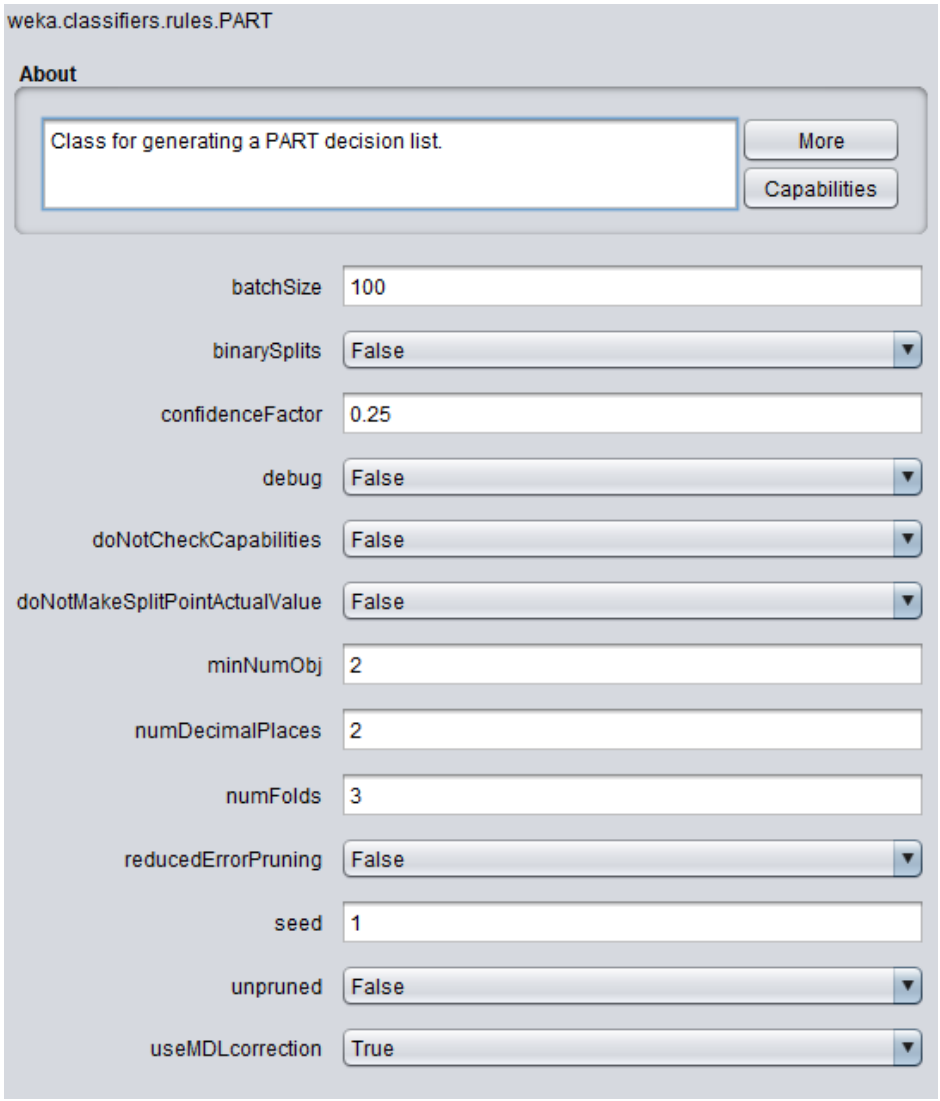
resultados, é possível identificar que em relação aos resultados obtidos na matriz confusão, em ambos os experimentos, as instâncias foram classificadas corretamente 90,11%.

Após concluir a análise dos resultados apresentados na heurística *Decision Table*, foi realizada a mineração de dados com os demais algoritmos.

4.3.4 PART

O algoritmo PART foi executado na heurística de regras. O algoritmo é definido pelo software Weka como: “classe para gerar uma lista de decisão PART. Constrói uma árvore de decisão C4.5 parcial em cada iteração e faz a "melhor" folha em uma regra”. Como parâmetros e valores padrão, o software apresenta os itens dispostos na Figura 28.

Figura 28 - Valores padrão dos parâmetros do algoritmo PART



FONTE: Dados da pesquisa utilizando o software Weka (2017)

O experimento 1 foi executado com os valores padrão para os parâmetros disponibilizados pelo Weka. Foram geradas 11 regras, disponibilizadas no Apêndice B, em 0,01 segundo. Para o experimento 2, foi alterado o valor mínimo de confiança para 0.90. O experimento 2 executou em 21,95 segundos e resultou em 24 regras, dispostas no Apêndice C.

A Tabela 8 apresenta de forma resumida a comparação entre os experimentos 1 e 2, no quesito de precisão.

Tabela 8 - Comparação dos resultados dos experimentos - PART

Parâmetro	Exp. 1	%	Exp. 2	%
Correctly Classified Instances	319	87,64%	305	83,79%
Incorrectly Classified Instances	45	12,36%	59	16,21%
Kappa Statistic	0,2406		0,1468	
Mean absolute error	0,0772		0,0893	
Root mean squared error	0,2314		0,2738	
Relative absolute error %	75,5711		88,6558	
Root relative squared error %	104,6262		123,8073	
Total Number of Instances	**364**		**364**	

FONTE: Elaborado pelo autor (2017)

A partir da análise dos resultados dispostos na Tabela 8, é possível identificar que o experimento 1 atingiu uma taxa de acerto de 87,64%, classificando corretamente 319 instâncias e incorretamente 45 instâncias. Logo, é possível verificar que o experimento 1 obteve resultados 3,85% melhores que o experimento 2, o qual atingiu 83,79% de taxa de acerto, classificando corretamente 305 instâncias e incorretamente 59 instâncias.

O Kappa Statistic (Estatística Kappa) obtido no experimento 1 foi de 0,2406, enquanto o experimento 2 resultou em 0,1468. Os valores são próximos, porém são classificados em categorias distintas. O experimento 1 se enquadra em grau de concordância razoável, enquanto o experimento 2 indica grau de concordância baixo. Logo, o experimento 1 obteve resultado mais satisfatório que o experimento 2 em relação ao grau de concordância.

O *Mean Absolute Error* (Erro Médio Absoluto) foi de 0,0772 no experimento 1 e 0,0893 no experimento 2. Levando em consideração esse parâmetro, o experimento 1 apresentou melhores resultados.

O *Root Mean Squared Error* (Erro Quadrado Médio) do experimento 1 foi de 0,2314 e 0,2738 no experimento 2. Considerando esse parâmetro, o experimento 1 apresentou menor erro entre os valores atuais e previstos, consequentemente, apresentou resultados melhores resultados.

O *Relative Absolute Error* (Erro Absoluto Relativo) do experimento 1 foi de 75,57% e do experimento 2 foi de 88,65%. Ambos apresentaram erro absoluto relativo

elevado. Todavia, o experimento 1 apresentou resultados mais eficazes nesse quesito.

O último quesito, o *Root Relative Squared Error* (Raiz do Erro Quadrado Relativo) apresentou 104,62% para o experimento 1 e 123,80% para o experimento 2. Isso indica que o experimento 1 apresentou resultados melhores em relação ao experimento 2.

As Figuras 29 e 30, respectivamente, apresentam os resultados no quesito de acurácia.

Figura 29 - Acurácia do experimento 1 PART

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,954	0,744	0,914	0,954	0,934	0,257	0,602	0,894	ALTA
	0,290	0,045	0,375	0,290	0,327	0,276	0,655	0,290	OBITO
	0,000	0,003	0,000	0,000	0,000	-0,007	0,419	0,027	TRANSF
	0,000	0,000	0,000	0,000	0,000	0,000	0,468	0,003	ND
Weighted Avg.	0,876	0,668	0,848	0,876	0,862	0,253	0,602	0,824	

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

Figura 30 - Acurácia do experimento 2 PART

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,914	0,769	0,908	0,914	0,911	0,148	0,509	0,879	ALTA
	0,226	0,078	0,212	0,226	0,219	0,144	0,482	0,153	OBITO
	0,143	0,008	0,250	0,143	0,182	0,177	0,420	0,063	TRANSF
	0,000	0,000	0,000	0,000	0,000	0,000	0,444	0,003	ND
Weighted Avg.	0,838	0,694	0,834	0,838	0,836	0,148	0,505	0,799	

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

Com relação a acurácia, o experimento 1 apresenta um resultado 0,40% maior que o experimento 2 para a classificação de status Alta do parâmetro TP Rate.

As Figuras 31 e 32 apresentam as matrizes confusão obtidas dos experimentos 1 e 2, respectivamente, demonstrando na diagonal principal as instâncias que foram corretamente classificadas.

Figura 31 - Matriz confusão do experimento 1 PART

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 310  15   0   0 |   a = ALTA
  21   9   1   0 |   b = OBITO
   7   0   0   0 |   c = TRANSF
   1   0   0   0 |   d = ND
```

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

Figura 32 - Matriz confusão do experimento 2 PART

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 297  25   3   0 |   a = ALTA
  24   7   0   0 |   b = OBITO
   5   1   1   0 |   c = TRANSF
   1   0   0   0 |   d = ND
```

FONTE: Resultado da pesquisa utilizando o software Weka (2017)

Em ambos os experimentos, a categoria com maior taxa de acerto foi “Alta”. No experimento 1, a taxa de acerto foi de 91,44%, enquanto no experimento 2 foi de 90,82%%. Além de apresentar melhores resultados na classificação da categoria Alta, o experimento 1 também apresentou melhores resultados que o experimento 2 na matriz confusão, com uma diferença de 3,85% na classificação correta das instâncias.

Após concluir a análise dos resultados apresentados na heurística PART, foi realizada a análise dos resultados obtidos com todos os experimentos.

4.3.5 Análise dos Resultados Obtidos

Analisando os resultados obtidos com os algoritmos J48, *Decision Table* e PART, foi possível identificar que, de maneira geral, o experimento 1 apresentou melhores resultados quanto a acurácia da base de dados, conforme apresentado na Tabela 9.

Tabela 9 - Comparação geral entre os experimentos 1 e 2

Parâmetros	**J48**		**Decision Table**		**PART**	
Experimento	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2
Correto	328	310	328	328	319	305
Incorreto	36	54	36	36	45	59
Tempo de Execução	0,01	32,02	0,15	0,13	0,01	21,95
Kappa Statistic	0,22	0,18	0,16	0,22	0,24	0,15
Mean absolute error	0,09	0,09	0,13	0,14	0,08	0,09
Root mean squared error	0,22	0,27	0,23	0,23	0,23	0,27
Relative absolute error (%)	87,00	85,81	127,50	134,92	75,57	88,66
Root relative squared error (%)	**78,43**	**121,61**	**103,95**	**105,81**	**104,63**	**123,81**

FONTE: Elaborado pelo autor (2017)

A Tabela 10 apresenta uma comparação entre o desempenho de classificação dos algoritmos no experimento 1. Na figura, é possível identificar os parâmetros de validação cruzada estratificada apresentada pelo sistema Weka.

Tabela 10 - Desempenho de Classificação - Experimento 1

Parâmetros	**J48**	**Decision Table**	**PART**	**Média**	**Desvio Padrão**
Correto	328	328	319	325	5,196152
Incorreto	36	36	45	39	5,196152
Tempo de Execução	0,01	0,15	0,01	0,0567	0,0808
Kappa Statistic	0,2219	0,1635	0,2406	0,208667	0,040217
Mean absolute error	0,0877	0,1285	0,0772	0,0978	0,0271
Root mean squared error	0,2161	0,2299	0,2314	0,2258	0,008434
Relative absolute error (%)	86,9959	127,5047	75,5711	96,69057	27,29038
Root relative squared error (%)	**78,4263**	**103,9544**	**104,6262**	**95,66897**	**14,93636**

FONTE: Elaborado pelo autor (2017)

Em relação ao tempo de execução, os resultados obtidos foram satisfatórios para todos os algoritmos. Entretanto, é possível identificar que o algoritmo *Decision Table* foi o que apresentou tempo maior de execução.

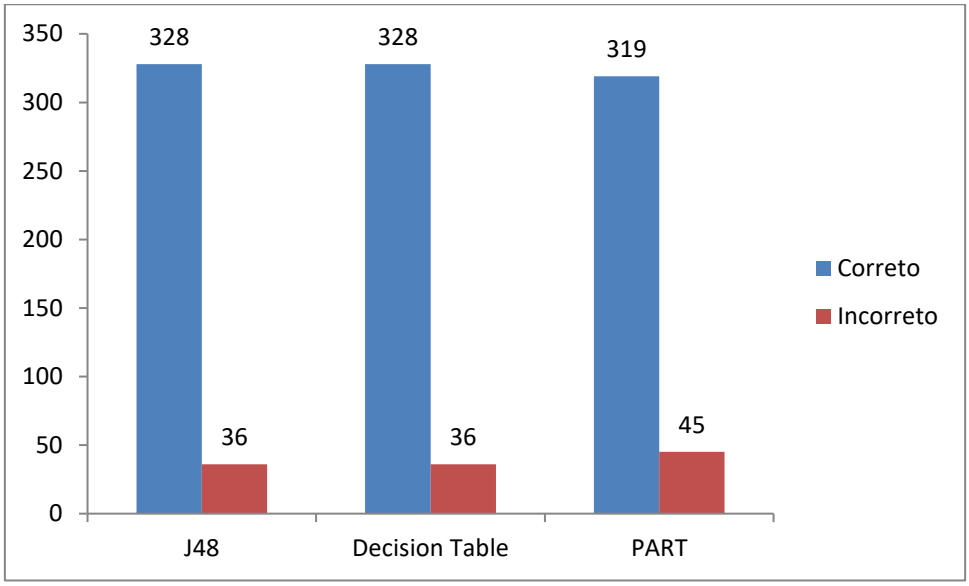
Quanto ao Kappa Statistic, tanto o algoritmo J48 quanto o PART foram classificados na categoria de correlação razoável, que corresponde entre 0.20 e 0.39. Já o algoritmo *Decision Table* foi classificado como correlação fraca, que corresponde ao intervalo de 0 a 0.19.

Quanto ao *Mean absolute error* (Erro absoluto médio), o algoritmo PART apresentou melhores resultados, considerando que classificou um número menor de instâncias do que os demais algoritmos. Quanto ao *Root mean squared error* (Erro quadrado médio), o algoritmo J48 apresentou menor erro entre os valores atuais e previstos.

Quanto ao *Relative absolute error* (Erro absoluto relativo), o algoritmo PART apresentou maior precisão para previsão do que os demais algoritmos. Por fim, quanto ao *Root relative squared error* (Raiz do Erro Quadrado absoluto), o algoritmo J48 obteve melhores resultados, pois apresentou menor erro que os demais algoritmos em relação ao experimento 1.

A gráfico 2 apresenta uma comparação entre os resultados dos algoritmos com base nas instâncias classificadas corretamente e incorretamente. Com base nesse critério, é possível observar que tanto o algoritmo J48 como o *Decision Table* obtiveram os mesmos resultados.

Gráfico 2 - Desempenho de classificação - Experimento 1



FONTE: Elaborado pelo autor (2017)

A Tabela 11 apresenta uma comparação entre o desempenho de classificação dos algoritmos no experimento 2. Na figura, é possível identificar os parâmetros de validação cruzada estratificada apresentada pelo sistema Weka.

Tabela 11 - Desempenho de Classificação - Experimento 2

Parâmetros	J48	Decision Table	PART	Média	Desvio Padrão
Correto	310	328	305	314,3333	12,09683
Incorreto	54	36	59	49,66667	12,09683
Tempo de Execução	32,02	0,13	21,95	18,0333	16,3018
Kappa Statistic	0,18	0,2219	0,1468	0,1829	0,037634
Mean absolute error	0,0865	0,1359	0,0893	0,1039	0,027748
Root mean squared error	0,269	0,234	0,2738	0,258933	0,021726
Relative absolute error (%)	85,8063	134,9196	88,6558	103,1272	27,56984
Root relative squared error (%)	**121,6101**	**105,8072**	**123,8073**	**117,0749**	**9,819733**

FONTE: Elaborado pelo autor (2017)

Em relação ao tempo de execução, os resultados obtidos foram satisfatórios para todos os algoritmos. Entretanto, é possível identificar que o algoritmo J48 foi o que apresentou tempo maior de execução.

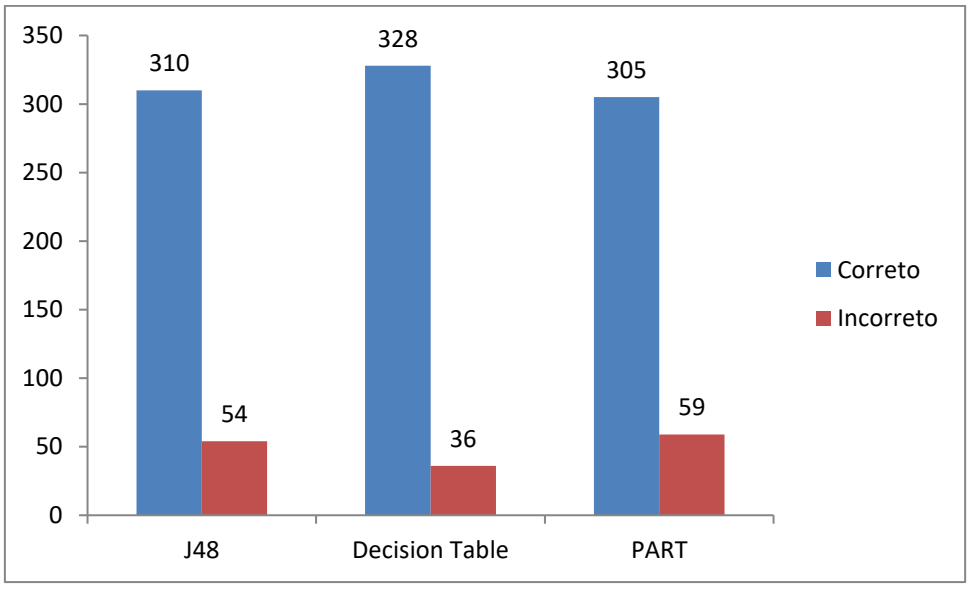
Quanto ao Kappa Statistic, tanto o algoritmo J48 quanto o PART foram classificados como correlação fraca, que corresponde ao intervalo de 0 a 0.19. Já o *Decision Table* foi classificado na categoria de correlação razoável, que corresponde entre 0.20 e 0.39.

Quanto ao *Mean absolute error* (Erro absoluto médio), o algoritmo J48 apresentou melhores resultados, considerando que classificou um número menor de instâncias do que os demais algoritmos. Quanto ao *Root mean squared error* (Erro quadrado médio), o algoritmo *Decision Table* apresentou menor erro entre os valores atuais e previstos.

Quanto ao *Relative absolute error* (Erro absoluto relativo), o algoritmo J48 apresentou maior precisão para previsão do que os demais algoritmos. Por fim, quanto ao *Root relative squared error* (Raiz do Erro Quadrado absoluto), o algoritmo *Decision Table* obteve melhores resultados, pois apresentou menor erro que os demais algoritmos em relação ao experimento 2.

A gráfico 3 apresenta uma comparação entre os resultados dos algoritmos com base nas instâncias classificadas corretamente e incorretamente. Com base nesse critério, é possível observar que tanto o algoritmo J48 como o *Decision Table* obtiveram os mesmos resultados.

Gráfico 3 - Desempenho de classificação - Experimento 2



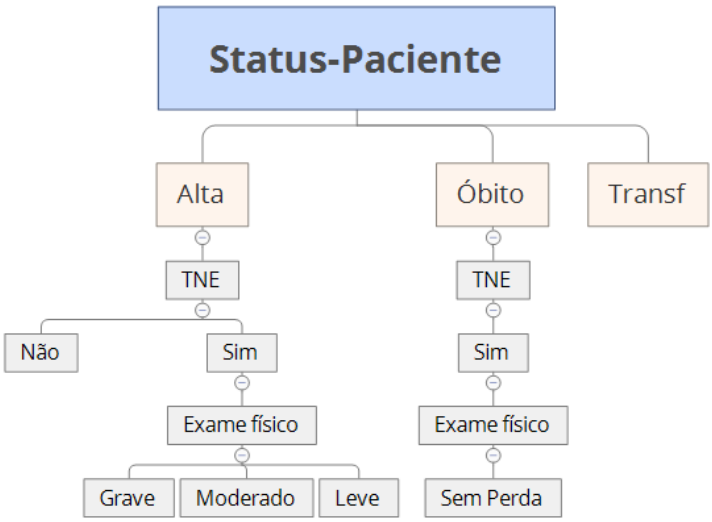
FONTE: Elaborado pelo autor (2017)

Analisando os resultados obtidos pelos algoritmos em ambos os experimentos 1 e 2, é possível identificar que os algoritmos *Decision Table* e J48 atenderam melhor as necessidades de classificação de acordo com a base de dados trabalhada. Ambos algoritmos apresentaram resultados relevantes para a análise e mantiveram a acurácia da base de dados, ainda que não tenham obtido altos valores da estatística Kappa, apresentando no máximo índices de correlações razoáveis.

A vantagem do algoritmo J48, para este estudo, é a apresentação da árvore de decisão, que facilita a análise das informações e posterior compreensão. Já o *Decision Table*, por ser uma heurística de regra, apresenta como resultado um conjunto de regras organizadas em uma tabela de decisão, que tende a tornar análise mais lenta.

Na árvore de decisão gerada pelo algoritmo J48, é possível identificar padrões referentes ao status dos pacientes. A Figura 33 apresenta um resumo dos dados apresentados pela árvore de decisão, agrupados pelo atributo Status-Paciente.

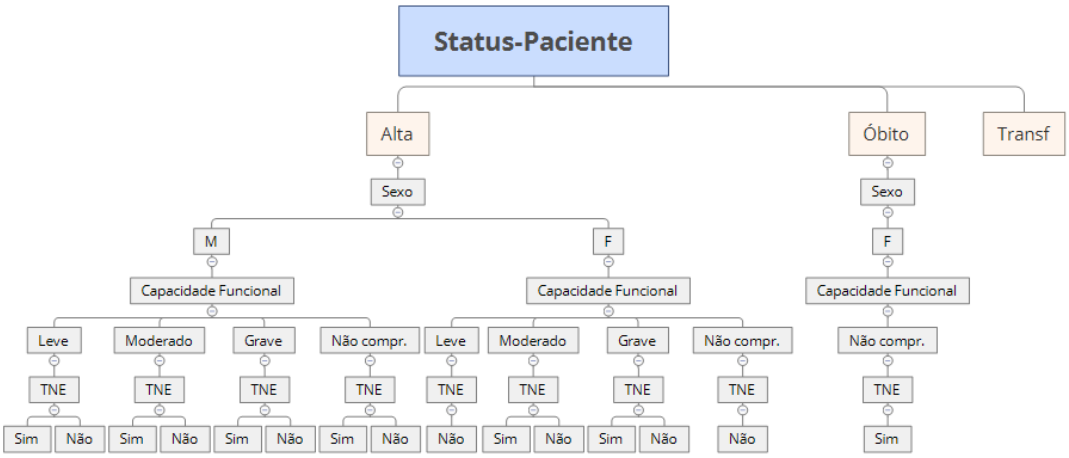
Figura 33 - Mapa Conceitual - Status Paciente – J48



FONTE: Elaborado pelo autor (2017)

Já a Figura 34 apresenta os resultados identificados na execução do algoritmo *Decision Table*, que gerou 15 regras. As regras foram organizadas em um mapa conceitual para facilitar a visualização.

Figura 34 - Mapa Conceitual - Status Paciente - Decision Table



FONTE: Elaborado pelo autor (2017)

Para apresentação dos resultados na ferramenta de Business Intelligence Power BI, foi necessário repassar os resultados obtidos no experimento 1 do algoritmo J48 e experimento 2 do algoritmo *Decision Table* para o Excel, conforme Figura 35 e 36.

Figura 35 - Transformação dos resultados obtidos J48

TNE	Exame Físico	Status-Paciente
NÃO		ALTA
SIM	SEM PERDA	ÓBITO
SIM	GRAVE	ALTA
SIM	MODERADO	ALTA
SIM	LEVE	ALTA

FONTE: Elaborado pelo autor (2017)

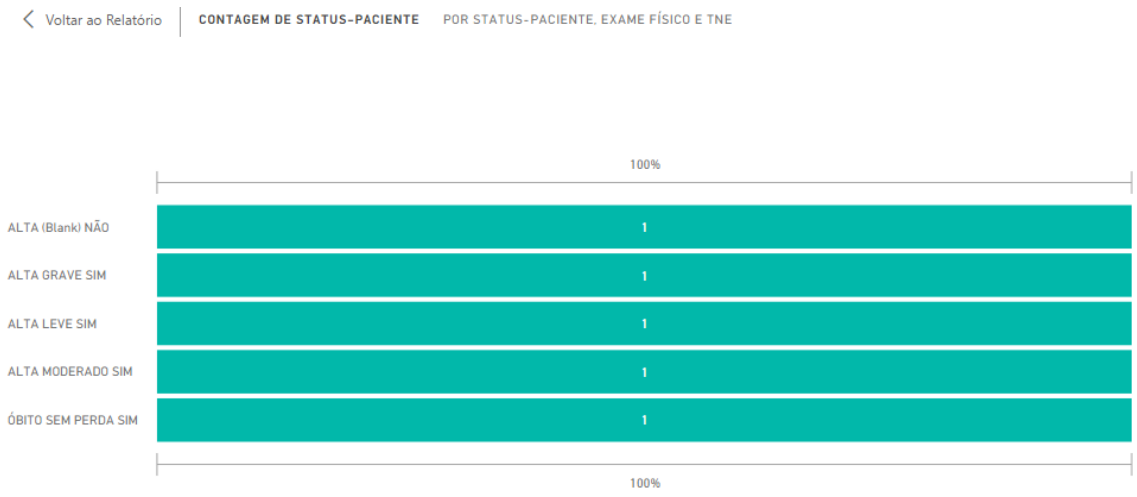
Figura 36 - Transformação dos resultados obtidos Decision Table

Sexo	Capacidade Funcional	TNE	Status-Paciente
M	LEVE	SIM	ALTA
M	MODERADO	SIM	ALTA
F	MODERADO	SIM	ALTA
F	GRAVE	SIM	ALTA
F	LEVE	NÃO	ALTA
M	GRAVE	SIM	ALTA
M	LEVE	NÃO	ALTA
F	NÃO COMPR.	SIM	ÓBITO
M	NÃO COMPR.	SIM	ALTA
F	MODERADO	NÃO	ALTA
M	MODERADO	NÃO	ALTA
M	GRAVE	NÃO	ALTA
F	GRAVE	NÃO	ALTA
F	NÃO COMPR.	NÃO	ALTA
M	NÃO COMPR.	NÃO	ALTA

FONTE: Elaborado pelo autor (2017)

A partir da transformação dos resultados dos algoritmos em um formato de registros, foi possível realizar a leitura das informações e gerar visualizações interativas, apresentadas nas Figuras 37 a 44.

Figura 37 - Gráfico de Funil J48

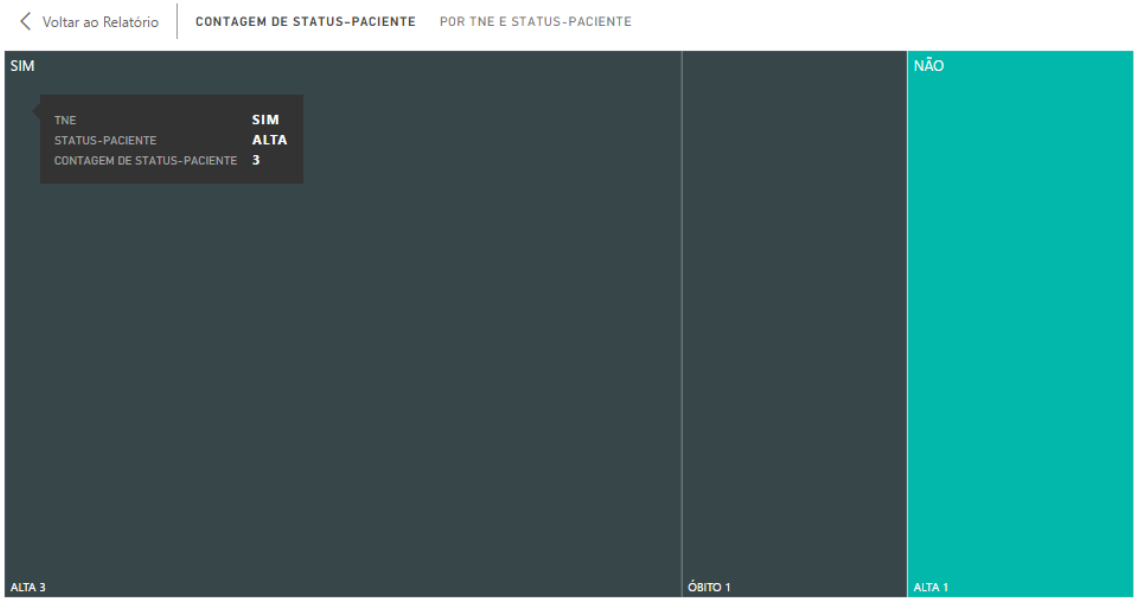


FONTE: Elaborado pelo autor na ferramenta Power BI (2017)

A Figura 37 apresenta os resultados obtidos do algoritmo J48 no gráfico de Funil, identificando por Status-Paciente, Exame físico e TNE a quantidade de

pacientes identificados. Por exemplo, a última linha representa que foi identificado um paciente com status de “Óbito”, Exame físico “Sem Perda” e TNE “Sim”.

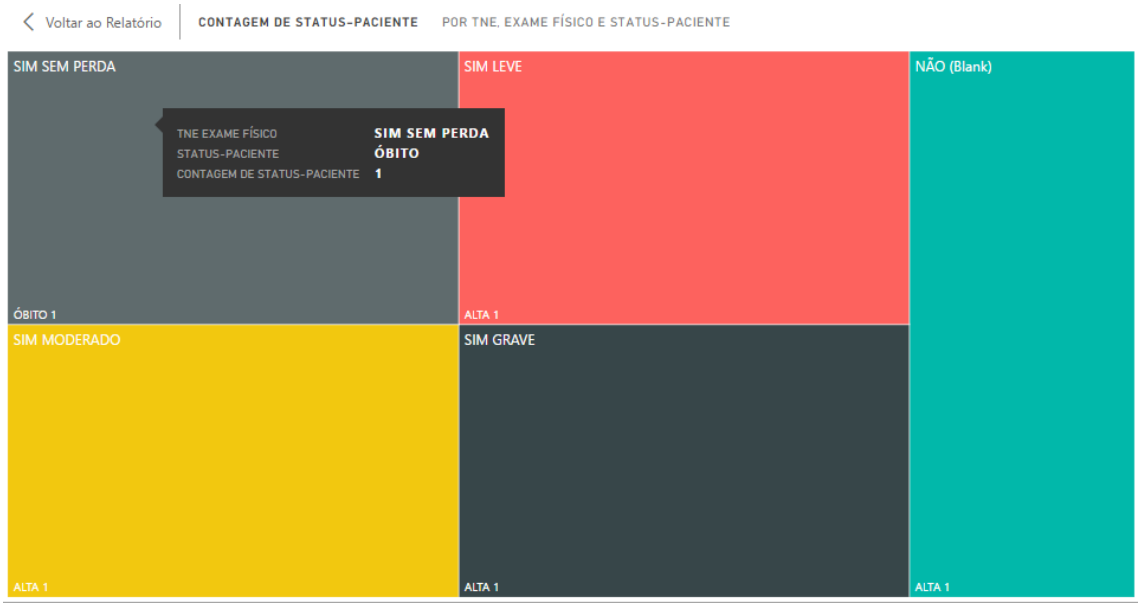
Figura 38 - Gráfico Treemap J48 - Detalhe 1



FONTE: Elaborado pelo autor na ferramenta Power BI (2017)

A Figura 38 representa os resultados obtidos do algoritmo J48 no gráfico Treemap (árvore), identificando a quantidade de pacientes por status no nível de TNE. A opção “Sim” para TNE representa a área cinza, da qual foram identificados 3 pacientes com Alta e 1 Óbito. Já a opção “Não”, representada pela cor azul claro, indica a identificação de 1 paciente com status de Alta.

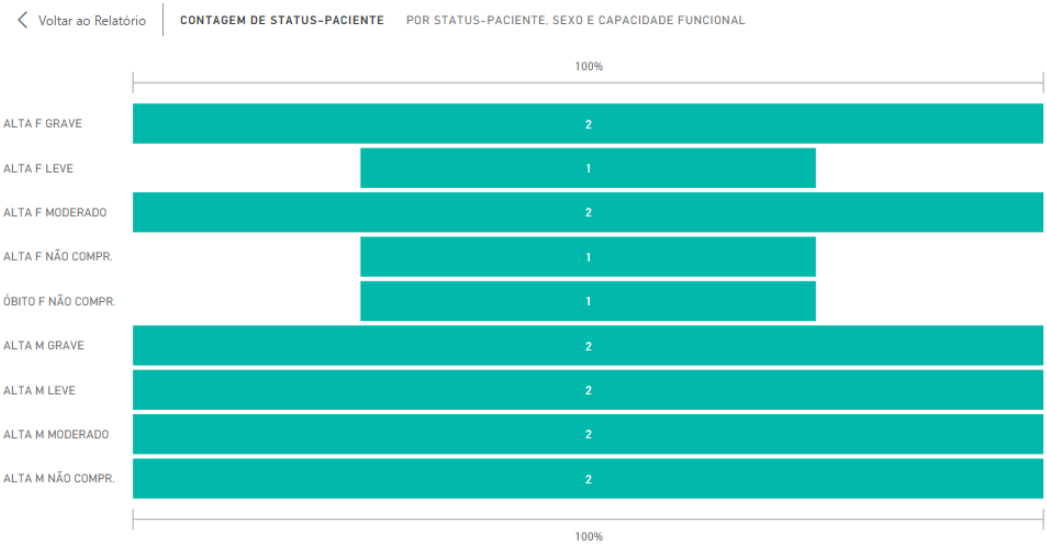
Figura 39 - Gráfico Treemap J48 - Detalhe 2



FONTE: Elaborado pelo autor na ferramenta Power BI (2017)

A Figura 39 apresenta os mesmos resultados do algoritmo J48 em um gráfico Treemap, bem como a Figura 38. A diferença é o nível de detalhe. Na Figura 39, é apresentado 1 nível a mais de detalhe (Exame Físico) que a Figura 38 (TNE). Por exemplo, na área representada pela cor amarela, foi identificado um paciente com status Alta, Exame físico “Moderado” e TNE “Sim”.

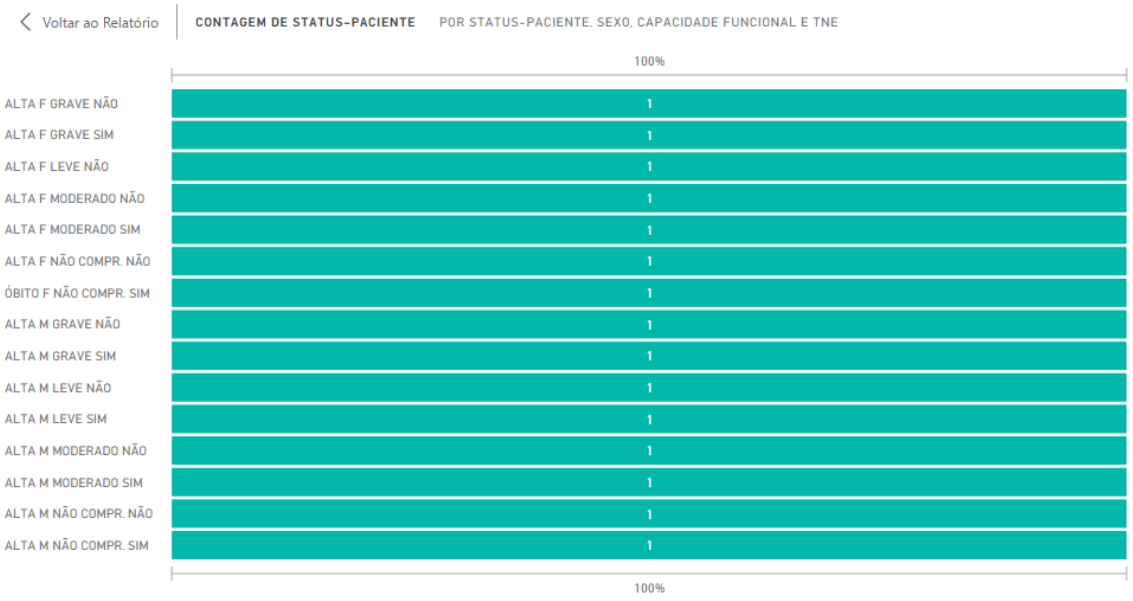
Figura 40 - Gráfico Funil Decision Table - Detalhe 1



FONTE: Elaborado pelo autor na ferramenta Power BI (2017)

A Figura 40 apresenta os resultados obtidos do algoritmo *Decision Table* no gráfico de Funil, identificando por Status-Paciente, Sexo e Exame Físico a quantidade de pacientes encontrados. Por exemplo, a primeira linha representa que foi identificado dois pacientes com status de “Alta”, Sexo “F” e Capacidade Funcional “Grave”.

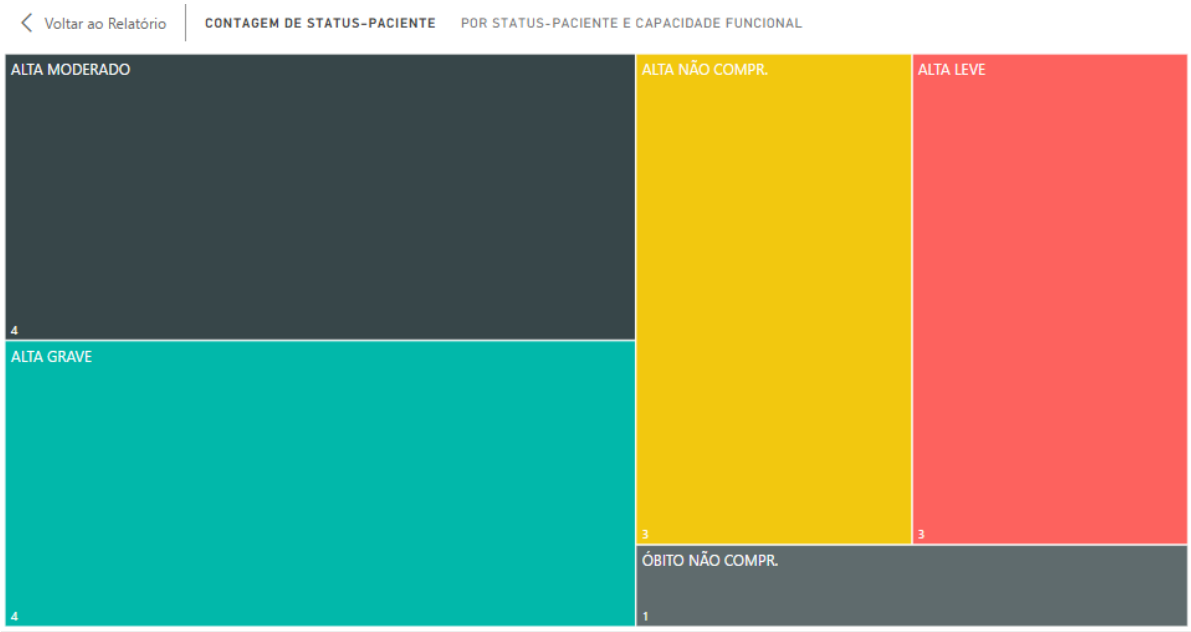
Figura 41 - Gráfico Funil Decision Table - Detalhe 2



FONTE: Elaborado pelo autor na ferramenta Power BI (2017)

A Figura 41 apresenta os mesmos resultados do algoritmo *Decision Table* em um gráfico Funil, bem como a Figura 40. A diferença é o nível de detalhe. Na figura 41, é apresentado 1 nível a mais de detalhe (TNE) que a Figura 40 (Capacidade Funcional). Por exemplo, na última linha, foi identificado um paciente de status Alta, Sexo “M”, Capacidade Funcional “Não Compr.” e TNE “Sim”.

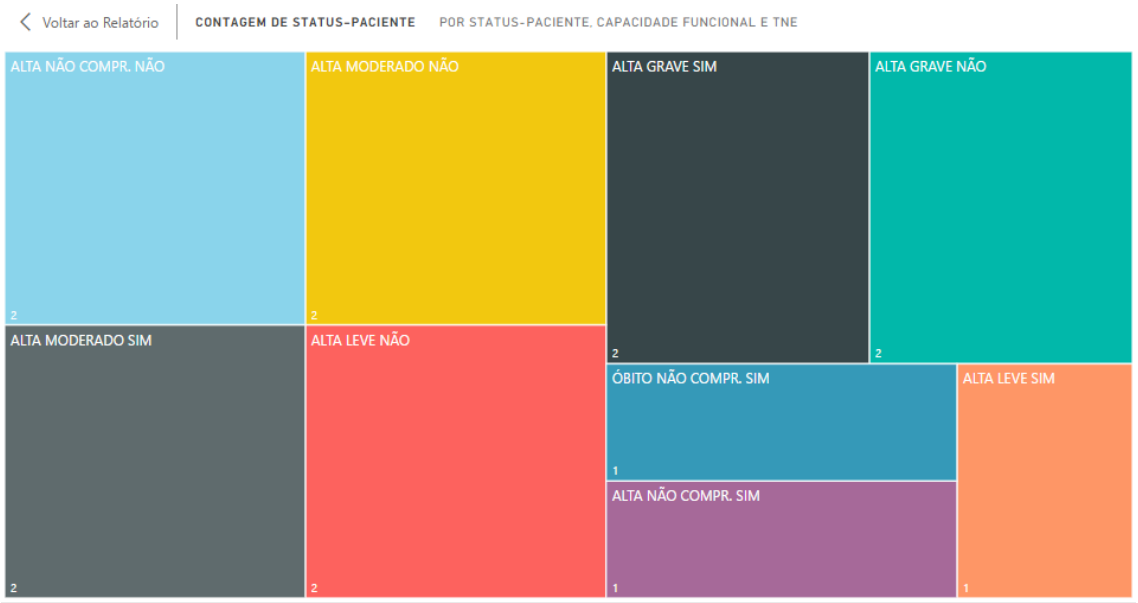
Figura 42 - Gráfico Treemap Decision Table - Detalhe 1



FONTE: Elaborado pelo autor na ferramenta Power BI (2017)

A Figura 42 representa os resultados obtidos do algoritmo *Decision Table* no gráfico Treemap (árvore), identificando a quantidade de pacientes por status no nível de Capacidade Funcional. Por exemplo, a área representada pela cor vermelha apresenta 3 pacientes com status “Alta” e Capacidade Funcional “Leve”.

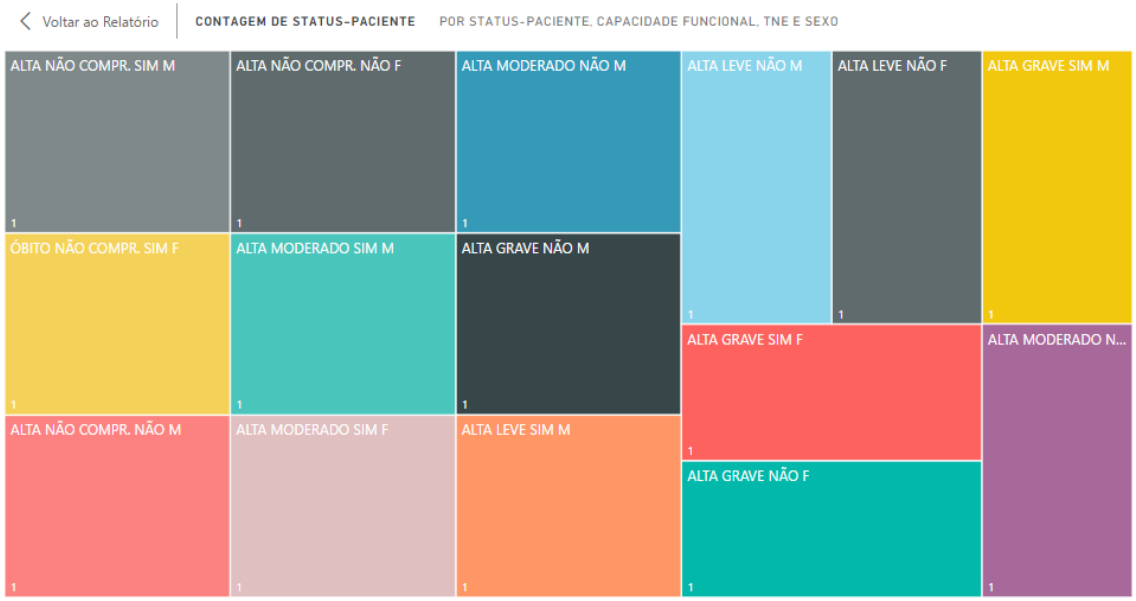
Figura 43 - Gráfico Treemap Decision Table - Detalhe 2



FONTE: Elaborado pelo autor na ferramenta Power BI (2017)

A Figura 43 apresenta os mesmos resultados do algoritmo *Decision Table* em um gráfico Treemap, bem como a Figura 42. A diferença novamente é o nível de detalhe. Na Figura 43, é apresentado 1 nível a mais de detalhe (TNE) que a Figura 42 (Capacidade Funcional). Por exemplo, na área de cor amarela, foi identificado um paciente de status Alta, Capacidade Funcional “Moderado” e TNE “Não”.

Figura 44 - Gráfico Treemap Decision Table - Detalhe 3



FONTE: Elaborado pelo autor na ferramenta Power BI (2017)

Por último, a Figura 44 apresenta 1 nível a mais de detalhe que a Figura 43. Na Figura 44, o nível de detalhe chega ao atributo Sexo, enquanto na Figura 43 o nível é TNE. Por exemplo, na área representada pela cor laranja, foi identificado um paciente de status “Alta”, Capacidade Funcional “Leve”, TNE “Sim” e Sexo “M”.

Analisando os resultados apresentados nos mapas conceituais e nas visualizações na ferramenta Power BI, é possível verificar que ambos algoritmos apresentaram dificuldades em encontrar padrões para o status “Transf”. Além disso, as visualizações geradas para o tipo de gráfico Funil não agregaram valor à representação das informações. Essa dificuldade é suportada pela distribuição da base de dados, a qual dos 364 registros, apenas 7 representam o status “Transf”, ou seja, apenas 1,92%. Os resultados obtidos pelo J48 apontaram o parâmetro TNE (Terapia Nutricional Enteral) como o principal influenciador do status do paciente. Em segundo, o parâmetro “Exame Físico” foi apontado como influenciador secundário. Assim como no J48, o algoritmo *Decision Table* também apresentou o TNE (Terapia

Nutricional Enteral) como fator de influência. Porém, entre as regras identificadas, o algoritmo também considerou outros fatores, como o Sexo e a Capacidade Funcional.

4.3.6 Validação dos Resultados

Conforme mencionado na metodologia, disposta na seção 3, para realizar a validação dos resultados foi realizada a mineração de dados e posteriormente, aplicado um questionário ao profissional nutricionista que disponibilizou a base de dados. A mineração de dados foi validada na seção 5.3.5, apresentando os resultados obtidos provindos da aplicação dos algoritmos de mineração de dados.

O questionário, que se encontra disponibilizado no Apêndice D, é composto por questões que buscam validar se os padrões encontrados a partir da aplicação de técnicas de mineração de dados são relevantes e contribuem de alguma forma no processo de tomada de decisão do profissional, bem como se os padrões identificados já estavam evidenciados. O questionário possui uma questão semiaberta, em que o profissional expressa se visão dele, faltou a identificação de algum padrão específico.

O questionário foi impresso e entregue ao profissional de nutrição responsável pela base. De acordo com as respostas, os resultados apresentados contribuíram com novos conhecimentos, bem como também evidenciaram conhecimentos já consolidados na área. O profissional acredita que é possível aplicar as técnicas de mineração de dados na área da saúde, e ficou satisfeito com a forma na qual os resultados foram apresentados. Na questão 5, o profissional expressa que os resultados contribuiriam de alguma forma para o processo de tomada de decisão nas atividades relacionadas à nutrição. Por último, na questão 6, o profissional respondeu que sentiu falta de alguma relação entre o IMC e a taxa de alta ou óbito. De forma verbal, o profissional salientou a importância de qualquer pesquisa que vise à melhoria da tomada de decisão que envolva o bem-estar dos pacientes.

5 CONSIDERAÇÕES FINAIS

O crescimento do volume de dados gerado e armazenado é uma realidade no século XXI. Cada vez mais, os grandes problemas do cotidiano são solucionados por meio da utilização de algum sistema de informação, o qual por sua vez, possui uma base de dados alimentada pelos usuários. Para isso, a mineração de dados aparece como uma ferramenta de auxílio à tomada de decisão, analisando grandes volumes de dados em busca de padrões relevantes.

A pesquisa bibliométrica permitiu validar a relevância científica e o ineditismo da pesquisa para o curso de Gestão da Informação. Com este levantamento, foi possível identificar a carência deste tipo de pesquisa na base Web of Science e no repositório de monografias do curso.

É necessário salientar a importância na escolha do método de mineração de dados, o que não é uma tarefa fácil devido a inexistência de um padrão para a escolha, a qual varia com base na composição dos atributos que compõem a base de dados. Dessa forma, a etapa de pré-processamento mostra-se relevante, a qual é responsável pelo tratamento e limpeza dos dados, optando por estratégias de como lidar com os campos em branco dentre outras. Os resultados obtidos após a fase de pós-processamento demonstraram a potência de contribuição desse tipo de trabalho para os profissionais da área de saúde, especificamente da área de nutrição.

5.1 VERIFICAÇÃO DOS OBJETIVOS PROPOSTOS

Para atingir o objetivo geral de aplicação de técnicas de mineração de dados em uma base de dados de nutrição, identificando padrões e apresentando-os de forma visual, foi necessário alcançar três objetivos específicos.

O primeiro objetivo específico - pesquisar e definir o(s) método(s) de mineração de dados que será(ão) utilizado(s) na base de dados de nutrição, - foi alcançado por meio de levantamento bibliográfico, com o objetivo de identificação dos principais métodos utilizados a literatura. Após o levantamento, de acordo com o objetivo, optou-se pela escolha de algoritmos de classificações e associações, visto que atendem melhor o resultado almejado com os experimentos. Em seguida, foram escolhidas as heurísticas. Na categoria de classificação, foram escolhidas as heurísticas de árvore

de decisão e regras, visto que apresentam resultados facilitam a compreensão de pessoas que não são familiarizadas com as práticas de mineração de dados. Para a categoria de associação, foi escolhido o algoritmo Apriori, por ser o mais utilizado de acordo com a literatura. Todavia, pelo algoritmo não apresentar a opção de atributo meta, acabou dificultando a análise dos resultados. Foi preciso o auxílio da ferramenta Excel para filtrar os resultados relacionados com o atributo meta. Os primeiros experimentos foram realizados com os valores padrão dos parâmetros disponíveis para cada algoritmo, os quais apresentaram melhores resultados em relação à apresentação e acurácia. Os demais experimentos foram realizados com o objetivo de alterar os valores padrão e buscar a apresentação de diferentes resultados em relação às regras identificadas e aos atributos de maior influência. Como contribuição, foi possível identificar uma grande diversificação de regras encontradas em relação aos primeiros experimentos.

O segundo objetivo - classificar os pacientes com base nos padrões de decisão identificados – foi atingido parcialmente, visto que não foi possível obter regras sólidas de classificação para o status “Transf”. Porém, para os status Alta e Óbito, foram encontrados padrões que se assemelham a realidade já conhecida pelos profissionais da nutrição, como por exemplo a influência do uso da sonda (representada pelo atributo “TNE”).

O terceiro objetivo – selecionar uma ferramenta de Business Intelligence para apresentação das informações encontradas – foi atingido na medida que foi possível transformar os resultados obtidos no sistema Weka em um formato em que foi possível realizar a leitura dessas informações na ferramenta Power BI e apresentá-las de forma visual.

Dessa forma, é possível concluir que o objetivo geral foi atingido a partir dos três objetivos específicos na medida em que na aplicação da pesquisa através do questionário, o profissional de nutrição indicou contribuição dos resultados apresentados para as atividades da área.

5.2 CONTRIBUIÇÕES

Os resultados passaram por avaliação de um profissional da área de nutrição, com o objetivo de validar a proposta. A partir da aplicação de um questionário composto por seis perguntas, cinco perguntas fechadas e uma pergunta semiaberta. De acordo com as respostas, os resultados apresentados contribuíram com novos conhecimentos, como a ausência da ascite no diagnóstico dos pacientes influencia indiretamente na alta dos pacientes.

O feedback foi interessante e contribuiu para o estudo, evidenciando o quão as tecnologias da informação aplicadas em diferentes áreas podem contribuir com diferentes fins, como na saúde da sociedade. Além disso, de acordo com o levantamento bibliográfico realizado, o qual evidenciou a carência de estudos desenvolvidos na área, os quais possibilitam ao profissional de nutrição tomar decisões mais assertivas. Um exemplo evidenciado foi a ausência do uso de sonda (TNE – Terapia Nutricional Enteral) como fator de influência para os pacientes que recebem alta.

O presente trabalho também evidenciou a possibilidade de utilização de uma ferramenta de Business Intelligence, nesse caso Power BI, para apresentação dos resultados obtidos na aplicação de técnicas de mineração de dados.

5.3 TRABALHOS FUTUROS

Para os trabalhos futuros, é sugerida a aplicação de técnicas de mineração de dados em outras bases de nutrição, com o objetivo de realizar uma comparação com o presente trabalho. Além disso, sugere-se a mineração de dados voltada para casos mais específicos, como por exemplo quais fatores influenciam na prescrição de suplementos dietéticos para os pacientes.

Ainda, é sugerido também um estudo sobre a viabilidade de implantação dessas técnicas em um sistema de saúde, de forma que dê autonomia para os profissionais da nutrição identificarem padrões por meio de um sistema e tomar

decisões pautadas em padrões encontrados na base de dados da instituição hospitalar.

Por fim, sugere-se a aplicação de outras técnicas de mineração de dados com objetivo de verificar a descoberta de novos conhecimentos.

REFERÊNCIAS

BLACKBURN, G. L.; THORNTON, P. A. Nutritional assessment of the hospitalized patients. **Medical Clinics of North America,** v. 63, p. 1103-115, 1979.

CARDOSO, Olinda Nogueira Paes. Recuperação de Informação. **INFOCOMP Journal of Computer Science**, v. 2, n. 1, p. 33-38, 2004.

CRN. Áreas de Atuação. Disponível em: <http://www.crn2.org.br/crn2/portal/default.php>. Acesso em: 14 jun. 2017.

DAVENPORT, T. H. **Ecologia da informação**: porque só a tecnologia não basta para o sucesso na era da informação. 4. ed. São Paulo: Futura, 2002.

DAVENPORT, T. H. **How strategists use "big data" to support internal business decisions, discovery and production**. Strategy and Leadership, 2014, 42(4), 45–50.

DEAN, Jared. **Big data, data mining, and machine learning**: value creation for business leaders and practitioners. John Wiley & Sons, 2014. Data Mining Know It All - Elsevier, 2009.

DE SORDI, J. O. **Administração da informação**: fundamentos e práticas para uma nova gestão do conhecimento. São Paulo: Saraiva, 2008.

DUARTE, C. A. **Dashboard visual**: uma ferramenta de Business Intelligence. 64 f. Dissertação (Mestrado Integrado em Engenharia Mecânica) – Faculdade de Engenharia, Universidade do Porto, Porto, 2012.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From data mining to knowledge discovery in databases**. AI magazine, v. 17, n. 3, p. 37, 1996. Disponível em:<http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf>. Acesso em: 03 jun. 2017.

FRAKES, William B.; BAEZA-YATES, Ricardo. **Information retrieval:** data structures and algorithms. 1992.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

LOBUR, M. et al. Some trends in knowledge discovery and data mining. In: **Perspective Technologies and Methods in MEMS Design, 2008. MEMSTECH 2008. International Conference on**. IEEE, 2008. p. 95-97.

MARCHIORI, P. **A ciência da informação**: compatibilidade no espaço profissional. Caderno de Pesquisas em Administração, São Paulo, v. 9, n.1, p. 91-101, jan./mar. 2002.

MARR, Bernard. **Big data:** 20 Mind-Boggling Facts Everyone Must Read. Disponível em: <https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#391aed4417b1>. Acesso em: 15 jun. 2017.

MARTINEZ, Luís; CASAL, Ricardo; JANEIRO, João. **Sistemas de apoio à decisão clínica**. Porto: Faculdade de Medicina do Porto, 2009. 16 p.

MCAFEE, A. & BRYNJOLFSSON, E. **Big data**: The management revolution. Harvard Business Review, 2012, 90(10), 4.

NOGUEIRA, E. D. A. **Avaliação de brand equity sob a perspectiva do consumidor nas mídias sociais por meio da mineração de opinião e análise de redes sociais**. 2015. 237 f. DISSERTAÇÃO (Mestrado) – Programa de Pós-graduação em Ciência, Gestão e Tecnologia da Informação, Ciências Sociais Aplicadas, Universidade Federal do Paraná.

PEREIRA, F. P. A. **Big data e data analysis**: visualização de informação. 79 f. Dissertação (Mestrado Integrado em Engenharia e Sistema de Informação) – Escola de Engenharia, Universidade do Minho, Braga, 2015.

SANTOS, J. S. D. **Proposição de uma interface de business intelligence para evasão escolar nos institutos federais de educação**. 2017, 125 f. Dissertação (Mestrado em Ciência, Gestão e Tecnologia) – Universidade Federal do Paraná.

SCHIEFERDECKER, Maria Eliana M. et al. **Criação de protocolo eletrônico para terapia nutricional enteral domiciliar**. ABCD arq. bras. cir. dig, v. 26, n. 3, p. 195-199, 2013.

SILVEIRA, M.; MARCOLIN, C. B.; FREITAS, H. M. R. Uso corporativo do big data: uma revisão de literatura. **Revista de Gestão e Projetos** - GeP, 2016, 6(3), 44-59. http://doi.org/10.5585/gep.v6i3.369.

YIN, R. K. **Estudo de caso**: planejamento e métodos. Bookman, 2005.

APÊNDICE A – RESULTADOS EXPERIMENTO 1 - DECISION TABLE

Rules:

Sexo	Perda peso NÃO=não SIM=sim	Ascite NÃO=não LEVE=leve MODERADO=mod GRAVE=grave	Altura real	TNE SIM=sim NÃO=não	Status-Paciente
M	NÃO	LEVE	C	NÃO	ALTA
F	NÃO	MODERADO	ND	NÃO	ALTA
F	SIM	LEVE	A	NÃO	ALTA
F	SIM	NÃO	A	SIM	ALTA
M	NÃO	NÃO	C	NÃO	ALTA
F	SIM	MODERADO	ND	NÃO	ALTA
M	SIM	NÃO	C	NÃO	ALTA
M	NÃO	LEVE	ND	NÃO	OBITO
F	NÃO	NÃO	ND	SIM	OBITO
M	NÃO	NÃO	ND	SIM	ALTA
M	NÃO	NÃO	A	NÃO	ALTA
F	SIM	LEVE	ND	NÃO	ALTA
M	SIM	LEVE	ND	NÃO	ALTA
M	SIM	NÃO	ND	SIM	ALTA
F	NÃO	NÃO	A	NÃO	ALTA
F	SIM	NÃO	ND	SIM	OBITO
M	SIM	MODERADO	B	NÃO	ALTA
M	NÃO	GRAVE	ND	NÃO	ALTA
M	SIM	NÃO	A	NÃO	ALTA
F	SIM	NÃO	A	NÃO	ALTA
M	NÃO	LEVE	B	NÃO	ALTA
F	SIM	GRAVE	ND	NÃO	ALTA
M	NÃO	NÃO	B	SIM	ALTA
M	SIM	LEVE	B	NÃO	ALTA
F	SIM	NÃO	B	SIM	ALTA
M	SIM	NÃO	B	SIM	ALTA
F	SIM	LEVE	B	NÃO	ALTA
F	NÃO	NÃO	ND	NÃO	ALTA
M	NÃO	NÃO	ND	NÃO	ALTA
F	NÃO	GRAVE	B	NÃO	ALTA
F	SIM	NÃO	ND	NÃO	ALTA
M	SIM	NÃO	ND	NÃO	ALTA
F	ND	NÃO	B	NÃO	ALTA
F	NÃO	NÃO	B	NÃO	ALTA
M	NÃO	NÃO	B	NÃO	ALTA
M	SIM	NÃO	B	NÃO	ALTA
F	SIM	NÃO	B	NÃO	ALTA

APÊNDICE B – RESULTADOS EXPERIMENTO 1 – PART

```
PART decision list
------------------

TNE SIM=sim NÃO=não = NÃO AND
NEVO SIM=sim NÃO=não = NÃO: ALTA (191.0/5.0)

TNE SIM=sim NÃO=não = NÃO AND
Edema 0=não 1=+ 2= ++ 3=+++ 4=++++ = NA AND
Exame físico 0= sem perda 1=leve 2=moderado 3=grave = SEM PERDA: ALTA (57.0/3.0)

Exame físico 0= sem perda 1=leve 2=moderado 3=grave = LEVE: ALTA (40.0/3.0)

Capacidade  funcional 1=leve 2=mod 3=grave = MODERADO: ALTA (19.0/1.0)

Altura utilizada 1=real 2=AJ 3=CH 4=refe = ALTURA DO JOELHO AND
1 =pastosa 2=liquida 3=jejum 4= TNE (inicial) = SEM ALTERACAO AND
Altura real = ND AND
Perda peso  NÃO=não SIM=sim = SIM AND
Último peso seco = ND: ALTA (11.0/4.0)

Peso usual = A AND
1= desn 2=eutr 3=sobre/obesid 4=risco = DESNUTRIDO AND
PCT = A AND
Sexo = M AND
CA = ND: ALTA (8.0/3.0)

TNE SIM=sim NÃO=não = NÃO AND
Ascite NÃO=não LEVE=leve MODERADO=mod GRAVE=grave = NÃO AND
Peso ideal = A: ALTA (11.0/1.0)

Ascite NÃO=não LEVE=leve MODERADO=mod GRAVE=grave = NÃO AND
PCR = ND: OBITO (5.0/1.0)

PCR = A AND
Idade = ADULTO: OBITO (11.0/2.0)

Perda peso  NÃO=não SIM=sim = SIM: ALTA (7.0/1.0)

: OBITO (4.0/1.0)
```

APÊNDICE C – RESULTADOS EXPERIMENTO 2 – PART

```
PART decision list
------------------

TNE SIM=sim NÃO=não = NÃO AND
NEVO SIM=sim NÃO=não = NÃO AND
Edema 0=não 1=+ 2= ++ 3=+++ 4=++++ = NA: ALTA (159.0/3.0)

TNE SIM=sim NÃO=não = NÃO AND
Exame físico 0= sem perda 1=leve 2=moderado 3=grave = LEVE AND
Peso Atual (real) = A AND
Altura utilizada 1=real 2=AJ 3=CH 4=refe = ALTURA DO JOELHO: ALTA (14.0)

TNE SIM=sim NÃO=não = NÃO AND
Altura real = B AND
1 =pastosa 2=liquida 3=jejum 4= TNE (inicial) = SEM ALTERACAO: ALTA (56.0/2.0)

Último peso seco = ND AND
Peso seco = B: ALTA (7.0/1.0)

Último peso seco = ND AND
1=  risco 2=aguda 3=cronica = NA AND
TNE SIM=sim NÃO=não = NÃO AND
Exame físico 0= sem perda 1=leve 2=moderado 3=grave = SEM PERDA AND
1= desn 2=eutr 3=sobre/obesid 4=risco = RISCO: ALTA (20.0)

Último peso seco = B AND
Peso usual = A: OBITO (3.0)

Último peso seco = A AND
Ascite NÃO=não LEVE=leve MODERADO=mod GRAVE=grave = NÃO AND
total de dias internados = D: ALTA (3.0)

Último peso seco = ND AND
1=  risco 2=aguda 3=cronica = AGUDA AND
Peso ideal = A AND
Altura estimada CH = ND: ALTA (12.0)
Último peso seco = ND AND
1=  risco 2=aguda 3=cronica = CRONICA AND
Sexo = F AND
1 =pastosa 2=liquida 3=jejum 4= TNE (inicial) = SEM ALTERACAO AND
IMC = Baixo Peso: ALTA (9.0)

GET = B AND
1=leve 2=mod 3=grave = MODERADO: ALTA (13.0/1.0)

IMC = Risco de Deficit: ALTA (4.0)

Edema 0=não 1=+ 2= ++ 3=+++ 4=++++ = MEMB INF E SUP AND
total de dias internados = C: ALTA (3.0)

AJ = A AND
1=  risco 2=aguda 3=cronica = NA: OBITO (2.0)

AJ = C AND
Peso usual = A: OBITO (5.0)

Edema 0=não 1=+ 2= ++ 3=+++ 4=++++ = NA AND
Altura real = ND AND
Cpanturilha = ND: ALTA (11.0/1.0)

1=  risco 2=aguda 3=cronica = NA AND
CMB = B: ALTA (10.0/1.0)

1=  risco 2=aguda 3=cronica = CRONICA AND
IMC = Baixo Peso: ALTA (8.0/1.0)

1=  risco 2=aguda 3=cronica = NA: OBITO (5.0)

IMC = Baixo Peso AND
total de dias internados = A: OBITO (2.0)
CH = B: OBITO (4.0/1.0)

CH = ND AND
IMC = Sobrepeso: ALTA (3.0/1.0)

Sexo = M: ALTA (4.0/1.0)

1=  risco 2=aguda 3=cronica = ND: OBITO (3.0)

: TRANSF (4.0/1.0)
```

APÊNDICE D – QUESTIONÁRIO DE VALIDAÇÃO DE RESULTADOS

Objetivo: avaliar os resultados apresentados no trabalho de conclusão “MINERAÇÃO DE DADOS EM BASES DE AVALIAÇÃO NUTRICIONAL”, validando a utilidade de técnicas de mineração de dados aplicadas a diferentes áreas.

1) Os resultados apresentados contribuíram com novos conhecimentos?

SIM ( ) NÃO ( )

2) Os resultados apresentados evidenciaram conhecimentos já consolidados na área?

SIM ( ) NÃO ( )

3) Você acredita que é possível aplicar as técnicas de mineração de dados na área da saúde?

SIM ( ) NÃO ( )

4) A forma que os resultados foram apresentados foi satisfatória?

SIM ( ) NÃO ( )

5) Os resultados apresentados contribuiriam para o processo de tomada de decisão de atividades da área da nutrição?

SIM ( ) NÃO ( )

6) Você sentiu falta de algum resultado em específico?

SIM ( ) NÃO ( )

QUAL:_____

_____

ANEXO A – RESPOSTA QUESTIONÁRIO DE VALIDAÇÃO

Objetivo: avaliar os resultados apresentados no trabalho de conclusão "MINERAÇÃO DE DADOS EM BASES DE AVALIAÇÃO NUTRICIONAL", validando a utilidade de técnicas de mineração de dados aplicadas a diferentes áreas.

1) Os resultados apresentados contribuiram com novos conhecimentos?

SIM (X) NÃO ( )

2) Os resultados apresentados evidenciaram conhecimentos ja consolidados na área?

SIM (X) NÃO ( )

3) Você acredita que é possível aplicar as técnicas de mineração de dados na área da saúde?

SIM (X) NÃO ( )

4) A forma que os resultados foram apresentados foi satisfatória?

SIM (X) NÃO ( )

5) Os resultados apresentados contribuiriam para o processo de tomada de decisão de atividades da área da nutrição?

SIM (X) NÃO ( )

6) Você sentiu falta de algum resultado em específico?

SIM (X) NÃO ( )

QUAL: Algum tipo de relação do IMC com alta/óbito.