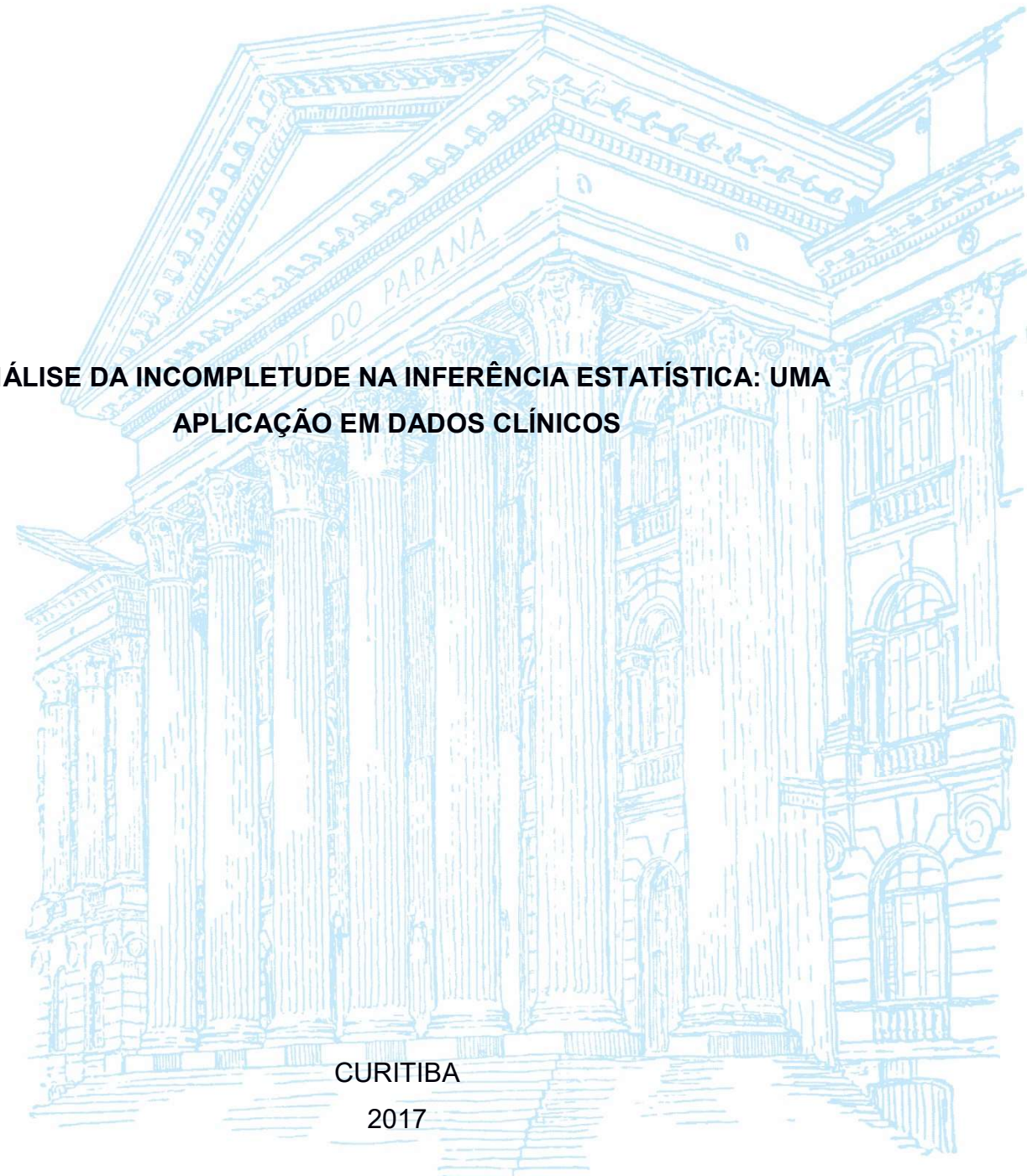
UNIVERSIDADE FEDERAL DO PARANÁ

MELISSA MELLO DE CARVALHO



# ANÁLISE DA INCOMPLETUDE NA INFERÊNCIA ESTATÍSTICA: UMA APLICAÇÃO EM DADOS CLÍNICOS

CURITIBA

2017

MELISSA MELLO DE CARVALHO

# ANÁLISE DA INCOMPLETUDE NA INFERÊNCIA ESTATÍSTICA: UMA APLICAÇÃO EM DADOS CLÍNICOS

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em bioinformática, no Curso de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica (SEPT), da Universidade Federal do Paraná.

Orientador: Prof. Dr. Geraldo Picheth
Coorientador: Prof. Dr. Paulo Afonso Bracarense da Costa

CURITIBA

2017

Carvalho, Melissa Mello de
C331 Análise da incompletude na inferência estatística: uma aplicação em dados clínicos / Melissa Mello de Carvalho. - Curitiba, 2017.
66 f.; il.

Orientador: Prof. Dr. Geraldo Picheth
Coorientador: Prof. Dr. Paulo Afonso Bracarense da Costa
Dissertação (Mestrado) Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática.
Inclui Bibliografia.

1. Redes neurais (computação). 2. Diabetes. 3. Bioinformática. I. Picheth, Geraldo. II. Costa, Paulo Afonso Bracarense da. III. Título. IV. Universidade Federal do Paraná.

CDD 570.285

TERMO DE APROVAÇÃO

**MELISSA MELLO DE CARVALHO**

"Análise da incompletude na inferência estatística: uma aplicação em dados clínicos"

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Dr. Geraldo Picheth
Universidade Federal do Paraná - UFPR

Dr. Paulo Afonso Bracarense Costa
Universidade Federal do Paraná - UFPR

Dr. João Artur de Souza
Universidade Federal de Santa Catarina - UFSC

Dr. Roberto Tadeu Raittz
Universidade Federal do Paraná - UFPR

Curitiba, 10 de março de 2017

## AGRADECIMENTOS

Agradeço primeiramente ao programa de pós-graduação de Bioinformática pela oportunidade e confiança.

Aos professores que tiveram contato comigo e com esse projeto, pelos seus conselhos e sugestões.

Agradeço ao meu orientador Prof.Dr.Geraldo Picheth e ao coorientador Prof.Dr.Paulo Afonso Bracarense da Costa pela oportunidade pelos conselhos, ensinamentos, incentivos a busca de conhecimento e ao questionamento.

Agradeço a CAPES pelo apoio ao projeto e concessão de bolsa.

Agradeço aos colegas de mestrado pelo apoio e amizade. Agradeço aos meus pais e marido pelo apoio e paciência.

Agradeço também aos muitos pesquisadores que trabalharam e trabalham questionando a incerteza provocada pela incompletude de conhecimento na inferência estatística. Pude aprender muito com suas produções científicas e suas propostas de quebras de paradigmas.

“A certeza é fatal. O que encanta
é a incerteza. A neblina torna as coisas
maravilhosas. ”

Oscar Wilde, “O retrato de Dorian
Gray” (1890).

## RESUMO

Lacunas são comuns a pesquisa empírica, em especial na área da saúde, onde a falta de dados é muitas vezes um fator inevitável devido à logística da captação de informações. Essa incompletude é danosa à análise de dados, tal como em um modelo de decisão ou de triagem clínica, pois a inferência estatística é afetada por incertezas da falta de conhecimento. As lacunas são empecilhos às análises paramétricas e a alguns softwares estatísticos. Assim, grande parte dos pesquisadores aplicam métodos de seleção, exclusão e imputação de informações faltantes. Entretanto, a prática de imputação pode não ser trivial, especialmente na presença atributos multivariados: a dificuldade de estimar valores adequados pode adicionar vieses e incertezas não desejáveis à análise de características e à decisão. Os dados analisados são provenientes do banco de dados do biorrepositório do laboratório de Bioquímica Clínica I e II da Universidade Federal do Paraná, sem incompletudes e em observação transversal em um modelo de decisão diagnóstico e de acompanhamento do *Diabetes mellitus* do tipo 2 (DM2). Todas as variáveis disponíveis à decisão têm apresentação multivariada. Para a discussão sobre os impactos e consequências da falta de dados são criadas, a partir da população, amostras com porcentagens de lacunas obtidas de maneira totalmente randômica (MCAR). A incompletude e suas incertezas são exploradas sem a imputação dos valores faltantes e os resultados comparados aos dados completos. A discussão de incertezas, vieses e distorções ocasionados pela incompletude e pelo método de análise caso completo são embasadas na aplicação da correlação τ de Kendall no software R, e análises classificatórias e preditivas com algoritmos de redes neuronais artificiais e algoritmos *fuzzy rough* do software WEKA. O τ de Kendall demonstrou ser um método de correlação robusto à análise de informações com incompletudes. As amostras com incompletudes não imputadas apresentaram poder de correlação com baixa variabilidade em relação à população do estudo e eficiente discernibilidade de características. Na classificação, o algoritmo *fuzzy rough Discernibility Classifier* demonstrou que a discernibilidade aliada ao método *fuzzy rough* é útil na classificação do modelo de decisão estudado, a alta cobertura de dados classificados demonstra sua capacidade em relação à incompletude de dados e à incerteza.



**Palavras-chave**: Dados faltantes, Correlação τ de Kendall, Conjuntos rugosos *fuzzy*, Classificadores, Redes neuronais artificiais, Diabetes.

## ABSTRACT

Gaps are common to empirical research, especially in the area of health, where lack of data is often an unavoidable factor due to the information gathering logistics. This incompleteness is damaging to data analysis, such as in a decision model or clinical screening, as statistical inference is affected by lack of knowledge uncertainties. The gaps are a hindrance to parametric analysis and some statistical software. Thus, most researchers apply selection, exclusion and imputation of missing information methods. However, the imputation practice may not be trivial, especially in the presence of multivariate attributes: the difficulty of estimating adequate values may add biases and undesirable uncertainties to the analysis of characteristics and decision. The analyzed data hails from the biorepository database of the Clinical Biochemistry I and II laboratory of the Federal University of Paraná, without incompleteness and cross-sectional observation in a model for the diagnosis and follow-up of *Diabetes mellitus* type 2 (DM2). All variables available to the decision have a multivariate presentation. For the discussion of the impacts and consequences of the lack of data, samples with percentages of gaps obtained in a totally random manner (MCAR) are created from the population. The incompleteness and its uncertainties are explored without imputation of the missing values and the results compared to the complete data. The discussion of uncertainties, biases and distortions caused by incompleteness and by the complete case analysis method is based on the application of τ-Kendall correlation in the R software, and classificatory and predictive analysis with algorithms of artificial neural networks and fuzzy rough algorithms of the WEKA software. Kendall's τ has been shown to be a robust method of correlation to incomplete information analysis. Samples with uncorrected incompleteness showed low variability correlation power in relation to the population of study and efficient characteristics discernibility. In the classification, the Fuzzy Rough Discernibility Classifier algorithm demonstrated that the discernibility allied to the fuzzy rough method is useful in the classification of the decision model studied. The high coverage of classified data demonstrates its capacity in relation to data incompleteness and uncertainty.



**Key-words:** Missing data, Correlation τ de Kendall, Fuzzy rough sets, Classifiers, Artificial neural networks, Diabetes.

## LISTA DE FIGURAS

## LISTA DE GRÁFICOS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

CDM - *Covariate dependent missingness* (covariável dependente da perda)

DM2 - *Diabetes mellitus* do tipo 2

MCAR - *Missing at complete random* (perda completamente randômica de dados)

MAR - *Missing at random* (perda randômica de dados)

MLP - *Multilayer Perceptron*

MNAR - *Missing at not random* (perda não randômica de dados)

**SUMÁRIO**

# 1 INTRODUÇÃO

Incompletude é um termo que designa a falta de dados em uma amostra. Conjuntos de dados incompletos, ou conjuntos com dados em falta, são onipresentes em todos os ramos da pesquisa empírica (MOLENBERGHS, 2009; ALLISON, 2001).

As lacunas de dados são resultado de diversos mecanismos que levam à falta de informação e trazem consequências aos processos de inferência da amostra, modificando irreversivelmente a análise, e à reprodução de resultados de maneira fidegna aos dados completos. A análise incompleta promove impactos e pode causar dependências que afetam, consequentemente, a observação de características da amostra e a tomada de decisão, gerando incertezas quanto aos resultados obtidos. Há diversos métodos estatísticos e não estatísticos que visam atenuar o entrave ao conhecimento provocado por dados faltantes, aproximando por estimativa, o resultado obtido ao resultado que seria observado na totalidade amostral. Esses métodos apresentam qualitativamente capacidade diversa de aproximação à resposta.

A complexidade, a multifatoriedade, a disponibilidade e a qualidade de informações torna a informática médica e biológica especiamente sensível à incompletude e à incerteza. A presença de lacunas em dados em saúde é por vezes inevitável. Entretanto as consequências da falta de informações são pouco conhecidas, exploradas e discutidas no contexto da pesquisa em saúde, em especial no contexto de triagem clínica diagnóstica. Há carência de estudos práticos e teóricos demonstrando os conceitos e impactos da incompletude, e esses estudos são pouco difundidos fora da estatística e pouco abordados na pesquisa científica em geral (STERNE, 2009).

Assim, a incompletude de dados é comumente trabalhada de modo equivocado por métodos de análise e por analistas. Geralmente, as lacunas tendem a ser ignoradas com a exclusão de instâncias inteiras, essa omissão pode gerar a exclusão de informações importantes para o processo de decisão, promovendo a captação de respostas que são erroneamente expandidas à totalidade da população, ou dependendo do mecanismo da perda de informação: falsas respostas, inferências inválidas, tendenciosas, enviesadas por características falsas ou distorcidas. A exclusão intencional de incompletudes sem a atribuição adequada de seus impactos é um fator muito comum em pesquisas médicas, que juntamente com a manipulação

inadequada e análises equivocadas, sugerem problemas sistemáticos na produção de evidências científicas com dados em saúde (WOOD, WHITE, THOMPSON, 2004).

Este trabalho tem o objetivo de discutir o impacto da incompletude na tomada de decisão em saúde e na inferência estatística, utilizando para isso uma amostra em triagem clínica diagnóstica e de risco do diabetes *mellitus*, dispondo para isso, de uma população amostral composta de dois grupos: um grupo sem diabetes, ou grupo controle, e o grupo portador do diabetes *mellitus* do tipo 2.

O Diabetes foi escolhido para modelo neste estudo por representar uma patologia relevante em termos de saúde pública, cuja frequência tem progressivamente aumentado em todo globo. Laboratórios clínicos, centros hospitalares e governo (Ministério da Saúde) detém bancos de dados com registros de múltiplos parâmetros de pacientes com diabetes. A análise destes bancos de dados, com objetivo de obter características para uso diagnostico ou prognóstico, com frequência colide com a incompletude de dados, o que enfraquece as análises estatísticas convencionais. Ademais, triagem clínica é um modelo de dados com características únicas e incertezas próprias onde, a complexidade dos atributos e o relacionamento multicritério entre atributos pode tornar difícil a captação de características precisos com sentido lógico.

O diferencial inovador proposto por esse trabalho está na aplicação da correlação de Kendall e de algoritmos *fuzzy rough* no estudo da incompletude em atributos multivariados sem a aplicação de imputação dos valores faltantes. Os resultados dos dados com incompletude são comparados aos resultados dos dados com completude. Dessa forma é possivel quantificar as variações, incertezas e viéses ocasionados pela incompletude, bem como a robutez do método de Kendall, e dos algoritmos empregados.

## 1.2 OBJETIVOS

### 1.2.1 Objetivo Geral

Estudar a incerteza provocada pela incompletude de dados e avaliar seus impactos na observação de características, tomada de decisão e na classificação em uma população de estudo para o *Diabetes mellitus* do tipo 2. A incerteza de decisão é abordada com o método não-paramétrico correlação τ de Kendall e a incerteza classificativa e preditiva é avaliada com algoritmos classificatórios. Os dados completos são utilizados para a criação de amostras com incompletude randomizada, que são analisadas por sua vez de duas formas: na forma de amostras com presença de lacunas e na forma de caso completo. Os dados completos são o parâmetro comparativo dos resultados obtidos.

### 1.2.2 Objetivos Específicos

a) Apresentar uma revisão conceitos da incerteza em modelos de decisão em saúde e seus tipos de incompletudes baseando-os na teoria da ignorância de Smithson (1988).
b) Enumerar características que contribuem para a incerteza e respostas enviesadas em modelos de decisão em saúde.
c) Revisar e aplicar conceitos do modelo difuso rugoso (*Fuzzy rough*) utilizada na classificação de dados com incompletudes

## 2 REVISÃO DE LITERATURA

O *Diabetes mellitus* do tipo 2 (DM2) em conjunto com doenças cardiovasculares, câncer e doenças respiratórias representam as maiores causas de mortalidade no mundo. O DM2 é uma patologia associada à resistência da ação do hormônio insulina, que promove como característica hiperglicemia crônica (SBD, 2015). É uma das principais causas de mortalidade no Brasil (VERAS, 2011), e 7 a cada 10 mortes nos Estados Unidos (BEHRA et Al. 2014). Na classificação do diabetes, o DM2 representa 90-95% dos diabéticos. A síndrome compromete a qualidade de vida de seus portadores e é um dos maiores custos à saúde pública mundial. As comorbidades associadas ao DM2 são relacionadas à hiperglicemia crônica, que acarreta lesões e disfunções em diferentes órgãos, principalmente: olhos, rins, nervos, coração e vasos sanguíneos (ADA, 2015).

As comorbidades do DM2 podem ser controladas e ter incidência evitável com a prática de medicina preventiva. Entretanto, o desenvolvimento lento, por vezes assintomático, e a grande complexidade, tornam o DM2 e outras patologias crônicas de difícil previsão (VERAS, 2011).

### 2.1 TRIAGEM CLÍNICA

As variações em práticas clínicas nas últimas décadas e os crescentes custos em cuidados de saúde deram início às implementações de triagens clínicas (SHIFFMAN, 1997). Triagem clínica é um termo restrito às áreas de bioestatística, biomatemática, epidemiologia e biometria (PIANTADOSI, 2005). Funciona como um modelo de decisão, um estudo comparativo empírico de condições distintas: as condições são grupos de dados ou classes, que podem ser avaliados por métodos estatísticos de não-estatísticos. A triagem clínica propõe a prestação personalizada de cuidados de saúde, com práticas médicas, testes, decisões e tratamentos adaptados ao nível individual do paciente, o que é útil na manutenção da saúde, prevenção e tratamento de doenças crônicas (RICH, CEFALU, 2016).

A metodologia se destaca como um bom método em pesquisa com drogas medicamentosas e na triagem de risco de patologias representando avanços importantes na qualidade de tratamentos e acompanhamentos em diferentes contextos em saúde. A triagem clínica eletrônica surgiu com o objetivo de facilitar o

rastreio de pacientes de risco à patologias e como facilitador do diagnóstico precoce de comorbidades separando perfis de fatores de risco específicos do histórico de saúde obtidos a partir do registro eletrônico (PIANTADOSI, 2005).

A triagem diagnóstica é o tipo de estudo de maior complexidade e custo dentre todos os tipos de triagem clínica. A complexidade ímpar da triagem preventiva é permeada por termos como risco benefício, julgamento de risco, epidemiologia, ética, risco individual, que contribuem para o aumento significativo de complexidade (PIANTADOSI, 2005).

De acordo com Piantadosi (2005), a triagem clínica diagnóstica tem três objetivos principais:

- Prevenção primária: intervenção das características percursoras da doença;
- Prevenção secundária: prevenção do agravamento de sequelas;
- Prevenção terciária: acompanhamento de condição clínica

A triagem clínica é adequada a um tamanho modesto de dados com variabilidade relativa e presença de viés (PIANTADOSI, 2005), que torna o possível o acompanhamento de doenças crônicas e a viável inclusão do DM2 a um programa de triagem populacional por ser passível de diagnóstico no período pré-sintomático e sintomático inicial, e por ter disponíveis exames de boa especificidade com pontos de corte precisos (NHS-UK, 2016).

A classificação dos diferentes grupos de dados é o objetivo de métodos classificadores, tais como algoritmos, a classificação promove poder preditivo a triagem clínica. A capacidade preditiva pode ser visualizada, do ponto de vista da medicina preventiva, como o poder de captar alterações condizentes a doenças até mesmo em nível assintomático e em comorbidades pouco sintomáticas (SOUZA, SCHWARTZ, GIUGLIANI, 2002), possibilitando que os acompanhamentos em saúde tenham maior precisão.

## 2.2 MODELOS DE DECISÃO

Bancos de dados com desfechos binários, como é a amostra de ensaio clínico escolhida para a prática desse trabalho, permitem o rastreio de pacientes com patologias possibilitando diagnósticos, achados epidemiológicos e identificação de grupos de risco no âmbito do diagnóstico precoce (PIANTADOSI, 2005).

A teoria da decisão descreve modelos que se agrupam com o objetivo de obter informações, nem sempre perceptíveis, de maneira rápida e dinâmica. As informações relevantes são selecionadas por associações classificativas, e tem aplicabilidade em diversas áreas do conhecimento (TCHEMRA, 2009). Métodos para a tomada de decisão dependem da complexidade de um problema de seus critérios e de suas alternativas de solução. Os critérios ou atributos, são entendidos como fatores que guiam a decisão e as alternativas são cursos de intenção de ações para resolver um problema (RAGSDALE, 2010).

A dinâmica e a complexidade dos processos decisórios exige o desenvolvimento de metodologias e de técnicas que auxiliem a tomada de decisão (TCHEMRA, 2009). Essa complexidade gera múltiplas associações entre dados que criam a necessidade de critérios para a definição coesa de uma decisão. A discernibilidade é o fator que define a decisão (NASIRI, MASHINCHI, 2009).

A triagem clínica de condições complexas que necessitam de muitos atributos para o acompanhamento e o diagnóstico, tal como a triagem do DM2, são modelos de decisão em multicritério. Os métodos multicritério utilizam técnicas que oferecem maior compreensão dos problemas de decisão com número finito de critérios e alternativas que apoiam a decisão nos processos decisórios de contextos multidisciplinares (TCHEMRA, 2009; FÜLÖP, 2005). Essa característica de multifatorial, campo da decisão multicritério, exige do decisor a avaliação de cenários possíveis e da avaliação de resultados quanto a sua em comunicação lógica com as hipóteses (GOMES, GOMES E ALMEIDA, 2002).

Modelos preditivos, triagem de fatores de risco, estudos de custo benefício obtidos por métodos estatísticos e não estatísticos são capazes de guiar decisões, mas se recomenda cautela na utilização de modelos de decisão, principalmente os computacionais aplicados sem o auxílio da estatística, já que dependendo da capacidade do método escolhido frente a características dos dados analisados tais como: complexidade da associação de informações e presença de dados faltantes,

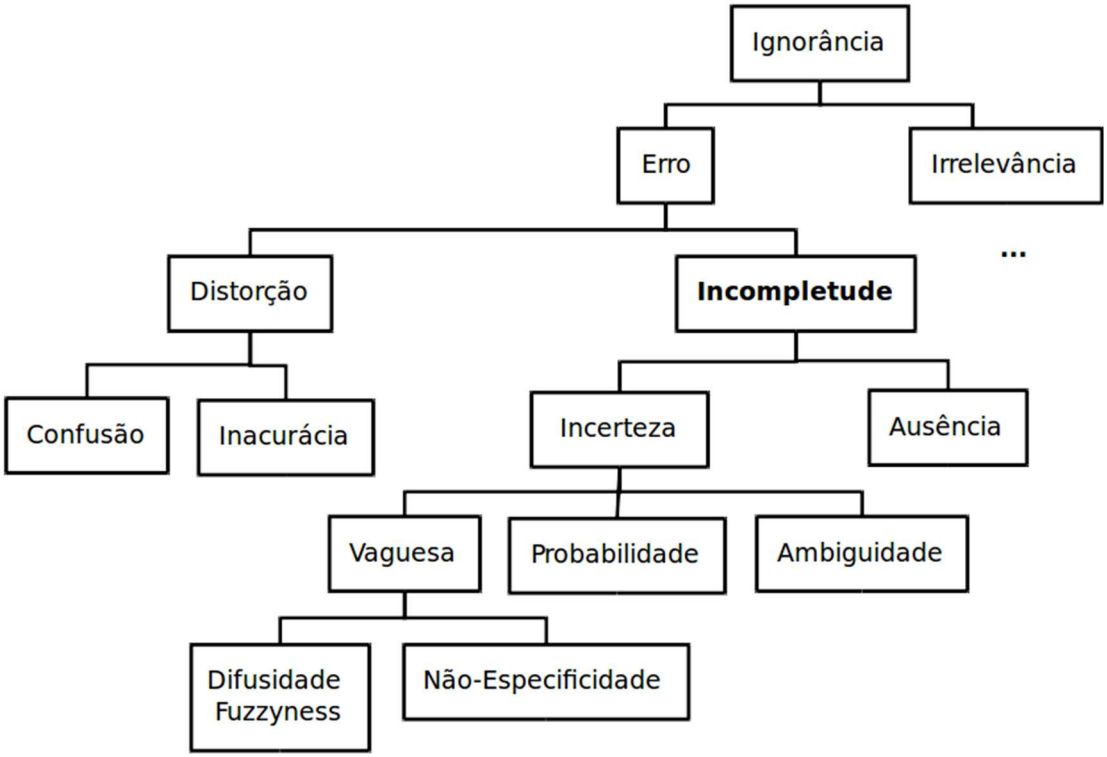
podem levar a erros na interpretação de resultados (WOOD, WHITE, THOMPSON, 2004). Os critérios podem ser conflitantes, gerando várias ações possíveis, o que resulta em dificuldades de identificação de impactos e incertezas.

## 2.3 INCERTEZA

Como dito anteriormente, triagem clínica é um modelo de decisão complexo e em multicritério. Sendo assim, é significativa a chance de conter erros, que podem ocorrer de diversas formas, sendo aferida por erros intrínsecos dos dados, erro do operador e erro do equipamento. A proporção da incerteza depende da natureza do mecanismo da falta de dados (MOLENBERGHS,2012).

O contexto de incerteza estuda neste trabalho é relacionada a ignorância de conhecimento proveniente da teoria da ignorância de Smithson (1988). De acordo com Smithson (1988), a incerteza provém da incompletude ocasionada pelo erro que é produto da ignorância do conhecimento. O autor elaborou uma taxonomia da ignorância de conhecimento, como demonstrado na FIGURA 1.

FIGURA 1 – TAXONOMIA DA INCERTEZA




FONTE: adaptado de SMITHSON (1988).

A incompletude explorada na classificação de Smithson (1988), parte do erro na obtenção do conhecimento, que pode ser puramente pela ausência de dados, pela inobservância de informação faltante, ou por erro de causa incerta. A incerteza pode estar presente tanto numa amostra completa quando em uma amostra incompleta, sendo que na amostra completa a incerteza é relacionada à qualidade da informação disponível e à informação deturpada: podem ser informações vagas, não-específicas, ou ambíguas, e na amostra incompleta é relacionada a informação desconhecida ou ausente. Os erros provenientes dessas condições atrapalham a captação de padrões e comprometem a classificação dos resultados. Byron e Brown (1980) descrevem a incerteza nas análises estatísticas como proveniente do *design* de triagens clínicas, para os autores, as controvérsias da incerteza estão relacionadas à escolha de informações e à forma com que são analisadas.

A incerteza por incompletude de informação representa um problema que pode causar viés ou levar a análises ineficientes. Dados incompletos são frequentes na pesquisa empírica, em ensaios clínicos randomizados em publicações médicas e biomédicas (WOOD, WHITE, THOMPSON, 2004), na investigação clínica e epidemiológica (SHIFFMAN, 1997), e em alguns tipos de modelos de decisão, o que ocorre devido à complexidade e ao multicritério do processo decisório. Nesse caso a escolha de informações específicas e coesas podem reduzir a incerteza e promover a decisão.

A ausência de dados, por sua vez, tem impactos relacionados à causa de sua falta e a mecanismos de dependência com informações presentes. As lacunas trazem problemas para a inferência estatística de conjuntos de dados por desequilibrar a matriz de dados (MOLENBERGHS, 2009), o desequilíbrio da matriz de dados, por sua vez, acresce níveis de incertezas relacionados a confiabilidade dos resultados observados, que é proporcional à dimensão da falta de dados.

Esses fatores, advindos da falta de dados, motivam muitos pesquisadores a evitar a abordagem de informações faltantes, mantendo na análise apenas os casos completos, há também a preferência por abordagens que ignoraram as informações faltantes com a utilização de sistemas tolerantes a incompletude. A principal razão dessa preferência, segundo Sterne et Al., (2009) está na dificuldade de acesso e de conhecimento de métodos estatísticos que possam resolver problemas decorrentes da falta de dados. O manejo inadequado ou equivocado de dados faltantes também é comum em pesquisas em saúde (WOOD, WHITE, THOMPSON, 2004), e como

resultado podem gerar análises com viés substancial em estimativas de ensaios clínicos.

A exclusão de dados em falta em diversas variáveis requer também a exclusão de dados presentes, como a exclusão de linhas inteiras, o que pode gerar um efeito cumulativo de exclusão com a consequente perda substancial de precisão e de poder estatístico (STERNE et Al, 2009). Essa exclusão proposital de dados consiste em uma metolodologia conhecida como análise de caso completo (MOLENBERGHS, 2004), ou análise de exclusão em lista (*Listwise*). Essa forma é muito comumente empregada no manejo de dados com lacunas, e será abordada ao longo desse trabalho.

A incompletude merece consideração com análises adequadas e descritivas, já que a simples exclusão de incompletudes pode gerar resultados tendenciosos. Assim, a exclusão de lacunas e de informações associadas as lacunas de uma amostra pode levar a danos na inferência estatística superiores aos de amostras onde as incompletudes são preservadas. Dados incompletos têm potencial, assim como os dados presentes, para minar a validade e quantificar resultados de investigações em saúde (STERNE et Al, 2009). A precisão das respostas das análises com dados faltantes e com dados não-faltantes pode ser avaliada por testes de sensibilidade. Os testes de sensibilidade dão validade a resultados e justificam seus achados (WOOD, WHITE, THOMPSON, 2004).

## 2.4 TAXONOMIA DA INCOMPLETUDE

Por muito tempo, a falta de dados poderia ser descrita como o "pequeno segredo" das estatísticas, um tabu na pesquisa empírica. Embora comum a todo tipo de amostra, há pouca literatura para orientação teórica ou prática. Como resultado, as razões obscuras dessa reticência criam um falso caráter de certeza em métodos (ALLISON, 2001). Como se vê, partindo dessa afirmação, os métodos mais utilizados para lidar com dados ignorando as lacunas podem ter sérias deficiências.

Como citado anteriormente, a incompletude por falta de dados tem diversas tipologias de distribuição na amostra (1), bem como causas e consequências diversas (2) à inferência e a obtenção de resultados. Essas faltas de dados podem ser nomeadas de forma taxonômica.

1 - Dados faltantes podem estar distribuídos na amostra da seguinte forma:

a) Falta não monótona: dados que estão faltando para algumas variáveis e para alguns casos.
b) Falta monótona: dados faltantes em uma variável em todos os casos, pode se tratar também de uma variável latente não observada.
c) Falta por unidade sem resposta: dados faltantes em todas as variáveis para alguns casos (ALLISON, 2001).

2 - Dados faltantes podem ter diferentes consequências e causas. Rubin et al. (1976), define três formas básicas de incompletude:

a) Incompletude completamente randômica (MCAR): onde a probabilidade de um valor estar faltando é independente de dados observados e não observados;
b) Incompletude randômica (MAR): gera um condicionamento nos dados observados, a probabilidade e os valores em perda são dependentes dos dados não observados;
c) Incompletude não randômica (MNAR): em que a probabilidade de valor estar ausente depende de dados observados e não observados.

A ausência de dados e suas diversas tipologias provocam diferentes impactos na resposta da análise. Portanto, o mecanismo que leva a ausência de informações em uma amostra precisa ser conhecido para que os resultados das análises possam ser interpretados. Para isso, como dito anteriormente, a verificação de respostas com base em testes de sensibilidade é fundamental. Em incompletude não aleatória, não randômica, por exemplo, só pode ter seus erros percebidos por análises de sensibilidade e com hipóteses sobre o mecanismo de dados em falta (STERNE et Al, 2009).

Nesse trabalho, a amostra foi randomizada de forma a apresentar distribuição da incompletude não-monotona totalmente randômica (MCAR) em variáveis, as consequências da incompletude realizada na amostra e suas incertezas serão analisadas e discutidas ao longo do trabalho.

## 2.5 MÉTODOS DE INFERÊNCIA ESTATÍSTICA NA INCOMPLETUDE

Métodos estatísticos convencionais e quase todos os softwares estatísticos presumem que todas as variáveis de um modelo de dados são medidas em todos os casos (ALLISON, 2001; WOOD, WHITE, THOMPSON, 2004). Assim, as lacunas produzem um empecilho à análise de amostras inteiras.

Para “contornar” o problema da falta de informação, existem alguns métodos aplicáveis à mineração de dados, são eles: o descarte de dados ligados as lacunas (análises por caso completo ou *Listwise* e *Pairwise*) e a criação de informações por meio de imputação de dados: imputação simples, múltiplas imputações, imputação por última observação realizada (*Last observation carried foward* - LOCF), entre outras (MOLENBERGHS, 2004; ALLISON, 2001; HORTON, KLEINMAN, 2012, SCHAFER, 1997).

Alguns métodos adequam-se em certa medida à falta de dados, vale listar: a estimativa maximizada, análise de máxima verossimilhança; análise probabilística; métodos que ignoram lacunas para fins de cálculo e algumas estatísticas não-paramétricas (SIEGEL, S. CASTELLAN, J, 2006; KARPENTER, J. KENWARD, 2007, HORTON, KLEINMAN, 2012, ALLISON, 2001).

Descartar dados implica no desperdício de recursos financeiros e de esforço no recolhimento de informações. Assim, muitos métodos para “recuperação" dos casos com dados incompletos se tornaram populares (ALLISON, 2001). Entretanto, essas abordagens são complicadas quando existem muitos padrões de valores faltantes ou quando estão envolvidas variáveis aleatórias categóricas e contínuas (HORTON, KLEINMAN, 2012, SCHAFER, 1997).

A imputação de dados faltantes aplicada à falta de dados do tipo não randômica (MNAR) pode levar a resultados enganosos por gerar conflitos com preditores. Quando os dados estão com incompletude randômica (MAR) o viés em análises baseadas em imputação múltipla pode ser tão grande quanto ou maior que o viés em análises em caso completo (STERNE et Al, 2009). Tendo essas ressalvas, métodos de imputação de dados precisam ser aplicados cuidadosamente de modo a evitar conclusões equivocadas.

Testes paramétricos não são aplicáveis à totalidade amostral perante a incompletude (STERNE et Al., 2009). Por que a aplicação de métodos de imputação ou de métodos que ignoram lacunas promove análises parciais e estimadas que

possivelmente conterão distorções. Essas distorções podem levar a resultados tendenciosos. Mesmo o preenchimento das lacunas com zeros podem refletir resultados distorcidos; os zeros computados na análise podem criar tendências ou padrões relacionado ao seu mecanismo de repetição, com grande quantidade de lacunas, por exemplo, o zero viria a ser uma moda.

A maioria dos métodos não-paramétricos não requerem completude para a demonstração de seus resultados, são flexíveis à presença de incompletudes e uma alternativa aos testes paramétricos. Dados faltantes podem ser melhor interpretados pela correlação de métodos não-paramétricos (BURKE, 2001), sendo ideais para explorar o conhecimento incerto sem a necessidade da criação de informações artificialmente. Entretanto, quando aplicados a incompletude do tipo randômica (MAR) constituem apenas estimativas que devem ser aparadas por testes de sensibilidade (KARPENTER, KENWARD, 2007).

Mesmo com a alternativa facilmente aplicável a dados com lacunas, estatísticas não-paramétricas ainda são pouco utilizadas por pesquisadores que preferem o uso de imputações de dados para a aplicação de métodos paramétricos.

### 2.5.1 Descarte de informações: análise de caso completo

A inviabilidade da análise de dados incompletos por alguns métodos estatísticos tradicionais e vários softwares estatísticos faz com que grande parte dos pesquisadores simplesmente retirem os casos de qualquer dado faltante das variáveis de interesse. Esse método é conhecido como exclusão de toda linha (*Listwise*) ou ainda aplicação de caso completo, ele é considerado o método convencional mais amplamente utilizado para a análise de dados com incompletude. Propõe uma análise de exclusão radical de informações da amostra (ALLISON, 2001).

Em uma análise de caso completo inclui-se apenas casos em que todas as medidas foram tomadas (MOLENBERGHS, et Al. 2004) omite aqueles para os quais os dados estão incompletos (MA, RAINA, BEYENE, THABANE, 2012). A análise de caso completo com a exclusão de instâncias com dados faltantes ocorre como demonstrado na FIGURA 2.

FIGURA 2 – APLICAÇÃO DO MÉTODO CASO COMPLETO

| N | Atributo A | Atributo B | Classe |
|---|---|---|---|
| **1** | 25 | -2 | T |
| **2** | 4 | -9 | F |
| **3** | 21 | -2 | T |
| **4** | 7 | 1 | T |
| **5** | 25 | _ | T |
| **6** | 6 | -9 | F |
| **7** | 29 | 12 | T |
| **8** | 15 | 3 | F |
| **9** | 22 | -8 | F |
| **10** | 10 | -4 | F |
| 11 | _ | -22 | T |

FONTE: a autora (2017).
LEGENDA: As linhas destacadas correspondem às instâncias excluídas na aplicação de caso completo.

Na estrutura deste método, os dados não diferem do experimento completo, então análises estatísticas podem ser realizadas. Mas, infelizmente, esse método possui diversas desvantagens. Em triagem clínica, quase sempre há uma perda substancial de eficiência, que resulta da remoção de informações de indivíduos parcialmente observados. A exclusão de linhas muitas vezes exclui uma grande fração da amostra, levando a perda de poder estatístico porque menos informação é utilizada. De acordo com Allison (2001) erros padrão tendem a ser maiores na amostra com dados eliminados em lista em comparação com dados submetidos a outras metodologias de análise.

Essa ineficiência do modelo forma um subconjunto potencialmente muito pequeno e que não representa o todo, o que inviabiliza a observação de padrões de variáveis que precisam ser comparadas para tal. Para Karpenter e Kenward (2007), a perda de precisão na aplicação do método caso completo é considerável, o que torna as estimativas tendenciosas e as inferências inválidas, além disso, a incerteza ocasionada pela exclusão de dados em lista acarreta erros que são manisfestados por desvios severos.

A excessão a esses efeitos danosos ocorre quando aplicada a uma incompletude do tipo completamente randômica (MCAR), e somente nesse tipo de incompletude: não há ocorrência de viés perceptível nos resultados. As exclusões do método levam a análises tendenciosas em incompletude do tipo randômica (MAR) devido à presença de dependências entre os dados ausentes e os dados presentes (MOLENBERGHS, et Al. 2004). Na incompletude do tipo randômica o nível de

incerteza é proporcional ao tamanho da incompletude ocasionada pela informação deletada (ALLISON, 2001). E na análise com incompletude não randômica (MNAR) revela-se como um modelo de análise incapaz de predizer, mesmo que grosseiramente, o todo dos dados (HOSSAIN, ORDAZ, BARTLETT, 2016; ALLISON, 2001; MA, BEYENE, THABANE, 2012). Portanto, a análise de caso completo não é adequada quando a incompletude gera dependências (MOLENBERGHS, et Al. 2004).

Ao se escolher a exclusão de linhas com incompletudes na amostra, na presença de incompletude com dependências, se despreza potencial estatístico que as instâncias incompletas teriam, produzindo maior quantificação de distorções, menores níveis de confiabilidade e maior distanciamento da análise global da amostra.

Frenquentemente triagens com aplicação de caso completo tem associado outros tipos de seleção e exclusão de informações, tais como a exclusão de *outliers*, extremos *outliers e* balanceamento quantitativo de agrupamentos. Essa exclusão reflete, mesmo que não intencionalmente, a preferência por dados centrais, uma análise de média que não pode ser expandida ao todo dos dados, tão somente um resultado parcial. Valores extremos podem sugerir populações diferentes, ou erro de medição. Com a possibilidade de erro de medição excluída se aconselha, para contornar esse problema, a análise de dados em grupos separados, tais como clusteres. Remover valores extremos automaticamente é prática comum, mas eles podem alterar as estatísticas calculadas, como diminir a estimativa da variância, e criar viés na média calculada (BURKE, 2001).

Com as considerações anteriormente apresentadas, por suas aplicabilidades amplas e por sua capacidade de aumentar distorções e comprometer a observação padrões e o caso completo é a metodologia de tratamento de dados faltantes escolhida nesse trabalho para a avaliação da capacidade de métodos de análise frente a esse acréscimo de incerteza.

Um método de exclusão sugerido como alternativa à exclusão em lista é a exclusão em pares (*Pairwise*) que tem a vantagem de excluir informações de forma a excluir também as dependências (ALLISON, 2001). A exclusão de pares não será abordada nesse trabalho.

### 2.5.2 Testes não-paramétricos

Nesse trabalho a escolha por demonstrar resultados pela aplicação do método de correlação τ Kendall (α). A correlação de Kendall e os testes de Spearman são uma alternativa não-paramétrica à coeficientes e correlações paramétricas (BURKE, 2001), tal como o correlação linear de Pearson.

A correlação τ Kendall atribui posições rankeadas ordem de preferência aos atributos que compõem uma amostra com grupos diferentes. Mede a correspondência entre rankings de dados e a distância entre os grupos da classe (ABDI, 2007). A classificação do τ kendall ocorre de forma exata, mesmo sem uma classificação objetiva prévia, aferindo a compatibilidade de duas classificações. Seus potenciais incluem a capacidade de rankeamento e de distinção de classificações mesmo com variáveis muito semelhantes. A significância de τ, expressa a compatibilidade das classificações. Se a correlação é negativa, significa que os dois rankings são significativamente incompatíveis (KENDALL, 1938).

Os aspectos da análise de amostras com dados faltantes por métodos de correleção são pouco conhecidos por pesquisadores e pouco estudados quanto ao "comportamento" da correlação na classificação de dados incompletos, e muitas vezes as informações incompletas são ignoradas com a aplicação do caso completo (ALVO, CABILIO, 1995). A correlação de Kendall generaliza informações faltantes em sua matriz de modo a possibilitar cálculos e estimativas de correlação com a presença da incompletudes. Os resultados generalizados pela falta de informação ocasionam perda de poder estatístico. Em consequencia disso, muitos pesquisadores que trabalham com a correlação de kendall em dados faltantes imputam as informações faltantes.

A aplicação do τ de Kendall de maneira semelhante à proposta foi realizada por outros pesquisadores, que aplicaram a correlação de Kendall a amostras com e sem prévia imputação de dados faltantes. Alvo e Cabilio (1995) propõem, que na presença de incompletude de dados, as medidas de correlação rankeada se baseiam em uma noção de distância entre rankings incompletos, e consideram que a aferição dessas distâncias possui potencial para um aumento significativo da eficiência em relação à abordagens que ignoram as observações faltantes como no caso completo.

Segundo Cabilio e Tilley (1999) análises que não ignoram as lacunas de dados apresentam vantagens em relação as que excluem lacunas da análise: as

estatísticas que utilizam as informações sobre as lacunas de dados parecem funcionar substancialmente melhores do que as estatísticas que ignoram a presença de observações faltantes. Assim, os dados faltantes geram diferenças pequenas no poder correlação de Kendall, o que torna o método confiável de acordo com os autores.

Ma (2012) aplicou a correlação de Kendall a um modelo de longitudinal de dados de estudo preventivo do HIV e considerou bons os resultados da correlação na incompletude do tipo MAR e MCAR. O autor sugere que na presença dados faltantes de origem multivariada, como é o caso de dados de expressão gênica e dados bioquímicos, a obtenção de valores adequados a uma imputação robusta pode ser muito dificultada, e nesses casos a generalização das lacunas pode ser uma alternativa de análise pelo τ de Kendall em corte longitudinal.

O método de Kendall utilizado no software R é flexível a incompletude, aceita os dados com incompletude excluindo os dados incompletos da matriz de correlação de dois atributos comparados, assim a incompletude é pontual a cada correlação e a cada tamanho de incompletude e afeta os resultados de correlação separadamente atributo a atributo. Como a amostra abordada neste trabalho possui atributos bioquímicos e antropométricos em forma multivariada, onde a imputação de valores faltantes seria difícil de se obter de forma lógica, optou-se por verificar e quantificar as diferenças de correlação atributo por classe de cada amostra com incompletudes comparada às respostas de correlações atributo por classe da população de estudo sem incompletude.

Essas características fazem do τ Kendall o método de escolha para a análise dos dados amostrais neste trabalho, as análises são realizadas sem a imputação de lacunas e respostas obtidas são analisadas juntamente com testes de sensibilidade p-valor unilateral.

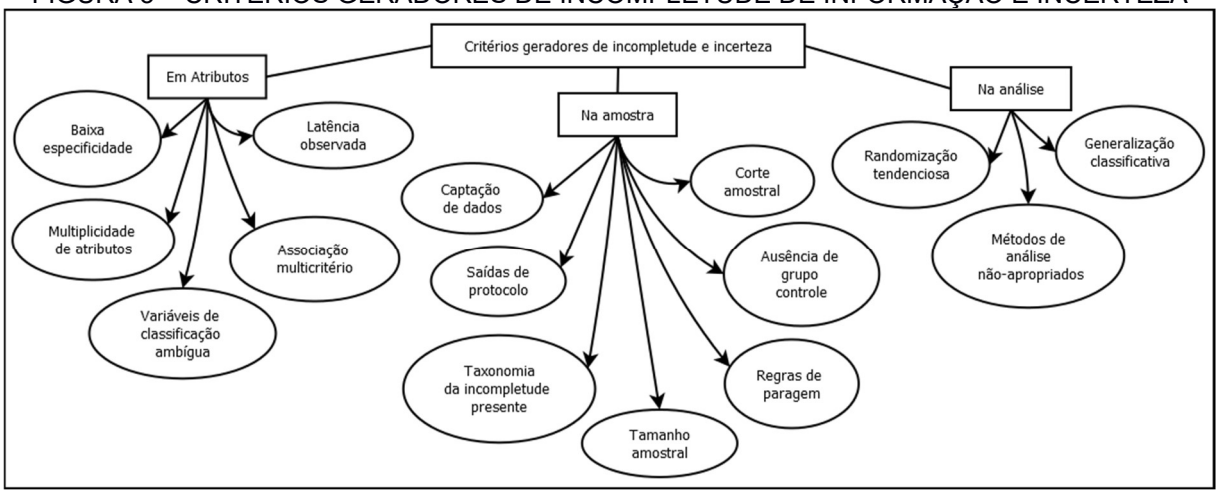
## 2.6 CRITÉRIOS GERADORES DE CONFUSÃO E INCERTEZA EM TRIAGEM CLÍNICA

Os critérios de incerteza e incompletude adotados nesse trabalho são referentes a visão da incerteza como fator da ignorância de conhecimento da classificação de Smithson (1988).

São vários os motivos para a existência de incertezas e incompletudes de informação nas diversas tipologias de triagem clínica. Segundo Armitage (1984) fatores como: randomização de amostra; saídas do protocolo; forma de registro do paciente; multiplicidade de exames; ensaios cruzados; multiplicidade de relações causa-alteração; combinação de padrões de resultado; e os modelos para ensaios clínicos, são fatores que promovem confusão na classificação em triagem clínica (ARMITAGE, 1984).

Com base nos fatores que ocasionam, influenciam e são influenciados pela incompletude e pela incerteza criou-se uma listagem de alguns de seus fatores geradores em atributos, na amostra e na análise, como demonstrado na FIGURA 3.

FIGURA 3 – CRITÉRIOS GERADORES DE INCOMPLETUDE DE INFORMAÇÃO E INCERTEZA




FONTE: a autora (2017).

## 2.6.1 Incompletude e incerteza em atributos

A escolha de atributos deve ser relacionada à sua importância classificatória, assim, a seleção de atributos é uma etapa importante para a tomada de decisão. A seleção de dados deve ter como objetivo a redução e a quantificação de dados não classificados, o que reduz a quantidade de ruídos e promove classificações mais concisas. Pode-se concluir assim, que exames pouco específicos empregados no diagnóstico podem gerar classificações falso positivas ou falso negativas.

A latência observada e a presença de atributos de classificação ambígua, ou irrelevantes também contribuem para a presença de elementos não-classificáveis. O exame glicose jejum, por exemplo, é ambíguo ao não ser capaz de discernir os grupos

diabética e não diabética com precisão absoluta: há presença de falsos positivos e falsos negativos, uma glicose jejum normal pode corresponder tanto a um paciente não diabético como a um paciente diabético em situação de bom controle glicêmico. Variáveis com classificação ambígua podem ser chamadas de variáveis de confusão.

A latência como termo estatístico consiste de duas tipologias: a latência observada e a latência não observada. A latência não observada pode ser tratada como um tipo de 'preditor' de grupos, e a latência observada corresponde a atributos de dados com associação a outras características ou com associação a outros atributos observáveis (LEE, 2006; MASTELLA, 2015). A latência observada relacionada a outros atributos também observáveis e tem tendência a gerar inconsistências classificatórias de dados (ALLISON, 2001). Atributos irrelevantes, atributos redundantes, e ambíguos são correlacionados parcial ou completamente, como resultado, afetam a precisão dos classificadores, portanto, devem ser eliminados (LEE, 2006), juntamente com atributos com latência observada, ambíguos e atributos irrelevantes à classificação. A seleção dos atributos nos dados de estudo foi realizada com a exclusão de elementos com latência observada como: PES (peso), ALT (altura), latentes ao IMC, NHDL (colesterol não HDL), e AIP (Relação triglicerídeos por HDL-colesterol).

A triagem do diabetes e suas comorbidades é um exemplo de triagem complexa: formada por múltiplos atributos que se relacionam em multicritério de forma a prover o diagnóstico. A necessidade de vários exames está relacionada as fases de diagnóstico inicial e a percepção de comorbidades, facilita o diagnóstico precoce e a prevenção de agravamentos da síndrome. Entretanto, alguns atributos podem provocar confusão classificatória ao não terem a capacidade de distinguir bem o grupo diabético do grupo não diabético. Uma coleção de variados tipos de atributos, uns mais fortes e outros mais fracos do ponto de vista classificatório aumentam a quantidade de desvios, ou ruídos (BRACARENSE COSTA, 1999). Estes atributos podem diminuir a discernibilidade quando se compara dois grupos muito distantes como o diabético grave pouco controlado com um grupo sem diabetes e com boa saúde, por exemplo.

Para evitar esse problema e manter um elevado número de atributos, se recomenda a seleção de grupos próximos como diabéticos com mal controle e diabéticos com bom controle para a visualização de exames úteis a aferição de

comorbidades, ou ainda, grupos diabéticos com o mesmo tipo de diabetes e tempo semelhante após o diagnóstico.

Ao se escolher atributos com melhor afinidade classificatória complexidade reduz com consequente diminuição do erro estatístico (BRACARENSE COSTA, 1999). A inclusão ou seleção de covariantes relevantes pode reduzir a heterogeneidade e permitir recomendações terapêuticas mais específicas (DERSIMONEAN, LAIRD, 1986). O DM2 é uma síndrome de complexidade relacionada a fatores ambientais e genéticos, além de relação de interdependências entre valores de exames. Assim, a diferença na quantidade de informação necessária para triar grupos a quantidade de informação disponibilizada pela amostra pode gerar dependências no processo decisório (LIPSHITZ, STRAUSS, 1997). A associação multicritério de exames e características antropométricas entre outros fatores promovem aumento da complexidade do sistema de triagem o que aumenta a incerteza da amostra. A multiplicidade da relação causa-alteração pode gerar uma complexidade difícil de ser elucidada principalmente se não há atributos fortemente específicos no estudo. Assim, a escolha de atributos específicos é essencial para a qualidade dos cuidados de saúde e para apoiar o uso eficiente de recursos limitados em saúde (TUNIS, STRYER, CLANCY, 2003).

### 2.6.2 Incerteza e incompletude na Amostra

Toda a amostra possui incertezas intrínsecas às suas informações e a incompletude promove aumentos da incerteza (BYRON, BROWN, 1980). O domínio do conhecimento requer a escolha adequada de processos para a modelagem e seleção de amostras representativas (SHIFFMAN, 1997). Assim, a captação de dados com quantidade e qualidade adequada com atributos com boa especificidade é fundamental a obtenção de resultados mais precisos a respeito da condição de interesse de estudo. A aquisição do conhecimento em problemas decisórios depende das informações coletadas que auxiliam na tomada de decisões (RAIFFA, 1977).

Outro fator importante que deve ser observado na captação de dados além da forma é a quantidade, o corte amostral define essa quantidade de informações suficientes para a representatividade da análise. A observação transversal, uma observação no tempo, é adequada com a disponibilização de exames de forte especificidade; já o longitudinal, várias observações no tempo, tem menos níveis de

incerteza e menor quantificação de erros, é adequado a diagnósticos complexos que exigem a aferição de variáveis em um determinado tempo para a sua confirmação.

Mais de uma observação no tempo de um mesmo indivíduo em triagem clínica requer o uso de estatísticas diferenciadas de conhecimento menos acessível, esse fator faz com que pesquisadores demonstram maior interesse por dados em corte transversal e com completude (HIRAKATA, 2009). O método de análise transversal de dados necessita, para ser coerente, de uma escolha de classe compatível formada com uma única observação no tempo. Para isso, as condições da classe devem ser completamente explicáveis pelos dados coletados não necessitando de mais coletas de dados para os mesmos atributos. A classe não pode ser ambígua para minimizar incertezas de classificação e ser tomada em condições concisas aos atributos. Deve haver pelo menos um atributo com poder de discernir as variáveis da classe, ou seja, um ou mais atributos com capacidade de formar a lógica de classificação das variáveis e o atendimento do objetivo de triagem. Quanto menor o poder discriminatório e classificatório que atributos isolados possuem para a condição estabelecida na triagem mais grosseira será a estimativa, e maior será o nível de incerteza dos resultados da amostra, portanto, menor confiabilidade. Além disso, medidas de erros são mais difíceis de serem estimadas em corte transversal.

A síndrome *diabetes mellitus* é um bom exemplo de condição onde o corte longitudinal é capaz de reduzir significativamente a incerteza classificatória quanto as suas alterações de longo prazo. Todavia, dados longitudinais podem ser difíceis de se obter, especialmente em condições de alta especificidade (ARMSTRONG & COLLOPY, 1992).

A triagem pode ser amparada por grupos de controle ou caso controle, o que fornece informações complementares aos resultados dos ensaios clínicos, e diminui a incerteza quanto à classificação de dados. Estudos de caso-controle podem ser o melhor cenário de estudo disponível para avaliar efeitos adversos raros e estudos de banco de dados grandes. Além de fornecerem informações sobre a extensão do alcance de efeitos esperados em ensaios clínicos randomizados. Entretanto, estudos de caso controle não são isentos da possibilidade de possuírem fatores de confusão por conceitos equivocados (KUNZ, OXMAN, 1998), sendo assim, o grupo controle deve demonstrar características que o classifiquem e que permita a classificação de outros casos. Nesse estudo está presente um grupo controle com DM2 que serve como base comparativa para a extração de características.

Outro fator ligado a incompletude informação na amostra são as dependências de dados ausentes ou insuficientes. A dependência entre várias variáveis e classe para a obtenção de resultado cria padrões, desta forma, quando há dados ausentes a dependência de informações gera desvios na acuidade classificatória, essa característica de relação de dependência é denominada de covariáveis dependentes da perda (CDM) (HOSSAIN, ORDAZ, BARTLETT, 2016).

### 2.6.3 Incerteza na análise

A randomização da amostra define a forma da triagem clínica (PIANTADOSI, 2005). E pode ser um fator tanto necessário como danoso à observação de resultados da amostra. A randomização cria duas consequências principais: alocação aleatória adequadamente escondida, e alocação aleatória inadequadamente escondida. Isso é, dependendo dos critérios utilizados à randomização, podem ocorrer a ocultação de informações essenciais ao mesmo tempo que ocorre a diminuição da força classificatória, essa é a dita randomização tendenciosa. As modificações de hipóteses produzidas pela randomização tendenciosa produz desequilíbrios nos resultados de prognósticos, por exemplo, além de possíveis subestimações da eficácia de intervenções práticas, viés de julgamento, falha de diagnóstico e desconexão ao grupo controle, fatores que criam incerteza e erro (KUNZ, OXMAN, 1998). Amostras pequenas randomizadas e não randomizadas tem um risco maior de resultados tendenciosos. A randomização deve ser realizada com cautela para evitar tendências. Qualificar e separar grupamentos de dados por semelhança não pode ser um processo automatizado, requer julgamentos e considerações apropriadas a construção de um problema formal da decisão para a avaliação de decisores e interpretação de resultados de análise (LIPSHITZ, STRAUSS, 1997). Mas, quando a ausência de dados é grande, a incompletude destrói justificativa de randomização do teste por afetar a representatividade de exames (MOLENBERGHS, et Al. 2004).

A não-randomização, defendida por alguns autores, pode: promover a validação de certos testes estatísticos, prover base para inferência, dificultar o mascaramento de informações e equilibrar grupos de comparação (PIANTADOSI, 2005), evitando o acréscimo de incertezas evitáveis.

Para concluir, a randomização é um retrato parcial da amostra, fruto de várias possibilidades de combinação de dados e um conhecimento que parte dessas

combinações. Se a amostra é novamente randomizada aleatoriamente outros dados são sorteados e poderão se combinar em padrões de forma diferente, já que a randomização pode criar tendências na amostra. Sendo assim, a randomização gera incertezas, e esse conhecimento parcial não pode ser expandido a totalidade da amostra.

Em séries temporais as generalizações de informações, de forma semelhante à randomização, podem facilitar ou dificultar a observação de características de dados, a qualidade da análise depende do grau de generalização empregado. A generalização é um método que visa aumentar a robustez da análise em amostras com dados que geram confusão, erro, ou distorção na percepção de padrões. Em generalizações excessivas podem levar a uma super ou a subestimação de resultados. Como dito por Kurt Gödel (1906 – 1978):

“Todas as generalizações – exceto esta – são falsas” (GÖDEL, K. 1906-1978).

Para concluir, sugerimos como alternativa a modelos de exclusão de dados e como forma de minimizar incertezas as seguintes etapas:

- Verificar se o objetivo classificatório é adequado ao tipo de corte de dados;
- Verificar prováveis incertezas na amostra por meio de estatística descritiva de dados;
- Analisar dados por padrões e separar os padrões muito distantes de modo a atender a análise global com menor quantificação de erros;
- Excluir de atributos e classes com latência observada e com ambiguidade;
- Dar preferência a exames com alta especificidade aos grupos da amostra.

## 2.7 MÉTODOS NÃO ESTATÍSTICOS

Os métodos testados nesse trabalho são apresentados na sequência. Métodos não-estatísticos são empregados através do uso de algoritmos com o objetivo de fornecer resposta ao modelo classificatório promovendo a decisão. Os métodos de classificação consistem em distribuir cada alternativa na categoria apropriada, ou grupo apropriado, de acordo com um valor associado à alternativa, o que é realizado pelas funções de funcionamento dos métodos. Escolhemos nessa etapa três algoritmos com funcionamentos distintos que serão aplicados de forma empírica às amostras com o objetivo de eleger o algoritmo com melhor capacidade de distinção de grupos classe e de inclusão dos dados ausentes a classificação dos grupos.

### 2.7.1 Teoria dos conjuntos rugosos

A teoria do conjunto áspero, ou conjunto rugoso (*Rough set Theory)* foi criada por Pawlak (1982), como método de aproximação formal da incerteza a um conjunto nítido de pares de conjuntos (PAWLAK, BUSSE, SLOWINSKI, ZIARKO, 1995). Quando a aproximação não é nítida, ou é vaga (SMITHSON, 1988) os conjuntos podem ser difusos ou *fuzzy* (ZIARKO, 1993). A teoria do conjunto áspero tem como base a noção de discernibilidade: a capacidade de distinguir entre objetos, com base em seus valores de atributos (RIZA et Al., 2014).

Algoritmos que aplicam o conjunto rugoso tem como objetivos: encontrar padrões ocultos, encontrar conjuntos mínimos, avaliar o significados dos dados e oferecer uma interpretação direta e simples dos resultados obtidos (PAWLAK, BUSSE, SLOWINSKI, ZIARKO, 1995). A teoria de conjuntos rugosos provou potencial para variados tipos de análises de dados (PAWLAK, 1998), é eficaz na medicina diagnóstica (SLOWINSKI, STEFANOWSKI, 1989), e para acomodar a inexatidão de dados em saúde (PATTARAINTAKORN, CERCONE, 2008). Shi (2002) descreveu o potencial da técnica no diagnóstico e reconhecimento de doenças pelo método de triagem (SHI, 2002). Dai e Xu (2013) demonstraram em um teste de classificação de tumores que o método *fuzzy* de conjuntos rugosos, é adequado e eficaz a classificação com grande quantidade de atributos.

#### 2.7.1.1 Algoritmos *Fuzzy rough*

Classificadores *fuzzy* tem se mostrado altamente úteis na redução da dimensionalidade dos dados não classificados (JENSEN, SHEN, 2009; NASIRI, MASHINCHI, 2008), são apontados como ferramentas com potencial de aplicação ao conhecimento incerto.

A necessidade de adequação a grande volume de dados e as lacunas do conhecimento gerada pelos dados ausentes objetivaram a criação do *Fuzzy rough NN* que é capaz de otimizar a classificação *fuzzy* de dados com incerteza. O *Fuzzy Rough NN* utiliza método de semelhança por vizinhança próxima (*Nearest Neighbours*) para a aproximação da decisão de classes, aproximando as instâncias próximas. É um algoritmo que trabalha com incerteza e vagueza na forma de associação *fuzzy* (JENSEN, CORNELIS, 2008).

O *Discernibility Classifier*, ou classificação por discernibilidade é um algoritmo *fuzzy rough* que tem o objetivo de demonstrar o poder de separação de classes dos atributos e seus pesos classificatórios correspondentes. Denota o peso dos padrões e a distância entre padrões e seus vizinhos (PANDA, HASSANIEN, ABRAHAM, 2017). Com a associação do critério de discernibilidade os problemas classificatórios são diminuídos sensivelmente, essa extensão *fuzzy* discerne mais nitidamente matrizes, o que reduz grandemente a dimensionalidade da incerteza preservando a precisão da classificação (JENSEN, SHEN, 2009).

A discernibilidade separa diferentes classes ou padrões de conjuntos de dados, nesse contexto, a remoção de redundâncias melhora a separação de padrões e melhora o desempenho em termos de complexidade. Com base nos vizinhos próximos para cada classe é produzida uma pontuação de classificação com saída provável. A relação de semelhança *fuzzy* deve ser idêntica à da abordagem de grau de dependência *fuzzy rough* para encontrar correspondentes reduções de elementos não classificados. A discernibilidade aplicada ao *fuzzy* tem a vantagem de fornecer uma computação mais eficiente para aquisição de conhecimento, especialmente em grandes sistemas incompletos (LEUNG, LI, 2003).

O *Fuzzy Rough NN* e o *Discernibility Classifier* são escolhidos para análises nesse trabalho, são comparados entre si de forma a demonstrar o melhor classificador *fuzzy* frente a incerteza e incompletude de informações.

### 2.7.2 Redes neuronais artificiais – algoritmo *Multilayer Perceptron* (MLP)

O MLP é o classificador de redes neuronais artificiais mais popular (HUSH, 1989), e o mais utilizado dentre os algoritmos de redes neurais artificiais, é capaz de resolver problemas complexos, é um preditor não linear (CHANG, 2006). A rede neuronal artificial *Multilayer Perceptron* é utilizada no reconhecimento de padrões, e utiliza *Backpropagation* para treinamento da rede neuronal. É um algoritmo supervisionado que utiliza o *imput* e o *output* no ajuste de pesos da rede. Esse método garante que a rede caminhe na direção de redução de erros (HECHT-NIELSEN, 1989). O *Multilayer Perceptron* (MLP) tem sido utilizado em triagem clínica com o objetivo de separar amostras de grupos de risco. Bounds et Al, 1988, aponta como promissor a triagem clínica de fatores de risco à dor lombar ciática separando sub-amostras que necessitam de cirurgia.

Rede neuronais artificiais são baseadas em modelagem estatística não-linear e são uma alternativa a regressão logística. São capazes de modelar e predizer resultados baseando-se em valores de variáveis preditoras. O mecanismo de convergência de informações é gerada pela rede neuronal minimiza funções de erros de mínimos quadráticos (TU, 1996), o que torna as redes neuronais artificiais tolerantes à falhas e lhe confere robutez classificativa. O algoritmo MLP é capaz de aceitar lacunas, entretanto, não consegue treinar com dados faltantes diretamente, necessita de mecanismos que promovem o preenchimento das informações faltantes, o que, se não for realizado de maneira criteriosa, pode provocar resultados tendenciosos (CHANG, 2006). O algoritmo MLP disponível no WEKA, substitui as lacunas de dados por zeros (WEKA, 2017 a.) para a promoção dos cálculos. A alternativa ao preenchimento de lacunas com zero é a utilização de filtros de imputação de dados.

Para comparações com os algoritmos *fuzzy rough: Fuzzy Rough NN e Discernibility Classifier*, o algoritmo MLP é escolhido neste trabalho.

# 3 MATERIAIS E MÉTODOS

## 3.1 AMOSTRA

A amostra utilizada no estudo contempla registros de banco de dados do biorrepositório do laboratório de Bioquímica Clínica I e II da Universidade Federal do Paraná. É composto por 271 registros de dados laboratoriais e antropométricos do sexo feminino com idades de 40 a 81 anos (média 55±10 anos), das quais 103 são portadoras de *Diabetes mellitus* tipo 2 (DM2) e 168 são saudáveis, sem a presença do DM2 (grupo controle). Os dois grupos amostrais selecionados são comparados com base em 11 exames laboratoriais: exames de perfil glicêmico, lipídico, analitos para estimar função renal e hepática, todos os exames realizados com amostras de sangue, e duas aferições antropométricas: idade e IMC de cada voluntária; totalizando 13 atributos numéricos e um atributo classe, que difere voluntárias com DM2 e voluntárias sem DM2. Amostra em corte transversal não possui incompletudes.

Os registros deste estudo foram obtidos após autorização do Comitê de Ética em Pesquisa com Seres Humanos do Setor de Ciências da Saúde da Universidade Federal do Paraná (CEP). A aprovação do projeto pelo CEP está sob o registro CAAE: 01038012.2.0000.0102.

Os dados completos são randomizados de forma “totalmente randômica” (MCAR) e dão origem a seis amostras com incompletudes percentuais com 5 a 50% de lacunas. As amostras incompletas são analisadas de duas formas: com a presença de lacunas e com a aplicação de caso completo.

## 3.2 CRITÉRIOS PARA A SELEÇÃO E CLASSIFICAÇÃO DE ATRIBUTOS

Foi adotado como critério para a seleção dos atributos na amostra a exclusão de elementos com latência observada como: PES (peso), ALT (altura), NHDL (colesterol não HDL), e AIP (Relação triglicerídeos por HDL-colesterol).

Os grupos, diabético (DM2) e sem diabetes foram classificados segundo os critérios da Sociedade Brasileira de Diabetes (SBD 2016), com base na glicemia de jejum. Resumidamente uma glicemia em jejum igual ou superior a 126 mg/dL foi considerada como critério de diabetes. Para o grupo sem diabetes (controle) a concentração máxima aceitável foi 100 mg/dL. A descrição dos exames utilizados no

modelo de decisão deste estudo e os valores de referência adotados encontram-se dispostos na TABELA 1.

TABELA 1. INTERVALO DE REFERÊNCIA DE DADOS ANTROPOMÉTRICOS E LABORATORIAIS

| EXAMES | MNEMÔNICO | INTERVALOS DE REFERÊNCIA* | FONTE |
|---|---|---|---|
| Índice de massa corpórea | IMC | 18,5 a 24,9Kg/m² | ABESO, 2016. |
| Ureia sérica | URE | 20 a 40mg/dL | SBN, 2011. |
| Creatinina sérica | CRE | 0,6 a 1,3mg/dL | SBN, 2011. |
| Proteína total | PT | 6,6 a 8,3g/dL | TIETZ, 1995. |
| Albumina sérica | ALB | 3,5 a 4,8 g/dL | TIETZ, 1995. |
| Colesterol total | COL | < 200mg/dL (Ótimo)<br>200-239 mg/dL (Limítrofe)<br>≥ 240 mg/dL (Alto) | SBC, 2015. |
| Triglicerídeos | TRI | < 150 mg/dL (Ótimo)<br>150-200 mg/dL (Limítrofe)<br>200-499 mg/dL (Alto) | SBC, 2015. |
| Colesterol HDL | HDL | > 60 mg/dL (Ótimo)<br>< 40 mg/dL (Baixo) | SBC, 2015. |
| Colesterol LDL | LDL | < 100 mg/dL(Ótimo)<br>100-129 mg/dL(Desejável)<br>130-159 mg/dL(Limítrofe)<br>160-189 mg/dL(alto) | SBC, 2015. |
| Ácido úrico (soro) | AU | 2,0- 5,0mg/dL | TIETZ, 1995 |
| Glicose Jejum | GLU | Normal < 100 mg/dL<br>Tolerância à glicose diminuída:<br>≥100 a < 126 mg/dL<br>Diabetes *mellitus*: ≥126 mg/dL | SBD, 2016 |
| Hemoglobina glicada | HbA1c | Normal ≤ 6,4%<br>Diabetes ≥ 6,5% | SBD, 2016 |

FONTE: a autora (2017).
NOTA: *intervalos de referência baseados na população de estudo: mulheres adultas não gestantes.

## 3.3 SORTEIO RANDOMIZADO

A randomização aplicada nesse trabalho foi realizada de forma a selecionar na matriz de dados um número percentual de intersecções de vetores: colunas e linhas, chamadas de células ou de casos, as informações selecionadas por meio de sorteio totalmente randômico sem repetição tiveram seus valores apagados com o intuito da formação de lacunas. As lacunas foram criadas com o auxílio de uma função de randomização em porcentagem desenvolvida na linguagem de programação R (3.1.1), o algoritmo de função do R está disponível no ANEXO deste trabalho.

Esse procedimento deu origem a amostras com incompletudes percentuais de 5% a 50% com N amostrais idênticos, as amostras com percentuais superiores de incompletude tiveram perdas de tamanho amostral devido a ocorrência de sorteios de todas as variáveis de alguns vetores linhas, a

substituição de toda a linha por lacunas inviabiliza a observação dessas linhas, e por tanto foram desconsideradas.

O sorteio de dados sem repetição é um modelo de incompletude totalmente randômica. Os atributos formadores da classificação prévia de grupos, o perfil glicêmico: glicose jejum e hemoglobina glicada (HbA1c) foram excluídos das amostras provenientes da randomização, com o objetivo de permitir a decisão e a classificação própria a partir de outros exames pela correlação τ de Kendall e pelos algoritmos *Multilayer Perceptron* (MLP) *Discernebility Classifier e Fuzzy Rough NN* disponíveis no *software* WEKA.

### 3.3.1 Amostras em caso completo

Nas amostras com incompletudes, não foram imputados valores às lacunas devido à complexidade e alta variabilidade dos dados numéricos dos exames bioquímicos. Nessas condições as imputações poderiam gerar uma quantificação de vieses a tal ponto de não ser vantajosa para a decisão. Então as amostras com lacunas preservadas são alvo de análises.

Com o intuito de comparar resultados das análises de amostras incompletas com as análises de dados oriundos de métodos que promovam a completude, foram criadas, a partir das amostras com incompletudes percentuais, amostras com completude pelo método caso completo. Utilizando a linguagem R (3.1.1) as amostras foram verificadas quanto as lacunas presentes com a função *complete.cases()* e tratadas com o caso completo com a função *na.omit()*.

Com a aplicação do caso completo foram obtidas 3 amostras com tamanho amostrais diferentes ($n$ = 162, 81, 23). A obtenção de amostras em caso completo a partir das amostras com lacunas foi considerada com até 20% de incompletude, incompletudes maiores geram um “*n”* amostral impossível de ser obtido pela inexistência de linhas totalmente preenchidas para aplicação do caso completo.

## 3.4 FERRAMENTAS UTILIZADAS NAS ANÁLISES

Foram utilizadas as seguintes ferramentas/*softwares*:

a) Análises estatísticas: linguagem de programação R (3.1.1) com o pacote Kendall e método de correlação rankeada de τ kendall na fórmula:

$$\tau = \frac{S}{D},$$

onde

$$S = \sum_{i>j}(sign(c[j] - x[i]).(sing(y[j] - y[i])), \quad (1)$$

$$D = \frac{n(n-1)}{2}. \quad (2)$$

Com a função: Kendall::kendall(*x,y*) onde *x* = exames e y = classe (classificação binária dos grupos “Diabetes” = 1 e “Sem diabetes” = 0). A análise não utilizou a aplicação de filtros de imputação das lacunas, nessas condições as lacunas são ignoradas na análise (MC LEOD, 2011). Ainda no software R, os resultados das correlações de exames por classe foram dispostos em forma vetorial, e os cálculos de média: *mean()*, variância amostral: *var(),* e desvio padrão amostral*: sd()* foram aplicados para possibilitar análises comparativas de dispersão das correlações de exames entre amostras com incompletudes, o mesmo foi feito para comparações entre as amostras com caso completo.

b) Análises não-estatísticas: Foram escolhidos três algoritmos classificadores disponíveis no software WEKA®: *Function Multilayer Perceptron* (MLP) e as extensões *fuzzy: Fuzzy Rough NN, e Discernibility Classifier* disponíveis com a atualização WEKA FULL (JENSEN, 2008). Na análise com o WEKA os grupos foram transformados de binários (0 e 1) para nominal (“Diabetes” e “Sem diabetes”). O MLP, *Fuzzy rough NN* e o *Discernibility classifier* foram testados e validados sem aplicação de filtros e sem qualquer forma de imputação de lacunas. A extensão WEKA FULL adicionou informações sobre a classificação, foram considerados na análise classificatória: Taxa de cobertura de casos  com valor de confiança fixo de 95%, e o Tamanho médio dos casos cobertos com confiança de 95%, ambos calculados pelo software WEKA.

As informações apresentadas a respeito do desempenho classificatório dos algoritmos consistem em:

- **Acurácia** (*Accuracy*): mede a proximidade entre o valor obtido na classificação e o valor verdadeiro da medição:

$$\frac{VP + VN}{VP + VN + FP + FN}$$

onde:
VP = verdadeiro positivo
VN = verdadeiro negativo
FP = falso positivo
FN = falso negativo (USFCA, 2017).

- **Estatística Kappa** (*Kappa Statistic*): teste não-paramétrico que mede o acordo da predição com a classe verdadeira, seus valores variam de 0 a 1, onde 1 significa acordo completo (USFCA, 2017).
- **Taxa de cobertura de casos** (0,95) (*Coverage of cases 0.95 level*): estimativa de intervalos de dados cobertos. É a largura de intervalos normalizados pelo intervalo de valores-alvo nos dados de treinamento. Para um estimador ter intervalo condiderado razoável, ou aceitável, deve exibir cobertura igual ou superior ao nível de 95%, produzindo intervalos estreitos entre os dados (WEKA, 2017 b).
- **Tamanho médio dos casos cobertos** (0,95) (*Mean rel. region size 0.95 level*): tamanho médio das regiões previstas em relação ao intervalo de dados de treinamento, com o nível de confiança especificado de 95% (WEKA, 2017 c).
- **Sensibilidade** (*Precision - sensibility*): classificação de verdadeiros positivos da amostra $\frac{VP}{VP+F}$ (USFCA, 2017). Corresponde ao grupo controle “Sem diabetes”.
- **Especificidade** (*Precision - specificity*): classificação de verdadeiros negativos da amostra $\frac{VN}{VN+FP}$ (USFCA, 2017). Corresponde ao grupo “Diabetes”.
- **Área ROC** (*AUC = Area Under the ROC Curve*): o WEKA utiliza o método estatístico Mann Whitney para calcular a área acima da curva ROC (WEKA, 2017 d).

Em todas as análises um valor de probabilidade menor que 5% foi considerado significativo (p-valor <0,05).

## 4 RESULTADOS

A primeira etapa da metodologia consiste na verificação de incertezas na amostra com completude. A estatística descritiva foi aplicada para o conhecimento das características amostrais. Na TABELA 2, com a estatística descritiva, se observam algumas características dos dois grupos da amostra.

TABELA 2 – ESTATÍSTICA DESCRITIVA DA AMOSTRA EM ESTUDO

| | | SEM DIABETES | | | COM DIABETES | | |
|---|---|---|---|---|---|---|---|
| **ATRIBUTOS** | **MNEMÔNICO** | | **μ** | **s** | | **μ** | **s** |
| Idade | IDA | 40 - 79 Anos | 52 | 10,17 | 42 – 81 Anos | 59 | 8,77 |
| Índice de massa corporal | IMC | 15 - 50Kg/m² | 24,9 | 5,58 | 21 - 55 Kg/m² | 35,8 | 5,45 |
| Ureia | URE | 11 – 68mg/dL | 28,9 | 8,67 | 11 – 86mg/dL | 37,4 | 12,15 |
| Creatinina sérica | CRE | 0,2 – 1,1mg/dL | 0,71 | 0,20 | 0,6 – 2,7mg/dL | 0,89 | 0,33 |
| Proteína total | PT | 5,1 – 9,1g/dL | 7,14 | 0,73 | 5,3 – 9,9g/dL | 7,53 | 0,70 |
| Albumina sérica | ALB | 2,3 – 5,2g/dL | 4 | 0,39 | 1,9 – 5,1g/dL | 4 | 0,42 |
| Colesterol total | COL | 98 - 374mg/dL | 196 | 42,58 | 94 - 444mg/dL | 192,5 | 52,76 |
| Triglicerídeos | TRI | 50 - 624mg/dL | 134 | 72,98 | 50 – 1013mg/dL | 205 | 155,14 |
| Colesterol HDL | HDL | 28 - 115mg/dL | 56,9 | 15,90 | 25 - 75mg/dL | 42,3 | 10,28 |
| Ácido úrico | AU | 2,2 -10,5mg/dL | 4,77 | 1,35 | 1,9 – 9,5mg/dL | 4,9 | 1,36 |
| Colesterol LDL | LDL | 41 - 296mg/dL | 112 | 36,73 | 31 - 351mg/dL | 110,2 | 42,03 |
| Glicose jejum | GLU | 49 - 120mg/dL | 88,7 | 10,32 | 59 - 468mg/dL | 162,6 | 75,86 |
| Hemoglobina glicada | HbA1c | 4,1 – 6,4% | 5,32 | 0,52 | 6,0 – 20,4% | 8,5 | 2,47 |

FONTE: a autora (2017).
LEGENDA: média (μ), desvio padrão amostral (s).

As características intrínsecas da amostra estudada podem ser observadas com a estatística descritiva e com a filtragem de dados. A resposta encontrada demonstra a capacidade que cada variável tem para discernir os grupos da amostra, ou de discriminar a condição diabetes da condição sem diabetes. Com o IMC é possível perceber que ≈ 38% das diabéticas são obesas (IMC>25), o dobro das não diabéticas ≈ 19% do total. Nos exames da função renal (creatinina sérica e ureia), 7 pacientes possuem comprometimento renal (CRE > 1,3 mg/dL e URE > 40 mg/dL). Não há casos semelhantes no grupo de não diabéticas. Quanto ao perfil lipídico: COL ≥ 240 mg/dL: sem diabetes ≈ 9%, diabetes ≈ 5%; LDL ≥ 160 mg/dL: sem diabetes ≈ 7%, diabetes ≈ 3%. Quanto ao colesterol total, colesterol HDL e colesterol LDL, percebeu-se que não são bons exemplos de exames para a distinção de grupos. A quantidade de valores alterados é pequena na população, e os valores que distinguem melhor os grupos populacionais dentre os exames do perfil lipídico são o HDL<40 mg/dL: sem diabetes ≈ 7%, diabetes ≈ 15%; e com o TRI ≥ 200 mg/dL: sem diabetes ≈ 8% e diabetes ≈ 12%, do quais, 4 pacientes com TRI ≥ 500 mg/dL. Com os exames colesterol HDL e triglicerídeos podemos perceber que o controle lipídico é pior, sugerindo maior risco de doença arterial coronária, no grupo diabetes em comparação

com o grupo sem diabetes. O exame ácido úrico (AU) se manifestou prevalente no grupo sem diabetes ≈ 24% enquanto que 14% das diabéticas tiveram valores alterados (AU > 5,0 mg/dL). Outros exames como Albumina (ALB) e Proteína total (PT) tiveram a ocorrências de valores alterados muito baixa para ser significativa (< 5 pacientes com valores alterados em cada grupo).

As alterações dos exames: hemoglobina glicada (HbA1c) e glicose jejum (GLU) foram associadas para a formação da lógica de classificação dos grupos. Cerca de 77% das pacientes diabéticas possuem glicose jejum controlada (GLU <126 mg/dL) e 22,5% apresentam mau controle glicêmico (GLU ≥ 126 mg/dL). O HbA1c apresentou média de ≈ 8,6% no grupo diabetes, e ≈ 12% de diabéticas apresentam HbA1c ≤ 7, sugerindo bom controle glicêmico nos últimos dois meses. Para concluir os exames que tem maior poder de discriminar o *diabetes mellitus* do tipo 2 na população de estudo, analisados na estatística descritiva são o IMC e perfil glicêmico. As diabéticas da população de estudo são mais obesas que as não diabéticas. A população de estudo possui em maioria bom controle lipídico, as alterações dos exames colesterol HDL e triglicerídeos são mais prevalentes no grupo das diabéticas.

Após a observação das características da amostra pela estatística descritiva e por filtragem de dados, foi aplicada a correlação tau de Kendall (α) a amostra com completude e às amostras com porcentagens de incompletudes (%) como exposto na TABELA 3.

TABELA 3 – CORRELAÇÃO DE KENDALL DE EXAMES POR GRUPO CLASSIFICADOR

| % | IDA | IMC | URE | CRE | PT | ALB | COL | TRI | HDL | AU | LDL |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0,27** | 0,577** | 0,327** | 0,251** | 0,239** | 0,086 | -0,059 | 0,27** | -0,396** | 0,051 | -0,045 |
| 5 | 0,243** | 0,575** | 0,347** | 0,269** | 0,215** | 0,078 | -0,083 | 0,286** | -0,391** | 0,047 | -0,032 |
| 10 | 0,277** | 0,572** | 0,34** | 0,248** | 0,263** | 0,093 | -0,082 | 0,291** | -0,396** | 0,043 | -0,049 |
| 20 | 0,285** | 0,574** | 0,291** | 0,245** | 0,229** | 0,092 | -0,084 | 0,243** | -0,414** | 0,076 | -0,063 |
| 30 | 0,261** | 0,581** | 0,366** | 0,259** | 0,289** | 0,090 | -0,006 | 0,284** | -0,411** | 0,099 | 0,046 |
| 40 | 0,296** | 0,553** | 0,335** | 0,212* | 0,24** | 0,079 | -0,024 | 0,255** | -0,387** | 0,061 | -0,086 |
| 50 | 0,293** | 0,562** | 0,4** | 0,147 | 0,241** | 0,163 | -0,089 | 0,241** | -0,374** | 0,053 | -0,033 |
| μ | 0,275 | 0,570 | 0,344 | 0,233 | 0,245 | 0,097 | -0,061 | 0,267 | -0,395 | 0,061 | -0,037 |
| s | 0,014 | 0,007 | 0,023 | 0,030 | 0,017 | 0,019 | 0,027 | 0,018 | 0,01 | 0,015 | 0,027 |
| $s^2$ | 4E-04 | 9,4E-05 | 0,001 | 0,002 | 6E-04 | 9E-04 | 0,001 | 4E-04 | 0,0001 | 4E-04 | 0,002 |

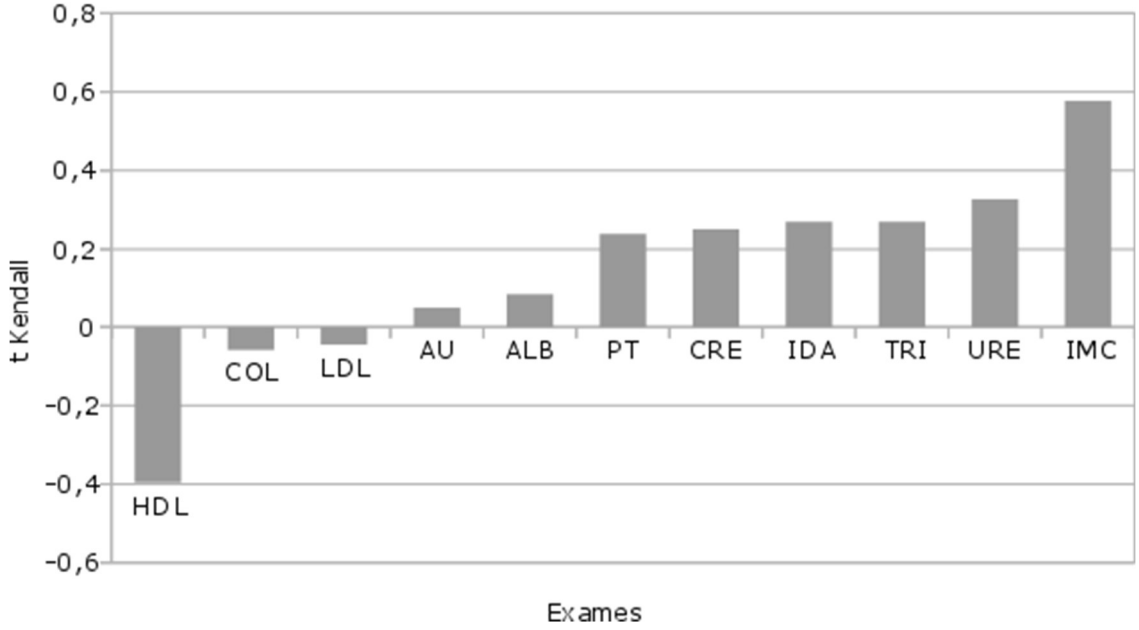
FONTE: a autora, (2017).
LEGENDA: Lacunas (%), média (μ), desvio padrão da amostra (s), variância da amostra ($s^2$). P-valor unilateral: ** ≤0,01, *≤0,05.

As correlações da amostra em completude, representada por 0% de incompletude, são utilizadas como base comparativa às amostras com porcentagens de incompletudes, e seus valores não são incorporados aos cálculos da média (μ) e da variância amostral ($s^2$), que correspondem à variância da correlação das amostras

em diferentes graus de incompletude. Comparando as correlações de Kendall de cada amostra incompleta à amostra sem incompletude, podemos observar que os exames das amostras incompletas que mais diferiram em poder de correlação, em relação à amostra completa, foram: creatinina sérica (CRE), colesterol LDL, ureia (URE), albumina (ALB), ácido úrico (AU), e colesterol total (COL).

Os valores das correlações das amostras com incompletudes em relação à amostra sem incompletude divergiram de: sem divergência observada na variável HDL (amostra com 10% de incompletude) à uma divergência de ≈ 0,104 no exame CRE (amostra com 50% de incompletude). Com esses resultados, consideramos que não houveram variações bruscas dos valores de correlação de Kendall entre cada amostra com incompletude em relação à amostra com completude. De fato, na forma de porcentagem, a divergência, em módulo, não superou 10,4% com incompletude de até 50%. Em relação à média de correlações, as amostras incompletas apresentaram respostas próximas dos valores das correlações da amostra completa: com diferença máxima encontrada no exame CRE ≈ 0,086, uma divergência de aproximadamente 8,6%. Os valores da correlação de Kendall (τ-α) de exames por grupos classificatórios da amostra em completude são observados no GRÁFICO 1.

GRÁFICO 1 – CORRELAÇÃO KENDALL DE EXAMES POR GRUPO CLASSIFICATÓRIO NA AMOSTRA EM COMPLETUDE




FONTE: a autora (2017).
LEGENDA: Exames: IDA(idade), IMC(índice de massa corporea), URE(uréia), CRE(creatinina sérica), PT(proteína total sérica), ALB(albumina sérica), COL(colesterol total), TRI(triglicerídeos), HDL(colesterol HDL), AU(ácido úrico), LDL(colesterol LDL).

De acordo com a escala de avaliação da força de correlação proposta por Bracarense Costa (2008), são observados os seguintes resultados referentes ao GRÁFICO 1:

a) Independente do grupo ou negativo ($r \leq 0$): HDL, COL, LDL
b) Correlação fraca ($0 < |r| \leq 0,5$): AU, ALB, PT, CRE, IDA, TRI, URE
c) Correlação moderada ($0,5 < |r| \leq 0,75$) : IMC

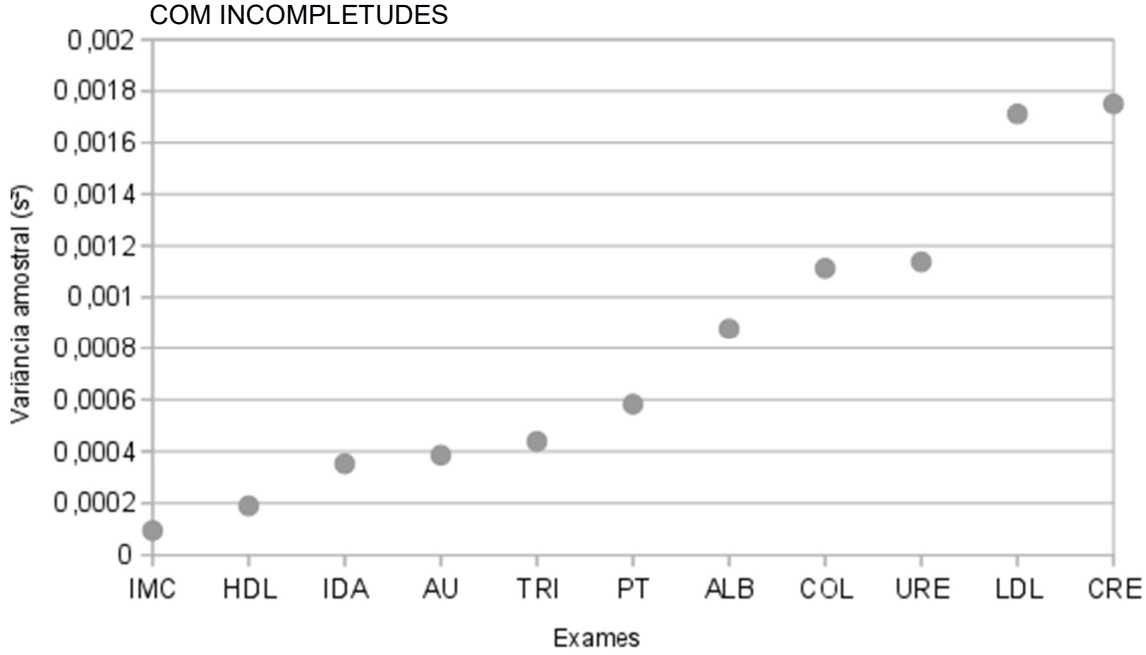
Os resultados encontrados pelo método de correlação τ de Kendall são complementares aos resultados observados na estatística descritiva dos dados. Parte do perfil lipídico (Colesterol total, Colesterol HDL e Colesterol LDL) é independente da classificação dos grupos diabetes e sem diabetes, assim como mostrado na estatística descritiva dos dados. Essas variáveis não demonstram poder de correlação para discernir grupos. Essa característica é própria do grupo populacional estudado, e de fato, menos de 30% das pacientes de ambos os grupos tem o perfil lipídico considerado elevado, o que inclui mais de 70% da população de estudo em situação de bom controle lipídico.

Os exames albumina, ácido úrico, creatinina, triglicerídeos, proteína total, ureia e a variável idade são correlacionados fracamente aos grupos. Isso significa que esses elementos distinguem a condição diabetes da não diabetes com pouca especificidade. O IMC é a variável com maior poder de correlação aos grupos e consequente maior capacidade de distinção de grupos. Não são observados valores de correlação forte ($0,75 < |r| \leq 0,9$), ou perfeita ($r = 1$). A maioria dos exames das amostras incompletas de 5 a 50% tiveram a mesma classificação na escala de avaliação de força de correlação. A exceção foi o exame LDL com 30% de incompletude que passou de independente ($r \leq 0$) a correlação fraca (0,0465).

As variações de resultados entre os dados completos e as amostras com incompletudes são devidas a duas causas: perda de poder estatístico gerado pelas lacunas, e pela randomização amostral. Para a observação das variações de resultados de correlações da amostra em completude às amostras com incompletudes é utilizado o cálculo de variância amostral. A variância é aplicada como critério de comparação e de confiabilidade de tendências indicadas pela correlação de Kendall nas amostras.

No GRÁFICO 2 se observa a variância da correlação de Kendall de exames por classe nas amostras com incompletudes partindo da amostra em completude (0 - 50%).

GRÁFICO 2 – VARIÂNCIA AMOSTRAL (S²) DE CORRELAÇÕES DE KENDALL DAS AMOSTRAS COM INCOMPLETUDES




FONTE: a autora (2017).

A diferença entre os valores encontrados nas correlações de Kendall da amostra em completude e das amostras com lacunas, calculadas subtraindo o valor de correlações da amostra em completude das amostras com incompletude, temos: o exame com menor divergência é o IMC (≈ 9,42E-05) e o com maior divergência é o CRE (≈ 0,00175), com esses valores consideramos que a variação de correlações entre a amostra completa até a amostra com o máximo de incompletude (50%) é considerada pequena.

A média de correlação das amostras com lacunas é próxima aos valores de correlação da amostra sem incompletude, como dito anteriormente, por isso a medida de dispersão da média das amostras com lacunas pode ser uma medida relevante ao estudo da dispersão na incompletude. A medida de dispersão dos desvios padrões em relação à média encontrada nas amostras com lacunas, pode ser demonstrada com o coeficiente de variação ( $cv = \frac{s}{\mu}.100$ ), nas amostras com lacunas a maioria dos exames não foi superou 15% indicando baixa dispersão (≤15%) nos resultados da correlação de Kendall em amostras com incompletudes. A dispersão é proporcional a força de correlação, a variável IMC (com correlação moderada aos grupos) teve a

menor dispersão em relação à média dentre as variáveis das amostras, *cv* ≈ 1,22%, as variáveis de maior dispersão são as variáveis de perfil lipídico COL e LDL. A distribuição de variáveis na escala de dispersão em relação à média (FONSECA; MARTINS, 2010), é descrita em ordem crescente de dispersão abaixo:

a) Baixa dispersão (*cv* ≤15%): IMC, HDL, IDA, URE, TRI, PT, CRE
b) Média dispersão (15% < cv < 30): ALB, AU
c) Alta dispersão (cv ≥ 30): COL, LDL

As amostras com incompletudes analisadas com o método de caso completo foram correlacionadas para fins de comparação de resultados. As correlações de Kendall (τ-α) nas amostras em caso completo (*N*) são expostas na TABELA 4.

TABELA 4 – CORRELAÇÃO KENDALL DE EXAMES E GRUPOS CLASSE EM CASO COMPLETO

| *N* | IDA | IMC | URE | CRE | PT | ALB | COL | TRI | HDL | AU | LDL |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 271 | 0,27** | 0,577** | 0,327** | 0,251** | 0,239** | 0,086 | -0,059 | 0,27** | -0,396** | 0,051 | -0,045 |
| 162 | 0,266** | 0,586** | 0,414** | 0,296** | 0,297** | 0,132 | -0,102 | 0,267** | -0,442** | 0,068 | -0,066 |
| 81 | 0,182 | 0,612** | 0,215 | 0,215 | 0,252* | 0,252* | -0,111 | 0,155 | -0,373 | 0,079 | -0,054 |
| 23 | 0,197 | 0,599** | 0,544** | -0,047 | 0,347 | 0,219 | -0,069 | 0,268 | -0,446* | -0,07 | -0,063 |
| μ | 0,28 | 0,593 | 0,375 | 0,178 | 0,283 | 0,172 | -0,085 | 0,24 | -0,414 | 0,032 | -0,057 |
| s | 0,045 | 0,015 | 0,14 | 0,154 | 0,05 | 0,08 | 0,025 | 0,056 | 0,035 | 0,068 | 0,009 |
| s² | 0,002 | 2E-04 | 0,019 | 0,023 | 0,002 | 0,006 | 6E-04 | 0,003 | 0,001 | 0,005 | 8,2E-05 |

FONTE: a autora (2017).
LEGENDA: média (μ), desvio padrão da amostra (s), variância da amostra (s²). P-valor unilateral: ** ≤0,01, *≤0,05.

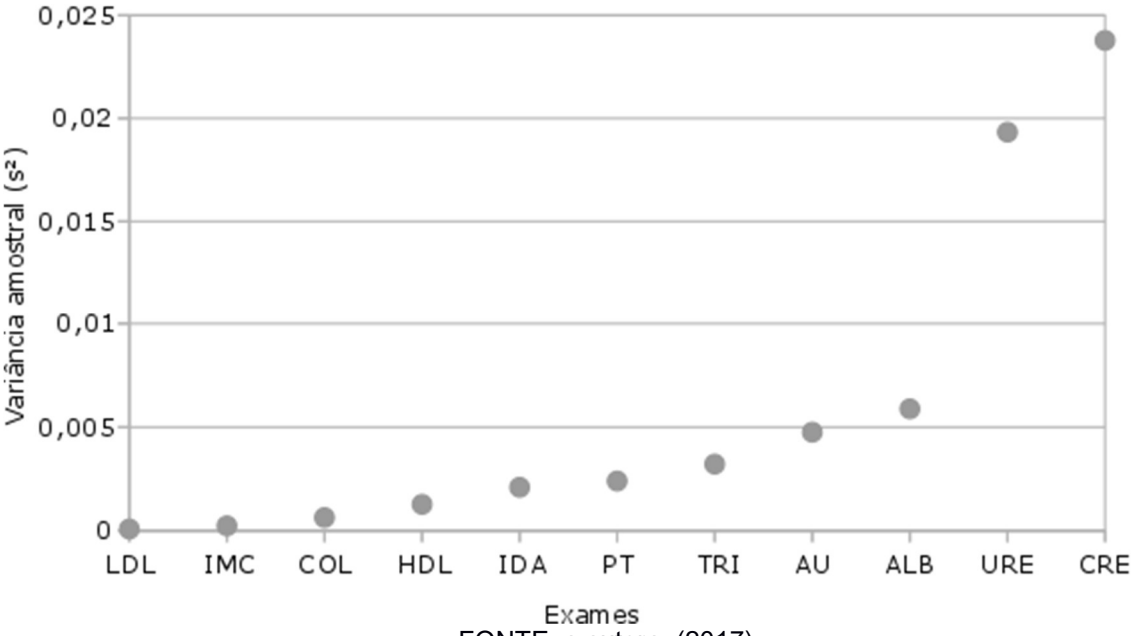
Como podemos visualizar na TABELA 4, a amostra com *n* = 23 o exame URE passou do fracamente correlacionado (0 < |r| ≤ 0,5) na amostra completa ao moderadamente correlacionado (0,5 < |r| ≤ 0,75) e os exames creatinina (CRE) e ácido úrico (AU) sem tornaram variáveis independentes dos grupos classificatórios, o que indica degeneração de características, ou ainda, distorção nessa amostra de caso completo.

Comparando as diferenças dos resultados de cada amostra em caso completo às respostas de correlações da amostra em completude (*n* = 271), calculadas subtraindo o valor da correlação dos dados em completude dos dados tratados com caso completo, temos que os exames: LDL (≈0,021) na amostra *n* = 23 e IMC (≈-0,035) na amostra *n* = 81 apresentaram menor divergência, e CRE (≈0,298) em *n* = 23, URE (≈-0,217) em *n* = 23, apresentaram maior divergência, convertendo em

porcentagem a divergência máxima no caso completo chegou a 29,8% na amostra com em *n* = 23.

As amostras tratadas com caso completo apresentaram em média uma divergência maior em comparação com os valores de correlações da amostra em completude em comparação com as amostras com incompletude (5 a 50%). Essa divergência em relação a amostra completa é mais que o dobro do valor máximo de divergência das correlações das amostras com incompletudes de 5 a 50% (≈ 10,4%) exposto anteriormente. As variações das respostas de correlações dos exames das amostras de caso completo partindo da amostra em completude (*n* = 271 a *n* = 23) podem ser vistas no GRÁFICO 3.

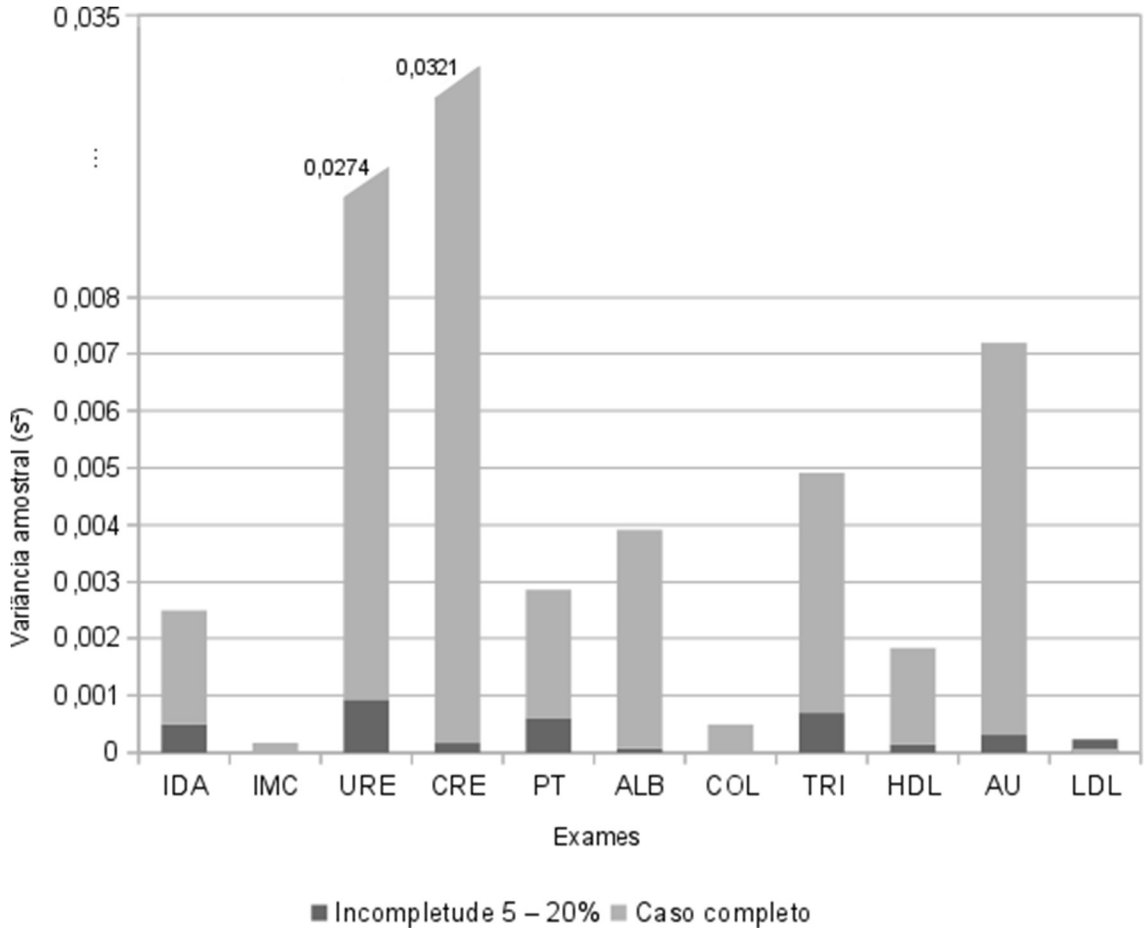
GRÁFICO 3 - VARIÂNCIA AMOSTRAL ($S^2$) DA CORRELAÇÃO KENDALL EM AMOSTRAS COM CASO COMPLETO




FONTE: a autora, (2017).

A resposta de maior variância entre as correlações de Kendall nas amostras tratadas com o método caso completo é apresentada pelo exame creatinina sérica (CRE) $s^2 \approx 0{,}02378$, enquanto que o mesmo exame nas amostras com incompletudes preservadas (até 20%) teve um CRE de $s^2 \approx 0{,}00012$. A variabilidade das correlações de Kendall, representada pela variância amostral ($s^2$) partindo da amostra em completude, é maior em amostras tratadas com caso completo em comparação com amostras com lacunas até 20%. A variabilidade das correlações nas amostras incompletas: com lacunas (5 a 20%) e com a aplicação de caso completo (*n* = 162 a *n* = 23) podem ser visualizadas no GRÁFICO 4.

GRÁFICO 4 - RESPOSTAS DA CORRELAÇÃO DE KENDALL COM AMOSTRAS COM LACUNAS 5 – 20% E AMOSTRAS EM CASO COMPLETO




FONTE: a autora (2017).

As variáveis índice de massa corporal (IMC) e colesterol (COL) apresentaram resposta de variação de correlações entre amostras com incompletudes não visualizável na escala do GRÁFICO 4: IMC $s^2 \approx 2{,}333E\text{-}06$ e COL $s^2 \approx 1{,}63E\text{-}06$. O LDL é o único exame que apresenta variância menor no caso completo $s^2 \approx 3{,}36E\text{-}05$.

Os exames com maior e menor variância de correlação nas amostras com lacunas não coincidem com os das amostras em caso completo, o que indica que os resultados da correlação de Kendall aplicada a lacunas difere da aplicada no caso completo levando a tendências diferentes. Na análise de amostras incompletas desse estudo é percebido que a falta de dados e a randomização geram tendências sutis. Já nas amostras analisadas com caso completo as tendências, vieses e distorções são devidos à seleção de informações ocasionada pela exclusão maciça de dados.

Para concluir, o exame mais discriminante do *diabetes mellitus* do tipo 2 nesse estudo é o IMC com um poder discriminante moderado ($\tau \geq 0{,}5$) em todas as amostras analisadas. Outros exames são considerados fracamente correlacionados.

Na segunda etapa de análise das amostras, foram aplicados os algoritmos *Multilayer Perceptron, Fuzzy Rough NN, Discernibility Classifier* do software WEKA. Os algoritmos foram treinados e validados (*10-fold cross-validation*), e os resultados se referem às validações cruzadas. As configurações dos testes foram as seguintes:

- *Multilayer Perceptron*: treinado e validado com 500 épocas, treinamento com formação de 7 nós. Sem uso de filtros;

- *Fuzzy Rough NN* e *Discernibility Classifier*: classificação por vizinhança próxima *(10 nearest neighbour(s) for classification)*:

Medida de similaridade: *Similarity1*: $1 - \frac{abs(a(x) - a(y))}{abs(a^{max} - a^{min})}$

*Implicador: Kleene-Dienes*

*T-Norm: Kleene-Dienes*

Relação de composição: Algébrica

Sem uso de filtros.

Os resultados observados da amostra completa e de amostras com incompletude em porcentagem (%) são descritos na TABELA 5:

TABELA 5 – VALIDAÇÃO DE AMOSTRA COM COMPLETUDE E AMOSTRAS COM INCOMPLETUDES COM ALGORITMOS (WEKA)

*Multilayer Perceptron* (MPL) (continua)

| % | *ACCURACY* | *KAPPA* | *COVERAGE* | *MEAN REL.* | *SENSIBILITY* | *SPECIFICITY* | *ROC* |
|---|---|---|---|---|---|---|---|
| 5 | 81,5498 | 0,6157 | 91,1439 | 62,7306 | 0,873 | 0,735 | 0,889 |
| 10 | 80,4428 | 0,5873 | 90,0369 | 61,6236 | 0,848 | 0,736 | 0,873 |
| 20 | 76,0148 | 0,492 | 88,5609 | 63,6531 | 0,808 | 0,683 | 0,84 |
| 30 | 70,4797 | 0,3736 | 83,7638 | 62,7306 | 0,762 | 0,612 | 0,762 |
| 40 | 68,6347 | 0,3406 | 84,1328 | 64,0221 | 0,755 | 0,583 | 0,777 |
| 50 | 66,0517 | 0,2823 | 87,4539 | 71,2177 | 0,729 | 0,552 | 0,74 |

*Fuzzy Rough NN*

| % | *ACCURACY* | *KAPPA* | *COVERAGE* | *MEAN REL.* | *SENSIBILITY* | *SPECIFICITY* | *ROC* |
|---|---|---|---|---|---|---|---|
| 5 | 76,3838 | 0,4893 | 99,631 | 86,1624 | 0,792 | 0,71 | 0,855 |
| 10 | 79,3358 | 0,5514 | 99,262 | 89,1144 | 0,811 | 0,758 | 0,863 |
| 20 | 70,8487 | 0,3825 | 99,262 | 96,1255 | 0,766 | 0,615 | 0,741 |
| 30 | 63,8376 | 0,2837 | 99,262 | 94,8339 | 0,769 | 0,518 | 0,74 |
| 40 | 58,6716 | 0,0816 | 99,631 | 99,0775 | 0,646 | 0,443 | 0,595 |
| 50 | 61,2546 | -0,0008 | 99,262 | 97,786 | 0,62 | 0,375 | 0,529 |

TABELA 5 – VALIDAÇÃO DE AMOSTRA COM COMPLETUDE E AMOSTRAS COM INCOMPLETUDES COM ALGORITMOS (WEKA)

*Discernibility Classifier* (conclusão)

| % | *ACCURACY* | *KAPPA* | *COVERAGE* | *MEAN REL.* | *SENSIBILITY* | *SPECIFICITY* | *ROC* |
|---|---|---|---|---|---|---|---|
| 5 | 83,3948 | 0,6496 | 100 | 99,8155 | 0,873 | 0,774 | 0,899 |
| 10 | 83,0258 | 0,6357 | 99,631 | 99,631 | 0,851 | 0,794 | 0,892 |
| 20 | 80,4428 | 0,5778 | 100 | 99,8155 | 0,825 | 0,766 | 0,884 |
| 30 | 78,9668 | 0,5425 | 100 | 99,4465 | 0,807 | 0,756 | 0,866 |
| 40 | 73,0627 | 0,4026 | 100 | 99,8155 | 0,749 | 0,688 | 0,802 |
| 50 | 61,2546 | 0,0513 | 99,262 | 98,7085 | 0,632 | 0,469 | 0,644 |

FONTE: a autora (2017).
LEGENDA: *Accuracy* (Acurácia) *Kappa* (Estatística Kappa), *Coverage: Coverage of cases 0.95 level* (taxa de cobertura de casos com 95%* de confiança), *Mean rel.: Mean rel. region size 0.95 level* (tamanho médio dos casos cobertos com 95%* de confiança), *Sensibility* (Sensibilidade) *Specificity* (Especificidade) *ROC: Area Under the ROC Curve* (área ROC). *Estabelecido pelo *software* WEKA.

Antes das explicações detalhadas quanto aos resultados classificatórios dos algoritmos, é preciso lembrar que o algoritmo MLP do *software* WEKA, não tem capacidade de trabalhar com lacunas de dados devido ao modelo de regressão logística que esse algoritmo adota na análise de informações: sem a aplicação de filtros que promovam a imputação de valores, as lacunas são substituídas por zero, como citado anteriormente. A substituição de lacunas por um valor, mesmo que esse valor seja zero, acrescenta informações e gera tendências que em geral melhoram os resultados observados em comparação com a análise da amostra com dados faltantes.

Quanto aos resultados observados na TABELA 5, podemos perceber que os valores de acurácia são levemente superiores nas amostras classificadas com o algoritmo *fuzzy Discernibility Classifier*, que teve a maior média de acurácia entre os testes (μ ≈ 76,691), MLP (μ ≈ 73,862), e *Fuzzy Rough NN* os piores resultados (μ ≈ 68,388). Os melhores resultados do *fuzzy Discernibility Classifier* também são observados na estatística Kappa: (μ ≈ 0,476), MLP (μ ≈ 0,448) e *Fuzzy Rough NN* (μ ≈ 0,297). As amostras com incompletudes registram uma concordância de κ que variou de substancial (0,61< κ < 0,80) a ligeira (0 < κ < 0,20) com o *fuzzy Discernibility Classifier*. A especificidade, condição de classificação da classe “Diabetes” é superior também com o *fuzzy Discernibility Classifier* indicando que esse algoritmo teve o melhor desempenho na classificação da diabetes. O MLP teve resultados melhores na amostra com maior incompletude (50%), isso pode ser devido ao “peso” dos zeros na análise. A diferença mais significativa de desempenho do algoritmo MLP e dos algoritmos *fuzzy rough*, está nos cálculos de cobertura de casos (*Coverage of cases*

*0,95 level)* e tamanho médio de regiões cobertas (*Mean rel. region size 0,95 level*), ambos definidos com 95% de confiança pelo WEKA. Assim a cobertura de casos deve ser ≥ 95% para que seja considerada adequada. A cobertura de casos máxima do MLP foi ≈ 91%, no *Discernibility Classifier* a cobertura mínima foi ≈ 99,2%, cobrindo dados de maneira adequada em todas as amostras. O *Fuzzy Rough NN* também teve uma cobertura de dados superior à do MLP com taxa mínima igual à do *Discernibility Classifier*.

Comparativamente, o tamanho médio de regiões cobertas é fortemente inferior no MLP em comparação com os dois algoritmos *fuzzy rough*: o MLP não foi superior a 71% (μ ≈ 64,3%), enquanto que foi superior a 99,6% no *Discernibility Classifier,* que classificou regiões previstas no treinamento quase a totalidade em todas as amostras. Comparativamente, o *Fuzzy Rough NN* teve resultados de tamanho médio de regiões cobertas acima de 95% em 3 das 5 amostras com incompletudes.

A seguir foram realizadas classificações com outras amostras em caso completo. Com essas amostras é possível visualizar a capacidade classificativa de algoritmos sem empecilhos das lacunas. As análises classificativas com amostras em caso completo *(N*) estão dispostas na TABELA 6.

TABELA 6 – VALIDAÇÃO DE AMOSTRAS EM CASO COMPLETO COM ALGORITMOS (WEKA)

*Multilayer Perceptron* (MPL)

| *N* | *ACCURACY* | *KAPPA* | *COVERAGE* | *MEAN REL.* | *SENSIBILITY* | *SPECIFICITY* | *ROC* |
|---|---|---|---|---|---|---|---|
| 162 | 85,1852 | 0,6919 | 93,2099 | 64,1975 | 0,895 | 0,791 | 0,904 |
| 81 | 81,4815 | 0,6151 | 92,5926 | 62,3457 | 0,82 | 0,806 | 0,898 |
| 23 | 82,6087 | 0,549 | 91,3043 | 58,6957 | 0,882 | 0,667 | 0,912 |

*Fuzzy Rough NN*

| *N* | *ACCURACY* | *KAPPA* | *COVERAGE* | *MEAN REL.* | *SENSIBILITY* | *SPECIFICITY* | *ROC* |
|---|---|---|---|---|---|---|---|
| 162 | 79,0123 | 0,5584 | 99,3827 | 89,5062 | 0,828 | 0,73 | 0,862 |
| 81 | 81,4815 | 0,6214 | 100 | 88,8889 | 0,848 | 0,771 | 0,9 |
| 23 | 73,913 | 0,2418 | 100 | 91,3043 | 0,789 | 0,5 | 0,784 |

*Discernibility Classifier*

| *N* | *ACCURACY* | *KAPPA* | *COVERAGE* | *MEAN REL.* | *SENSIBILITY* | *SPECIFICITY* | *ROC* |
|---|---|---|---|---|---|---|---|
| 162 | 85,1852 | 0,6901 | 100 | 100 | 0,887 | 0,8 | 0,884 |
| 81 | 85,1852 | 0,6983 | 100 | 100 | 0,889 | 0,806 | 0,925 |
| 23 | 78,2609 | 0,4041 | 100 | 100 | 0,833 | 0,6 | 0,882 |

FONTE: a autora (2017).
LEGENDA: *Accuracy* (Acurácia) *Kappa* (Estatística Kappa), *Coverage: Coverage of cases 0.95 level* (taxa de cobertura de casos com 95%* de confiança), *Mean rel.: Mean rel. region size 0.95 level* (tamanho médio dos casos cobertos com 95%* de confiança), *Sensibility* (Sensibilidade) *Specificity* (Especificidade) *ROC: Area Under the ROC Curve* (área ROC). *Estabelecido pelo *software* WEKA.

A acurácia, estatística Kappa, e Curva ROC apresentaram resultados semelhantes nos algoritmos MLP e *Discernibility Classifier* diferindo na amostra com *n* = 23, onde o MLP teve melhores resultados. A especificidade a classificação da diabetes também é semelhante entre esses dois algoritmos. Novamente, o tamanho médio das regiões cobertas é superior nos algoritmos *fuzzy rough,* em especial no *Discernibility Classifier* que teve cobertura de 100% em todas as amostras. O MLP apresentou cobertura de casos máxima ≈ 93,2%, e tamanho médio de regiões não superior ≈ 64%. O *Fuzzy Rough NN* teve na análise de amostras em caso completo desempenho inferior ao *Discernibility Classifier.*

Há diferenças sutis entre os resultados das validações nas amostras com lacunas e nas amostras com caso completo, dentre essas diferenças está a pequena melhora de desempenho do MLP em acurácia, cobertura de regiões e tamanho médio de regiões cobertas nas amostras tratadas com caso completo.

## 5 CONCLUSÕES

O modelo de triagem clinica adotado nesse trabalho se mostrou adequado para a discussão da complexidade da triagem clínica diagnóstica e dos diferentes tipos de impactos da ignorância de conhecimento e da incerteza na inferência estatística.

A manifestação de vieses provenientes da randomização é uma característica comum em modelos de decisão diagnósticos com grande quantidade de variáveis com afinidades diferentes ao diagnóstico. O viés está diretamente relacionado a afinidade dos exames ao diabetes. Exames que apresentaram baixa afinidade diagnóstica na população estudada, como os exames de perfil lipídico que se encontram em situação de bom controle global, são mais sensíveis à randomização, já que a quantidade de valores alterados para promover a separação dos grupos na amostra é pequena. O IMC é o exame com o melhor poder discriminante do *diabetes mellitus* do tipo 2, já que a obesidade se manifesta mais em diabéticas, aproximadamente o dobro, em comparação com as não diabéticas na população de estudo.

O viés classificativo é observado mais fortemente em amostras tratadas com caso completo, onde foram observadas variações bruscas de poder de correlação, condizente com distorções, e aumentos significativos da variância e desvio padrão em comparação com as respostas das amostras com lacunas.

A perda de poder de correlação nas amostras com lacunas não imputadas foi pequena em comparação com a força de correlação observada nos dados completos. Houve aumentos de correlação em algumas variáveis o que condiz com a manifestação de vieses e dependências que são características da falta de dados do tipo randômico (MAR), embora as incompletudes sejam originárias de processo totalmente randômico (MCAR). Em relação às correlações de Kendall nos dados completos, os aumentos de correlação observados nas amostras com até 50% de lacunas também foram pouco significativos, não houve variações bruscas de correlação, o que indica que a adaptabilidade do método de Kendall na presença de lacunas não imputadas, boas estimativas com pouca variabilidade de resultados e boa confiabilidade no modelo de incompletude randômica não monótona testado.

Concluindo, a quantificação da incerteza é menor ao observar parcialmente um indivíduo na população em comparação com a vertente de não o observar com a aplicação do caso completo.

Quanto aos algoritmos aplicados para a classificação e predição, é possível perceber seus potenciais e limitações quanto a incompletude de informações. O *Multilayer Perceptron*, algoritmo mais comumente utilizado para classificar problemas em saúde, mostrou-se limitado ao lidar com dados faltantes. O *Discernibility Classifier* apresentou os melhores resultados, e notável cobertura de casos e de tamanho médio de regiões cobertas, classificando os dados de forma abrangente: próxima da totalidade de regiões previstas, é assim o algoritmo com melhor adaptabilidade a condição de falta de dados dentre os testados. Comparando os dois algoritmos *fuzzy rough,* o *Discernibility Classifier* teve resultados melhores, o que indica a importância da discernibilidade deste método a dados com variáveis de distinção difícil e presença de viés, como é o caso das amostras estudadas. O desempenho superior do *Discernibility Classifier* nas áreas de informações classificadas pode indicar que o método de discernibilidade tem capacidade de incluir o conhecimento incerto representado pelas lacunas em suas classificações de maneira superior nessas amostras, em comparação com o *Fuzzy Rough NN*.

Para concluir é preciso lembrar que não há método, seja de inferência estatística ou não estatístico, capaz de predizer todas as características que dados incompletos, tal como teriam na completude. Embora seja possível observar características por meio de amostras parciais da população com baixo viés como observado nos resultados deste trabalho. A correlação de Kendall, assim é considerada adequada e robusta a tangência da incompletude com capacidade de discernir de maneira eficaz atributos pouco específicos em relação a tomada de decisão. E a capacidade do algoritmo *Discernibility Classifier* em classificar dados se aproximando de uma totalidade de cobertura classificatória mesmo em dados pouco diferenciáveis destaca a discernibilidade de características aliada ao método *fuzzy rough* como um bom decisor em ensaios clínicos e modelos diagnósticos de saúde.

## 5.1 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

O conceito de incerteza e de incompletude de dados em todos os seus âmbitos requer mais estudos, e divulgação no meio da pesquisa científica com o tocante à estatística, mineração, processamento, e extração de características de dados biológicos. Percebe-se a necessidade de expor os impactos possíveis da incerteza e da incompletude de informação na tomada de decisão em ensaios clínicos, de forma a orientar pesquisadores e quebrar paradigmas da manipulação de dados equivocada e ineficaz.

Como recomendação futura salientamos a importância do fomento à discussão de métodos com maior eficácia preditiva do conhecimento incerto na incompletude, estudando métodos com capacidade de aproximação ao conhecimento ignorado, bem como suas falhas e/ou limitações.

# REFERÊNCIAS


ABDI, H. **The Kendall Rank Correlation Coefficient**, ed, Neil Salkind, Encyclopedia of Measurement and Statistics, Thousand Oaks (CA), Sage, p. 1-7, 2007.

**ABESO – Sociedade Brasileira para o Estudo de Obesidade e Síndrome Metabólica**: Diretrizes Brasileiras de Obesidade, 2016, 4° edição, p.16. Disponível em: http://www.abeso.org.br/uploads/downloads/92/57fccc403e5da.pdf.

**ADA, Standards of Medical Care in Diabetes—2015, 2**. Classification and Diagnosis of Diabetes, American Diabetes Association, Diabetes Care, 38(Supplement 1): S8-S16, Jan 2015.

ALLISON, PD. **Missing data. Design and Inference**. Sage University Papers Series on quantitative applications in the social sciences. Thousand Oaks (CA). Cap. 4, p.72-89, 2001.

ALVO, M.; CABILIO, P. Rank correlation methods for missing data. **The Canadian Journal of Statistics** / La Revue Canadienne de Statistique. v. 23, n. 4, p. 345-358, Dec, 1995.

ARMITAGE, P. Controversies and Achievements in Clinical Trials. Elsevier Science Publishing Co. **Controlled Clinical Trials.** Department of Biomathematics. University of Oxford. England, n.5, p.67-72, 1984.

BEHRA, R. et Al. Predictive Modeling for Wellness and Chronic Conditions. **IEEE International Conference on Bioinformatics and Bioengineering (BIBE)**. p.394-398, 10-12 Nov, 2014.

BOUNDS, D.; LLOYD, P.; MATHEW, B.; WADDEL, G. A multilayer perceptron network for the diagnosis of low back pain. **IEEE International Conference on Neural Networks**. v.2, p.481-489, 24-27 Jul, 1988.

BRACARENSE COSTA, P. **Um enfoque segundo a teoria de conjuntos difusos para a meta-análise**. f.155. Tese (Doutorado em Engenharia de Produção) – Universidade Federal de Santa Catarina (UFSC), Florianópolis, 1999.

BRACARENSE COSTA, P. **Métodos Quantitativos para a Tomada de Decisão**. IESDE (ed.digital), 2008.

BURKE, S. Missing values. outliers. robust statistics & non-parametric methods. LC-GC Europe Online Supplement**, Statistics & Data Analysis**. Buckinghamshire, UK. v.2, n.0, p.19-24, 2001.

BYRON, W.; BROWN, J. Statistical Controversies in the Design of Clinical Trials Some Personal Views. **Controlled Clinical Trials**, v.1, n.1, p.13-27, Mai, 1980.

CABILIO, P.; TILLEY, J. Power calculation for tests of trend with missing observations **Environmetrics**, n.10, p.803–816, 1999.

CHANG, W. Missing data handling in multi-layer perceptron. Proceedings of the 10th **WSEAS. international conference on Computers**, Research Gate, july 2006.

CIOS, K.; MOORE, G. Uniqueness of medical data mining. **Artificial Intelligence in Medicine**, Medical Data Mining and Knowledge Discovery, v. 26, n.1-2, p. 1-24, Set/Out 2002.

DAI, J.; XU, Q. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. **Applied Soft Computing**, v.13, n.1, p.211-221, Aug. 2013.

DERSIMONEAN, R.; LAIRD, N. Meta-analysis in clinical trials. Elsevier. **Controlled Clinical Trials**, v.7, n.3, p.177–188, set. 1986.

FONSECA, J.; MARTINS, G. **Curso de Estatística**. Ed. Atlas, 6°ed., p. 147-148, 2010.

FÜLÖP, J. **Introduction to decision making methods**. Laboratory of Operations Research and Decision Systems, Computer and Automation Institute, Hungarian Academy of Sciences, In: BDEI-3 workshop. Washington. p. 1-15, 2005.

GOMES, L.; GOMES, C.; ALMEIDA, A. **Tomada de Decisão Gerencial – Enfoque em Multicritério**, ed. Atlas, São Paulo, p.1-264. 2002.

HECHT-NIELSEN, R. Theory of the backpropagation neural network. **International Joint Conference on Neural Networks (UCNN)**, v. 1, p. 593-605, 1989.

HIRAKATA, V. Estudos Transversais e Longitudinais com Desfechos Binários: Qual a melhor medida de efeito a ser utilizada?. **Revista HCPA**, v. 29, n. 2. p. 174-176, 2009.

HORTON, N.; KLEINMAN, K. Much Ado About Nothing. A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. **The American Statistician**, v. 61, v. 1, p. 79-90.

HOSSAIN, A.; ORDAZ, K.; BARTLETT, J. Missing continuous outcomes under covariate dependent missingness in cluster randomized trials. SAGE. **Statistical Methods in Medical Research**, v.0, n.0, p.1–16, 2016.

HUSH, D. Classification with neural networks: a performance analysis. **IEEE International Conference on Systems Engineering**, 24-26 Aug. 1989.
JACQUEZ, G.; WALLER, L. **Quantifying Spatial Uncertainty in Natural Resources. The effects of Uncertain Locations on Disease Cluster Statistic**, Ann Arbor Press. cap.5, p.53-63, 1996.

JAIN, A.; DUIN, R.; MAO, J. Statistical pattern recognition: a review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n.1, p.4-37. Aug. 2002.

JENSEN, R. Fuzzy-rough data mining with Weka. Tutorial for WEKA. Documentação técnica do software WEKA, 2008.

JENSEN, R.; CORNELIS, C. A New Approach to Fuzzy-Rough Nearest Neighbour Classification. **In: 6th International Conference on Rough Sets and Current Trends in Computing**, Volume 5306 of the book series Lecture Notes in Computer Science (LNCS), p. 310-319, 2008.

JENSEN, R.; SHEN, Q. New Approaches to Fuzzy-Rough Feature Selection. **IEEE Transactions on Fuzzy Systems**, v.17, n.4, p. 824 – 838, Aug. 2009.

KARPENTER, J.; KENWARD, M. **Missing data in randomized controlled trials** — a practical guide, London School of Hygiene & Tropical Medicine, London, UK, Spring, 2007.

KENDALL, M. A New Measure of Rank Correlation. Oxford University **Press on behalf of Biometrika Trust**, v.30, n.1-2, p.81-93, Jun. 1938.

KUNZ, R.; OXMAN, A. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. **The BMJ**, p.317:1185, Out. 1998.

LEE, H. **Seleção de Atributos Importantes para a Extração de Conhecimento de Bases de Dados**. f.154. Proceedings of the International Joint Conference IBERAMIA/SBIA/SBRN. 2006. thesis contest (CTDIA'2006). Tese (Doutorado em Ciências de Computação e Matemática Computacional) Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP), São Paulo, 2005.

LEUNG, Y.; LI, D. Maximal consistent block technique for rule acquisition in incomplete information systems. **Information Sciences**, v.153, p.85–106, Jul. 2003.

LIPSHITZ, R.; STRAUSS, R. Coping with Uncertainty: A Naturalistic Decision-Making Analysis. **Organizational behavior and human decision processes**, v.69, n.2, p.149–163, fev.1997.

MA, Y. On Inference for Kendall's τ within a Longitudinal Data Setting**. Journal of Applied Statistics**, v.39, n.1, p.2441–2452, dec. 2012.

MA, J.; RAINA, P.; BEYENE, J.; THABANE, L.; Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: a simulation study, **Open Access Medical Statistics**, v.2, p.93–103, dec. 2012.

MASTELLA, J. **Análise de Classes Latentes: da Teoria à Prática**. f.53. Monografia (Bacharelado em Estatística). Departamento de Estatística. Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, 2015.

MC LEOD, A. Kendall: Kendall rank correlation and Mann-Kendall trend test. R package version 2.2, 2011.

MIHELIČ, J.; MAHJOUB, A.; RAPINE, C.; ROBIČ, B. Two-stage flexible-choice problems under uncertainty. **European Journal of Operational Research**, v. 201, n.2, p.399-403, 2010.

MOLENBERGHS, G. Incomplete Data in Clinical Studies: Analysis. Sensitivity. and Sensitivity Analysis. Data Management-Missing Data. **Drug Information Journal**, Therapeutic Innovation & Regulatory Science, v. 43, n. 4, p. 409 – 429, jul. 2009.

MOLENBERGHS et Al. Analyzing Incomplete Longitudinal Clinical Trial Data. Biostatistics. **Biometrika Trust**, v.5, n.3, p.445-64, jul, 2004.

NASIRI, JH.; MASHINCHI, M. Rough Set and Data Analysis in Decision Tables. **Journal of Uncertain Systems**, v.3, n.3, p.232-240, Aug. 2009.

**NHS-UK. 2016,** SCREENING IN THE UK: MAKING EFFECTIVE RECOMMENDATIONS 2015 to 2016, National Screening Committee, First report of the UK National Screening Committee, Department of Health, London, UK.

PANDA, M.; HASSANIEN, A.; ABRAHAM**,** A. **Biometrics: Concepts. Methodologies. Tools and applications,** Hybrid Data Mining Approach for Image Segmentation Based Classification Information, Resources Management Association (USA), ed. IGI Global, Cap. 64, p.1549-1550, 2017.

PATTARAINTAKORN, T.; CERCONE, N. Integrating rough set theory and medical applications. **Applied Mathematics Letters**, v.2, n.4, p.400–403, abr.2008.

PAWLAK, Z. Rough Sets. **International Journal of Computer & Information Sciences**, v.11, n.5, p.341–356, sep.1982.

PAWLAK, Z. **Decision Tables - A Rough Set Approach**, EATCS, p. 85-95, 1987.

PAWLAK, Z.; BUSSE, J.; SLOWINSKI, R.; ZIARKO, W. Rough Sets. **Communication of the ACM**, v.38, n.11, p.88-95, nov.1995.
PAWLAK, Z. Rough set theory and its applications to data analysis. **Cybernetics and Systems an International Journal**, v.29, n.7, p.661-688, 1998.

PAWLAK, Z. **Rough sets and decision tables**. Computation Theory, Volume 208 of the series Lecture Notes in Computer Science, cap.187-196, 2005.

PIANTADOSI, S. **Clinical Trials: A methodologic Perspective**. Wiley series in probability and statistics, 2ed, 2005.

RAGSDALE, C. **Spreadsheet modeling and decision analysis: a practical introduction to management science**. Cengage Learning, ed.6, p.1-792, 2010.

RAIFFA, H. **Teoria da decisão: aulas introdutórias sobre escolhas em condições de incerteza**. Ed.Vozes, São Paulo, p.1-346, 1977.

RICH, S.; CEFALU, W. The Impact of Precision Medicine in Diabetes: A Multidimensional Perspective. **Diabetes Care**, v. 39, n.11, p.1854-1857, nov. 2016.

RIZA, L. et. Al. Implementing algorithms of rough set theory and fuzzy rough set theory in the R package ‘‘RoughSets’’, **Information Science**, v. 287, p.68-89, dec. 2014.

ROY, B. **Multicriteria methodology goes decision aiding.** Kluwer Academic Publishers, Springer, p. 269-276, 1999.

RUBIN et Al. Inference and missing data. **Biometrika**, n.63, p.581-592, dec. 1976.

**SBC. Sociedade Brasileira de Cardiologia**. Diretrizes da Sociedade Brasileira de Cardiologia 2013-2015, V – Diretriz Brasileira de Dislipidemia e Prevenção da Aterosclerose, p.468. Disponível em: http://www.smc.org.br/diretrizes_sbc.html.

**SBD. Sociedade Brasileira de Diabetes** - Diretrizes da Sociedade Brasileira de Diabetes 2014 - 2015. São Paulo - SP: AC farmacêutica. 2015. Disponível em: http://www.diabetes.org.br/images/2015/area-restrita/diretrizes-sbd-2015.pdf.

**SBD – Sociedade Brasileira de Diabetes**. Diretrizes da Sociedade Brasileira de Diabetes. 2015-2016. Métodos e critérios para o diagnóstico. p.11. Disponível em: http://www.diabetes.org.br/profissionais/images/pdf/DIRETRIZES-SBD-2015-2016.pdf.

**SBN – Sociedade Brasileira de Nefrologia**. e-Book: Biomarcadores em Nefrologia, 2011. Disponível em: http://arquivos.sbn.org.br/pdf/biomarcadores.pdf.

SCHAFER, J. **Analysis of incomplete multivariate data**. Series: Chapman & Hall/CRC Monographs on Statistics & Applied Probability CRC PRESS, ed. Chapman and Hall, p. 1-448, 1997.

SHI, K. S-Rough sets and its applications in diagnosis recognition for disease. **Proceedings of the First International Conference on Machine Learning and Cybernetics**, Beijing, 4-5 nov, 2002.

SHIFFMAN, R. Representation of Clinical Practice Guidelines in Conventional and Augmented Decision Tables. Model Formulation, **Journal of the American Medical Informatics Association**, v. 4, n.5, p.382 – 393, Set/Out, 1997.

SIEGEL, S.; CASTELLAN, J. **Estatística Não-Paramétrica para Ciências do Comportamento, Métodos de Pesquisa**, Artmed, Bookman, 2º ed, p. 287. 294, p. 318, 325-326, 2006.

SLOWIŃSKI, K.; SLOWIŃSKI, R.; STEFANOWSKI, J. Rough sets approach to analysis of data from peritoneal lavage in acute pancreatitis. **Medical Information**, v.13, n.3, p.143-59, jul/set,1988.

SMITHSON, M. **Ignorance and Uncertainty, Emerging Paradigms.** Springer-Verlag, New York, p. 1-40, 1988.

STERNE, J. et Al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. **The BMJ**, 338: b 2393, jun. 2009.

TCHEMRA, A. **Tabela de Decisão Adaptativa na Tomada de Decisão Multicritério**. Tese (Doutorado em Sistemas Digitais), Escola Politécnica, Universidade de São Paulo (USP), São Paulo, 2009.

TIETZ, N.W. **Clinical guide to laboratory tests**. 3ed. Sauders: Philadelphia, 1995.

TU, J. Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. **Journal Clinical Epidemiology**, v.49, n.11, p.1225-1231, nov. 1996.

TUNIS, S.; STRYER, D.; CLANCY, C. Practical Clinical Trials: Increasing the Value of Clinical Research for Decision Making in Clinical and Health Policy. **JAMA**, v. 290, n.12, p.1624-1632, set. 2003.

**USFCA – University of San Francisco**, Department of computer science - WekaDataAnalysis, 2017. Disponível em: http://www.cs.usfca.edu/~pfrancislyon/ courses/640fall2014/WekaDataAnalysis.pdf.

VERAS, R. Estratégias para o enfrentamento das doenças crônicas: um modelo em que todos ganham. **Revista Brasileira de Geriatria e Gerontologia**, Artigo de opinião, v.14, n.4, p.779-786, 2011.

**WEKA. 2017, a.** PENTAHO FORUMS – Missing Values in Multilayer Perceptron. Disponível em: http://forums.pentaho.com/showthread.php?96337-Missing-Values-in-Multilayer-Perceptron-(MLP-Neural-Networks)

**WEKA. 2017, b.** WEKA FORUM - Confidence Interval. Disponível em: http://weka.8497.n7.nabble.com/confidence-interval-td25141.html

**WEKA. 2017, c.** WEKA FORUM - Rel. Region. Disponível em: http://weka.8497.n7.nabble.com/Rel-Region-td24399.html#a24400

**WEKA. 2017, d.** WEKA WIKISPACE - Area under the curve. Disponível em: https://weka.wikispaces.com/Area+under+the+curve

WOOD, A.; WHITE, I.; THOMPSON, S. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. **SAGE Clinical Trials**, n.1, p.368-76, ago. 2004.

ZIARKO, W. Variable precision rough set model**. Journal of Computer and System Science**, n. 46, p.39-59, fev.1993.

## ANEXO – SCRIPT DA FUNÇÃO LACUNAS EM PORCENTAGEM

```
# R - INSERÇÃO DE LACUNAS EM PORCENTAGEM NO DATAFRAME #

Lacunas <- function(df, prop = .1){  #porcentagem:.1 = 10%#
    n <- nrow(amostra)
    m <- ncol(amostra)
    num.to.na <- ceiling(prop*n*m)
    id <- sample(0:(m*n-1), num.to.na, replace = FALSE)
    rows <- id %/% m + 1
    cols <- id %% m + 1
    sapply(seq(num.to.na), function(x){
            df[rows[x], cols[x]] <<- NA
        }
    )
    return(df)
}
```