UNIVERSIDADE FEDERAL DE JUIZ DE FORA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Guilherme Coelho Neves

QUALIDADE DE DADOS ATRAVÉS DE ÁRVORES DE DECISÃO

Uma Aplicação a Dados de triagem de Pacientes com Suspeita de Tuberculose

JUIZ DE FORA

2014

Guilherme Coelho Neves

Qualidade de Dados através de Árvores de Decisão

Uma Aplicação a Dados de triagem de Pacientes com Suspeita de Tuberculose

Monografia apresentada ao curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a obtenção do grau de Bacharel em Estatística.

Orientador: Ronaldo Rocha Bastos

PhD em Urban and Regional Planning-Liverpool University

JUIZ DE FORA

2014

Qualidade de Dados através de Árvores de Decisão

Monografia apresentada ao curso de Estatística da Universidade Federal de Juiz de Fora, como requisito para a obtenção do grau de Bacharel em Estatística.

Aprovada em 3 de Fevereiro de 2014.

BANCA EXAMINADORA

_____

Ronaldo Rocha Bastos (orientador)

PhD em Urban and Regional Planning-Liverpool University

_____

José Márcio Faier

Doutor em Engenharia Elétrica- COPPE/UFRJ

_____

Augusto Carvalho Souza

Mestre em Estudos Populacionais e Pesquisas Sociais – ENCE/IBGE

Neves, Guilherme Coelho – Juiz de Fora, 2014

Qualidade de Dados através de Árvores de Decisão - Uma Aplicação a Dados de triagem de Pacientes com Suspeita de Tuberculose/ Guilherme Coelho Neves - 2014

60.p

Monografia – Universidade Federal de Juiz de Fora - Instituto de Ciências Exatas

Orientador: Ronaldo Rocha Bastos

À minha família.

"Eu escolho viver, não apenas existir."

James Hetfield

AGRADECIMENTOS

Primeiramente agradeço a Deus por ter me ajudado a superar todas as dificuldades encontradas pelo caminho, a superar os obstáculos e sempre me indicar o caminho certo.

Agradeço á meus pais José e Cirene por toda educação e incentivo que me deram sempre, foram fundamentais para esta conquista.

Ao meu irmão Rafael por estar sempre comigo e por toda ajuda nesta conquista.

À toda minha família, em especial às Tias Marisa e “Katinha”, por todo o apoio neste período e sempre.

Aos companheiros tiroleses: Rafael (“Tião”), Fabrício (“Vô”), Tharlles (“Tatu”), Ruy, Ednilton (“Ed”) e Coutinho pelo período de convivência na república mais hilária de Juiz de Fora.

Aos colegas do curso de Exatas: Leandro, Ruy e Daniel por se tornarem mais que amigos, sendo minha segunda família em Juiz de Fora.

Aos colegas do curso de Estatística: Manoel, Thalita e Amanda por estarem sempre dispostos a ajudar.

Aos amigos Robert e Albert, que mesmo longe sempre estiveram presentes.

À minha namorada Ana Carolina por me apoiar, compreender nos momentos difíceis e sempre acreditar nesta conquista. À dona Sílvia, seu Ricardo, Ana Lívia e Ana Júlia por todo apoio, já sendo parte da família.

Ao meu orientador Ronaldo, por ter me orientado nestes quase quatro anos de Iniciação Científica, por me aconselhar nas dúvidas profissionais e compreender nos momentos de dificuldade. Certamente aprendi muito com seus ensinamentos, e não conseguiria chegar até aqui sem os mesmos.

Ao José Márcio pelo apoio e por ter me dado a oportunidade de me aprofundar neste tema, além de ter aberto as portas para o mestrado na COPPE.

Aos Professores do departamento, pelos ensinamentos passados, pela paciência e pelas aulas ministradas. Em especial ao professor Marcel que além de ser um ótimo professor, sempre apoiou nossas iniciativas em prol da constante melhoria do curso. Ao professor Clécio pelos conselhos, dicas e por “tomar conta” da gente nos congressos. Ao professor Augusto, por sempre estar disposto a ajudar, principalmente durante minha Iniciação Científica. Ao professor Lupércio por compartilhar suas experiências

profissionais, nos preparando para o mercado. À professora Camila pelas excelentes aulas ministradas. À professora Ângela pela paciência e por analisar meus créditos, senão eu não formaria. Aos professores Joaquim e Márcio pelas conversas, conselhos e ensinamentos, e ao professor Ronaldo por ser o orientador deste trabalho, e também pelos grandes conselhos dados que vão muito além de meros estudos.

Aos meus alunos de aulas particulares e a todos que contribuíram nesta caminhada.

A todos vocês, muito obrigado.

RESUMO

Após a implementação de um programa baseado em redes neurais artificiais para a triagem de pacientes quanto ao diagnóstico de tuberculose em Unidades de Saúde do Brasil, foi identificado que os dados coletados poderiam estar “viesados”. Esta suspeita ocorreu ao comparar os dados e saídas obtidas pela rede neural com o diagnóstico dos médicos.

Através de entrevistas e investigações feitas por parte dos responsáveis pelo *software* implantado, foi identificado que os usuários do programa, muitas vezes pressionados pelas condições de trabalho, poderiam estar alterando os dados de entrada dos pacientes para que obtivessem como possível diagnóstico pela triagem uma alternativa que não a tuberculose.

Neste trabalho, é proposta uma metodologia que busca identificar a confiabilidade da informação processada pelo sistema, bem como aplicar técnicas que retardam e desmotivam o preenchimento tendencioso do questionário por parte do usuário. Utilizando de técnicas de agrupamento de dados (árvores de decisão) e de visualização de dados (Análise De Componentes Principais - para dados categóricos -Análise de Correspondência), pretende-se não só qualificar os dados, como também quantificar a confiabilidade dos mesmos.

Sumário

Lista de Figuras

Lista de Tabelas

1- Introdução

Através de um projeto de pesquisa na área de Tuberculose Pulmonar, onde foi desenvolvido um sistema de assistência ao diagnóstico médico, uma análise de qualidade dos dados de entrada foi realizada.

Este sistema que auxilia o diagnóstico médico (“NeuralTb-SAPEM”, Boletim Faperj-2012) é baseado em redes neurais artificiais e está dividido em dois subsistemas. O primeiro faz a classificação do grupo de risco ao qual o paciente pertence: Baixo, Médio ou Alto. O segundo determina se o paciente é portador ou não tuberculose.

Utilizando-se de uma base de dados de triagem de pacientes com suspeita de Tuberculose, podemos identificar através de classificações obtidas por Árvores de Decisão variáveis que afetam de forma significativa os resultados obtidos pelo sistema. Além disso, também podemos analisar como variáveis faltantes ou errôneas podem influenciar nos resultados. Os dados utilizados neste trabalho são apenas para corroborar a proposta da metodologia, não sendo interpretados seus resultados. Sendo a variável “Desfecho” , o resultado obtida através da rede neural e não do diagnóstico médico.

Através da definição de critérios que compõem a qualidade de dados para o objetivo em questão, podemos definir um “score” que classifique os dados de entrada, dimensionando a qualidade da informação.

O objetivo deste trabalho é apresentar a metodologia dos métodos de classificação, tais como a Árvore de Decisão e de Análise Multivariada e posteriormente fazer uma aplicação dos mesmos aos dados de pacientes com suspeita de tuberculose, criando “scores” e classificando os dados quanto à sua confiabilidade.

Este trabalho será apresentado da seguinte forma: além da introdução, o capítulo 2 apresenta a teoria sobre métodos de classificação, em especial a Árvore de Decisão; no capítulo 3 será feita uma breve revisão sobre a metodologia de Análise Multivariada, em especial a Análise de Componentes Principais e algumas de suas propriedades; no capítulo 4 abordaremos o tema Qualidade de Dados; no capítulo 5 será feita uma aplicação das metodologias dentro do problema estudado e apresentação dos principais resultados; o capítulo 6 apresenta as conclusões e pesquisas futuras.

2- Árvores de Decisão

Neste capítulo serão apresentados os conceitos de métodos de classificação, focado no método de árvores de decisão.

2.1- Introdução

Em geral, as aplicações que compreendem a área de Inteligência Artificial estão relacionadas à criação de um ou mais modelos computacionais de classificação. Na elaboração de tais modelos, a associação entre as classes e o conjunto de atributos que caracterizam os objetos a serem classificados pode se dar de formas variadas, empregando processamento simbólico e/ou numérico. Segundo BASGALLUP (2010), a construção dos modelos computacionais de classificação geralmente emprega um dentre dois paradigmas alternativos:

- "Top-down": obtenção do modelo de classificação a partir de informações fornecidas por especialistas;
- "Bottom-up": obtenção do modelo de classificação pela identificação de relacionamentos entre variáveis dependentes e independentes em bases de dados rotulados (que possuem variáveis categóricas). O classificador é induzido por mecanismos de generalização fundamentados em exemplos específicos (conjunto finito de objetos rotulados), apesar de existir propostas também para dados não-rotulados.

As árvores de decisão, foco deste trabalho, estão baseadas no paradigma "Bottom-up" e possuem seguinte pré-requisito: Toda informação sobre cada objeto (caso) a ser classificado deve poder ser expressa em termos de uma coleção fixa de propriedades ou atributos. Dessa forma, objetos distintos não podem requerer coleções distintas de atributos. Bases de dados que atendem a este requisito são denominadas "*flat files*", segundo BASGALLUPP (2010).

O processo para obtenção do número de classes pode ser fruto de um treinamento supervisionado ou não. O número de classes pode ser definido a priori, o que transforma a modelagem num processo de treinamento supervisionado, ou então será definido automaticamente a partir dos dados disponíveis, o que caracteriza um processo de treinamento não-supervisionado. Pode ainda haver duas possibilidades de classes: discreta e contínua.

Necessita-se de um número bem maior de observações (objetos) do que classes, inclusive para permitir a aplicação de testes estatísticos. A quantidade adequada de objetos vai depender do número de atributos, do número de classes e da complexidade intrínseca ao modelo de classificação. A tarefa de classificação deve poder ser implementada de forma lógica, ou seja, empregando uma base de regras de decisão. Assim, a classificação de cada objeto pode ser descrita por uma expressão lógica. Em contrapartida a este requisito, podemos mencionar a classificação por operações aritméticas, empregada em discriminantes lineares, por exemplo, que realizam a classificação por uma combinação linear dos atributos, seguida da comparação com um pré-determinado limiar.

Assim, cada objeto da base de dados a ser estudada deverá ser constituído por um conjunto de atributos (propriedades, características), sendo que a cada objeto deve ser associada uma classe, dentre um conjunto de classes possíveis. Os atributos são variáveis observáveis e independentes, que assumem valores em variados domínios, podendo ser Contínuos, Categóricos ordinais e Categóricos não-ordinais (nominais).

Árvores de decisão são geralmente aplicadas a grandes bases de dados. Para tanto regularidades implícitas presentes na base de dados devem ser descobertas automaticamente e expressas, predominantemente, na forma de regras. Conhecimentos de Inteligência Artificial e Estatística são comumente empregados para a geração das árvores de decisão.

2.2- Classificação em Árvores de Decisão

Uma das aplicações de classificação por árvores de decisão é o diagnóstico médico. No mesmo são definidos e analisados atributos (variáveis) para cada paciente analisado. Neste caso, a tarefa do responsável pela classificação é mapear estes atributos (sexo, idade, febre, etc.) para um diagnóstico. De acordo com TAN *et al.* (2005), a classificação pode ser utilizada para os seguintespropósitos: modelagem descritiva e modelagem preditiva. Os autores ainda distinguem estes propósitos da seguinte forma:

Na **modelagem descritiva**, um modelo de classificação é utilizado para diferenciar exemplos de classes diferentes. Como exemplo, um médico poderia utilizar um modelo de classificação descritiva para identificar quais são os principais fatores associados a uma determinada enfermidade, como por exemplo, a Tuberculose e ainda verificar, por exemplo, que a maioria dos pacientes com Tuberculose apresentou sudorese noturna e tosse contínua. Já na **modelagem preditiva**, o modelo pode ser utilizado para classificar unidades amostrais

cujas classes são desconhecidas, que não fizeram parte da elaboração do modelo. Assim, podemos ter como exemplo, um médico que já tenha construído um modelo de classificação para identificação de pacientes com tuberculose com dados de pacientes atendidos por ele e queira através destes dados, construir um modelo para classificar novos pacientes. Temos na **Figura 1** que segue a diagramatização do processo indutivo de um classificador e sua respectiva utilização.

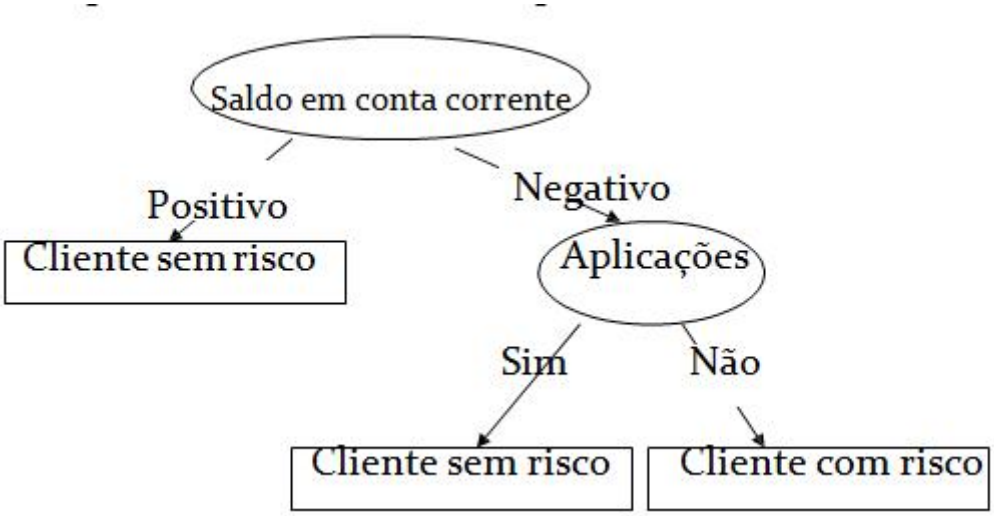
Fonte: Notas de aula dos Profs. Fernando J. Von Zuben & Romis R. F. Attux DCA/FEEC/Unicamp

Figura 1 – Indução de um classificador e dedução para novas observações.

Na **figura 1**, temos em um primeiro momento, o conjunto de treinamento com classificações conhecidas, de onde se define um algoritmo para a construção de um modelo. Após a elaboração desse modelo, esse classificador pode ser utilizado para predizer rótulos de classes desconhecidas de um conjunto de teste. Outra grande vantagem da árvore de decisão é sua estrutura de fácil interpretação, onde temos uma seqüência de perguntas com suas possibilidades de respostas respeitando uma hierarquia organizada em “ramos” e “nós” aparentando uma árvore. No exemplo dos diagnósticos, conseguiríamos notar esta estrutura, tal como diz ZUBEN *(2013)* onde é possível utilizar uma árvore de decisão para classificar um novo paciente como saudável ou doente. Para isso, basta partir do nó raiz da árvore e ir percorrendo-a, através das respostas aos testes dos nós internos, até chegar a um nó folha, o

qual indica a classe correspondente do novo paciente. A trajetória percorrida até o nó folha representaria uma "regra" o que facilitaria a interpretação da árvore.

Como exemplo para "vizualização" da árvore de decisão, e interpretação de seus nós e folhas (ramos), ZUBEN (2013) apresenta em suas notas de aula um exemplo para tomada de decisão sobre liberação ou não de crédito para clientes de uma instituição financeira. Observemos a **figura 2** em seguida:



Fonte: Notas de aula dos Profs. Fernando J. Von Zuben & Romis R. F. Attux DCA/FEEC/Unicamp

Figura 2 – Exemplo fictício de árvore de decisão, tomando atributos de clientes de alguma instituição financeira.

Para um dado conjunto de dados, podemos ter diversas árvores de decisão, basta variar os caminhos percorridos e a classificação-alvo (que nesse exemplo é o risco dos clientes). Podemos interpretar a árvore acima, da seguinte forma: Iniciamos pela raiz, aplica-se um teste de decisão, obtemos uma classificação. A partir daí repete-se o procedimento para as sub-árvores em seqüência. Ao construir uma árvore de decisão, procura-se associar a cada nó de decisão o atributo "mais informativo" entre aqueles ainda não utilizados no caminho desde a raiz da árvore. Para cada árvore temos um algoritmo distinto, onde cada algoritmo tem a sua própria metodologia para distinguir o atributo mais informativo, fazendo com que a topologia da árvore e a qualidade da mesma variem em função do algoritmo utilizado.

2.3- *Top-Down Induction of Decision Tree* (TDIDT)

O TDIDT é um algoritmo que serve de base para a construção de diversos outros algoritmos comumente utilizados para a definição de critérios em árvores de decisão, tais como ID3 (QUINLAN, 1986), C4.5 (QUINLAN, 1993) e CART (BREIMAN *et al.*, 1984).

De acordo com BRAMER (2007), o TDIDT é baseado no processo de particionamento recursivo, onde são feitas, para cada unidade observada, sucessivas divisões de seus atributos preditivos para a produção de regras de decisão. Ainda, segundo BRAMER (2007), dado que tenhamos uma base de dados (em treinamento) "T" possuindo C_k classes, o TDIDT seria baseado em 3 possibilidades:

1. *T* contém um ou mais objetos, sendo todos da classe C_j. Assim, a árvore de decisão para *T* é um nó folha que identifica a classe C_j.
2. *T* não contém objetos. A árvore de decisão também é um nó folha, mas a classe associada deve ser determinada por uma informação externa. Por exemplo, pode- se utilizar o conhecimento do domínio do problema.
3. *T* contém exemplos pertencentes a mais de uma classe. Neste caso, a idéia é dividir *T* em subconjuntos que são, ou tendem a dirigir-se para, coleções de exemplos com classes únicas. Para isso, é escolhido um atributo preditivo *A*, que possui um ou mais possíveis resultados O_1, O_2,..., O_n. *T* é particionado em subconjuntos T_1, T_2,..., T_n, onde T_i contém todos os exemplos de *T* que têm resultado O_i para o atributo *A*. A árvore de decisão para *T* consiste de um nó de decisão identificando o teste sobre o atributo *A*, e uma aresta para cada possível resultado, ou seja, *n* arestas. No lugar de um único atributo *A*, pode também ser considerado um subconjunto de atributos.

De forma simplificada, analisando todo o conjunto de dados (nó raiz) o algoritmo TDIDT procura, sobre um conjunto de atributos, aqueles que "melhor" dividem o conjunto de exemplos em subconjuntos. Para cada nó é escolhido um atributo preditivo para representar esta divisão (como por exemplo "sim" para a divisão de uma classificação se o paciente tosse ou não) e assim é feito para as seguidas divisões e seus respectivos subconjuntos. Repetimos o algoritmo até que todas as observações estejam classificadas ou até que todos os atributos preditivos (classes) já tenham sido utilizados. Neste trabalho utilizaremos o algoritmo CART, principalmente por nossa variável de interesse ser binária (Apresentar ou não Tuberculose), um dos requisitos para a utilização deste algoritmo. Este algoritmo pode ser encontrado em BREIMAN *et* al.(1984), que explicita todas suas definições e pré-requisitos de utilização.

2.4- Escolha dos atributos preditivos para os nós da árvore

Existem diferentes tipos de critérios de seleção e estes podem variar com o algoritmo utilizado. O critério é o que irá definir qual atributo preditivo será utilizado em cada nó da árvore decisão. Segundo TAN *et al.* , (2005) [cada nó interno da árvore é dividido de acordo com um único atributo (divisão uni variável), buscando o melhor atributo para realizar esta divisão.]

Os critérios de seleção podem ser baseados em diferentes medidas, tais como distância, dependência e impureza, sendo esta última a mais utilizada. Utilizando-se do critério de impureza, os algoritmos subdividem o nó pai- originador de outros nós - de forma que os nós filhos – originários de um nó pai - sejam mais puros possíveis, ou seja, que a distribuição de classes esteja o mais balanceado (bem distribuído) possível. A definição de impureza pode ser mais bem interpretada ao analisar as classes de um determinado nó; assim a impureza é nula se todos os exemplos do nó pertencer à mesma classe. Analogamente, o grau de impureza é máximo no nó se houver o mesmo número de exemplos para cada classe possível. Uma das medidas comumente utilizadas para a avaliação da melhor divisão é o *Gini*, a qual emprega um índice de dispersão estatística proposto em 1912 pelo estatístico italiano *Corrado Gini*, ver TAN *et al.* , (2005).

2.5- Índice Gini para avaliação da melhor divisão

Sendo o índice Gini uma das medidas comumente utilizadas para a avaliação da melhor divisão dos dados. Este foi utilizado neste trabalho como critério de avaliação, pois ele é utilizado no algoritmo CART (BREIMAN et al., 1984). O *gini-index* é definido pela equação (1):

$$gini - index(nó) = 1 - \sum_{i=1}^{c} p(i\ /nó) \qquad (1)$$

Assim, basta calcular a diferença entre o *gini-index* do nó antes e depois e teremos o índice Gini, que pode também ser representado pela equação (2):

$$Gini = gini - index(pai) - \sum_{j=1}^{n} \left[\frac{N(v_j)}{N} gini - index(v_j)\right] \qquad (2)$$

onde *n* é o número de valores do atributo, ou seja, o número de nós-filhos, *e* N é o número total de objetos do nó-pai e $N(v_j)$ é o número de exemplos associados ao nó-filho(v_j). Assim, selecionamos o atributo que gerar um maior valor para o índice gini.

O modo como são representados os nós (e a construção da árvore) pode influenciar na interpretação da mesma. Existem diferentes tipo de representação dos nós para o particionamento dos dados. Entre os diversos métodos, temos que os mais conhecidos, segundo TAN *et al.* , (2005) são:

- Um ramo para cada valor de atributo;
- Solução de Hunt;
- Atributos categóricos ordinais;
- Agrupamento de valores em dois conjuntos;
- Agrupamento de valores em vários conjuntos.

Durante a construção de árvores de decisão, muitas vezes encontramos o problema de "sobreajuste", que é quando os ramos das sub-árvores podem apresentar erros ou ruídos, ou seja, significa um aprendizado muito específico do conjunto de treinamento. Para resolver este problema, temos o mecanismo de poda, que detecta e exclui estas sub-árvores com "sobreajuste" e repetindo este processo sucessivamente, geramos um conjunto de árvores podadas. Por fim, para cada uma delas é calculada a acurácia na classificação de um dado conjunto de dados, sendo que a acurácia do classificador representa a porcentagem de observações do conjunto de teste que são corretamente classificadas por ele. Caso a acurácia seja alta, o modelo de classificação é considerado eficiente e pode ser utilizado para classificar novos casos. Assim, a árvore que obtiver a melhor acurácia será a escolhida. Neste trabalho,

utilizaremos a árvore de decisão para identificar as variáveis que mais influenciam a variável desfecho (Apresentar ou não Tuberculose). Aplicando a metodologia em uma base de dados de aproximadamente 1064 pacientes, pretende-se além de identificar tais variáveis, analisar as diversas combinações das mesmas que levem a um desfecho positivo para a presença de Tuberculose. Identificando ainda a probabilidade de ocorrência daquela combinação (caminho da árvore) e a certeza da árvore sobre a classificação da variável desfecho.

3-Análise de Componentes Principais

No capítulo que segue apresentaremos as definições e conceitos a respeito de um dos métodos de Análise Multivariada para variáveis numéricas, a Análise de Componentes Principais.

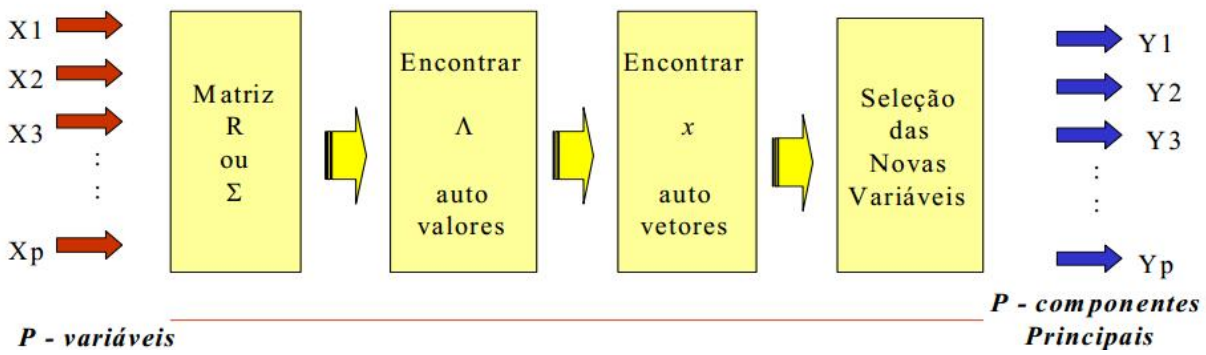
3.1- Introdução

A Análise de Componentes Principais (ACP) é uma técnica que possibilita a extração de informações de um dado conjunto de dados (variáveis numéricas) através de um outro conjunto de dados, fruto de uma combinação linear das variáveis do conjunto original, chamados de componentes principais. Assim, a ACP resulta em uma redução do número de variáveis do conjunto de dados original, com a menor perda possível de informações, identificando as medidas responsáveis pelas maiores variações entre os resultados. A técnica foi desenvolvida a partir da necessidade de se analisar conjuntos de dados com um número elevado de variáveis correlacionadas.

A chave para o estudo de dados utilizando a ACP é sem dúvida a determinação de autovalores e autovetores das matrizes de covariância para os dados originais, e correlação para dados padronizados. A interpretação de tais medidas é peça – chave para obtenção e determinação do método. A ACP é muito utilizada quando se estuda fenômenos ou processos em que diversas características devem ser observadas ao mesmo tempo.

3.2- O método

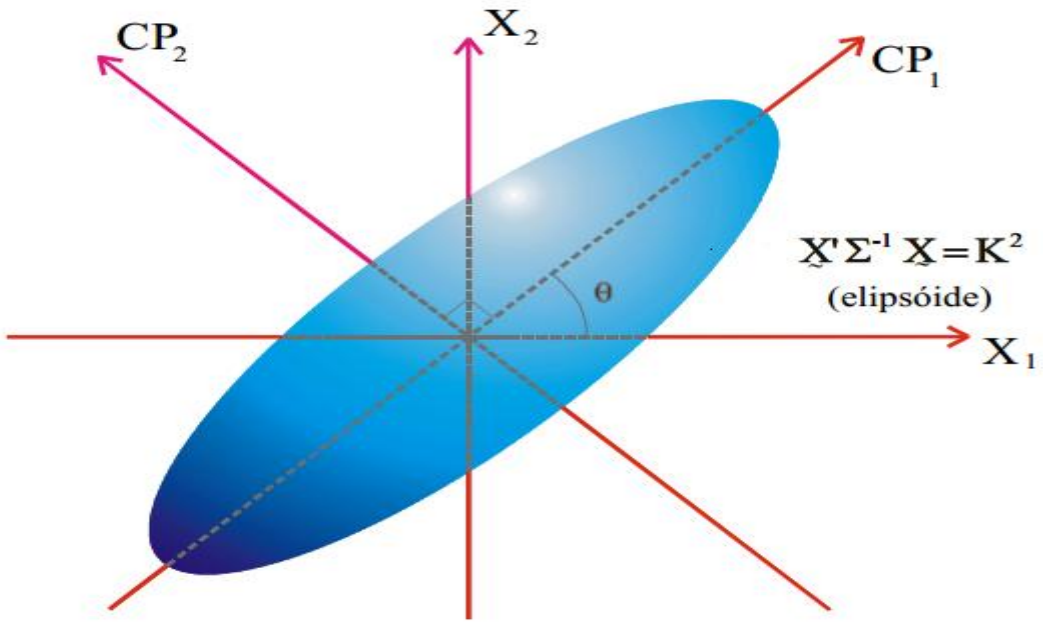
O principal ideal por trás da ACP é a redução de dimensionalidade do conjunto de dados, agrupando indivíduos similares em gráficos de fácil interpretação, mantendo ao máximo a variabilidade do mesmo e com a menor perda possível de informação. Essa redução de dimensão é chamada de transformação de *Karhunnen-Loéve* (conhecida como Transformada de Hotelling em outros campos de pesquisa), no qual as componentes principais são os autovalores ou vetores da matriz a ser considerada (covariância ou correlação). Através da ACP, substitui-se o conjunto de variáveis original, que estão correlacionadas, por novas variáveis não correlacionadas e ortogonais, chamadas de Componentes Principais (“CP”). A figura que segue apresenta o esquema de aplicação da ACP:



Fonte: SOUZA (2000, p.25)

Figura 3– Esquema de aplicação de análise de componentes principais.

Seguindo a figura acima, vimos que é necessário o cálculo da matriz de covariância (∑), ou a matriz de correlação (R), encontrar os autovalores e os autovetores correspondentes e, por fim, escrever as combinações lineares, que serão as novas variáveis, denominadas de componentes principais, sendo cada componente principal uma combinação linear de todas as variáveis originais, independentes entre si e estimadas com o propósito de reter, em ordem de maior variação, a variação total contida nos dados iniciais. (REGAZZI, 2001). Suponha ainda, que estejamos analisando um conjunto de dados com apenas duas variáveis X1 e X2, conforme a figura que segue; podemos observar um elipsóide com densidade de probabilidade constante:



Fonte: LOPES (2001, p.31)

Figura 4– Elipsóide de densidade constante.

Analisando a figura anterior, temos que o eixo de maior explicação (maior variância), o maior da elipse (CP1) é proporcional à raiz quadrada do autovalor correspondente aquele eixo ($\sqrt{\Lambda_1}$). Da mesma forma, o eixo de menor explicação (menor variância), definido por CP2 é perpendicular à CP1. O CP2 é proporcional a $\sqrt{\Lambda_2}$). Quando fazemos a análise de ACP, deslocamos os eixos X1 e X2 para a direção de maior variabilidade. (JOHNSON & WICHERN, 1992).

Quando olhamos para a disposição dos dados a serem analisados por meio da ACP, necessitamos que eles estejam dispostos preferencialmente na forma de uma matriz n x p, onde X1, X2,..., Xp serão as variáveis a serem analisadas; denominamos este conjunto de dados a serem analisados de matriz X, que deverá ser convertida em uma matriz quadrada de correlação ou covariância para ser tratada pela ACP. Segue abaixo um exemplo da matriz de dados comentada acima:

Indivíduos	Variáveis							
	X_1	X_2	X_3	X_4	...	X_j	...	X_p
1	X_{11}	X_{12}	X_{13}	X_{14}	...	X_{1j}	...	X_{1p}
2	X_{21}	X_{22}	X_{23}	X_{24}	...	X_{2j}	...	X_{2p}
3	X_{31}	X_{32}	X_{33}	X_{34}	...	X_{3j}	...	X_{3p}
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	...	X_{ij}	.	X_{ip}
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
n	X_{n1}	X_{n2}	X_{n3}	X_{n4}	...	X_{nj}	...	X_{np}

Fonte: REGAZZI (2001)

Tabela 1– Matriz de dados de n indivíduos e p variáveis.

Preferencialmente, desejamos um número sempre alto de variáveis e que elas sejam as mais interdependentes possíveis, pois assim será mais fácil comparar indivíduos,

representando esta interdependência pela matriz de correlação R ou pela matriz de variância-covariância ∑.

A matriz ∑ está associada a um vetor aleatório $\vec{X}$=[X1, X2, ..., Xp] , onde ($\widehat{\Lambda_1}$, X1) , ($\widehat{\Lambda_2}$, X2), ... ($\widehat{\Lambda_p}$, Xp) são pares de autovalores e autovetores estimados de uma determinada amostrada a ser analisada.

Temos ainda que $\widehat{\Lambda_1} \geq \widehat{\Lambda_2} \geq \cdots \geq \widehat{\Lambda_p} \geq 0$, e fornecendo o i-ésimo componente principal dado por:

$$Y_I = \vec{x}_i X = \vec{x}_{1i} X_1 + \vec{x}_{2i} X_2 + \cdots + \vec{x}_{Pi} X_P, onde\ i = 1,2,\ldots,p. \quad (3)$$

Definindo também:

$$Var(Y_I) = x_i' \sum x_i = \widehat{\Lambda_1}, onde\ i = 1,2,\ldots,p. \quad (4)$$

$$Cov(Y_I, Y_k) = x_i' \sum x_k = 0 \ \ onde\ i = 1,2,\ldots,p. \quad (5)$$

Dadas as definições acima, temos que os componentes principais (CP), são não correlacionados e possuem variâncias iguais ao autovalor de ∑. (JOHNSON & WICHERN, 1992).

Segundo REIS (1997), a correta aplicação da ACP deverá incluir:

- As variáveis incluídas na análise;
- O valor percentual das variâncias explicadas por cada um dos componentes principais;
- O número de componentes retidos e a proporção de variância total por eles explicada;
- Uma tabela com a contribuição de cada variável para cada componente (*loadings*), antes e depois de ser aplicado um método de rotação de fatores.
- Fazer a interpretação de cada componente principal retido.

Assim definido, neste trabalho será utilizado primeiramente a Análise de Correspondência múltipla (ACM) para confirmar ou não as diversas teorias de que apenas 10 variáveis (Sudorese Noturna, Tosse, Febre e etc) são responsáveis por influenciar a variável Desfecho. Sendo a ACM uma “ACP para dados categóricos”

GREENACRE (2008), onde os vetores e autovetores estudados são extraídos da matriz de resíduos padronizados ao quadrado dos dados analisados, trabalharemos com todas as variáveis disponíveis para, através de uma Análise Exploratória, confirmar ou não por ACM se as 10 variáveis são responsáveis por toda a variabilidade da variável Desfecho.

Em um segundo momento, definiremos os critérios utilizados para a classificação de cada observação de entrada no sistema de triagem em “Confiável” ou “Não Confiável” através da teoria de Qualidade de Dados que será vista no próximo capítulo. Definiremos a princípio 5 critérios, que serão calculados para cada observação de nossa base de dados (pacientes que passaram pelo sistema de triagem do sistema “NeuralTB/SAPEM”) e teremos uma nova base de dados , apenas com estes 5 critérios para cada observação , que será analisada por ACP afim de se descobrir a carga(fator) ou *loadings* de cada um desses 5 critérios em cada componente. Em seguida aplicaremos estas cargas em cada observação, obtendo um valor que após a comparação com um limiar pré determinado poderá classificar cada observação como “Confiável” ou “Não Confiável”.

4-Qualidade de Dados

No capítulo que segue discutiremos o termo Qualidade de dados, suas definições e contexto.

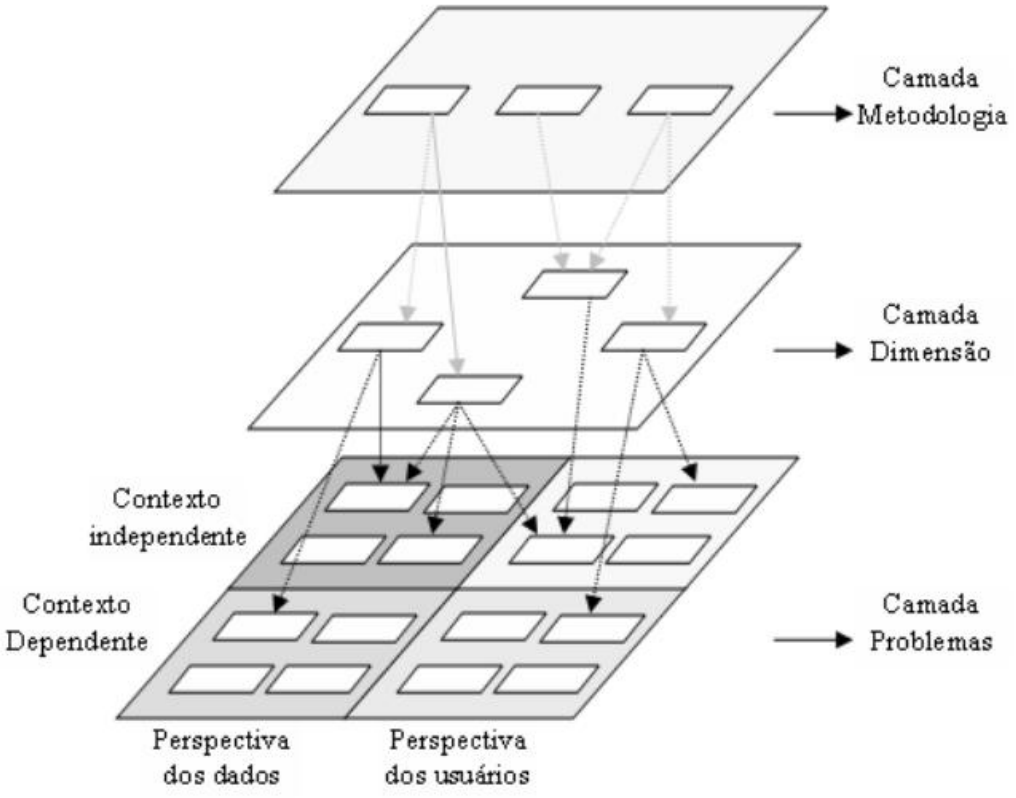
4.1- Introdução

Segundo FAIER (2011), o tema qualidade de dados é um assunto abordado em diferentes áreas do conhecimento, e vem tomando enorme notoriedade a partir da década de 90. Qualidade de Dados (QD) e Qualidade da Informação (QI) muitas vezes é tratada de forma igualitária, mas diferem por se tratar de objetos diferentes no quesito qualidade, visto que a palavra “dado” pode ser definida como um “conjunto de medidas, caracteres ou símbolos” e a palavra informação como “dados processados que transmitem uma mensagem”; por sua vez, neste trabalho não faremos distinção de ambas, optando por usar o termo Qualidade de Dados (QD) . Temos ainda, poucas referências, em português, sobre o tema.

A definição de QD é vista sob duas perspectivas: dos usuários e dos dados. FAIER (2011) argumenta que sob a perspectiva dos dados, o termo qualidade tem sido definido como uma conformidade com as especificações ou uma “conformidade ao uso” e que sob a perspectiva do usuário, QD tem sido definida como uma “conformidade ao uso pelos usuários da informação”. Alguns autores argumentam que são os usuários que definem se a informação ajusta-se ao uso ou não e, portanto, são eles que definem o que é qualidade. FAIER (2011) diz que podemos analisar e classificar a QD sob três aspectos: avaliação, gestão e contexto, que serão vistos em seqüência.

4.2- Avaliação da QD

A avaliação da QD é definida por Faier (2011) como sendo um processo que determina valores numéricos para as dimensões da QD e pode ser encontrada sob três diferentes aspectos: problemas de QD, dimensões e metodologia, com o objetivo de facilitar o entendimento deste conceito.



Fonte: FAIER (2011)

Figura 5– Avaliação da Qualidade de Dados.

- **Problemas de QD:** Os problemas em QD são discutidos em duas óticas, como vemos na Figura 5: a dos dados e a do usuário analisando a dependência ou independência do contexto. Segundo FAIER (2011), em uma análise independente do contexto, sob a ótica dos dados, a falta de dados poderia ser relacionada à completude da base de dados; já sob olhares dos usuários, o problema poderia ser visto como uma impossibilidade dos usuários acessarem os dados, sendo analisada do aspecto da dimensão acessibilidade. Analisando a perspectiva contexto-dependência, e do ponto de vista dos dados, o problema poderia significar dados que violem as regras do negócio, refletindo a dimensão de acurácia. E por último, sob olhares do usuário, o problema poderia estar relacionado a dados irrelevantes para o trabalho em questão, sendo esta a dimensão relevância.

Problema: falta de dados		
	Perspectiva dos dados	Perspectiva dos usuários
Independência do contexto	*base de dados incompleta (Completude)*	*informação inacessível para os usuários (Acessibilidade)*
Dependência do contexto	*dados violam as regras do negócio (Acurácia)*	*Informação irrelevante (Relevância)*

Fonte: FAIER (2011)

Tabela 2– Problema de QD sob várias perspectivas.

- **Dimensões de QD:** Podemos identificar as dimensões através de três abordagens, como vemos na Tabela acima: intuitiva, teórica e empírica. FAIER (2011) argumenta que a abordagem intuitiva define dimensões a partir da perspectiva dos dados. Por exemplo, em FAIER (2011), a completude é definida de forma objetiva como a medida da quantidade de dados faltantes. A abordagem teórica define as dimensões a partir da perspectiva de uma situação ideal. Em FAIER (2011), completude é definida genericamente como a habilidade do sistema de informação representar um estado real. Já a abordagem empírica define as dimensões a partir da perspectiva do usuário. Por exemplo, em FAIER (2011), a completude é definida subjetivamente como sendo i nível para quais os dados são suficientes para a tarefa dos usuários. Podemos classificar as dimensões de vários modos, seguindo diversos autores. Alguns classificam como intrínseca, contextual, representacional e de acessibilidade e outros como sintática semântica e representacional por exemplo. O fato é que uma dimensão pode ser classificada de diferentes maneiras, dependendo da perspectiva, e que alguns casos são classificados de forma ambígua.

- **Metodologia de QD:** A metodologia de QD pode ser classificada como objetiva e subjetiva. Sendo a objetiva relativa a problemas no conjunto de dados e a subjetiva relativa à necessidade e à experiência do usuário. Em Faier (2011) é apresentado um quadro para indicar as diferenças entre essas duas metodologias. Segue o quadro abaixo:

	Objetivo	Subjetivo
Ferramenta	*Software*	*Questionário*
Alvo da medida	*Dado*	*O que a informação representa*
Padrão da medida	*Padrões, regras*	*Satisfação do usuário*
Processo	*Automação*	*Usuário envolvido*
Resultado	*Único*	*Múltiplo*
Local armazenada	*Banco de dados*	*Contexto do negócio*

Fonte: FAIER (2011)

Tabela 3– Diferenças entre a avaliação objetiva e subjetiva da QD.

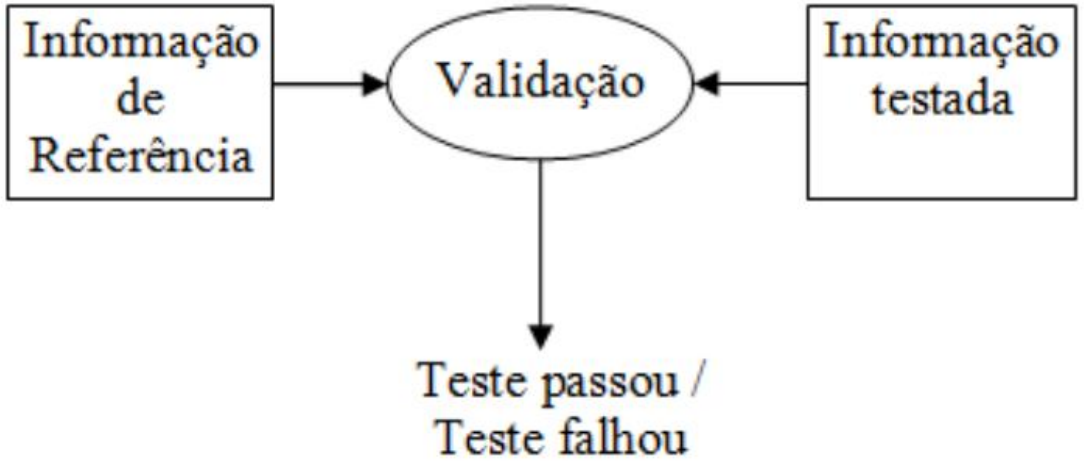
4.3- Gestão da QD

Segundo FAIER (2011), a gestão da qualidade de dados pode ser descrita em três grandes grupos: gestão da qualidade, gestão da informação e gestão do conhecimento. Analisando a gestão da qualidade, conceitos como Gestão Total da Qualidade de Dados (TDQM) buscam direcionar o gerenciamento da qualidade analogamente ao gerenciamento da qualidade de um produto manufaturado. Sob a perspectiva da gestão da informação, FAIER (2011) argumenta que princípios como a integração de dados e contextualização são fundamentais para transformar dados em informações relevantes. Já sob a ótica da gestão de conhecimento, princípios como “Know –what”, “Know-how” e “Know-why” são abordados com o objetivo de melhora da qualidade da informação e tornar o conhecimento explícito para criar conhecimento organizacional.

4.4- Contexto da QD

Segundo FAIER (2011), devido a grande especificidade que o tema Qualidade de Dados pode ter em determinados contextos, a aplicação das técnicas acaba seguindo a sua lógica própria na definição da qualidade de dados. Assim, falar de todos os contextos nos quais a QD está inserida seria quase impossível. Neste trabalho, é apresentado um método específico para medir QD através da realização da definição de alguns critérios e suas respectivas contribuições para a variável “Desfecho”. O dado

testado é comparado a uma informação de referência (FAIER, 2011). O processo genérico é apresentado na figura abaixo.



Fonte: FAIER (2011)

Figura 6– Testes de QD.

Assim, FAIER (2011) discute que são reportadas dois tipos de informação de referência:

- Meta-Informação: independente dos dados, mas com relações rígidas sobre eles. Por exemplo, o formato dos campos de uma tabela.
- Modelos estatísticos: obtidos a partir dos dados livres de erros e tendo a forma de relações aproximadas com os dados como, por exemplo, média e desvio-padrão.

Para o caso de modelos estatísticos, técnicas de modelagem com variáveis simples e múltiplas são utilizadas para o monitoramento da QD e sob variáveis únicas, por meio do conhecimento a priori da Função Densidade de Probabilidade, classificamos as amostras como corretas (prováveis) ou incorretas (improváveis). O monitoramento pode ocorrer por meio de predições com redes neurais e também a utilização de técnicas de reconhecimento de padrões e "clusterização". (FAIER, 2011).

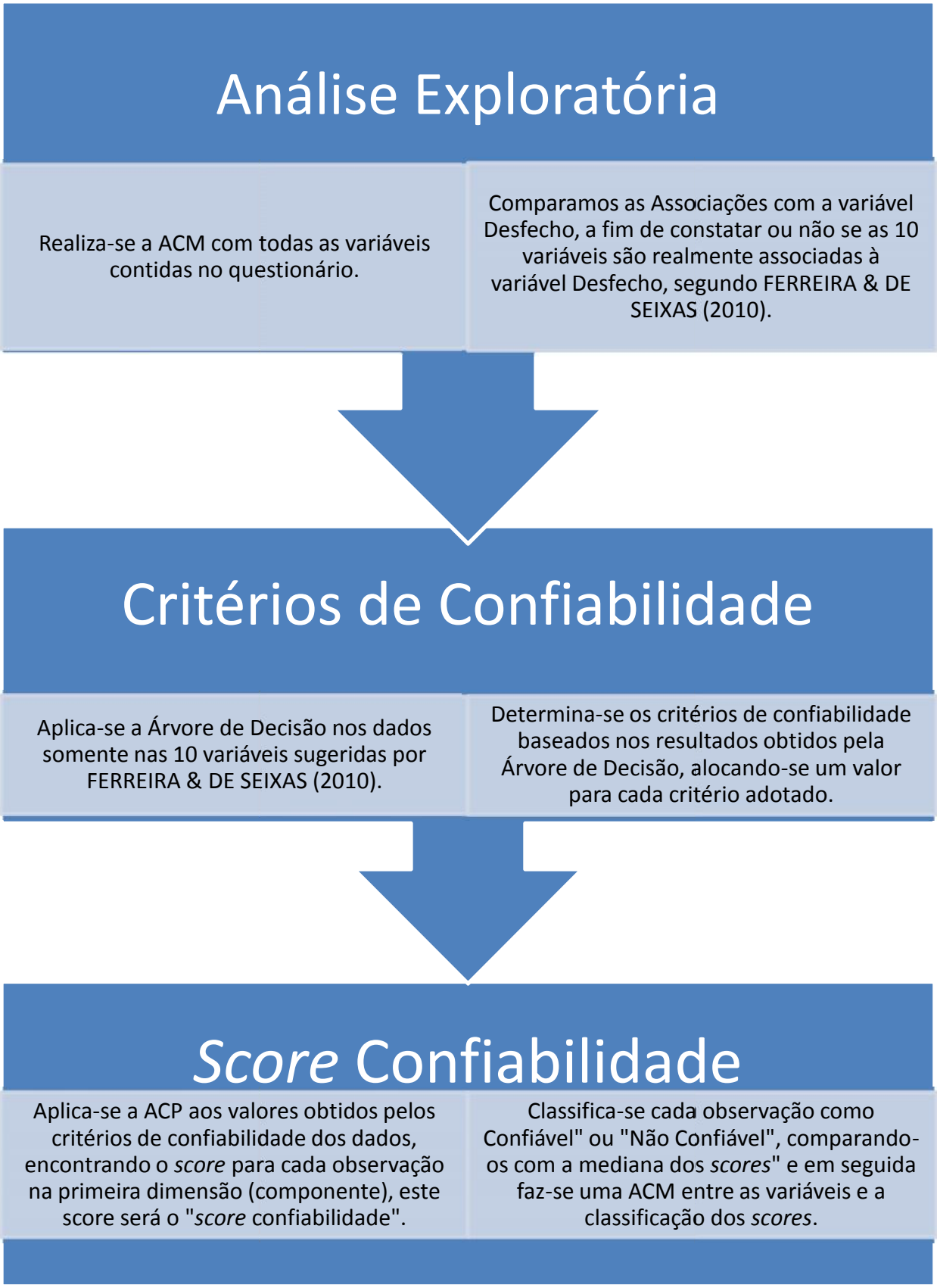
No presente trabalho, utilizamos a teoria de Qualidade de Dados para definir critérios que possam colaborar para a criação de um valor numérico "*score*" e assim classificar as observações (dados de entrada no sistema de triagem para a Tuberculose) como "Confiável" ou "Não Confiável" e assim corroborar o Resultado da Triagem do sistema, o tornando mais confiável.

5-Aplicação

5.1- Introdução

Partindo-se da premissa que apenas pacientes classificados pelo sistema de triagem NeuralTB/SAPEM com diagnóstico positivo para Tuberculose são encaminhados para a consulta médica, um diagnóstico negativo errôneo (não confiável) pelo sistema (sendo o erro contido no preenchimento dos dados e não no sistema de triagem) para algum paciente implica na exposição deste á sociedade, podendo contaminar novas pessoas, disseminando a doença.

A falta de confiabilidade dos dados de entrada foi detectada pelos autores do classificador através de entrevistas e investigações, identificando que os usuários do programa NeuralTB/SAPEM, muitas vezes pressionados pelas condições de trabalho (falta de médicos, enfermeiras, materiais, etc), estariam alterando os dados de entrada dos pacientes para que obtivessem como possível diagnóstico pela triagem uma alternativa que não a tuberculose (pois para um resultado de diagnóstico positivo, muitas vezes a unidade de saúde não ofereceria condições de tratá-lo ou de encaminhamento adequado. Assim, pacientes diagnosticados com ausência de Tuberculose, mas classificados como “Não Confiável” quanto à entrada dos dados, também seriam submetidos ao atendimento médico, podendo evitar que pacientes com Tuberculose e erroneamente classificados com diagnóstico negativo para doença não sejam tratados. A seguir apresentamos um esquema (fluxograma) que relaciona todas as metodologias aqui utilizadas, com os procedimentos propostos para se chegar ao “score confiabilidade”.



Fonte: O autor.

Figura 7– Fluxograma da Metodologia.

5.2- Análise Exploratória

Nossa base de dados apresenta um número significativo de variáveis (22), assim foi feita uma ACM com estas variáveis, tratando a variável Desfecho como suplementar, a fim se confirmar ou não se estas variáveis sugeridas por FERREIRA & DE SEIXAS (2010) são as mais associadas à variável Desfecho. Seguem as variáveis analisadas:

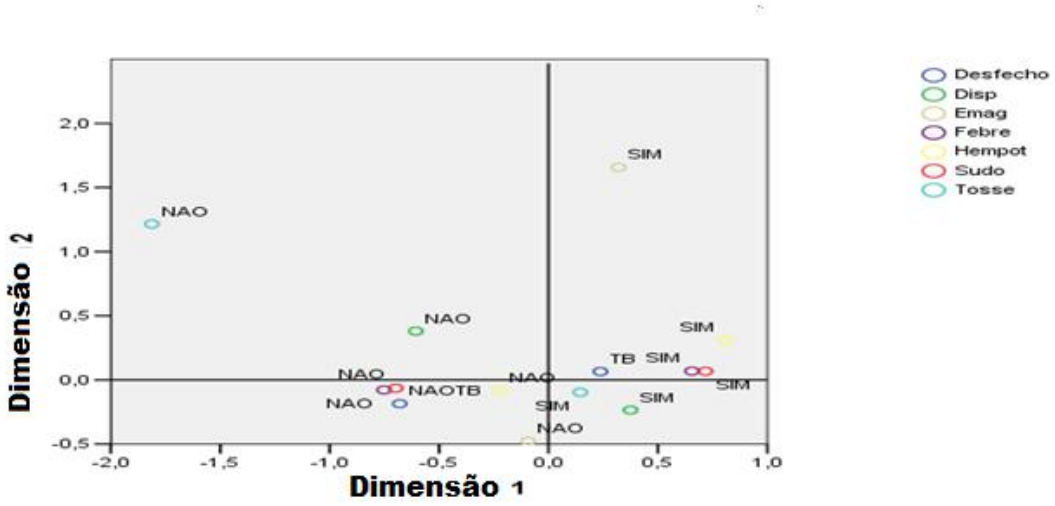
- Região: Unidade de Saúde (A, B,..., K);
- Cor (Auto-declaração);
- Estado Civil (Solteiro, Casado, Viúvo ou divorciado);
- Grau de instrução (Até qual série cursou dos níveis primário, fundamental e médio ou se fez/faz curso superior)
- Expectoração (Apresenta ou não o sintoma);
- Classe Social (A, B, C, D ou E);
- Perda de apetite (Apresenta ou não o sintoma);
- Chiado Peitoral (Apresenta ou não o sintoma);
- Espirros freqüentes (Apresenta ou não o sintoma);
- Tratamento anterior para Tuberculose (Apresenta ou não o sintoma);
- Contato com pessoas Tuberculose (sim ou não);
- Consumo de Álcool (sim ou não).

Somam- se a essas, as 10 variáveis propostas por FERREIRA & DE SEIXAS (2010) são:

- Tosse (Apresenta ou não o sintoma);
- Fumante (Fuma, é Ex-fumante ou Não fuma);
- Emagrecimento (Apresenta ou não o sintoma);
- Idade (Originalmente estava em anos, porém foi categorizado em classes de 20 anos);
- Sudorese Noturna (Apresenta ou não o sintoma);
- Febre (Apresenta ou não o sintoma);
- Sexo (Masculino ou Feminino);
- Dispnéia (Apresenta ou não o sintoma);
- Internação Hospitalar (Foi ou Não internado anteriormente);
- Hemoptóico (Apresenta ou não o sintoma).

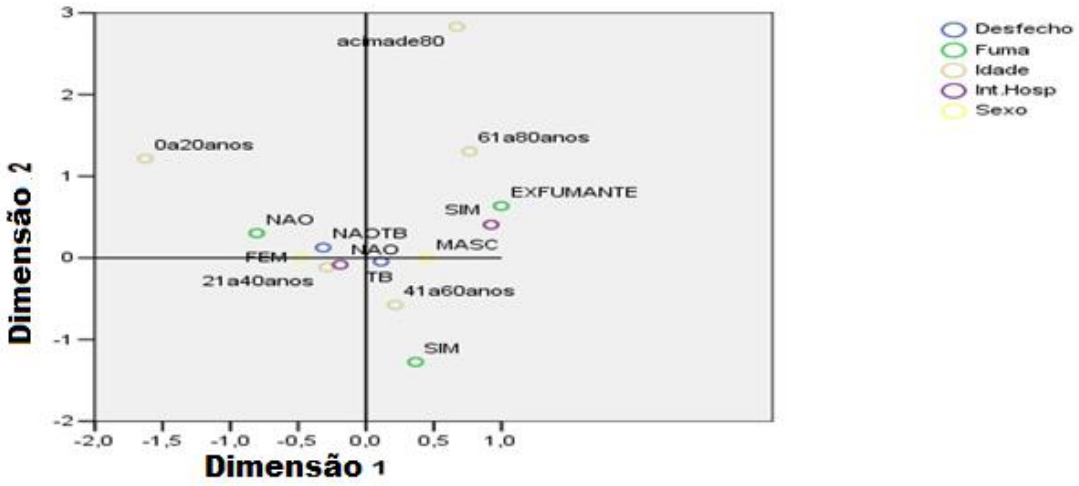
Segue a ACM para as 10 variáveis sugeridas por FERREIRA & DE SEIXAS (2010). Os dados foram divididos em dois gráficos de ACM, um para os sintomas e outros para os perfis de pacientes, ambos com a variável Desfecho como ponto suplementar. Através dos gráficos, podemos ver claramente como é o perfil das pessoas que

apresentam Diagnóstico positivo para Tuberculose, bem como os sintomas mais impactantes.



Fonte: O autor.

Figura 8– ACM Sintomas e Diagnóstico como alvo.



Fonte: O autor.

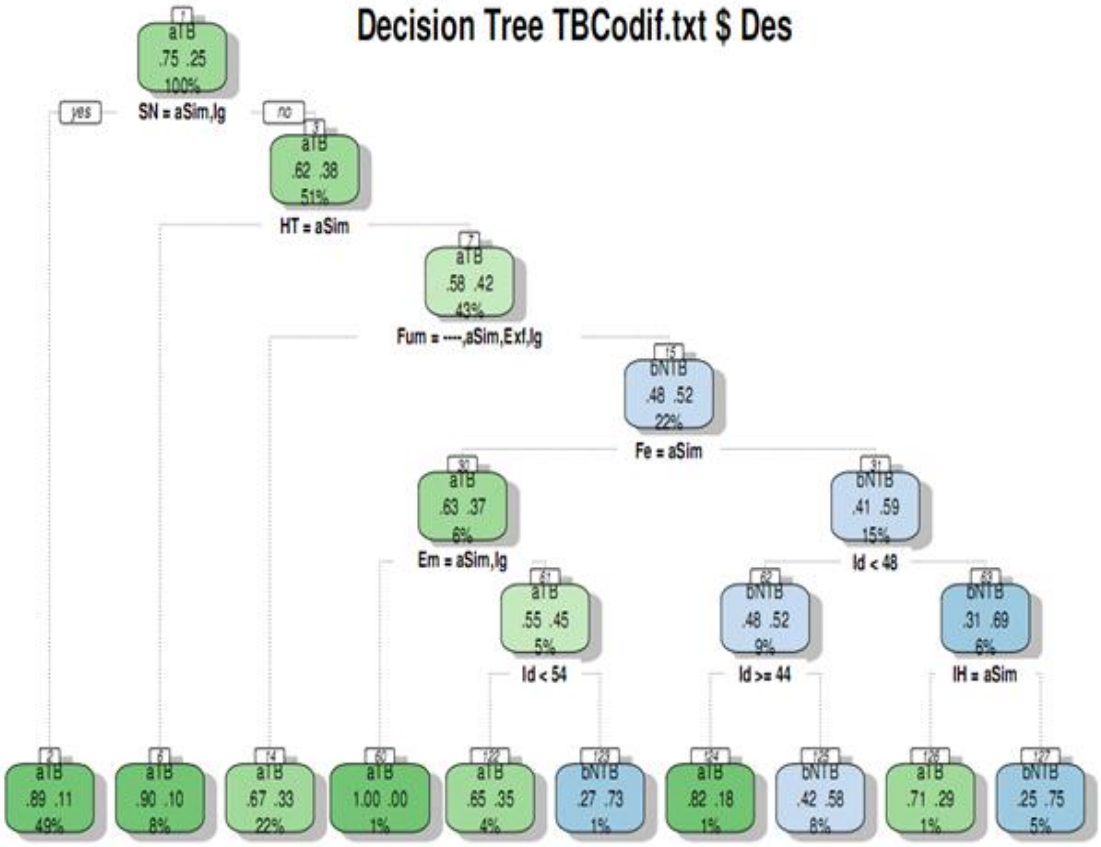
Figura 9– ACM Pefis de pacientes e variável Diagnóstico como alvo.

Analisando os gráficos acima e olhando as contribuições no Apêndice, podemos ver claramente que as variáveis sugeridas por FERREIRA & DE SEIXAS (2010) são influentes para a variável Desfecho.

Apesar de encontrarmos mais variáveis que poderiam estar associadas ao desfecho pela ACM, vamos nos limitar às 10 variáveis descritas por FERREIRA & DE SEIXAS (2010) para a elaboração da árvore de decisão, por se tratar apenas de ilustração da metodologia

5.3- Critérios de Confiabilidade

Aplicamos a técnica de árvores de decisão em nossa base de dados, com o objetivo de identificar, dentre as 10 variáveis sugeridas por FERREIRA & DE SEIXAS (2010), quais são mais determinantes para alterar o resultado da variável Diagnóstico, com base nos critérios de classificação da árvore de decisão utilizando o algoritmo CART e assim determinar os critérios de qualidade que originarão os *scores* de confiabilidade. Assim, após aplicação do método, chegamos à seguinte árvore:



Fonte: O autor.

Figura 10– Árvore de Decisão com a variável Diagnóstico como alvo.

Temos então que apenas as variáveis contidas na árvore de decisão podem alterar o resultado da variável Desfecho (podemos até ter outras, mas podem estar correlacionadas com alguma já selecionada pela árvore). As variáveis selecionadas pela árvore foram renomeadas para uma melhor visualização da árvore. Seguem as variáveis e suas respectivas codificações:

- Idade= “id”;
- Fumante = “Fum”, obs.: categoria 2= “Ex Fumante”;
- Emagrecimento = “Em”;
- Sudorese Noturna = “SN”;
- Febre = “Fe”;
- Internação Hospitalar =“IH”;
- Hemoptoico = “HT”.

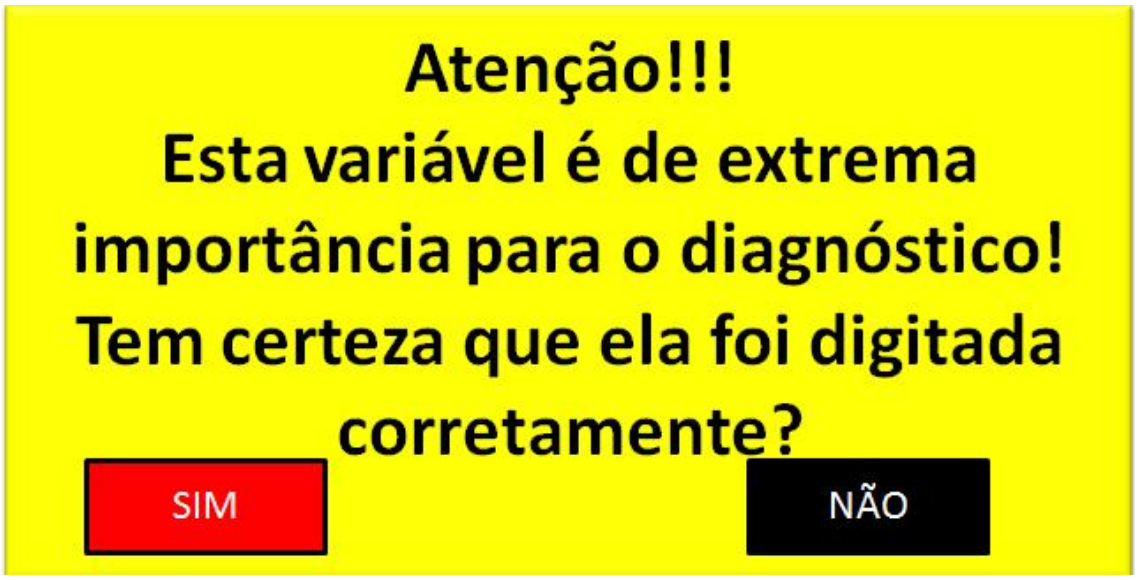
A seguir, as informações contidas na árvore poderão ser interpretadas durante a definição dos critérios de qualidade a serem utilizados para o cálculo do *score* confiabilidade, que seguem:

- **Critério 1:** Certeza do Digitador;
- **Critério 2:** Probabilidade de ocorrência daquela combinação de variáveis(caminho da árvore de decisão);
- **Critério 3:** Fator Região - Prevalência da doença na região;
- **Critério 4:** Certeza de Classificação da árvore;
- **Critério 5:** Completude- Número de Variáveis Contidas na árvore não respondidas.

Todos os critérios acima estarão na mesma escala (de zero a um) e a definição de cada um segue abaixo:

-Critério 1: Certeza do Digitador

Vimos acima, que das variáveis propostas por FERREIRA & DE SEIXAS (2010), somente sete realmente influenciam no resultado da classificação da variável Desfecho. O critério 1- certeza do Digitador é compreendido do seguinte modo: Durante o preenchimento do questionário online no sistema NeuralTb/SAPEM , cada vez que o digitador preencher umas das sete variáveis selecionadas pela árvore, aparecerá em sua tela uma mensagem semelhante à da figura abaixo:



Fonte: O autor.

Figura 11– Janela de alerta do critério 1.

Sendo assim, para cada janelinha, como a mostrada acima, que aparecer e o usuário trocar a resposta (sim para não e vice versa), a resposta será penalizada da seguinte forma, por exemplo, se no caminho percorrido aparecerem 3 janelas(Considerando que as 4 outras variáveis não foram preenchidas, e em 2 janelas o usuário mudou a resposta). Nesse caso, consideramos o total de janelas que apareceram (3) iguais a 100% e subtraímos a proporção do número de variáveis não respondidas (2), ou seja, 67%, logo o valor para este critério nessa situação específica será de 33%. Para nossa base de dados, foram calculados os valores deste critério em todos os pacientes.

-Critério 2: Probabilidade de ocorrência daquela combinação de variáveis(caminho da árvore de decisão)

O critério 2 pode ser encontrado na própria árvore. Ele exibe a probabilidade de ocorrência daquela combinação de variáveis, situa-se nos últimos ramos (quadros para o exemplo abaixo) na árvore, localizado na parte central inferior do ramo. No exemplo na página seguinte, as setas indicam a opção selecionada pelo digitador para cada variável preenchida. Temos que foi digitado para as variáveis:

-Sudorese Noturna=Não (Ausente), a árvore caminhou á direita;

-Hemoptóico=Não (Ausente), a árvore caminhou à direita;

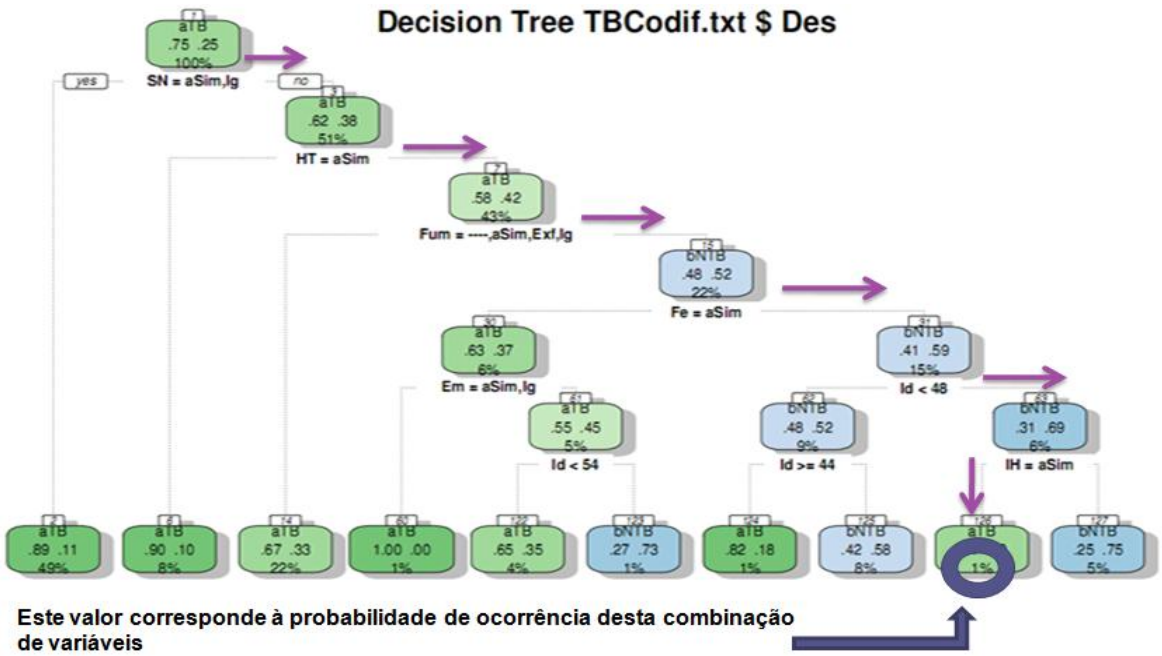
-Fumante=Não (Ausente), a árvore caminhou à direita;

-Febre=Não (Ausente), a árvore caminhou à direita;

Idade<48=Não, a árvore caminhou à direita;

-Internação Hospitalar=Sim, a árvore já foi capaz de classificar O Desfecho como paciente com Tuberculose;

Na figura abaixo, vemos todo o caminho percorrido com esta combinação de variáveis e obtemos como probabilidade de ocorrência desta combinação igual a 0.01 ou 1%, este será o valor do critério 2 para este exemplo.



Fonte: O autor.

Figura 12– Esquema explicativo do critério 2.

Critério 3: Fator Região

O fator região calcula a prevalência da doença na Unidade de Saúde na qual foi feita a triagem no paciente. A título de exemplificação, suponhamos que nossa base de dados seja a retratada na tabela a seguir:

Unidade de saúde	Desfecho
Policlínica Augusto Amaral Peixoto Guadalupe	provavelTB
Policlínica Augusto Amaral Peixoto Guadalupe	naoTB
Policlínica Augusto Amaral Peixoto Guadalupe	provavelTB
Policlínica Augusto Amaral Peixoto Guadalupe	naoTB
Policlínica Augusto Amaral Peixoto Guadalupe	provavelTB
Policlínica Augusto Amaral Peixoto Guadalupe	provavelTB
Policlínica Augusto Amaral Peixoto Guadalupe	provavelTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	provavelTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	provavelTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	provavelTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	provavelTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	provavelTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	provavelTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	provavelTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	provavelTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	provavelTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	naoTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	provavelTB
Hospital Universitário Clementino Fraga Filho (HUCFF/UFRJ)	naoTB
Instituto Brasileiro para Investigação da Tuberculose	provavelTB
Instituto Brasileiro para Investigação da Tuberculose	provavelTB
Instituto Brasileiro para Investigação da Tuberculose	naoTB

Fonte: O autor.

Tabela 4– Base de dados critério 3.

De posse dos dados acima, seleciona-se um paciente para o cálculo do critério 3. Em nosso exemplo, o paciente solicitado será da Policlínica Augusto Amaral Peixoto Guadalupe e seu diagnóstico será o de presença de Tuberculose. Calculamos a proporção de diagnósticos com tuberculose ou não para as diferentes localidades, e atribuímos ao paciente selecionado a prevalência da doença em sua localidade, como vemos na tabela abaixo:

Região	Prob TB	Prob n TB
Poli. A.A.P. Guada.	71%	29%
H.U.	85%	15%
I.B.I.T.	67%	33%

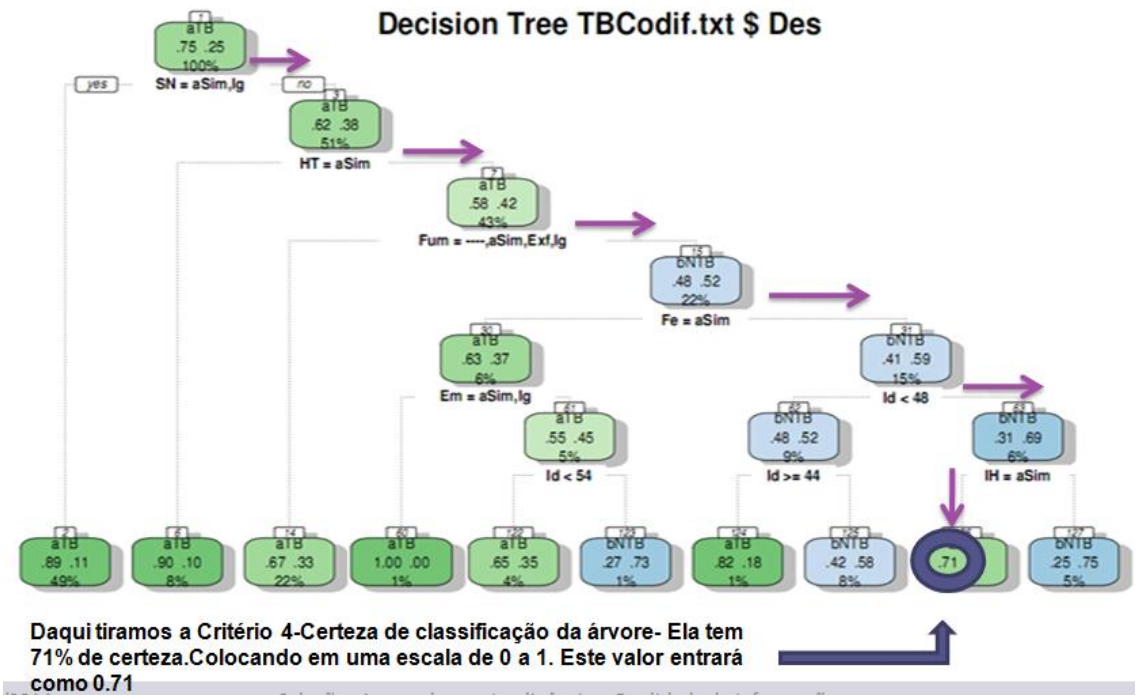
Fonte: O autor.

Tabela 5– Prevalência critério 3.

Assim, pela tabela 5, o valor do critério 3 para o paciente do exemplo será de 71% ou 0.71.

Critério 4: Certeza de classificação da árvore.

Além de exibir a probabilidade de ocorrência de um determinado caminho, a árvore de decisão ainda apresenta a certeza desta classificação. Este valor também é encontrado nas últimas folhas (nós) da árvore, na parte superior esquerda. Para os mesmos dados de entrada do paciente do exemplo do critério 2, podemos localizar a certeza da árvore para o Desfecho Tb Positiva. Vejamos:



Fonte: O autor.

Figura 13– Esquema explicativo do critério 4.

Logo, para o paciente do exemplo acima, o valor do critério 4 será de 71% ou 0.71.

Critério 5: Completude- Número de Variáveis Contidas na árvore não respondidas.

A completude contabiliza o número de variáveis totais, considerando inicialmente o valor do critério igual a 100% e o penaliza proporcionalmente com o número de variáveis não respondidas (ignoradas). Vejamos um exemplo de preenchimento de um paciente no sistema de triagem abaixo:

Variável	Digitado
Sud. Not.	Não
Dispnéia	Ignor.
Hemop.	Não
Fuman.	Não
Tosse	Ignor.
Febre.	Não
Idade	54
Emagr.	Sim
Sexo	M
Intern.	Sim

Fonte: O autor.

Tabela 6– Exemplo critério 5.

Assim, tendo 10 variáveis disponíveis estas correspondem a 100% ou 1 do valor do critério, como temos 2 não respondidas, ou seja 20% dos dados, o valor do critério 5 será 100% -20% =80% =0,8 para o exemplo acima.

Após a definição de todos os critérios, foi feito o respectivo cálculo para todas as observações (pacientes) disponíveis. Na tabela a seguir, podemos ver alguns dos pacientes com os valores de seus respectivos critérios calculados.

Paciente	Critério1	Critério2	Critério3	Critério4	Critério 5
1	0.33	0.49	0.94	0.89	1.0
2	0.33	0.49	1.0	0.89	1.0
3	0.33	0.49	0.71	0.89	1.0
4	0.66	0.49	0.71	0.89	1.0
5	0.66	0.49	0.71	0.89	1.0
6	0.33	0.49	0.71	0.89	1.0
7	0.33	0.49	0.71	0.89	1.0
8	1.0	0.49	0.71	0.89	1.0
9	0.33	0.49	0.71	0.89	1.0
10	0.66	0.49	0.71	0.89	1.0
12	0.66	0.49	0.71	0.89	1.0
13	0.33	0.49	0.71	0.89	1.0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1054	1.0	0.49	0.71	0.89	1.0

Fonte: O autor.

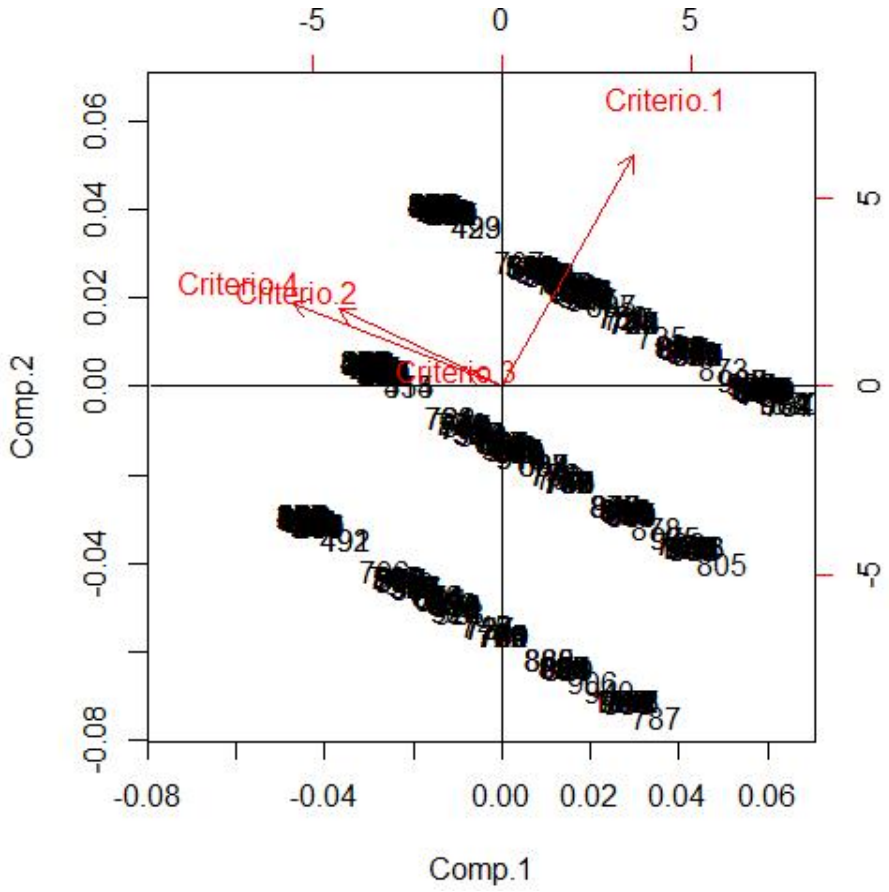
Tabela 7– Critérios Calculados.

Neste ponto, temos todos os critérios calculados, mas não sabemos seus coeficientes (cargas) para o cálculo do *score* confiabilidade. No próximo tópico será feito uma ACP com os critérios, encontrando seus respectivos coeficientes (cargas), sendo possível o cálculo do *score.*

5.4- *Score* Confiabilidade

Para o cálculo do *score* confiabilidade foi usada a ACP para os dados (critérios de 1 a 5). Em uma análise preliminar, decidimos retirar o critério 5 (completude), pois mais de 90% das amostras (pacientes) possuíam o mesmo valor para este critério(igual a 1), logo ele não seria capaz de discriminar os dados. Assim, prosseguimos a ACP com apenas os 4 critérios já definidos. Seguem os resultados obtidos:

-Gráfico para as primeiras 2 componentes principais:

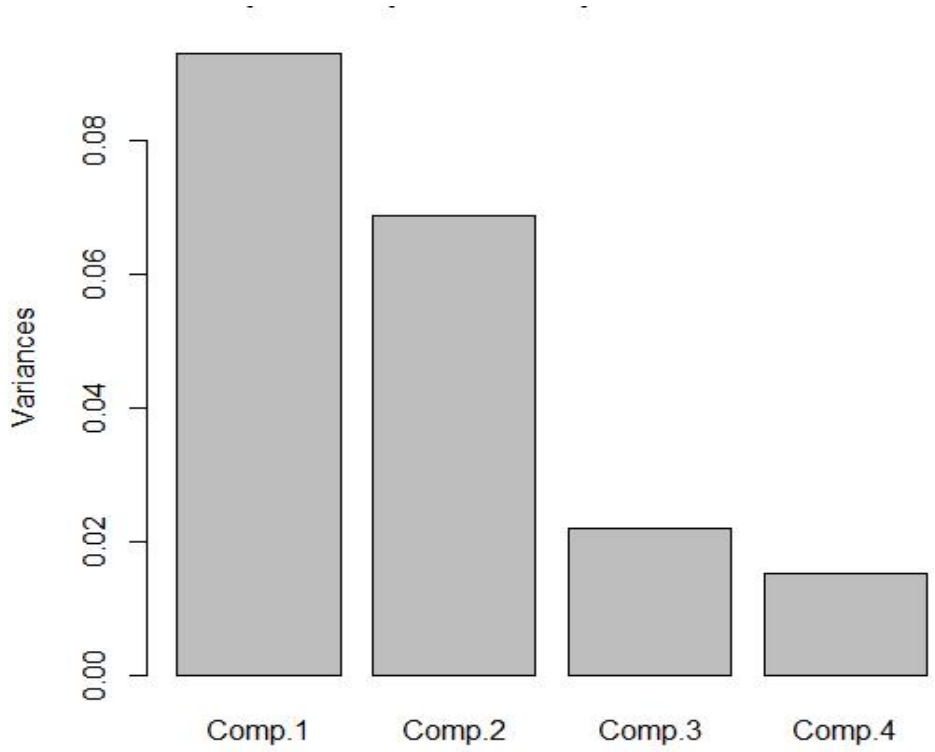


Fonte: O autor.

Figura 14– Gráfico de ACP para os dois primeiros componentes principais.

Pela figura 14, constatamos que os critérios 2, 3 e 4 estão bem correlacionados, sendo o critério 1 o único que destoa dos demais. Os dados estão segregados em 3 filas devido ao critério 1, em que ocorreram apenas 3 categorias para esta variável.

-Gráfico de importância para os componentes principais:



Fonte: O autor.

Figura 15– Importância dos componentes principais.

-Porcentagem da variância explicada por cada componente:

Componente 1	Componente 2	Componente 3	Componente 4
0.4670328	0.3459939	0.1099649	0.07700835

Fonte: O autor.

Tabela 8– Porcentagem da variância explicada em cada componente.

Como vimos no gráfico e na tabela acima, o primeiro componente principal concentra 46.7% da variabilidade dos dados. É neste componente que concentraremos nossos estudos. A seguir calculamos o coeficiente (carga) de cada critério no componente 1:

Critério	Componente1	Correlação
Critério.1	0,44	0,49
Critério.2	-0,55	-0,804781526
Critério.3	-0,12	-0,241027194
Critério.4	-0,7	-0,877336861

Fonte: O autor.

Tabela 9– Cargas e correlação de critério no primeiro componente principal.

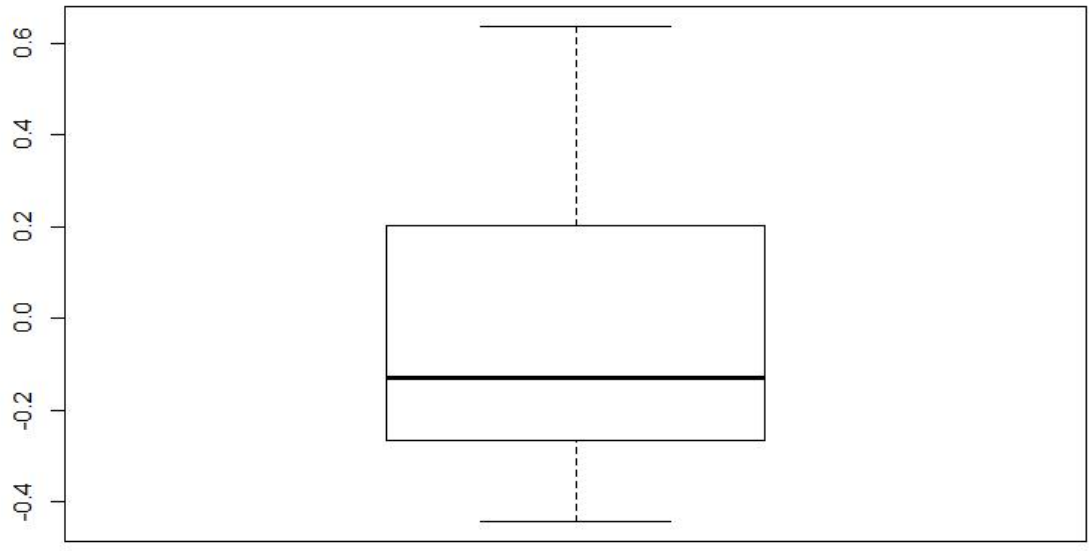
Após o cálculo das cargas dos critérios no primeiro componente, calculamos o *score* de confiabilidade que será o *score* da ACP para os dados (critérios de 1 a 4). Vejamos na tabela abaixo alguns desses *scores* (confiabilidade) calculados:

***Scores* Confiabilidade**
-0.4343997853
-0.4415114473
-0.4071384140
-0.2741470646
-0.2741470646
-0.4071384140
-0.4071384140
-0.1371256742
-0.4071384140
.
.
.
-0.2741470646

Fonte: O autor.

Tabela 10– *Scores* Confiabilidade.

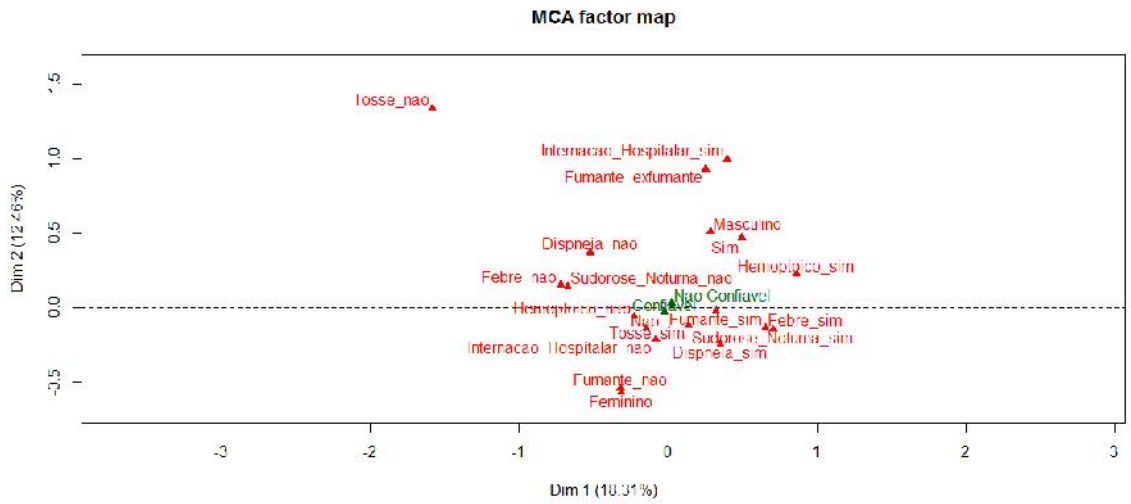
Fazendo uma análise exploratória dos *scores*, através do boxplot a seguir, vemos que em sua maioria seus valores são negativos e que os dados possuem certa assimetria à direita.



Fonte: O autor.

Figura 16– Boxplot para o *score* confiabilidade.

Através dos *scores* de confiabilidade, podemos classificar cada amostra (pacientes) quanto à confiabilidade da informação. Calculamos a mediana destes *scores*, amostras com *scores* abaixo da mediana serão classificadas como “Confiável” e acima como “Não Confiável” (Para este exemplo específico, onde as variáveis que mais se correlacionam com o componente 1 apresentam coeficientes negativos). A título de análise exploratória, o próximo passo de nossa análise será verificar quais variáveis estão associadas a dados com classificação “Confiável” e quais estão relacionadas com “Não Confiável”. Para isso, foi realizada uma ACM entre as variáveis de entrada e a variável que classifica os *scores*. Vejamos:



Fonte: O autor.

Figura 17– Gráfico ACM para variáveis e *score* confiabilidade.

Pelo gráfico acima e pelas tabelas A3 e A4 localizadas no Apêndice, notamos que apesar das classificações do *score* confiabilidade estarem bem próximos do centróide, elas discriminam as demais classificações, estando em quadrantes opostos. Assim, concluímos que, em geral a ausência dos sintomas associados a um paciente do sexo feminino, classifica a informação como “Confiável”. Devido a poucas informações sobre a base de dados (não se sabe como as amostras foram selecionadas, por exemplo), não é possível interpretar os resultados. Porém, os dados aqui utilizados foram extremamente úteis para a elaboração desta metodologia.

6- Considerações Finais

O objetivo inicial deste trabalho foi desenvolver uma metodologia que fosse capaz de identificar e classificar as informações inseridas no sistema NeuralTb/SAPEM quanto à sua confiabilidade.

Além da introdução, no capítulo 2 foi feita uma apresentação da teoria de Árvores de Decisão e no capítulo 3 foram apresentados os conceitos envolvendo a temática de Qualidade de Dados. No capítulo 4 foi feita uma breve explanação da teoria sobre Análise de Componentes Principais.

No capítulo 5 foi descrita a metodologia através de um exemplo, verificando que ela pode vir a ser útil, porém, requer ainda muito trabalho e pesquisa para que seja comprovada sua real eficiência. Neste trabalho, não tivemos acesso aos reais diagnósticos, fornecidos pelos médicos. A ausência desta informação não nos permite comprovar se o modelo proposto neste trabalho teve grande eficiência. Assim, o próximo passo deste trabalho será fazer tal comparação, para que possíveis ajustes sejam feitos no método.

O desenvolvimento da metodologia proposta neste trabalho, corroborada com os reais diagnósticos médicos, pode ser extremante útil na triagem de pacientes com suspeita de Tuberculose. O que se propõe aqui é que, após as classificações dos pacientes quanto à confiabilidade de sua informação, se reduza o número de pacientes com a presença da Doença sem o tratamento médico adequado.

Referências

BASGALUPP, M.P. 2010. LEGAL-Tree: Um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão. Tese de Doutorado, ICMC-USP, São Carlos.

BRAMER, M. 2007. Principles of data mining. Springer, London.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., & STONE, C. J. 1984. Classification and Regression Trees. Wadsworth.

FAIER, J. M. 2011. Análise de componentes independentes para a monitoração da qualidade de dados em séries temporais. Tese (Doutorado em Engenharia Elétrica) – Universidade Federal do Rio de Janeiro. Rio de Janeiro.

FAPERJ. 2012. Consultado no site http://www.faperj.br/boletim_interna.phtml?obj_id=7807 .

FERREIRA, D. D. & DE SEIXAS, J. M. 2010. Qualidade de Dados via Árvores de Decisão em Apoio ao Diagnostico da Tuberculose Pulmonar, XVIII Congresso Brasileiro de Automática, Bonito-MS.

GREENACRE, M. 2008. **La Práctica del Análisis de Correspondencias**. Fundación BBVA, Madrid (Spanish translation of Correspondence Analysis in Practice, Second Edition), Fundación BBVA.

JOHONSON, R.A.; WICHERN, D.W. 1992. **Applied multivariate statistical analysis**. 3. ed. New Jersy: Prentice-Hall.

LOPES, L. F. D. 2001. Análise de componentes principais à confiabilidade de sistemas complexos. Tese (Doutorado em Engenharia de Produção) –Universidade Federal Santa Catarina.

QUINLAN, J. R. 1986. Induction of decision trees. Machine Learning, 1(1):81-106.

QUINLAN, J. 1988. Decision trees and multivalued attributes. Machine Intelligence, 11:305-318.

QUINLAN, J. R. 1993. C4.5: **Programs for machine learning**. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

REGAZZI, A. J. 2001. Análise multivariada. Apostila de disciplina. Universidade Federal de Viçosa, Centro de Ciências Exatas e Tecnológicas. Departamento de Informática, 166p. Viçosa-MG.

REIS, E. 1997. Estatística multivariada aplicada. Lisboa.

SOUZA, A. M. 2000. Monitoração e ajuste de realimentação em processos produtivos multivariados. Tese (Doutorado em Engenharia de Produção) – Universidade Federal Santa Catarina.

TAN, P.-N., STEINBACH, M., & KUMAR, V. 2005. **Introduction to Data Mining**, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

ZUBEN & ROMIS. 2013. Notas de Aula –Árvores de Decisão, DCA/FEEC/Unicamp, Campinas.

Apêndice

Classificações de variáveis	Contrib. Dim1	Contrib. Dim2
0a20anos	0.794419533	1,05E+06
21a40anos	0.951007311	3,37E+05
41a60anos	0.014675117	7,03E+04
61a80anos	0.788472322	8,11E+05
Acimade80	0.464016165	7,32E+05
Fem	0.821934949	1,57E+06
Masc	0.751589165	1,44E+06
Emagrecimento_nao	0.170382403	2,59E+06
Emagrecimento_Sim	0.591854662	9,00E+06
Tosse_nao	6.009546752	1,70E+06
Tosse_sim	0.481069181	1,36E+05
Hemoptoico_nao	0.866353650	3,88E+04
Hemoptoico_sim	3.157693357	1,41E+05
Sudorese_nao	6.365328597	1,28E+05
Sudorese_sim	6.547629178	1,32E+05
Febre_nao	5.87118892	5,47E+04
Febre_sim	5.121119879	4,77E+04
Dispneia_nao	3.601517970	8,44E+05
Dispneia_sim	2.229511125	5,22E+05
Inter.Hosp._nao	0.069665359	4,56E+05
Inter.Hosp._sim	0.338717779	2,22E+06
Fumante_exfumante	0.337279446	4,02E+06
Fumante_nao	2.202237737	2,29E+06
Fumante_sim	1.959574495	5,27E+03
ConsomeAlcool_jaconsumiu	1.012731632	2,70E+06
ConsomeAlcool_nunca	0.940601409	9,56E+05
ConsomeAlcool_sim	0.013376144	5,79E+05
U.Saude_A	2.015375909	2,28E+06
U.Saude_C	3.268095657	1,92E+04
U.Saude_D	0.002874782	8,60E+05
U.Saude_E	0.006770823	3,47E+06
U.Saude_F	0.013837885	2,71E+05
U.Saude_G	1.903764082	9,38E+05
U.Saude_H	0.881401364	3,53E+06
U.Saude_I	0.110996992	7,42E+06
U.Saude_J	0.478536186	1,35E+01
U.Saude_K	0.148850710	3,02E+04
U.Saude_L	0.147334844	6,17E+05
U.Saude_M	0.482366314	1,33E+05
amarela	0.255839704	2,09E+05
branco	0.624714592	2,30E+06

Ind¡gena	0.050671719	2,71E+05
parda	0.407233994	3,83E+05
preta	0.022101912	3,32E+05
amigado	0.071158128	6,74E+02
casado	0.054349476	9,42E+03
divorciado	0.300427335	9,44E+04
separado	0.034452262	2,54E+05
solteiro	0.218637643	1,93E+04
viuvo	0.944230369	7,08E+02
colegialCompleto	1.913415797	1,54E+06
ginasioOuFundamentalCompleto	0.061245949	2,03E+06
naoFrequentaEscola	0.041355944	5,92E+06
posGraduacao	0.053538965	2,50E+05
primarioCompleto	1.021562923	5,37E+03
primarioIncompleto	0.162085563	3,55E+06
superiorCompleto	0.466428048	5,63E+05
C.Social_A	0.109095604	2,55E+04
C.Social_B	0.030437324	2,51E+06
C.Social_C	0.136138208	1,99E+06
C.Social_D	0.052048060	2,46E+06
C.Social_E	0.700298080	7,92E+06
Expectoracao_nao	6.367741804	1,06E+06
Expectoracao_sim	1.662793112	2,77E+05
Perda.de.Apetite_nao	7.055484583	1,64E+04
Perda.de.Apetite_sim	5.208650173	1,21E+04
ChiadoPeitoral_nao	2.966542287	7,24E+05
ChiadoPeitoral_sim	3.739736819	9,12E+05
Espirrosfreq_nao	0.491191864	1,86E+06
Espirrosfreq_sim	0.579354506	2,20E+06
TratAntTb_nao	0.342598582	1,35E+06
TratAntTb_sim	1.050744932	4,13E+06
ContatoTB._nao	0.264070583	2,15E+05
ContatoTB._sim	0.605929400	4,93E+05

Fonte: O autor.

Tabela A1– Contribuições da ACM para todas as variáveis disponíveis

Classificações de variáveis	Coord. Dim1	Coord. Dim2
0a20anos	-0.57308707	-0.571615088
21a40anos	0.29077770	-0.149979974
41a60anos	0.03078906	0.058419085
61a80anos	-0.40885713	0.359243386
Acimade80	-1.07284710	1.167691338
Fem	-0.22409233	-0.268351628
Masc	0.20491325	0.245384597
Emagrecimento_nao	-0.08002242	-0.270370304
Emagrecimento_Sim	0.27797262	0.939181057
Tosse_nao	-1.53823207	0.708587519
Tosse_sim	0.12313675	-0.056723016
Hemoptoico_nao	-0.17949649	-0.032900145
Hemoptoico_sim	0.65423038	0.119914738
Sudorese_nao	-0.60526151	0.074355775
Sudorese_sim	0.62259597	-0.076485296
Febre_nao	-0.60643812	0.050706599
Febre_sim	0.52896364	-0.044228663
Dispneia_nao	-0.52429098	0.219925879
Dispneia_sim	0.32456109	-0.136144592
Inter.Hosp._nao	-0.04950911	-0.109812183
Inter.Hosp._sim	0.24071673	0.533914407
Fumante_exfumante	0.18631924	0.557739019
Fumante_nao	-0.37188172	-0.328682588
Fumante_sim	0.47659039	-0.021427482
ConsomeAlcool_jaconsumiu	0.28793678	0.407814996
ConsomeAlcool_nunca	-0.30130583	-0.263203541
ConsomeAlcool_sim	-0.03382500	-0.192843565
U.Saude_A	0.67110090	-0.617921940
U.Saude_C	-0.43021109	0.028557402
U.Saude_D	-0.10901797	1.633756529
U.Saude_E	0.05683166	1.114188712
U.Saude_F	0.15658210	0.601003794
U.Saude_G	1.19624703	0.727605611
U.Saude_H	0.43413934	-0.752835448
U.Saude_I	0.33186151	2.351730102
U.Saude_J	0.66305078	-0.003053394
U.Saude_K	-1.92152807	-0.750133011
U.Saude_L	-0.38234376	-0.677844490
U.Saude_M	2.44593418	1.115074467
amarela	-0.67327305	-0.527265153

branco	0.28483635	-0.473242888
Ind¡gena	0.45769743	0.917753584
parda	-0.16797776	0.141206868
preta	0.04417057	0.148443144
amigado	0.11228442	-0.009469834
casado	-0.07757914	-0.027986089
divorciado	-0.43162939	0.209707347
separado	0.19709191	0.463691628
solteiro	0.11899655	-0.030627396
viuvo	-0.75581992	0.017937648
colegialCompleto	-0.49462389	-0.384552315
ginasioOuFundamentalCompleto	0.08386545	-0.417994246
naoFrequentaEscola	-0.12968070	1.344241767
posGraduacao	0.57620400	-1.078278598
primarioCompleto	0.34172281	0.021466675
primarioIncompleto	0.16946477	0.687775160
superiorCompleto	-0.80172856	-0.763473226
C.Social_A	1.1632476	0.487525159
C.Social_B	-0.10695536	-0.841199094
C.Social_C	-0.08449494	-0.280179391
C.Social_D	0.06914992	0.411871078
C.Social_E	0.66739115	1.945224375
Expectoracao_nao	-0.94734391	0.334939839
Expectoracao_sim	0.24737764	-0.087462035
Perda.de.Apetite_nao	-0.69627548	-0.029055972
Perda.de.Apetite_sim	0.51401932	0.021450319
ChiadoPeitoral_nao	-0.39401012	0.168640794
ChiadoPeitoral_sim	0.49670425	-0.212595043
Espirrosfreq_nao	-0.16274886	0.274618616
Espirrosfreq_sim	0.19196019	-0.323909137
TratAntTb_nao	-0.11514234	-0.197801101
TratAntTb_sim	0.35313990	0.606653138
ContatoTB._nao	-0.10518903	0.082190218
ContatoTB._sim	0.24136397	-0.188591509

Fonte: O autor.

Tabela A2– Coordenada da ACM para todas as variáveis disponíveis

Classificações de variáveis	Contrib. Dim1	Contrib. Dim2
Tosse_nao	1.09220374	1.1416106922
Tosse_sim	0.9458259	0.988611321
Fumante_exfumante	0.9738081	1.9375879349
Fumante_nao	2.6135534	1.1078680453
Fumante_sim	1.3923164	0.008949905
Emagrecimento_Nao	0.9009189	1.180899503
Emagrecimento_Sim	3.0556167	4.005217480
Sudorese_Noturna_nao	1.26918958	0.802949013
Sudorese_Noturna_sim	1.32330231	0.837183273
Febre_nao	1.35247146	0.882863244
Febre_sim	1.22064211	0.796807982
Feminino	2.5823958	1.2243664700
Masculino	2.3306821	1.1050238899
Dispneia_nao	5.9699158	4.411441984
Dispneia_sim	4.0020196	2.957274104
Internacao_Hospitalar_nao	0.3228635	2.967184652
Internacao_Hospitalar_sim	1.5165860	1.3937748449
Hemoptoico_nao	2.2780077	0.222905433
Hemoptoico_sim	8.5373981	0.835393334

Fonte: O autor.

Tabela A3– Contribuições da ACM para variáveis e *score* confiabilidade.

Classificações de variáveis	Coord. Dim1	Coord. Dim2
Tosse_nao	-1.5840.440	1.33586891
Tosse_sim	0.1371749	-0.11568349
Fumante_exfumante	0.2519683	0.92710586
Fumante_nao	-0.3185610	-0.54101623
Fumante_sim	0.3208513	-0.02121942
Emagrecimento_Nao	-0.1461456	-0.13801896
Emagrecimento_Sim	0.4956771	0.46811429
Sudorese_Noturna_nao	-0.6747256	0.13999010
Sudorese_Noturna_sim	0.7034930	-0.14595867
Febre_nao	-0.7224938	0.15226714
Febre_sim	0.6520702	-0.13742522
Feminino	-0.3157050	-0.56704194
Masculino	0.2849323	0.51177070
Dispneia_nao	-0.5218780	0.37005355
Dispneia_sim	0.3498485	-0.24807076
Internacao_Hospitalar_nao	-0.0846749	-0.21174193
Internacao_Hospitalar_sim	0.3977432	0.99461480
Hemoptoico_nao	-0.2298640	-0.05931208
Hemoptoico_sim	0.8614724	0.22228670

Fonte: O autor.

Tabela A4– Coordenadas da ACM para variáveis e *score* confiabilidade.