**UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC**

**CURSO DE CIÊNCIA DA COMPUTAÇÃO**

**LEANDRO LUIZ MAZZUCHELLO**

# ACURÁCIA DOS CLASSIFICADORES BAYESIANOS NAIVE BAYES, BAYES NET E AODE NO APOIO AO DIAGNÓSTICO DE OSTEOPOROSE E OSTEOPENIA

**CRICIÚMA**

**2013**

**LEANDRO LUIZ MAZZUCHELLO**

# ACURÁCIA DOS CLASSIFICADORES BAYESIANOS NAIVE BAYES, BAYES NET E AODE NO APOIO AO DIAGNÓSTICO DE OSTEOPOROSE E OSTEOPENIA

Trabalho de Conclusão de Curso, apresentado para obtenção do grau de Bacharel no curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC.

Orientadora: Profa Dra Priscyla Waleska Targino de Azevedo Simões

**CRICIÚMA**

**2013**

LEANDRO LUIZ MAZZUCHELLO

# ACURÁCIA DOS CLASSIFICADORES BAYESIANOS NAIVE BAYES, BAYES NET E AODE NO APOIO AO DIAGNÓSTICO DE OSTEOPOROSE E OSTEOPENIA

Trabalho de Conclusão de Curso aprovado pela Banca Examinadora para obtenção do Grau de Bacharel, no Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense, UNESC, com Linha de Pesquisa em Informática em Saúde.

Criciúma, 25 de novembro de 2013.

**BANCA EXAMINADORA**

. Priscyla Waleska Targino de Azevedo Simões – Doutora - (Universidade do Extremo Sul Catarinense) - Orientadora

Prof. Paulo João Martins – Mestre - (Universidade do Extremo Sul Catarinense)

Profa. Viviane Raupp Nunes de Araújo – Mestre - (Universidade do Extremo Sul Catarinense)

## Dedicatória

Dedico está minha conquista a minha família pelos anos de luta, a equipe do projeto que se fez presente no desenvolvimento do mesmo.

A CEPAE, por ter me ajudado incondicionalmente, sendo estas pessoas, Janaina Vitório, Daniel Preve, entre outras pessoas que me motivaram para que eu pudesse chegar onde cheguei.

## AGRADECIMENTOS

Agradeço primeiramente a Deus, por ter me ajudado a nunca desistir dos meus objetivos, por não me deixar desmotivar nas minhas dificuldades, por sempre me acompanhar em meus caminhos.

Aos meus pais Ana Maria Boaventura Luiz Mazzuchello e Antônio Albertino Mazzuchello, por sempre acreditarem em mim, sempre me apoiarem durante a graduação.

Também ao Professor Paulo João Martins por ter me orientado durante o TCC I e II, devido ao afastamento da minha orientadora por causa do seu doutorado.

Aos meus colegas e amigos do grupo de pesquisa em Tecnologia da Informação e Comunicação na Saúde, Lucas Carlessi, Everson Bigaton e Ramon Vensom, por me ajudarem nos momentos mais difíceis deste projeto, e por estarem sempre me apoioando.

Ao Maicon Bastos Palhano por esses anos de amizade, por sempre me ajudar nos momentos mais difíceis da graduação.

E quero agradecer a Deus ter colocado uma orientadora competente e grandiosa que é a Professora Priscyla Waleska Targino de Azevedo Simões, por ter me ajudado muito neste projeto, por ter acreditado em mim, por ter extraído o máximo de conhecimento que eu obtive, por cobrar as coisas que eram para serem cobradas, acho que sem ela na orientação o trabalho não sairia como o desejado, pois foi ela que me indicou o assunto no qual eu me propus ir a fundo.

.

**“Você não é derrotado quando perde, você é derrotado quando desiste”**
**Dr. House**

## RESUMO

Buscando tratar a incerteza inerente ao diagnóstico biomédico, os classificadores bayesianos são baseados em modelos estatísticos tendo o diferencial em relação aos classificadores clássicos, em determinar a classe a que pertence determinado registro tendo-se como base a probabilidade de um elemento pertencer a determinada classe. Assim, o diagnóstico de doenças como a osteoporose e osteopenia pode se tornar mais rápido e preciso, e particularmente, nos casos em que o estado do paciente piora a cada dia agravando o quadro clínico. Nesse contexto, esta pesquisa buscou avaliar a acurácia dos classificadores bayesianos Bayes Net, Naive Bayes e AODE no apoio ao diagnóstico de osteopenia e osteoporose. É um estudo de natureza aplicada (tecnológica). No desenvolvimento desta pesquisa foram realizadas as seguintes etapas metodológicas: levantamento bibliográfico, pré-processamento, mineração de dados e pós-processamento. Os experimentos foram realizados na shell Weka versões 3.6.9 e 3.7.8, pois o algoritmo AODE só está implementado na versão 3.6.9, e a versão 3.7.8 é mais completa nas medidas de avaliação para os demais algoritmos se comparada com a 3.6.9. Nos experimentos foram utilizados os três algoritmos supracitados, sendo geradas 327 minerações, as quais foram utilizadas no pós-processamento e resultaram na análise da acurácia. Entre as medidas estatísticas analisadas nos experimentos (instâncias classificadas corretamente e incorretamente, estatística Kappa, Erro médio absoluto, sensibilidade, especificidade, precisão recall, medida-F, e AUC), o algoritmo AODE foi considerado o mais acurado, apesar de um pouco mais lento em relação ao tempo de execução. A partir dos experimentos realizados sugere-se como trabalhos futuros: utilizar uma maior quantidade de ferramentas comparando algoritmos Bayesianos com outros algoritmos de classificação; e minerar a base de dados buscando caracterizar o conhecimento implícito por meio de árvores de decisão, J48, entre outros.



**Palavras-chave:** Inteligência Computacional. Informática em Saúde. Mineração de Dados. Classificação Bayesiana. Osteoporose.

## ABSTRACT

Seeking to address the referring to uncertainty in biomedical diagnosis, bayesian classifiers are based on statistical models with the differential compared to classical classifiers, to determine the class to which it belongs given record taking as basis the probability of an element belonging to a certain class. Therefore, the diagnosis of diseases such as osteoporosis and osteopenia can become faster and more accurate, and particularly in cases where the patient's condition every day worsens symptoms. In this context, this study sought to evaluate the accuracy of bayesian classifiers Bayes Net, Naive Bayes and AODE to support diagnosis of osteopenia and osteoporosis. It is a study of applied nature (technology). Literature survey, pre-processing, data mining and post-processing: In the development of this research the following methodological steps were performed. The experiments were performed in Weka shell versions 3.6.9 and 3.7.8, for the AODE algorithm is only implemented in version 3.6.9, and version 3.7.8 is the most complete evaluation measures for the other algorithms compared with 3.6.9. In the experiments the above three algorithms, mining 327 being generated, which were used in post-processing and analysis resulted in accuracy were used. Among the statistical measures analyzed in the experiments (instances classified correctly and incorrectly, Kappa statistic, mean absolute error, sensitivity, specificity, recall, accuracy, F-measure, and AUC), the AODE algorithm was considered the most accurate, though a little more slow relative to the runtime. From the experiments it is suggested as future work: using a larger quantity of shells comparing algorithms with other Bayesian classification algorithms, and mine the database trying to characterize the implicit knowledge via decision trees, J48, among others.



**Keywords:** Computational Intelligence. Medical Informatics. Data Mining. Bayesian Classification. Osteoporosis.

## LISTA DE TABELAS

## LISTA DE ILUSTRAÇÕES

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| ACC | Taxa de Acurácia |
| AODE | Averaged One-Dependence Estimators |
| AUC | Area Under Curve |
| cm² | Centímetro Quadrado |
| DCBD | Descoberta de Conhecimento em Base de Dados |
| DMO | Densidade Mineral Óssea |
| DP | Desvio Padrão |
| EUA | Estados Unidos da América |
| FN | Falso Negativo |
| FP | Falso Positivo |
| GPL | General Public License |
| GPS | Global Positioning System |
| IMC | Índice de Massa Corporal |
| IOF | International Osteoporosis Foundation |
| KNN | K-Nearest Neighbor |
| MD | Mineração de Dados |
| NB | Naive Bayes |
| NBD | Naive Bayes Supervised Discretization |
| NBK | Naive Bayes Kernel Estimator |
| RB | Rede Bayesiana |
| ROC | Receiver Operating Characteristic |
| SEN | Sensibilidade |
| SMO | Sequential Minimal Optimization |
| SPE | Especificidade |
| SPSS | Statistical Package for Social Sciences |
| SVM | Support vector machine |
| VN | Verdadeiro Negativo |
| VP | Verdadeiro Positivo |

**SUMÁRIO**

# 1 INTRODUÇÃO

Ao longo dos anos a Tecnologia da Informação vem contribuindo gradativamente para a saúde, com prioridade à melhoria do acesso aos recursos, fato que foi um dos principais passos para que a saúde e a computação caminhassem juntas, buscando oferecer informações e diagnósticos eficientes por meio do uso das ferramentas tecnológicas (CRAIG; PATTERSON, 2005, tradução nossa).

Segundo Shortliffe e Cimino (2006, tradução nossa), dentre as principais áreas de atuação da Informática em Saúde tem-se: os sistemas de informação em saúde; os sistemas de registro eletrônico em saúde; a telessaúde; os sistemas de apoio à decisão; a Mineração de Dados (MD), entre outras.

A MD é entendida como uma das principais etapas do processo de Descoberta de Conhecimento em Bases de Dados (DCBD) e tem sido amplamente usada na busca de novos padrões, novas hipóteses e conhecimento oriundos de dados biomédicos. Devido ao seu poder preditivo, como por exemplo, em estabelecer se uma pessoa está doente ou não, as tarefas da mineração de dados têm sido muito utilizadas no apoio ao diagnóstico médico e em aplicações de assistência à saúde (CHEN et al., 2005, tradução nossa).

Dentre as tarefas relacionadas à MD, tem-se a classificação que visa estabelecer uma função que mapeie um conjunto de elementos em certo número de classes, e tem representado nos últimos anos a tarefa mais utilizada na saúde (DREISEITL et al., 2001, tradução nossa; GOLDSCHMIDT; PASSOS, 2005). Por exemplo, um estudo realizado por Kandaswamy et al. (2004, tradução nossa), resultou em uma aplicação que atua no diagnóstico médico a partir da classificação de sinais sonoros do pulmão.

Buscando tratar a incerteza inerente ao diagnóstico biomédico, os classificadores bayesianos são baseados em modelos estatísticos tendo o diferencial em relação aos classificadores clássicos, em determinar a classe a que pertence determinado registro tendo-se como base a probabilidade de um elemento pertencer a determinada classe. Assim, o diagnóstico de doenças como a osteoporose e osteopenia pode se tornar mais rápido e preciso, e particularmente,

em que o estado do paciente piora a cada dia agravando o quadro clínico (CHANDRA; PANDAV, 1987, tradução nossa; CHEN et al., 2005, tradução nossa).

Como medida de avaliação do conhecimento gerado por um classificador bayesiano tem-se a acurácia que indica a capacidade de um sistema em oferecer um diagnóstico próximo ao real, sendo calculada considerando a sensibilidade e a especificidade. A sensibilidade indica o poder de um algoritmo ou software em classificar como doentes aqueles que realmente são doentes; em contrapartida, a especificidade representa o poder de um aplicativo em sugerir como não doentes, os indivíduos que realmente não apresentam determinada doença. Logo, um software com a alta sensibilidade pode ter baixa especificidade, gerando falsos-positivos levando, por exemplo, indivíduos que não possuem determinada doença a realizar um tratamento desnecessário, o que resultaria em um custo alto e o uso de métodos invasivos de tratamento. Assim, é ideal que um sistema para ser usado na prática clínica apresente acurácia alta (próxima de 100%), resultante de altas taxas de sensibilidade e especificidade combinadas (FLETCHER; SUZANNE, 2006, tradução nossa).

Nesse contexto, pretende-se nesta pesquisa avaliar a acurácia de alguns classificadores bayesianos no apoio ao diagnóstico de osteoporose e osteopenia.

## 1.1 OBJETIVO GERAL

Avaliar a acurácia de alguns classificadores bayesianos no apoio ao diagnóstico de osteopenia e osteoporose.

## 1.2 OBJETIVOS ESPECÍFICOS

Os objetivos deste projeto consistem em:

a) compreender sobre os Classificadores Bayesianos e suas aplicações;
b) conhecer as medidas de avaliação de testes diagnósticos;
c) atender aos aspectos éticos e legais pertinentes à Informática em Saúde;
d) conhecer alguns classificadores bayesianos;

e) conhecer sobre osteopenia e osteoporose, e seus aspectos epidemiológicos;
f) oferecer alguns experimentos por meio da utilização de alguns classificadores bayesianos;
g) oferecer a acurácia do conhecimento gerado e medidas associadas por meio da classificação bayesiana em osteoporose e osteopenia.

## 1.3 JUSTIFICATIVA

A capacidade computacional crescente das máquinas em processar, manipular dados e formalizar a lógica diagnóstica tem permitido o desenvolvimento de programas específicos com boa acurácia diagnóstica, e auxiliado aos profissionais da saúde no processo de tomada de decisão. Nesse contexto, a MD tem sido considerada bem sucedida em diversas aplicações voltadas à saúde com destaque na saúde pública (MOON; KWON; LEE, 2001, tradução nossa; YANG; WANG, 2006, tradução nossa).

Técnicas matemáticas e estatísticas clássicas, como o Teorema de Bayes implicitamente presente nos classificadores bayesianos têm apresentado acurácia alta em diversos estudos, como o realizado nos Estados Unidos da América (EUA) por Shabbeer et al. (2012, tradução nossa) que revelou acurácia maior que 95% em uma aplicação voltada à classificação de genótipos do complexo da Tuberculose. Outro estudo dos EUA realizado por Nolen et al. (2012, tradução nossa) apresentou acurácia superior a 80% em uma aplicação voltada a classificação de imagens tomográficas para o diagnóstico de câncer de ovário.

Por outro lado, a osteoporose apesar de considerada um processo normal do envelhecimento, afeta 8 milhões de mulheres e 2 milhões de homens nos EUA. No Brasil há uma prevalência de 15% nas mulheres e 13% nos homens em indivíduos com 40 anos ou mais (AMED et al., 2012). Como consequência destacam-se as fraturas as quais diminuem a qualidade de vida e contribuem para aumentar a morbimortalidade devido às lombalgias, alterações circulatórias, respiratórias e tromboembólicas (COSTA-PAIVA et al., 2007). Enfatiza-se a fratura de quadril a qual apresenta taxa de mortalidade de 30% em 1 ano, sendo que 46% dos indivíduos desenvolverão dependência de cuidados de terceiros para as

atividades da vida diária (SZEJNFELD et al., 2007). Tudo isto gera um alto custo e diminuição da qualidade de vida.

Assim, devido a inexistência de estudos comparativos sobre a acurácia de classificadores bayesianos voltados ao diagnóstico de osteopenia e osteoporose, e à baixa utilização da MD nessa área de conhecimento, buscamos com essa pesquisa explorar a acurácia de alguns classificadores bayesianos, a fim de oferecer um método diagnóstico precoce e menos invasivo.

## 1.4 TÓPICOS DO TRABALHO

No primeiro capitulo, é apresentada uma breve introdução, os objetivos, e justificativa, logo em seguida no segundo capitulo, é detalhada a descoberta de conhecimento em base de dados sendo uma das principais etapas desse processo, a mineração de dados.

O terceiro capítulo descreve a classificação bayesiana e seus métodos de avaliação, tendo em vista que eles são implementados em ferramentas, descritas no capitulo 4.

Já no quinto Capitulo acrescenta-se uma breve introdução sobre osteopenia e osteoporose, bem como seu surgimento, fatores de risco, e principais fraturas. O capítulo 5 descreve alguns trabalhos correlatos.

No sexto capítulo a metodologia do estudo é detalhada e os resultados apresentados. O sétimo capítulo apresenta as principais conclusões obtidas, limitações da pesquisa e sugestões para trabalhos futuros.

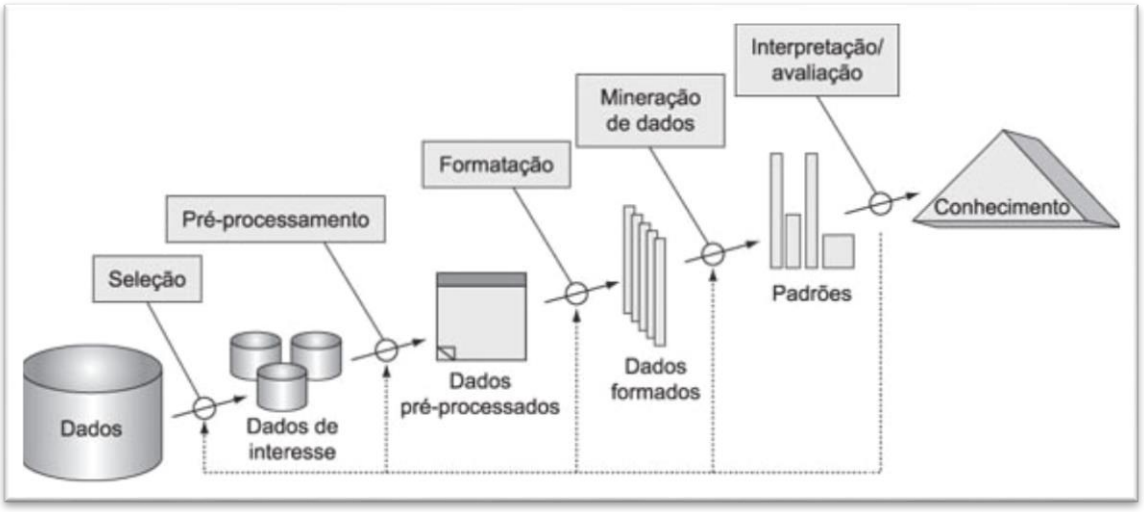
# 2 DESCOBERTA DO CONHECIMENTO EM BASE DE DADOS

O processo *de* descoberta de conhecimento em bases de dados, surgiu em 1989, onde foi realizado o primeiro workshop onde o propósito deste era que produto final fosse o conhecimento, já em 1995 realizou-se a primeira conferência mundial, onde o assunto principal era o DCBD, e no ano seguinte teve sua segunda versão, onde muitos pesquisadores desta área participaram, onde apresentaram trabalhos desenvolvidos afim que pudessem dar ênfase a essa área (ROMÃO, 2002).

O principal objetivo do DCBD é a extração de informações relevantes a partir de grandes bases de dados, passando por todas as etapas do processo (MELLO, 2001).

Segundo Manarin (2004), para que se possa aplicar o DCBC, é essencial que se conheça a base de dados também é preferível que a base contenha muitos dados, assim é possível obter resultados satisfatórios, a Figura 1 mostra quais as etapas do DCBD são essências para a obtenção do conhecimento.

Figura 1 - Etapas do processo de DCBD




Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

As principais etapas do DCBD consistem em:

a) **limpeza e integração dos dados**: Consiste em tratar os valores e a inconsistência dos dados afim de que possa coletar só as informações mais importantes (ESTEVAM JUNIOR, 2003);
b) **seleção e transformação dos dados:** escolhe-se a base onde será aplicado o DCBD, transformando na formas apropriadas, utilizando as técnicas de data mining (ESTEVAM JUNIOR, 2003);
c) **mineração de dados (MD*)***: é a etapa onde, efetivamente ocorre a busca por padrões e relacionamentos nos dados, por meio de algoritmos específicos de *data mining* (GOLDSCHMIDT; PASSOS, 2005);
d) **avaliação e apresentação**: este é um critério no qual vai se verificar quais os resultados são relevantes e quais não são. Considerando esses fatores o conhecimento é obtido e apresentado ao usuário (ESTEVAM JUNIOR, 2003).

## 2.1 MINERAÇÃO DE DADOS

A Mineração de Dados é referenciada como sendo o processo onde é possível obter conhecimentos relevantes e implícitos em um grande conjunto de dados, também conhecidos como padrões encontrados nos conjuntos de dados, o processo pode ser de dois tipos, semiautomático ou automático (WITTEN; FRANK; HALL, 2011, tradução nossa).

Os padrões em MD são eventos/características, ou ainda uma combinação de eventos em uma base de dados, no qual, quando os padrões são descobertos, maiores são as chances de se fazer predições acuradas aos dados futuros, porém, podem existir problemas com os padrões se a qualidade dos dados for ruim, onde, é aconselhável que sejam utilizados algoritmos robustos e que os mesmos possam trabalhar com dados imperfeitos (REZENDE, 2003).

### 2.1.1 Histórico

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996, tradução nossa), a descoberta de conhecimento teve vários nomes antes de ser denominada de

descoberta de conhecimento, diversas nomenclaturas foram estabelecidas: extração do conhecimento, descoberta da informação, colheita da informação, arqueologia de dados e processamento de padrões de dados, até que, a definição mais atribuída foi a mineração de dados.

A MD é uma área interdisciplinar, tendo como base: banco de dados, estatística, redes neurais, aprendizado de máquina, reconhecimento de padrões, visualização de dados e sistemas distribuídos (REZENDE, 2003).

### 2.1.2 Categorias

Segundo Han e Kamber (2006, tradução nossa), a mineração de dados é dividida em categorias, sendo elas descritivas e preditivas, onde, na descritiva se apresenta um conjunto de dados de maneira concisa e sumarizada apresentando as propriedades mais interessantes dos dados, já na preditiva analisam-se os dados com finalidade de construir um modelo que busque prever o comportamento dos conjuntos de dados obtidos pela MD.

Normalmente um conjunto de dados contém muitos detalhes, mas geralmente os usuários preferem visualizar os dados da maneira mais simples possível, ou seja, sem muito detalhamento, portanto é necessário que essas visões devam ser flexibilizadas e tratadas (REZENDE, 2003).

### 2.1.3 Área de aplicação

Segundo Cheng, Bell e Liu (2001, tradução nossa), a MD pode ser utilizada em muitas áreas, como:

a) **mineração de dados comercial**: destina-se obter vantagens competitivas no mercado através de extração do conhecimento;
b) **mineração de dados na Internet:** o objetivo principal é melhorar a pesquisa e extração de conhecimento na rede, além disso, o DM pode ser utilizado em segurança de redes onde é possível detectar invasão;
c) **mineração de dados científicos:** o propósito inicial do MD científico é enfatizar a detecção de padrões e construir modelos capazes de imitar os comportamentos de fenômenos, seja ele físico, químico ou

biológico, no qual, é possível desenvolver sistemas especialistas, afim de ajudar nos diagnósticos em diversas áreas inclusive a medicina.

### 2.1.4 Aplicações científicas

Segundo Politi (2006), houve uma rápida evolução na computação nessas duas décadas, com uma grande mudança, alto desempenho e várias outras ferramentas que buscam obter o conhecimento ajudando assim várias áreas distintas, esses dados podem ser obtidos através de vários equipamentos como radares, GPS, sensores, instrumentos ópticos, sondas, satélites, entre outros.

Este capítulo permitiu conhecer sobre o processo de descoberta de conhecimento que possibilita obter conhecimentos relevantes muitas vezes implícito mas de difícil visualização. O capítulo a seguir apresenta a classificação bayesiana.

# 3 CLASSIFICAÇÃO BAYESIANA

Uma das tarefas que existem na MD é a Classificação, que segundo Witten, Frank e Hall (2011, tradução Nossa), consiste em gerar modelos que descrevam uma classe de dados, sendo que estes modelos são baseados tendo-se como base um conjunto de dados de treinamento, onde a classe que estão armazenados os dados seja conhecida. Dentre as técnicas de classificação, destacam-se as redes neurais, K-Nearest Neighbor *(*KNN*),* árvores de decisão e classificadores bayesianos.

A classificação em MD acontece em duas etapas:

a) **treinamento**: é utilizado um conjunto de dados, sendo assim, constrói-se um modelo baseado a partir dos atributos dos dados, cada registro de dados pertence a uma determinada classe, onde é descrita por um dos seus atributos, sendo esse denominado de atributo da classe. Nesta etapa de treinamento é construído o modelo, sendo a classificação uma técnica de aprendizagem supervisionada, onde se conhecem os valores de atributos de classe (BERRY; LINOFF;1997);
b) **teste**: o modelo gerado na etapa de treinamento é utilizado para predizer o valor de um atributo de classe de um conjunto de dados de teste, devendo-se na etapa de testes agir com precaução, pois o classificador pode apresentar certas anomalias que estão contidas no conjunto de dados de treinamento, mas que podem não estar contidas no conjunto de dados gerais (FONSECA, 2007).

Segundo Witten, Frank e Hall (2011, tradução nossa), depois de finalizada a classificação, para cada registro dos dados, a classe predita é comparada com a classe real. A acurácia representa então a proporção de registros que foram classificados corretamente.

Já a classificação bayesiana foi descrita pela primeira fez por Cestnick, revelando alguns vantagens em relação a classificação clássica, em que a estrutura do gráfico é fixa, sendo que, os valores são independentes incondicionalmente um

do outro, assim a forma pra se aprender o classificador se torna simples, pois as ordens de dependências são fixas, e reduzidas em duas variáveis, sendo que apenas os parâmetros numéricos possam ser aprendidos (SANTOS, 2007).

De acordo com Curotto e Ebecken (2006, tradução nossa) o classificador bayesiano possui eficiência em termos de velocidade e acurácia. Segundo Sahami (2004, tradução nossa), a eficiência de um classificador bayesiano é boa, sendo utilizado para resolver muitos problemas de classificação, fazendo que se obtenham melhores resultados se comparados a outros classificadores.

Estes classificadores bayesianos são algoritmos estatísticos que preveem a probabilidade de associação a uma determinada classe (HAN; KAMBER, 2006, tradução nossa).

Segundo Tan, Steinbach e Kumar (2009) são utilizados com frequência para lidar com situações de incerteza por aleatoriedade, sendo que é preciso combinar o conhecimento prévio acerca de uma hipótese com novas evidências extraídas dos dados que se pretende analisar, tornando os cálculos envolvidos mais compreensíveis e obtendo um menor tempo no processo de aprendizagem.

## 3.1 REDES BAYESIANAS

O classificador bayesiano é originado do conceito de rede bayesiana (RB) que é um grafo orientado, onde cada nó possui informações de probabilidade quantitativas (RUSSEL; NORVIG, 2004), sendo que cada nó raiz pode influenciar seus nós filhos.

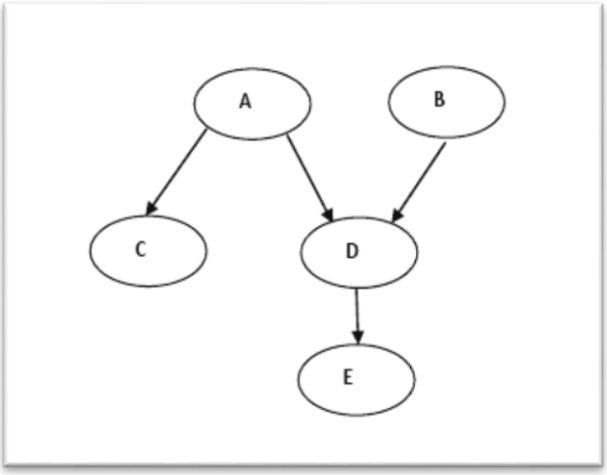
As redes bayesianas são definidas por dois componentes fundamentais, sendo eles: quantitativo e qualitativo, onde, o quantitativo é formado pela distribuição de probabilidades condicionais e o qualitativo é representado pela topologia do grafo (NEAPOLITAN, 2004, tradução nossa).

A parte qualitativa é dada pelo seguinte grafo $G = (V, E)$, onde $V$ representam as variáveis, e estas variáveis podem ser discretas ou continuas dependendo da ocasião, enquanto $E$ representa os arcos da rede bayesiana, e tais arcos representam a dependência dos nós. Os nós podem conter variáveis de múltiplos estados, e os arcos fornecem a relação de dependência condicional, tal que, para cada arco construído de uma variável $X$ em relação a uma variável $Y$, tal

dueto irá representar uma dependência direta de $Y$ em $X$, e isso posteriormente gera uma relação de causa e efeito (BOBBIO et al., 2001, tradução nossa).

Bobbio et al. (2001, tradução nossa) sugere ainda que quando um arco é construído de uma variável $X$, então esta variável é denominada de pai, sendo que $Y$ será seu filho, conforme ilustra a Figura 2.

Figura 2 - Estrutra gráfica de uma redes bayesiana




Fonte: adaptado de Ara-Souza (2010).

Na Figura 2, A é pai de $C$ e D, B é pai de D e D é pai de E, onde A e B são nós raízes, ou seja, não têm pais.

Segundo Sigurdsson, Walls e Quigley (2001, tradução nossa), a probabilidade condicional é determinada pra cada um dos nós da RB que não seja classificado como nó pai (nesse caso ele possui probabilidades incondicionais *a priori*), sendo que tal probabilidade condicional representa a influência de um nó pai sobre determinado nó filho, e o valor das probabilidades *a priori*, representam o componente quantitativo de uma RB.

## 3.2 AVALIAÇÃO DE CLASSIFICADORES BAYESIANOS

Segundo Witten, Frank e Hall (2011), a eficiência do classificadores é medida com sua taxa de erro, o classificador descreve cada registro dos dados, se estiver correto a classificação é correta, senão incorreta, para se medir a eficiência de uma classificação temos índice instâncias classificadas corretamente e

incorretamente, estatística Kappa, Erro médio absoluto, sensibilidade, especificidade, precisão recall, medida-f, e AUC. Para cada tipo de classificação, podem existir métodos diferentes para o cálculo em relação ao erro.

Assim, pode-se utilizar na avaliação dos resultados gerados por um classificador bayesianos, os testes diagnósticos que buscam medir a eficiência de um Classificador Bayesiano. Assim, depois de treinar um classificador, é possível testar seu desempenho, realizando a comparação dos resultados do classificador, com um resultado real, que pode ser um padrão ouro, ou mesmo o diagnóstico de um especialista.

Os resultados de cada registro (instância) devem ser apresentados em uma tabela de contingência 2x2 (Figura 3).

Figura 3 - Tabela de contingência 2x2

| | | Padrão ouro + | Padrão ouro - | |
|---|---|---|---|---|
| Teste em avaliação | + | a=VP | b=FP | P |
| | - | c=FN | d=VN | N |
| | | D | ND | T |

**Sendo:**

**a=VP**: número de pessoas com resultado positivo no teste em avaliação e para o padrão ouro: verdadeiros positivos (VP).
**b=FP**: número de pessoas com resultado positivo no teste em avaliação e negativo para o padrão ouro: falsos positivos (FP).
**c=FN**: número de pessoas com resultado negativo para o teste em avaliação e positivo para o padrão ouro: falsos negativos (FN).
**d=VN**: número de pessoas com resultado negativo para teste em avaliação e negativo para o padrão ouro: verdadeiros negativos (VN).
**P**: número total de pessoas com resultado positivo.
**N**: número total de pessoas com resultado negativo.
**D**: número total de pessoas com a doença.
**ND**: número total de pessoas sem a doença.
**T**: número de pessoas no estudo.

Fonte: Adaptado de Zamora et al. (2006)

Existem quatros tipos de resultados em teste, estes são:

a) **verdadeiro positivo (VP):** consiste em um teste com o resultado positivo para a pessoa que apresente a doença;

b) **verdadeiro negativo (VN):** trata-se de um teste com o resultado negativo para a pessoa que não apresenta a doença;
c) **falso positivo (FP):** é o resultado positivo para um individuo que não tem a doença;
d) **falso negativo (FN):** esse é o resultado negativo para um individuo que apresente a doença.

A acurácia pode ser estimada pela combinação da sensibilidade e especificidade ou das razões de verossimilhança. A sensibilidade (Fórmula 2) representa o poder do teste em classificar como doentes aqueles que realmente são doentes (ZAMORA et al., 2006).

Tal medida é essencial para se constatar se os resultados da classificação estão corretos ou não.

### 3.2.1 Sensibilidade

Segundo Massad et al. (2004), indica a proporção de verdadeiros positivos entre todos os doentes, sendo que quanto mais próximo de 1, maior a chance do indivíduo apresentar a doença.

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \quad (1)$$

### 3.2.2 Especificidade

Segundo Massad et al. (2004), revela a proporção de verdadeiros negativos entre todas as pessoas sadias, sendo que quanto mais próximo de 1, maior a chance do indivíduo não apresentar a doença.

$$\text{Especificidade} = \frac{VN}{FP + VN} \quad (2)$$

### 3.2.3 Acurácia

A acurácia expressa a chance de ter um resultado positivo pelo teste em avaliação nos doentes se comparado com um teste negativo (ZAMORA et al., 2006).

$$Acurácia = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} \times \frac{\frac{d}{b+d}}{\frac{c}{a+c}} = \frac{a \times d}{b \times c} \quad (3)$$

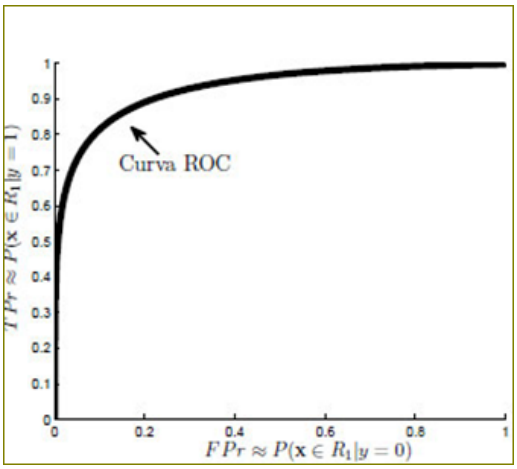
### 3.2.4 Curva ROC

É possível analisar a acurácia pela construção de um gráfico da curva ROC que possibilita avaliar os experimentos individuais, incluindo as medidas estatísticas consideradas, as quais podem destacar uma covariação entre a sensibilidade e especificidade (ZHOU et al., 2002; GATSONIS; PALIWAL, 2006).

Por definição, uma curva ROC é a representação gráfica dos pares sensibilidade ou VP (ordenadas) e 1-especificidade ou FP (abcissas), resultante da variação do valor de corte ao longo de um eixo de decisão, x. A representação gráfica assim é designada por curva ROC no plano unitário.

Segundo Massad et al. (2004), a curva ROC é um diagrama que integra a sensibilidade e a especificidade, permitindo observar como os atributos do diagnóstico se relacionam em relação ao valor do corte. Sendo que, quanto mais próxima de 1 for a sua área, mais preciso é o diagnóstico.

Figura 4 - Exemplo de Curva ROC




Fonte: adaptado de Castro e Braga (2011)

A Figura 4 ilustra os valores de sensibilidade e especificidade para alguns experimentos. Uma classificação acurada deve posicionar os pontos associados à sensibilidade e especificidade de cada instância no canto superior esquerdo do gráfico onde a sensibilidade e especificidade são próximas de 1 (LEEFLANG et al., 2008).

Assim, a área sob a curva (AUC) ROC relaciona a taxa de acerto com a taxa de falsos alarmes, e tornou-se uma medida padrão para avaliação de classificadores. A AUC é uma estimativa da probabilidade de que um classificador irá classificar um exemplo positivo escolhido aleatoriamente maior do que um exemplo negativo escolhido aleatoriamente.

$$Area \approx \int_a^b f(x)dx \tag{4}$$

### 3.2.5 Kappa

A estatística Kappa é uma medida de concordância inter-observadora que mede o grau de concordância além do que seria esperado pelo acaso. Esta medida de concordância tem como valor máximo 1, onde este valor 1 representa total concordância e os valores próximos, contudo os valores próximos de 0 indicam nenhuma concordância, ou que a concordância não foi exatamente a esperada pelo acaso. Um eventual valor de Kappa menor que zero, ou seja, negativo indica que a concordância encontrada foi menor do que aquela esperada pelo acaso (CAMPOS, 2004).

Sua intepretação pode ser realizada de acordo com os critérios apresentados na Tabela 1.

Tabela 1 – Interpretação da estatística Kappa

| Valor do kappa | Concordância |
|---|---|
| **0** | Pobre |
| **0 – 0,20** | Ligeira |
| **0,21 – 0,40** | Considerável |
| **0,41 – 0,60** | Moderada |
| **0,61 – 0,80** | Substancial |
| **0,81 – 1** | Excelente |

Fonte: Campos (2004)

### 3.2.6 Medida-F

De acordo com Hripcsak e Rothschild (2005), a Medida-f é definida como uma média harmônica de precisão (P), e Recall (R), sendo melhor quando seu valor estiver próximo de 1, sua formula é descrita a seguir.

$$F = \frac{2PR}{P + R} \tag{5}$$

### 3.2.7 Precisão

Segundo Hripcsak e Rothschild (2005), a precisão é a proximidade entre os valores obtidos pela repetição do processo de mensuração, sendo que mais próximo de 1.

$$P = \frac{TP}{TP + FP} \tag{6}$$

### 3.2.8 Recall

Recall a probabilidade de que um documento relevante (escolhidos aleatoriamente) é recuperado em uma pesquisa, sendo que quanto mais o recall for perto de 1 melhor serão os resultados o recall é dado pela fórmula abaixo (HRIPCSAK; ROTHSCHILD, 2005).

$$R = \frac{TP}{TP + Fn} \tag{7}$$

### 3.2.9 Erro Absoluto Médio

Segundo Witten, Frank e Hall (2011) é uma medida que indica a média do afastamento de todos os valores fornecidos pelos classificadores e o seu real valor. Pode ser calculada utilizando a Equação abaixo, na qual $n$ é o número de amostras, $x_i$ é o valor fornecido pelo classificador para a i-ésima amostra e $\bar{x}$ é a média dos valores de todas as amostras.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right| \tag{8}$$

### 3.2.10 Instâncias Classificadas Corretamente

As Instâncias Classificadas Corretamente representam a proporção de registros classificados corretamente, determinada pela proporção de acertos entre os registros classificados correta e incorretamente (WITTEN; FRANK; HALL, 2011).

### 3.2.11 Instâncias Classificadas Incorretamente

As Instâncias Classificadas Incorretamente representam a proporção de registros classificados incorretamente, determinada pela proporção de acertos entre os registros classificados correta e incorretamente (WITTEN; FRANK; HALL, 2011).

Pode-se inferir que os classificadores bayesianos são adequados para a resolução de diversos problemas, pois trabalham com probabilidades, o que sugere seu uso em problemas que tratam a incerteza por aleatoriedade. Tais classificadores estão também implementados em shells, e algumas são apresentadas no próximo capítulo.

# 4 FERRAMENTAS PARA CLASSIFICAÇÃO BAYESIANA

Nos dias atuais, a mineração de dados tem sido importante na busca de conhecimentos implícitos em grandes bases de dados em que a elaboração de probabilidades manuais seria inviável, pois a rapidez não seria a mesma se comparada ao uso de uma ferramenta, contudo, foram desenvolvidas ferramentas denominadas de *shells* que possibilitam usar a MD, incluindo várias tarefas como a classificação, clusterização, entre outras, e que oferecem diversos algoritmos. Nesse capítulo veremos três dessas ferramentas que disponibilizam classificadores bayesianos: Weka, Tanagra e Knime.

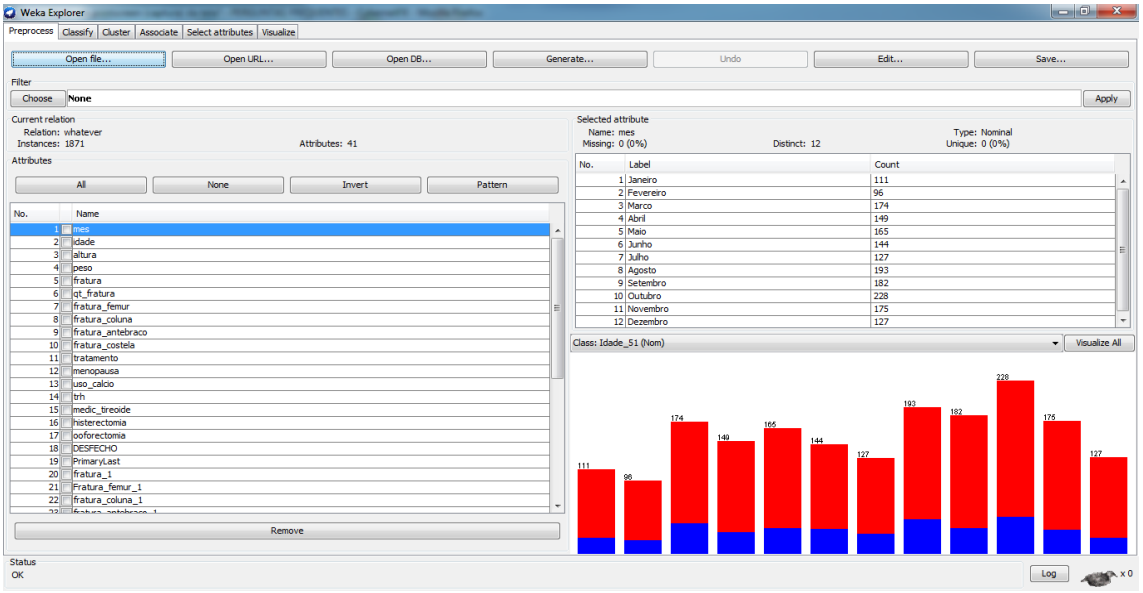
## 4.1 WEKA

O Weka é uma ferramenta desenvolvida pela universidade de Waikato, localizada na Nova Zelândia, sendo open source com licença GPL, apresentando diversos algoritmos incluindo os bayesianos e também algumas tarefas de Mineração de Dados, dentre elas, a classificação. Esta ferramenta vem sendo amplamente utilizada na comunidade científica, pois possui uma interface amigável que permite aos seus usuários um rápido aprendizado.

Segundo Witten, Frank e Hall (2011, tradução nossa) e Hall et al. (2009, tradução nossa), o WEKA surgiu através da percepção da necessidade de um software unificado que permitisse aos pesquisadores a fácil utilização de técnicas em aprendizagem de máquina. No momento do início do projeto, em 1992, os algoritmos de aprendizagem estavam disponíveis em vários idiomas, para uso em diferentes plataformas e já funcionavam com diversos formatos de dados. Foi previsto que o WEKA não só iria fornecer um conjunto de ferramentas de Mineração de Dados, mas também uma estrutura no interior da qual os desenvolvedores poderiam programar novos algoritmos.

Hoje em dia, o WEKA é reconhecido como um ponto de referência para a Mineração de Dados e aprendizado de máquina (WITTEN; FRANK; HALL, 2011, tradução nossa). A Figura 5 ilustra o pré-processamento na ferramenta Weka.

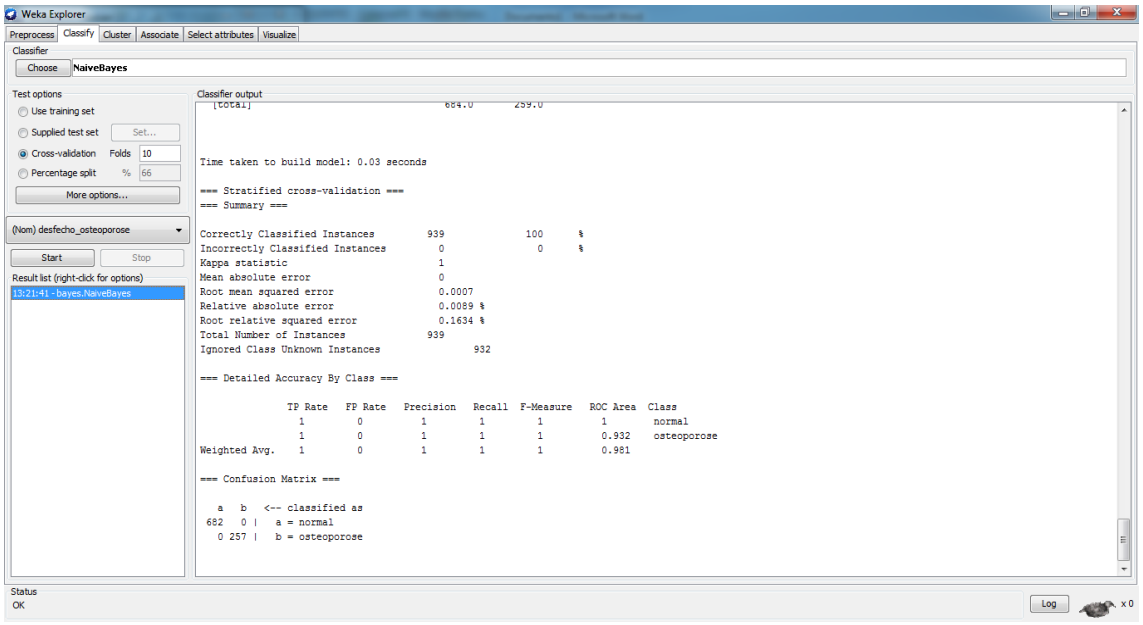
Figura 5 - Aba de pré-processamento do Weka.




Fonte: Weka (2013).

Na figura 6 é possível visualizar alguns resultados obtidos pelo WEKA, oriundos da tarefa de classificação bayesiana pelo algoritmo Naive Bayes.

Figura 6 - Resultados obtidos utilizando o Naive Bayes




Fonte: Weka (2013).

Um tutorial detalhado sobre a utilização do algoritmo Naive Bayes na ferramenta Weka é apresentado no Apêndice A.

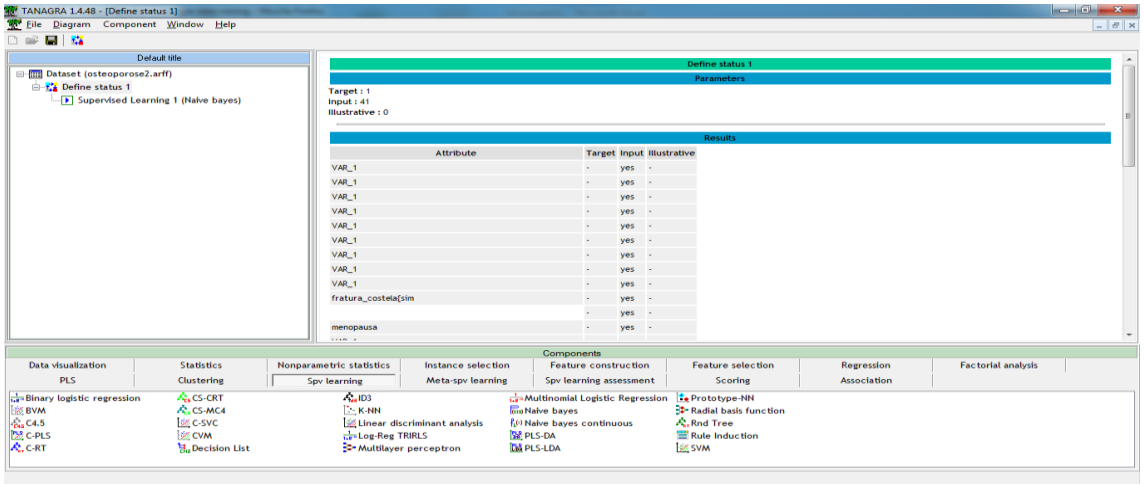
## 4.2 TANAGRA

Segundo os desenvolvedores dessa ferramenta, o Tanagra é um software que apresenta interface simples e de fácil utilização, o que simplifica seu uso para usuários menos experientes. Possui licença open source para fins acadêmicos e de pesquisa. Apresenta vários métodos de mineração de dados a partir da análise exploratória de dados, aprendizagem estatística, e aprendizado de máquina (TANAGRA, 2013).

O principal objetivo desta ferramenta é oferecer aos pesquisadores e estudantes um software de mineração de dados de facil utilização, e que esteja em conformidade com as normas de desenvolvimento de software neste domínio (especialmente relacionadas ao design de interface gráfica e a maneira de usá-lo) (TANAGRA, 2013).

Por ser open source, tal ferramenta possibilita o livre acesso ao código-fonte, perimitindo analisar como o software é construído, os problemas para evitar, as principais etapas do projeto e quais ferramentas e bibliotecas de código devem ser utilizadas. Desta forma, o TANAGRA pode ser considerado uma ferramenta pedagógica por viabilizar um instrumento que pode ser utilizado no processo de ensino-aprendizagem de técnicas de programação. A Figura 7 a seguir ilustra a tela principal do TANAGRA.

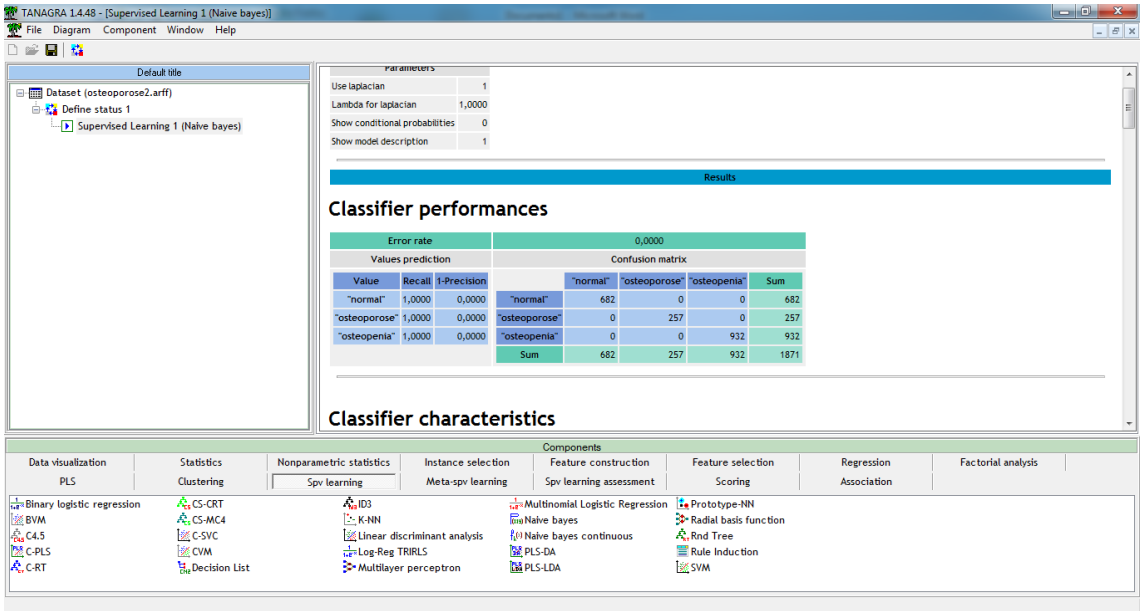
Figura 7 - Ferramenta TANAGRA




Fonte: Tanagra (2013)

Na figura 8 pode-se observar como os resultados são gerados pelo algoritmo Naive Bayes.

Figura 8 - Resultados do TANAGRA utilizando o Naive Bayes.




Fonte: Tanagra (2013)

Um tutorial detalhado sobre a utilização do algoritmo Naive Bayes na ferramenta Tanagra é apresentado no Apêndice B.

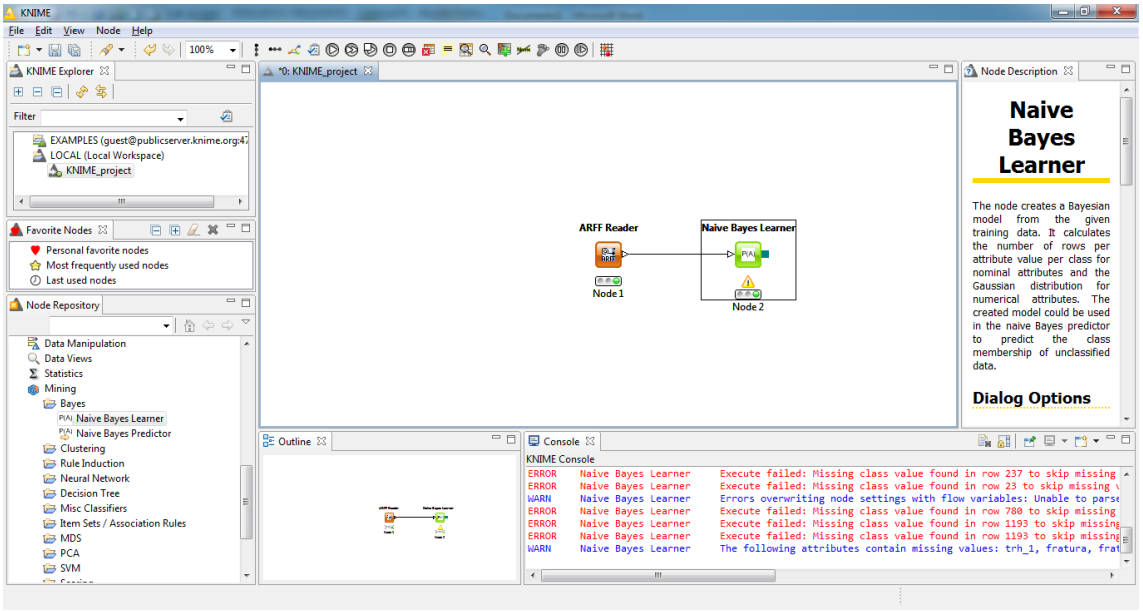
## 4.3 KNIME

No início de 2004 uma equipe de desenvolvedores originada de uma empresa de software do Vale do Silício especializada em aplicações farmacêuticas começou a trabalhar na Universidade de Konstanz (Alemanha) em uma nova plataforma open source como uma colaboração de uma ferramenta de pesquisa (KNIME, 2013).

Este produto teria que processar e integrar grandes quantidades de dados, assim, os desenvolvedores aderiram a rigorosos padrões de engenharia de software para criar uma robusta, modular e escalável plataforma, abrangendo vários tipos de arquivos e conexões com bancos de dados, análises e modelos de exploração visual (KNIME, 2013).

Quando a primeira versão do KNIME foi lançada em 2006, várias empresas farmacêuticas começaram a usá-lo e, logo em seguida, os fornecedores de software começaram a construir aplicativos baseados nesta ferramenta. Nos dias atuais os usuários desta ferramenta podem ser encontrados em empresas de grande escala através de uma ampla gama de indústrias, incluindo médica, a de serviços financeiros, editores, distribuidores e varejistas on-line, empresas de consultoria, governamentais e de pesquisa em mais de 50 países. O desenvolvimento é open source, com código escrito em Java e baseado em Eclipse (KNIME, 2013). A Figura 9 ilustra a execução do algoritmo Naive Bayes Learner no KNIME.

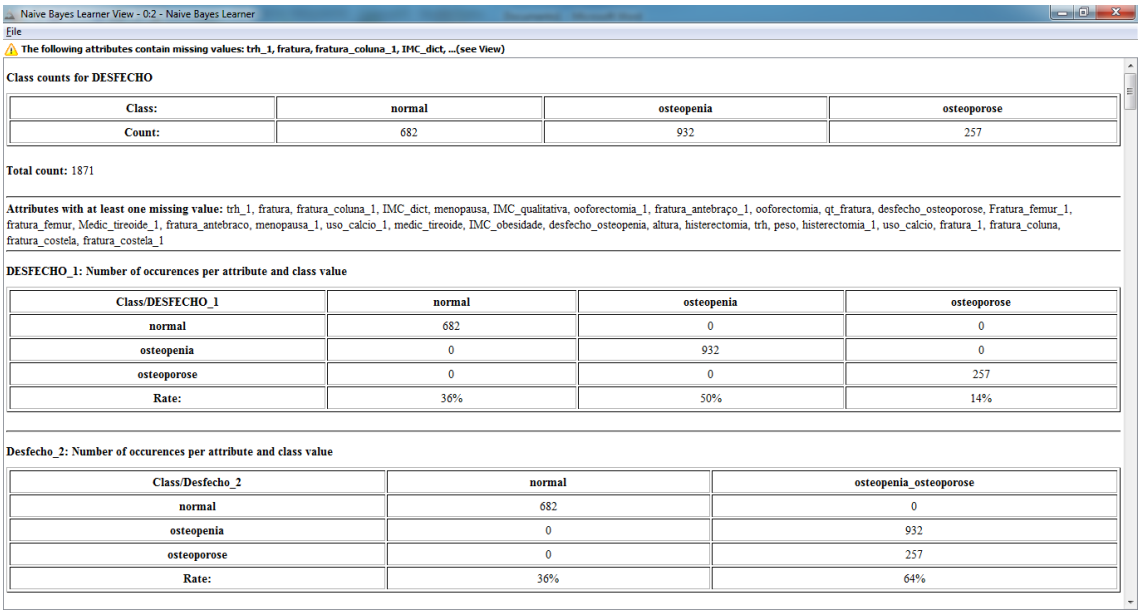
Figura 9 - Knime Data Mining




Fonte: Knime (2013).

Logo depois desse processo são obtidos os resultados com o Naive Bayes Learner, como vemos na figura 10.

Figura 10 - Resultados utilizando o algoritmo Naive Bayes Learner



Naive Bayes Learner View - 0:2 - Naive Bayes Learner

File

**The following attributes contain missing values: trh_1, fratura, fratura_coluna_1, IMC_dict, ...(see View)**

**Class counts for DESFECHO**

| Class: | normal | osteopenia | osteoporose |
|---|---|---|---|
| Count: | 682 | 932 | 257 |

**Total count:** 1871

**Attributes with at least one missing value:** trh_1, fratura, fratura_coluna_1, IMC_dict, menopausa, IMC_qualitativa, ooforectomia_1, fratura_antebraço_1, ooforectomia, qt_fratura, desfecho_osteoporose, Fratura_femur_1, fratura_femur, Medic_tireoide_1, fratura_antebraco, menopausa_1, uso_calcio_1, medic_tireoide, IMC_obesidade, desfecho_osteopenia, altura, histerectomia, trh, peso, histerectomia_1, uso_calcio, fratura_1, fratura_coluna, fratura_costela, fratura_costela_1

**DESFECHO_1: Number of occurences per attribute and class value**

| Class/DESFECHO_1 | normal | osteopenia | osteoporose |
|---|---|---|---|
| normal | 682 | 0 | 0 |
| osteopenia | 0 | 932 | 0 |
| osteoporose | 0 | 0 | 257 |
| Rate: | 36% | 50% | 14% |

**Desfecho_2: Number of occurences per attribute and class value**

| Class/Desfecho_2 | normal | osteopenia_osteoporose |
|---|---|---|
| normal | 682 | 0 |
| osteopenia | 0 | 932 |
| osteoporose | 0 | 257 |
| Rate: | 36% | 64% |

Fonte: Knime (2013).

Um tutorial detalhado sobre a utilização do algoritmo Naive Bayes Learner na ferramenta Tanagra é apresentado no Apêndice C.

Essa pesquisa considerou o estudo das ferramentas Weka, Knime e Tanagra, pois elas possuem algoritmos bayesianos. A Tabela 1 a seguir apresenta um comparativo entre as três ferramentas.

Tabela 1 - Descrição das shells e seus respectivos algoritmos

| Ferramenta | Fabricante | Licença | Pais | Download | Algoritmo bayesiano |
|---|---|---|---|---|---|
| **Tanagra** | Tanagra | Open source | França | http://eric.univ-lyon2.fr/~ricco/tanagra/ | NaiveBayes e NaiveBayes contínuos |
| **Knime** | Knime | Open source | Alemanha | http://www.knime.org/about/news/knime-version-273-released | Naive Bayes Learner, Naive Bayes Pedictor |
| **Weka** | Waikato | GNU | Nova Zelândia | www.cs.waikato.ac.nz/ml/weka/ | Naive Bayes, Averaged One-Dependence Estimators (AODE),Bayes Net |

Fonte: Dados da pesquisa.

Pela análise das ferramentas, optou-se por utilizar nessa pesquisa somente a shell Weka, que oferece mais medidas de avaliação, se comparada com o Tanagra e o Knime.

As ferramentas apresentadas nesse capítulo facilitam o processo de mineração pelos algoritmos bayesianos, que foram utilizados nessa pesquisa em experimentos voltados ao diagnóstico de osteopenia e osteoporose.

# 5 OSTEOPENIA E OSTEOPOROSE

Segundo Portugal (2012), a perda de massa óssea pode trazer duas doenças parecidas, sendo que para identificar uma da outra é preciso de um diagnóstico com base na densidade mineral óssea.

De acordo com Holick e Chen (2006), osteopenia é a denominação que foi designada a baixa densidade óssea, tendo como consequência a osteoporose, ou seja, a osteopenia é o inicio de uma osteoporose, sendo assim, a osteopenia nem sempre é associada à doença, pois existem pessoas que possuem baixa densidade óssea.

Segundo Raisz (2005), a osteoporose e caracterizada por uma síndrome esquelética, onde por sua vez acontece a perda de massa óssea e alteração do tecido ósseo, afetando dessa forma a massa óssea e tornando-a suscetível a fraturas.

Segundo Bishop (2010), a osteoporose pode ocorrer pela desmineralização óssea grave, pela falta de absorção de cálcio durante a fase de crescimento (SIMS; GOOI, 2008, tradução nossa).

A osteoporose é mais prevalente em mulheres que já passaram da menopausa, e em idosos (TARANTNO et al., 2007).

## 5.1 CLASSIFICAÇÃO DA OSTEOPOROSE

A osteoporose pode ser classificada em duas categorias, sendo elas primária e secundária. A primária afeta as pessoas com idade avançada, com progressão da doença com o passar do tempo, com isso a pessoa vai perdendo massa óssea e sua deteorização (PORTUGAL, 2012). Esta categoria pode ser subdividida em dos tipos (USDHHS, 2004, tradução nossa): o tipo 1 e o tipo 2.

a) **Tipo 1**: O tipo 1 está associado à menopausa, afetando principalmente mulheres com idade depois dos 50, variando de pessoa para pessoa, considerando que tal doença é um fenômeno biológico, sendo que após a menopausa o corpo das mulheres passam por transformações resultando na deficiência nos níveis de

estrogênio e de outros hormônios, e tais deficiências podem resultar no risco de osteoporose (USDHHS, 2004, tradução nossa).

b) **Tipo 2**: Está relacionada com o envelhecimento, sendo mais frequente no gênero feminino, contudo este tipo só tende a aparecer depois dos 65 anos de idade e pode ocasionar principalmente as fraturas vertebrais e de fêmur, sendo que a principal causa deste tipo é o hiperparatiroidismo secundário, pois se sabe que depois dos 65 anos a absorção de cálcio diminui (USDHHS, 2004, tradução nossa).

Segundo Usdhhs (2012, tradução nossa), a osteoporose secundária é tratada como patologia, sendo a perda de massa óssea associada a uma doença secundária, oriunda de abusos de remédios inadequados, sendo que este tipo de osteoporose tende mais ao sexo masculino, com chances no sexo feminino, neste caso quando a mulher entra na pré-menopausa.

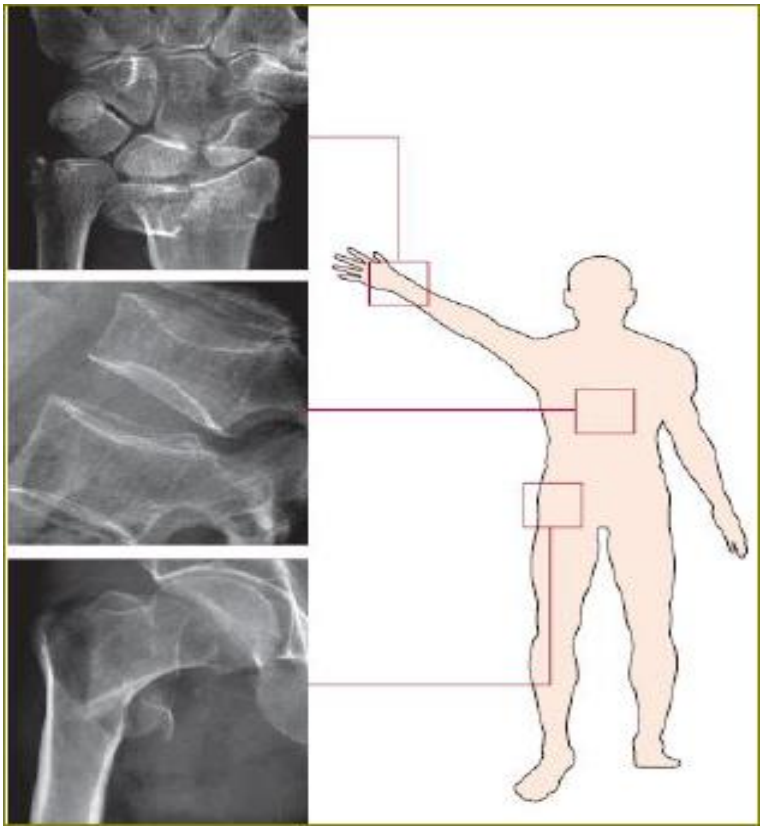
As principais causas da osteoporose secundária estão relacionadas a doenças endócrinas, reumáticas, gastroenterólogicas e hematológicas (USDHHS, 2004, tradução nossa).

## 5.2 FRATURAS OSTEOPORÓTICAS

As fraturas e suas respectivas complicações são a principal fonte de manifestação de osteopenia e osteoporose. A definição é complexa mesmo utilizando métodos eficazes, pois as opiniões são diferente quando o assunto é a localização das fraturas, pode-se definir a fratura como o resultado de um traumatismo baixo ou moderado, sendo que quase sempre a consequência é uma queda sobre o osso fragilizado (PETERS; MARTINS, 2010).

As fraturas ocorrem principalmente na coluna vertebral e na extremidade discal do antebraço, apesar desses três lugares serem os mais destacados a osteoporose pode aparecer em outros lugares, como por exemplo, nas costas. A figura 11 ilustra essas três localizações.

Figura 11 - Principais fraturas osteoporóticas



Fonte: Adaptado de Rachner, Khosla e Hofbauer (2011).

Segundo Johnell e Kanis (2006, tradução nossa), em 2000 foram estimadas em todo mundo 8,959 milhões de fraturas, sendo que 3,1 milhões ocorreram na Europa atingindo o sexo masculino e feminino com a faixa etária igual ou superior a 50 anos, sendo que destas, 620.000 tiveram fratura de fêmur, 490.000 de coluna vertebral e 574.000 no antebraço discal.

### 5.2.1 Fratura de Fêmur

De acordo com Portugal (2012), essa fratura é oriunda da osteoporose, e acarreta muita preocupação tanto nos níveis sócios econômico, quanto na saúde pública.

No ano 2000, Johnell e Kanis (2006, tradução nossa), obtiveram uma estimativa na população mundial ficou em torno de 1,6 milhões, com pessoas com mais de 50 anos que continham osteoporose, mas que segundo a International Osteoporosis Foundation (IOF) deverá passar entre 4,5 e 6,3 milhões.

Tal fratura com mais frequência em mulheres entre os 60 e 85 anos (IP; LEUNG; KUNG, 2010, tradução nossa). Portugal (2012) sugere que as pessoas que tem essa fratura ficam afetadas psicologicamente e a nível motor, com dor são crônica e constante.

### 5.2.2 Fraturas vertebrais

Segundo Kondo (2008, tradução nossa) as fraturas vertebrais constituem 27% de todas as fraturas osteoporóticas em ambos os sexos, estima-se que anualmente deve ocorrer 550 mil a 770 mil fraturas vertebrais.

É um dos tipos de sinais típicos da osteoporose, e as pessoas que apresentam esse tipo de fraturas perdem funcionalidades, tais como descer escada, cozinhar, dormir na horizontal, entre outras. Também pode resultar em sequelas tais como: cifose, diminuição da função pulmonar, distensão abdominal (PAPAIOANNOU et al., 2002, tradução nossa).

### 5.2.3 Fratura da extremidade do antebraço

Essas fraturas são originadas por um traumatismo com intensidade baixa, sendo prevalente em mulheres com mais de 50 anos, e altamente frequente em mulheres com mais de 75 anos, sendo que 85% apresentam baixo índice de DMO e 51% sofrem de osteoporose (BLAKENEY, 2010, tradução nossa).

## 5.3 FATORES DE RISCO

Na osteoporose e osteopenia, ainda existem os fatores de risco que são de dois tipos: fatores modificáveis; e não modificáveis. Os modificáveis consistem na ideia de modificar o padrão de vida, e os não modificáveis resultam de ações biológicas, ou seja, sem alterações. A Tabela 2 a seguir ilustra os fatores de risco modificáveis e não modificáveis.

Tabela 2 - Fatores modificáveis e não modificáveis

| **Fatores modificáveis** | **Não modificáveis** |
|---|---|
| Menopausa | Idade |
| Hipogonadismo/amenorréia prolongada (> 1 ano) | Sexo |
| Redução da ingestão de cálcio, principalmente durante a infância e a puberdade. | Raça |
| Baixo índice de massa Corporal / anorexia nervosa | Fatores Hereditários |
| Consumo excessivo de álcool / cafeína | |
| Tabagismo | |
| Sedentarismo | |
| Quedas | |
| Iatrogenia resultante do uso prolongado de corticoterapia, quimioterapia, antineoplasica, hormônios para tireóide | |
| Terapia crônicas com anticoagulante, antiepiléticos e anticonvulsivantes | |
| Situações anteriores de gastrectomia, artrite reumatóide, hiperparatiroidismo e doenças hepáticas | |

Fonte: Adaptado Portugal (2012).

Tal capítulo possibilitou conhecer os aspectos epidemiológicos, fatores de risco e demais questões associadas ao diagnóstico de osteoporose e osteopenia, morbidades essas ainda não utilizadas em experimentos sobre a acurácia de classificadores bayesianos, assim, o próximo capítulo apresenta alguns trabalhos correlatos sobre estudos de acurácia nos algoritmos considerados em nossa pesquisa e aplicados a diversas áreas de conhecimento.

# 6 TRABALHOS CORRELATOS

Este capítulo apresenta alguns trabalhos correlatos que utilizam alguns algoritmos de classificação bayesiana como o Naive Bayes, AODE e Bayes Net. Tais pesquisas abordam também estudos de análise da acurácia dos algoritmos supracitados, comparados com diversos outros algoritmos e aplicados a diversas áreas de conhecimento.

## 6.1 TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS NA PREVISÃO DE EVASÃO ACADÊMICA

Este trabalho divulgado no ano de 2008 (AMORIM; BARONE; MANSUR, 2008) no Centro Federal de Educação Tecnológica de Campos (RJ), demonstra a acurácia do uso das técnicas de aprendizado de máquina aplicadas à previsão da evasão acadêmica, ou seja, considera os principais aspectos que possam levar um aluno que ainda não concluiu o seu curso, a trancá-lo ou abandoná-lo. Tal estudo considerou uma base de dados com 8.703 registros provenientes de matrículas de uma Instituição de Ensino Superior Particular, localizada no município de Campos dos Goytacazes (RJ), com 4 cursos de graduação plena.

A classificação foi realizada na ferramenta Weka e o banco de dados gerado em formato arff. Em tal ferramenta foram usados os algoritmos J48 (baseado em árvores de decisão), no SMO (que considera máquina de vetores de suporte) e o Bayes Net, também utilizado em nossa pesquisa.

Os algoritmos acima foram treinados com o banco de dados de matrículas de um semestre e testados com o banco de dados de matrículas do semestre seguinte.

Na análise dos resultados (Tabela 3), verificou-se que o algoritmo SMO foi o que revelou maior acurácia dentre os analisados.

Tabela 3 - Análise da acurácia do estudo de Amorim, Barone, Mansur (2008)

| | Bayes Net | SMO | J48 |
|---|---|---|---|
| Classificacao Correta | 89,7084% | 91,2521% | 89,6512% |
| Classificacao Incorreta | 10,2916% | 8,7479% | 10,3488% |

Fonte: Amorim, Barone, Mansur (2008)

## 6.2 TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL APLICADAS AO DESENVOLVIMENTO DE UM SISTEMA DE SUPORTE A DECISÃO PARA O DIAGNÓSTICO DE DOENÇA CELÍACA

Este trabalho divulgado no ano de 2011 na Universidade Federal de São Paulo, demonstra o apresenta o desenvolvimento de um sistema de suporte a decisão clínica integrado a um classificador aplicado ao diagnóstico de doença celíaca (TENÓRIO et al., 2011).

Foi desenvolvido um sistema Web que considerou um banco de dados originado de um estudo retrospectivo que incluiu 178 casos clínicos, utilizados no treinamento. Os testes foram realizados com um banco de 270 casos na ferramenta Weka 3.6.1 a partir de 5 técnicas de Inteligência Artificial (árvores de decisão, redes Bayesianas, algoritmo do vizinho mais próximo, máquinas de suporte vetorial, e redes neurais). Nesse estudo foram avaliadas a acurácia, sensibilidade, especificidade, e a área da curva ROC, critérios utilizados na decisão para selecionar o algoritmo que seria usado no Sistema de Apoio a Decisão.

Dentre os algoritmos analisados o Averaged One-Dependence Estimator (AODE), um dos classificadores bayesianos incluídos nas análises (e também utilizado em nossa pesquisa) foi o que apresentou maior acurácia, conforme ilustra a Tabela 4.

Tabela 4 - Análise da acurácia do estudo de Tenório et al. (2011)

| Algorithm | ACC (%) | SEN | SPE | AUC |
|---|---|---|---|---|
| LBR[a] | 76.9 | 0.79 | 0.75 | 0.84 |
| AODE-F1[a] | 76.3 | 0.78 | 0.75 | 0.84 |
| AODEsr[a] | 75.9 | 0.78 | 0.74 | 0.84 |
| NaiveBayes[a] | 76.0 | 0.77 | 0.75 | 0.84 |
| NaiveBayesSimple[a] | 76.0 | 0.77 | 0.75 | 0.84 |
| KStar-B60 | 76.2 | 0.76 | 0.77 | 0.84 |
| BayesNet A0.75 | 76.0 | 0.78 | 0.74 | 0.83 |
| IBk-k16 | 73.3 | 0.80 | 0.67 | 0.82 |
| MLP-L0.4-M0.2[b] | 73.4 | 0.70 | 0.76 | 0.79 |
| LibSVM-C1.0-G0.0 | 77.3 | 0.76 | 0.78 | 0.77 |
| J48-C0.7-M9 | 72.7 | 0.77 | 0.69 | 0.76 |
| ADTree[a] | 71.5 | 0.70 | 0.72 | 0.76 |
| SimpleCart[a] | 71.9 | 0.72 | 0.72 | 0.73 |

ACC: Taxa de acurácia; SEN: Sensibilidade; SPE: Especificidade;
AUC: Área sob a curva ROC.
[a] Algoritmos com parâmetros padrão.
Kernel usado no Support vector machine (SVM): sigmoid.
[b]Camadas escondidas: 22.
Fonte: Tenório et al. (2011)

.

## 6.3 DESENVOLVIMENTO E AVALIAÇÃO DE UM CLASSIFICADOR DE PADRÕES PARA ANÁLISE DO CRESCIMENTO FACIAL PELO MÉTODO DE MATURAÇÃO VERTEBRAL CERVICAL

Este trabalho desenvolvido no ano de 2011 na Universidade Federal de São Paulo apresenta um classificador de padrões baseado no método Cervical Vertebral Maturation, que auxilia o profissional ortodontista na determinação do período ideal para o tratamento de uma série de deformidades dentofaciais (BAPTISTA et al., 2011).

Para o desenvolvimento do classificador, foram coletadas 187 imagens de radiografias cefalométricas laterais, cuja idade óssea foi avaliada manualmente por um especialista. Em um software desenvolvido pelos autores, outro especialista marcou os pontos de interesse da vértebras nas imagens, originando a base de dados utilizada nos experimentos que foram realizados na ferramenta Weka 3.6.1.

Foram utilizados três algoritmos baseados no Naive Bayes (NB) (também utilizado em nossa pesquisa), O Naive Bayes original, o Naive Bayes – use Kernel Estimator (NBK) e o Naive Bayes – use Supervised Discretization (NBD) (DUDA; HART; STORK, 2000).

Foram então realizados os experimentos dos algoritmos acima citados, sendo avaliada a Área sob a curva ROC de cada estágio cervical conforme apresenta a Tabela 5.

Tabela 5 - Resultados dos testes com os classificadores de Baptista et al. (2011)

| Classificador | CS1 | CS2 | CS3 | CS4 | CS5 | CS6 |
|---|---|---|---|---|---|---|
| NB | 0,85 | **0,74** | 0,82 | 0,75 | 0,79 | **0,92** |
| NBK | 0,85 | 0,73 | **0,86** | 0,75 | 0,8 | 0,91 |
| NBD | **0,88** | 0,72 | 0,84 | **0,76** | **0,82** | 0,91 |

Fonte: Baptista et al. (2011)

Segundo os autores, os valores da AUC apresentados pelos classificadores foram satisfatórios, sugerindo que cada algoritmo possui um grau de certeza adequado para diferenciar cada estágio cervical dos demais (BAPTISTA et al., 2011).

Este capítulo mostrou a utilização de classificadores bayesianos em diversas áreas do conhecimento, assim, mediante estudo realizado, o próximo capítulo aborda a metodologia dessa pesquisa e os resultados obtidos.

# 7 TRABALHO DESENVOLVIDO

Este trabalho teve como objetivo avaliar a acurácia de alguns classificadores bayesianos no apoio ao diagnóstico de osteopenia e osteoporose.

No desenvolvimento desta pesquisa foram realizadas as seguintes etapas metodológicas: levantamento bibliográfico, pré-processamento, mineração de dados e pós-processamento dos dados, sendo finalizada com a apresentação dos experimentos realizados.

## 7.1 LEVANTAMENTO BIBLIOGRAFICO

O levantamento bibliográfico contemplou toda a pesquisa e foi realizado durante o desenvolvimento do projeto que norteou este trabalho, permitindo a compreensão dos conceitos sobre Descoberta do Conhecimento em Base de Dados (incluindo nesse item o Data Mining e a Classificação Bayesiana), Classificadores Bayesianos, algumas ferramentas para Classificação Bayesiana, e trabalhos correlatos.

A partir do levantamento realizado realizaram-se os experimentos de acurácia bayesiana utilizando-se para isso uma base de dados de osteoporose desenvolvida no estudo de Silva (2013).

As referências utilizadas para o desenvolvimento deste trabalho consideraram trabalhos de conclusão de curso, dissertações, teses, artigos científicos indexados, trabalhos publicados em eventos científicos, alguns livros e os sites das ferramentas Weka, Tanagra e Knime.

Após o levantamento bibliográfico inicial necessário para o entendimento dos conceitos supracitados, realizou-se o pré-processamento da base de dados.

## 7.2 PRÉ-PROCESSAMENTO

O pré-processamento foi iniciado com a definição do banco de dados que foi submetido ao processo de descoberta de conhecimento em base de dados, sendo realizada a seleção de atributos, limpeza dos dados, binarização e transformações de variáveis. O pré-processamento é fundamental pois aperfeiçoa a

etapa central da DCBD que é a mineração de dados (WITTEN; FRANK; HALL, 2011, tradução nossa).

### 7.2.1 Definição da Base de Dados

A base de dados utilizada nos experimentos foi oriunda de um estudo transversal realizado como Trabalho de Conclusão do Curso de Medicina da Universidade do Extremo Sul Catarinense (SILVA, 2013) com 1871 mulheres submetidas ao exame de densitometria óssea num serviço especializado do sul de Santa Catarina no período de janeiro de 2012 a dezembro de 2012, e aprovado pelo Comitê de ética em Pesquisa da Universidade do Extremo Sul Catarinense sob o protocolo 82939/2012 (Anexo A).

Nesse estudo que originou o banco de dados, o diagnóstico de osteopenia e osteoporose foi realizado através de densitometria de dupla energia de Raios X (DEXA), permitindo medir a densidade mineral óssea (DMO) com o uso do equipamento GE Lunar Prodigy Primo com software Encore versão 13.20.

A densitometria é considerada o padrão ouro para medir a DMO e fazer o diagnóstico de osteopenia/osteoporose. Os resultados de densitometria óssea são apresentados através de:

a) **Valores absolutos da DMO (g/cm2)**: Os valores absolutos são importantes, pois são os utilizados para monitorar as mudanças da DMO ao longo do tempo.
b) **T-score**: Calculado, em desvios padrão (DP), tomando como referência a DMO média do pico da massa óssea em adultos jovens. Os critérios diagnósticos propostos pela Organização Mundial de Saúde, em 1994, baseiam-se nestes dados: até -1,0 DP  normal, de -1,1 a -2,49 DP  osteopenia e abaixo de -2,5 DP  osteoporose (KANIS; WHO GROUP, 1994).

Os dados gerais foram retirados do laudo densitométrico e incluíram a medida do peso, da altura, idade, Índice de Massa Corpórea, fraturas prévias, uso de cálcio, medicação para tireóide, menopausa, terapia de reposição hormonal, sintomas de climatério, histerectomia e ooforectomia prévias.

A idade foi categorizada em percentis:

a) **percentil 25**: idade até 51 anos;
b) **percentil entre 25-50**: idade entre 52-57 anos;
c) **percentil entre 50 e 75**: idade entre 58 a 65 anos;
d) **percentil acima de 75**: idade igual ou maior de 66 anos.

As demais variáveis qualitativas foram dicotomizadas. O índice de massa corporal (IMC) foi calculado pela fórmula peso (kg) / altura$^2$ (m). Dados densitométricos coletados incluíram os valores da DMO (g/cm$^2$) do colo do fêmur preferencialmente direito, do fêmur total, e valor médio das vértebras lombares (L1-L4).

### 7.2.2 Seleção dos atributos

Na seleção de atributos, a quantidade de variáveis que constituíram o banco de dados foi reduzida de 60 para 39, sendo que as mais associadas ao diagnóstico de osteopenia e osteoporose foram selecionadas (SILVA, 2013). Tal seleção foi feita considerando o objetivo do estudo de Silva (2013) que buscou pesquisar associações multifatoriais a osteoporose e osteopenia entre as variáveis selecionadas.

A caracterização do bando de dados completo é apresentada na Tabela 6.

Tabela 6 - Características do banco de dados

| Nome dos atributos | Atributos do banco | Descrição | Numeração atributos |
|---|---|---|---|
| Numero de identificação | Protocolo | Numero de identificação | - |
| Prontuário | Código_URC | Número do prontuário | - |
| Mês | Mês | Sem descrição | -- |
| Tempo da menopausa | tempo_menop | tempo da menopausa | - |
| Coluna lombar | CL_DMO | DMO coluna lombar | - |
| T score da coluna lombar | CL_T_score | T-score coluna lombar: | - |
| Percentual coluna lombar | CL_T_percentual | % T-score coluna lombar: % T-score | - |
| Z score coluna lombar | CL_Z_score | Z-score coluna lombar: | - |
| Percentual da coluna lombar | CL_Z_percentual | % Z-score DO coluna lombar | - |
| Dmo | CF_DMO | DMO colo femoral | - |
| T score do colo femural | CF_T_score | T-score colo femoral | - |
| Percentual T score do colo femural | CF_T_percentual | % T-score colo femoral: | - |
| Z score do colo femural | CF_Z_score | Z-score colo femoral | - |
| Percentual do Z-score do colo femural | CF_Z_percentual | % Z-score colo femoral: | - |
| Total DMO de fêmur | Femur_total_DMO | DMO fêmur total | - |
| Total T-score Fêmur | Femur_total_T_score | T-score fêmur total | - |
| Percentual total de T score de fêmur | Fêmur_total_T_percentual | % T-score fêmur total | - |
| Total de Z score de fêmur | Fêmur_total_Z_score | Z-score fêmur total: Z-score | - |
| Percentual de Z score fêmur total | Fêmur_total_Z_percentual’ | % Z-score fêmur total | - |
| Transformações | TRANSFORMACOES | SEM DESCRIÇAO | - |
| Índice de massa corporal contínuo | IMC_cont | IMC continuo | - |
| Índice de massa corporal qualitativa | IMC_qualitativa | classificação do IMC | - |
| Diagnostico | DESFECHO_1 | normal osteopenia e osteoporose | 1 normal, 2-osteopenia e osteoporose |
| Idade até 49 | idade_50 | Idade até 49 | |
| Diagnóstico de osteoporose | desfecho_osteoporose | Osteoporose | - |
| Diagnóstico de osteopenia | desfecho_osteopenia | Osteopenia | 1-Osteopenia, 2-normal |
| Percentual do grupos de idades | NIdade | Percentual do grupo de idades | - |
| Idade até 51 | Idade_51 | Sem descrição | - |
| Fratura | fratura_1 | fratura dicotômica | - |
| Fratura do fêmur | Fratura_femur_1 | FEMUR_ dicotômica | - |
| Fratura de coluna | Fratura_coluna_1 | coluna dicotômica | - |
| Fratura do antebraço | Fratura_antebraço_1 | antebraço dicotômica | - |
| Fratura da costela | Fratura_costela_1 | costela dicotômica | - |

| | | | |
|---|---|---|---|
| Menopausa | Menopausa_1 | Menopausa dicotômica | - |
| Uso de cálcio | Uso_calcio1 | Cálcio dicotômico | - |
| Tratamento hormonal | Trh_1 | TRH dicotômico | - |
| Ooforectomia | Ooforectomia_1 | ooforectomia dicotômica | - |
| Histerectomia | Histerectomia_1 | histerectomia dicotômica | - |
| Idade dicotomizada | Idade_1 | idade_dicotomica_media | - |
| IMC dicotomizado | IMC_dict | IMC dicotomizado até 25 e 25 ou mais | - |
| Medicamento tireoide | Medic_tireode_1 | medicamento para tireoide | 1-sim, 2 não |
| Diagnostico dicotomizado | Desfecho_2 | desfecho dicotômico | |
| Idade | idade | Idade dos indivíduos por faixa etária | - |
| Altura | Altura | Altura dos individuo por faixa etária | - |
| Peso | Peso | Media de peso por faixa etária | - |
| Fratura | Fratura | Contém fratura | 1-Sim,2-não |
| Quantidade de fratura | Qt_fratura | Quantidades de fraturas prévias | - |
| Fratura de fêmur | Fratura_ fêmur | Fratura prévia em colo do fêmur | 1-Sim,2-não |
| Fratura de coluna | Fratura_coluna | Fratura prévia na coluna | 1-Sim,2-não |
| Fratura de antebraço | Fratura_antebraço | Fratura prévia em antebraço | 1-Sim,2-não |
| Fratura_costela | Fratura_costela | Fratura prévia em costela | 1-Sim,2-não |
| Tratamento | tratamento | Uso de medicamentos | 1-Sim,2-não |
| Menopausa | menopausa | Menopausa | 1-Sim,2-não |
| Uso de cálcio | Uso_calcio | Uso de Cálcio | 1-Sim,2-não |
| Terapia de reposição | Trh | Terapia de reposição hormonal | 1-Sim,2-não |
| Medicamento | Medic_tireoide | Uso de medicamentos para tireóide | 1-Sim,2-não |
| Histerectomia | histerectomia | Histerectomia | 1-Sim,2-não |
| Ooforectomia | ooforectomia | Ooforectomia | 1-Sim,2-não |
| Diagnóstico | Desfecho | Diagnostico | - |
| Indicador de cada ultimo caso de osteoporose | PrimaryLast | Indicador de cada último caso | - |
| Índice de massa corporal obesidade | IMC_obesidade | IMC normal sobrepeso e 3 graus de obesidade | |

Fonte: Dados da pesquisa.

Os atributos excluídos os experimentos são descritos na Tabela 7 a seguir.

Tabela 7 - Atributos não utilizados

| Variável | Descrição | Códigos dos atributos do banco original | Código dos atributos pré processadas |
|---|---|---|---|
| Protocolo | Numero de identificação | - | - |
| Código_URC | Número do prontuário | - | - |
| Mês | Sem descrição | 1,2.3,4,5,6,7,8,9,10, 11,12 | 1-janeiro, 2-fevereiro, 3-marco, 4-abril, 5-maio, 6-junho, 7-julho, 8-agosto, 9-setembro, 10-outubro, 11-novembro, 12-dezembro |
| Tempo_menop | Tempo da menopausa | 1,2,3, 5,...,54, 98-sem respostas,99-não se aplica | 1,2,3, 5,..., 54, 98-sem_respostas,99-não_aplica |
| Cl_dmo | Dmo coluna lombar | - | - |
| CL_T_score | T-score coluna lombar: | - | - |
| CL_T_percentual | % T-score coluna lombar: % T-score | - | - |
| CL_Z_score | Z-score coluna lombar: | - | - |
| CL_Z_percentual | % Z-score DO coluna lombar | - | - |
| Cf_dmo | Dmo colo femoral | - | - |
| CF_T_score | T-score colo femoral | | - |
| CF_T_percentual | % T-score colo femoral: | - | - |
| CF_Z_score | Z-score colo femoral | - | - |
| CF_Z_percentual | % Z-score colo femoral: | - | - |
| Femur_total_DMO | DMO fêmur total | - | - |
| Femur_total_T_score | T-score fêmur total | - | - |
| Fêmur_total_T_percentual | % T-score fêmur total | - | - |
| Fêmur_total_Z_score | Z-score fêmur total: Z-score | - | - |
| Fêmur_total_Z_percentual’ | % Z-score fêmur total | - | - |
| Transformações | Sem descrição | - | - |
| IMC_cont | IMC continuo | - | - |

Fonte: Dados da pesquisa.

### 7.2.3 Limpeza dos dados e binarização

Durante a limpeza dos dados, os registros com informações inconsistentes foram retificados ou excluídos, além disso, verificou-se que algumas variáveis (associadas ao diagnóstico) já estavam dicotomizadas, assim, não havendo a necessidade da binarização.

No estudo de Silva (2013) foram dicotomizadas as variáveis com mais de duas respostas, assim optamos por manter a binarização destas variáveis. Para finalizar esta etapa foi realizada a transformação das variáveis, onde as possíveis respostas de cada atributo foram convertidas de nominal para numérico.

### 7.2.4 Transformação das Variáveis

Na etapa de transformação das variáveis, campos que continham espaços foram substituídos underline, e campos vazios foram substituídos por pontos de interrogação. Tal funcionalidade foi necessário em virtude das ferramentas utilizadas necessitando de tal ajuste.

### 7.2.5 Escolha da ferramenta

A partir do estudo das ferramentas Weka, Tanagra e Knime apresentadas anteriormente, optou-se por utilizar nos experimentos somente a shell Weka, que oferece mais medidas de avaliação, se comparada com o Tanagra e o Knime, ou seja, instâncias classificadas corretamente e incorretamente, estatística Kappa, Erro médio absoluto, sensibilidade, especificidade, precisão recall, medida-f, AUC, entre outras.

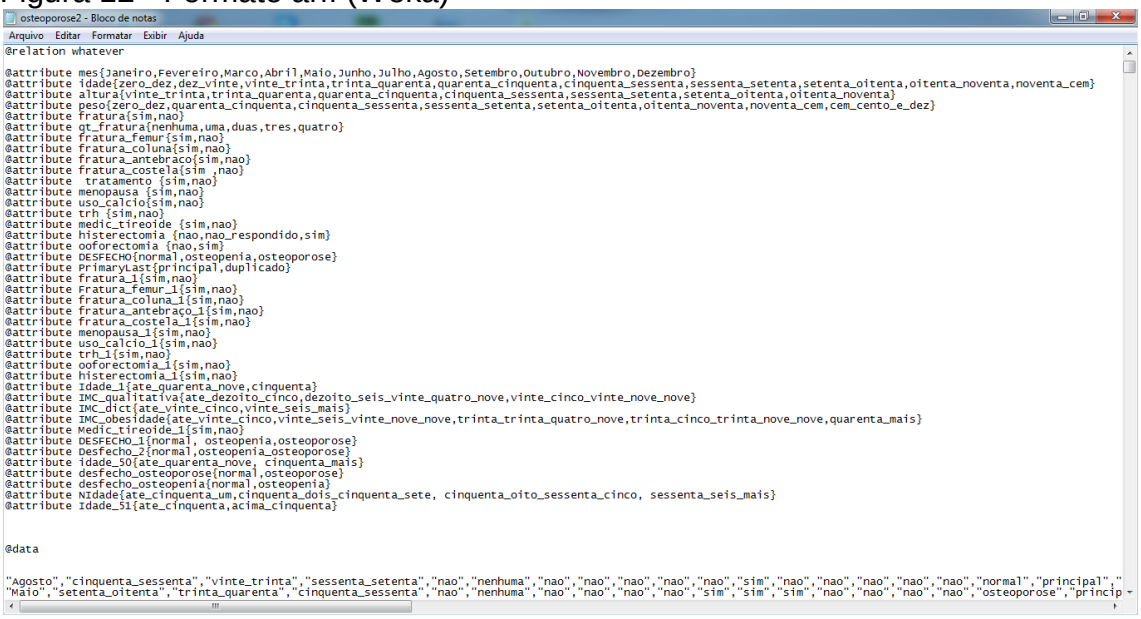
No Tanagra existem poucas medidas de avaliações sendo elas, o Recall e a 1- Precision e no KNIME existe apenas o negative, positive e o rate.

### 7.2.6 Modelo ARFF

A base de dados original foi elaborada no software Statistical Package for Social Sciences (SPSS) versão 18.0 (formato .sav), e em seguida exportada para.csv buscando realizar o pré-processamento que se deu no software Excel.

Após finalizada a etapa de pré-processamento, a base de dados (.csv) foi exportada para o formato.arff, utilizado na ferramenta onde foram realizados os experimentos de acurácia (Weka). Tal formato (.arff) é dividido em dois cabeçalho principais, @whatever onde são definidas as variáveis e seus atributos, e @data onde são tabulados os dados, sendo que cada informação disponibilizada no @data deve estar de acordo com sua variável correspondente no cabeçalho @whatever, conforme ilustra a figura 12.

Figura 12 - Formato arff (Weka)




Fonte: Dados da pesquisa.

Após o pré-processamento seguiu-se com a mineração de dados que contemplou a utilização dos algoritmos bayesianos Naive Bayes, Bayes Net e AODE.

## 7.3 MINERAÇÃO DE DADOS

A Mineração de Dados foi iniciada pela seleção das variáveis, seguindo-se da execução dos algoritmos, etapa essa realizada na shell Weka versões 3.6.9 e 3.7.8, pois o algoritmo AODE só está implementado na versão 3.6.9, e a versão 3.7.8 é mais completa nas medidas de avaliação para os demais algoritmos se comparada com a 3.6.9.

### 7.3.1 Seleção das variáveis

Para a mineração de dados utilizou–se apenas as variáveis abordadas no estudo de Silva (2013) fortemente associadas com a osteoporose e osteopenia ilustradas na Tabela 8 e Tabela 9.

### 7.3.2 Algoritmos utilizados

Optou–se por usar os algoritmos bayesianos disponibilizados na ferramenta Weka: o NaiveBayes, Bayes Net, e Aode.

O algoritmo Naive Bayes é considerado um simples classificador bayesiano, sendo seu modelo constituído por um conjunto de probabilidades, que são estimadas pela frequência de cada valor de característica para as instâncias dos dados de treino, sendo que uma nova instância é selecionada e o classificador calcula a sua probabilidade estimando a qual classe pertence essa instância. O Naive Bayes utiliza o teorema de bayes para calcular as probabilidades existentes e por seus atributos serem independentes, considerando o valor ao qual classe pertença.

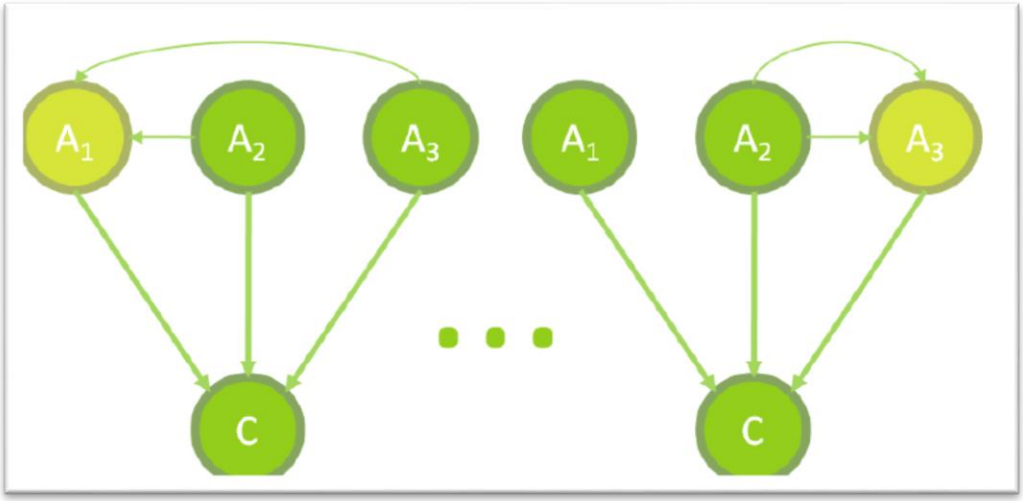
O Bayes Net utiliza diversos algoritmos de busca e medidas de qualidade fornecendo estruturas entre os dados (CHICKERING et al., 2004).

Segundo Sahami (2004, tradução nossa) o AODE tem outra abordagem para minimizar a dependência dos atributos. Estendendo a estrutura do Naive Bayes, inclui também a dependência de cada atributo com outro atributo. O AODE também constrói diversos modelos, mais precisamente **n** modelos, onde **n** é o

número de atributos. Em cada modelo, existe um atributo alvo **A** e as probabilidades consideradas são as condicionais desse alvo (Figura 13).

Segundo Sahami (1996, tradução nossa) este algoritmo é usado para classificação, e produz uma tabela tridimensional de frequência conjunta, indexada na primeira dimensão aos valores da classe e nas duas dimensões restantes aos valores dos atributos. Tal algoritmo permite a falta de valores de atributos, mas não permite a falta de valores de classe. Para aplicar o algoritmo permitindo a falta de valores de classe, é possível adicionar um pré-processamento que remova todos os objetos.

Figura 13 - Estrutura do AODE




Fonte: Webb, Boughton e Wang (2005).

Segundo Webb, Boughton e Wang (2005, tradução nossa) o AODE visa melhorar a relação entre o custo computacional e a precisão do Naive Bayes  mas um pequeno aumento pode enfraquecer a hipótese de independência, e com isso apresentar maior precisão.

### 7.3.3 Descrição dos experimentos

Considerando o estudo prévio que originou a base de dados de nossa pesquisa (SILVA, 2013), foram realizados 327 experimentos: com as variáveis não dicotomizadas e com as dicotomizadas.

Os experimentos com as variáveis dicotomizadas consideraram as variáveis apresentadas na Tabela 8 a seguir, que não foram transformadas pela especialista do domínio de aplicação de 2 categorias (apesar de algumas variáveis já possuírem *a priori* duas categorias originais).

Tabela 8 - Variáveis não dicotomizadas

| Variáveis | Descrição | Atributos - banco original | Atributo - banco pré - processado |
|---|---|---|---|
| **idade** | Idade dos indivíduos por faixa etária | - | - |
| **Altura** | Altura dos individuo por faixa etária | - | - |
| **Peso** | Media de peso por faixa etária | - | - |
| **Fratura** | Fratura prévia | 0<br>1 – sim<br>2 – não<br>98 | 1-Sim, 2-Não |
| **Qt_fratura** | Quantidades de fraturas prévias | - | - |
| **Fratura_ fêmur** | Fratura prévia em colo do fêmur | 0<br>1-Sim, 2-Não<br>98 | 1-Sim, 2-não |
| **Fratura_coluna** | Fratura prévia na coluna | 0<br>1-Sim, 2-Não<br>98 | 1-Sim, 2-não |
| **Fratura_antebraço** | Fratura prévia em antebraço | 0<br>1-Sim, 2-Não<br>98 | 1-Sim, 2-não |
| **Fratura_costela** | Fratura prévia em costela | 0<br>1-Sim, 2-Não<br>98 | 1-Sim, 2-não |
| **tratamento** | Uso de medicamentos | 1-Sim, 2-Não | 1-Sim, 2-não |
| **menopausa** | Menopausa | 0<br>1-Sim, 2-Não<br>5<br>98 | 1-Sim, 2-não |
| **Uso_calcio** | Uso de Cálcio | -1<br>1-Sim, 2-Não<br>7<br>98 | 1-Sim, 2-não |
| **Trh** | Terapia de reposição hormonal | 1-Sim, 2-Não<br>88<br>98 | 1-Sim, 2-não |
| **Medic_tireoide** | Uso de medicamentos para tireóide | -1<br>1-Sim, 2-Não<br>98<br>99-não se aplica | 1-Sim, 2-não |
| **histerectomia** | Histerectomia | 1-Sim, 2-Não<br>93<br>98-não respondido | 1-Sim, 2-não |
| **ooforectomia** | Ooforectomia | 1-Sim, 2-Não<br>98 | 1-Sim, 2-não |
| **Desfecho** | Diagnostico | 1-normal<br>2-osteopenia<br>3-osteoporose<br>4-normal para idade cronológica | 1-normal<br>2-osteopenia<br>3-osteoporose<br>4-normal<br>5- |

| | | | |
|---|---|---|---|
| | | 5-densidade abaixo do esperado para idade cronológica | |
| **PrimaryLast** | Indicador de cada último caso correspondente como principal | 0 - caso duplicado<br>1 - caso principal | 0 - caso duplicado<br>1 - caso principal |
| **IMC_obesidade** | IMC normal sobrepeso e 3 graus de obesidade | 1 - até 25<br>2 - 26-29,9<br>3 - 30-34,9<br>4 - 35-39,9<br>5 - 40 ou mais | 1 - até 25<br>2 - 26-29,9<br>3 - 30-34,9<br>4 - 35-39,9<br>5 - 40 ou mais |

Fonte: Dados da pesquisa.

Os experimentos realizados com as variáveis dicotomizadas apresentadas na Tabela 9 a seguir, consideraram as variáveis transformadas pela especialista do domínio de aplicação em 2 categorias, quando necessário (apesar de algumas variáveis já possuírem *apriori* duas categorias originais). Tal transformação se deu originalmente no estudo de Silva (2013) pois, além da análise descritiva das variáveis do banco de dados, a pesquisadora realizou também uma análise multivariada por meio da regressão logística, com o objetivo de revelar as variáveis mais fortemente associadas à osteopenia e osteoporose.

Tabela 9 - Variáveis dicotomizadas utilizadas nos experimentos

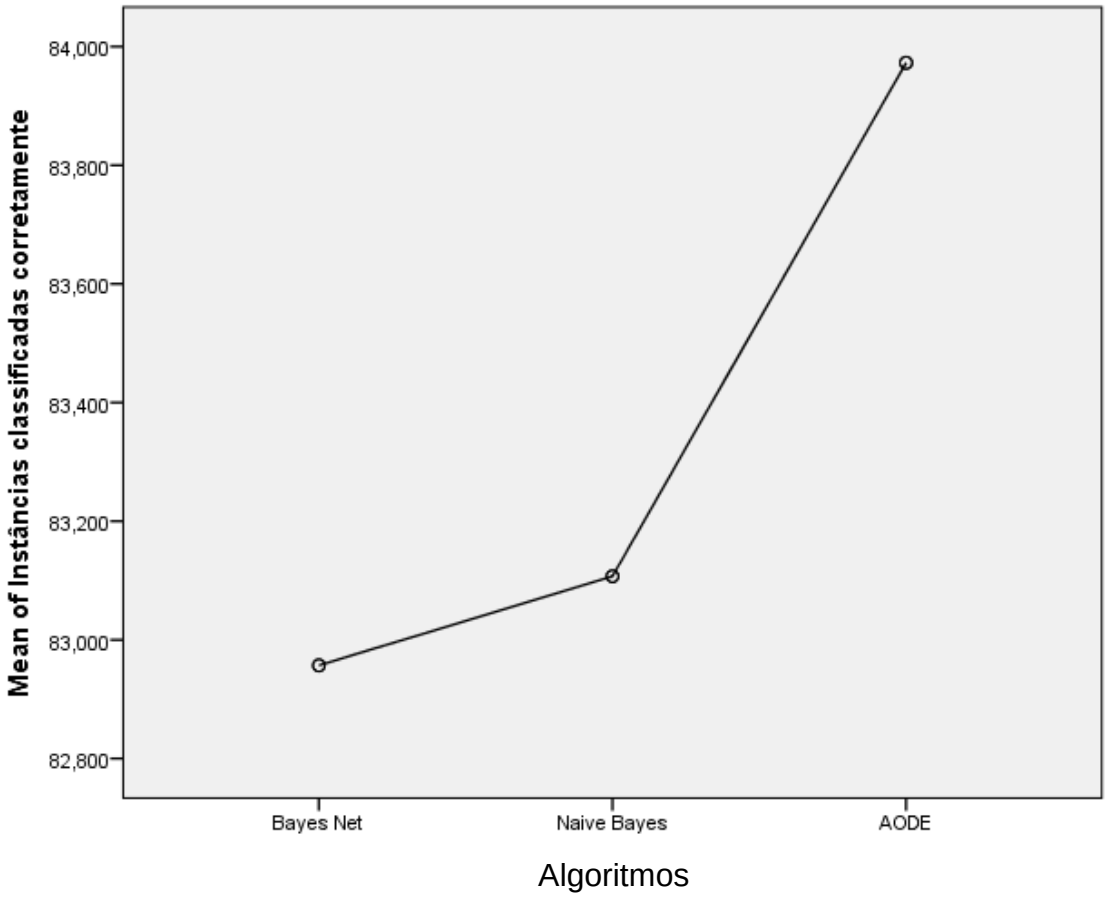
| **Variáveis** | **Descrição** | **Atributos - banco original** | **Atributo - banco pré - processado** |
|---|---|---|---|
| Fratura_1 | Fratura dicotômica | 1-Sim,2-não | 1-Sim, 2-não |
| Fratura_femur_1 | FEMUR_ dicotômica | 1-Sim,2-não | 1-Sim, 2-não |
| Fratura_coluna_1 | Coluna_dicotomica | 1-Sim,2-não | 1-Sim, 2-não |
| Fratura_antebraço_1 | Antebraço dicotômica | 1-Sim,2-não | 1-Sim, 2-não |
| Fratura_costela_1 | Costela dicotômica | 1-Sim,2-não | 1-Sim, 2-não |
| Menopausa_1 | Menopausa dicotômica | 1-Sim,2-não | 1-Sim, 2-não |
| Uso_calcio_1 | Cálcio dicotômico | 1-Sim,2-não | 1-Sim, 2-não |
| Trh_1 | TRH dicotômico | 1-Sim,2-não | 1-Sim, 2-não |
| Ooforectomia_1 | Ooforectomia dicotômica | 1-Sim,2-não | 1-Sim, 2-não |
| Histerectomia_1 | Histerectomia dicotômica | 1-Sim,2-não | 1-Sim, 2-não |
| Idade_1 | Idade_dicotomica_media | 1- 50<br>2 - até 49 | 1- cinquenta<br>2 – ate_quarenta_nove |
| IMC_dict | IMC dicotomizado até 25 e 25 ou mais | 1- 26 ou mais<br>2- até 25 | 1-vinte_seis_mais<br>2- ate_vinte_cinco |
| Medic_tireode_1 | Medicamento para tireoide | 1-Sim,2-não | 1-Sim, 2-não |
| Desfecho_2 | Desfecho dicotômico | 1-osteopenia e osteoporose<br>2-normal | 1 – osteopenia_osteoporose<br>2- normal |

Fonte: Dados da pesquisa.

## 7.4 RESULTADOS OBTIDOS

Nesta etapa foram realizados aproximadamente 327 experimentos que foram selecionados para avaliação. Na avaliação utilizou-se também a ferramenta Weka sendo consideradas as instâncias classificadas correta e incorretamente, o coeficiente kappa, o erro absoluto médio, a sensibilidade, a especificidade, a área da curva ROC, e a acurácia.

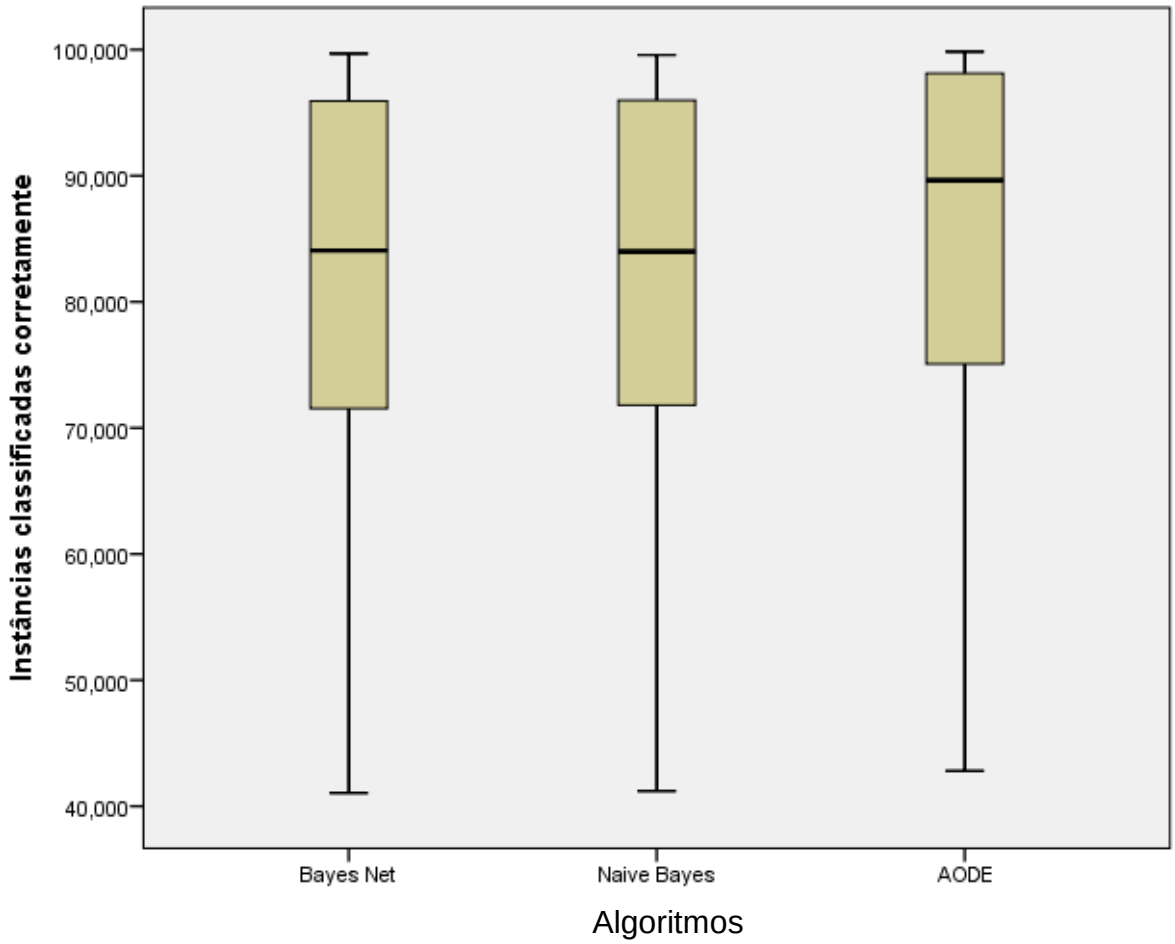
Figura 14 - Média de instâncias classificadas corretamente




Fonte: Dados da pesquisa

Conforme ilustra a Figura 14 e a Figura 15, a média de instâncias classificadas corretamente foi de 82,54 (±14,76), no algoritmo Bayes Net; 82,84 (±14,42) no Naive Bayes; e 84,52 (±14,08) no AODE.

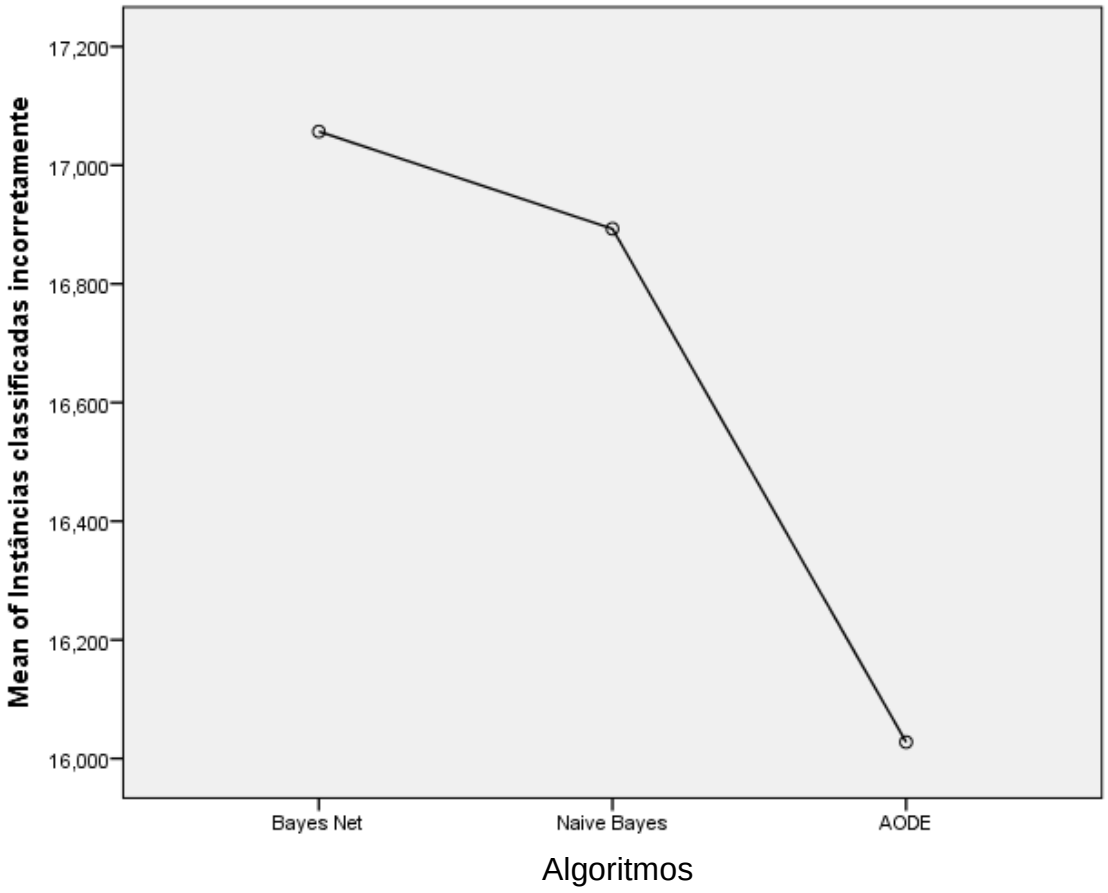
Figura 15 - Box plot das instâncias classificadas corretamente




Fonte: Dados da pesquisa

A Figura 15 apresenta a mediana de Instâncias Classificadas Corretamente, que foi de 84,06 (±26,76), variando de 41,04 a 99,67 no algoritmo Bayes Net. No algoritmo Naive Bayes a mediana foi de 83,99 (±26,61) com variação de 41,20 a 99,57. No AODE a mediana foi de 89,65 (±23,08), oscilando de 42,81 a 99,84.

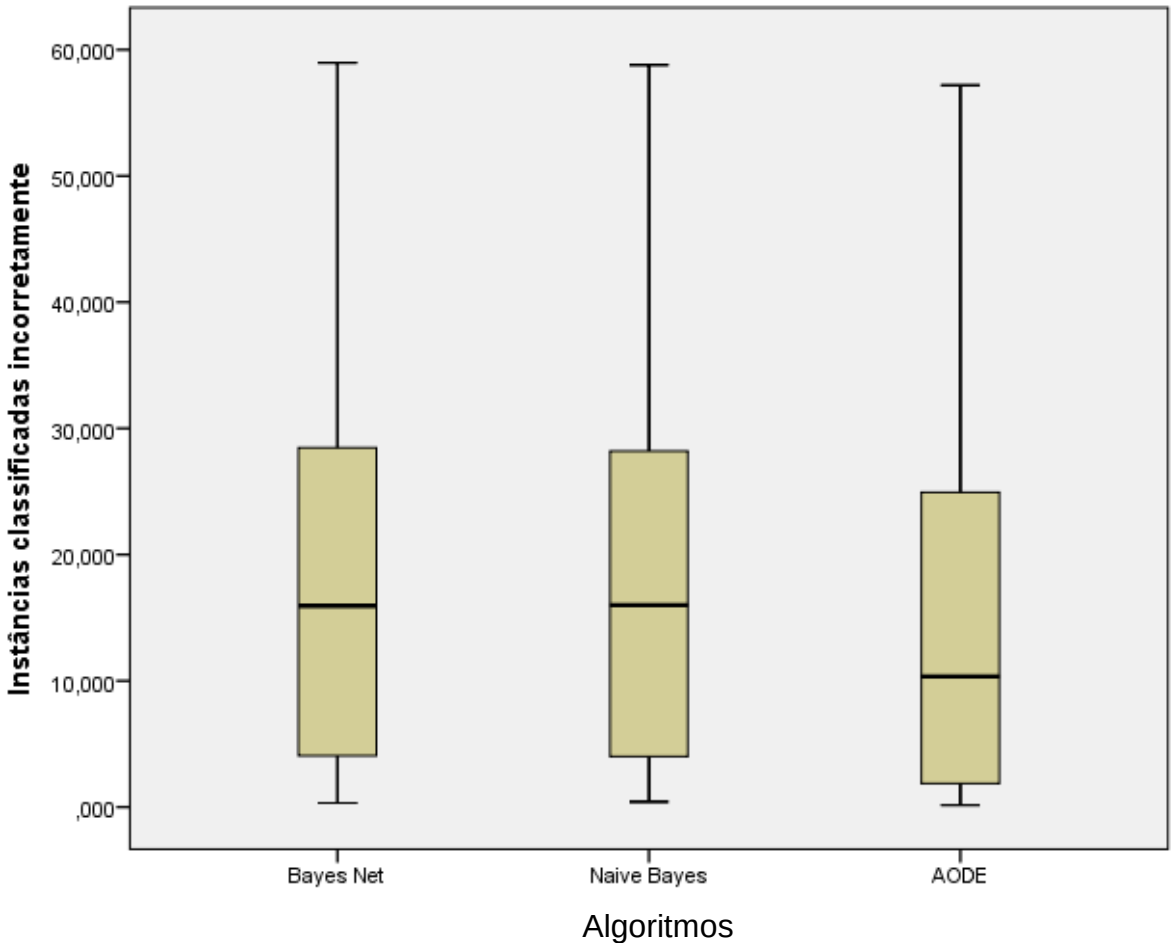
Figura 16 - Média de instâncias classificadas incorretamente




Fonte: Dados da pesquisa

Conforme ilustra a Figura 16 e a Figura 17, a média de instâncias classificadas incorretamente foi de 17,46 (±14,76), no algoritmo Bayes Net; 17,15 (±14,42) no Naive Bayes; e 15,48 (±14,08) no AODE.

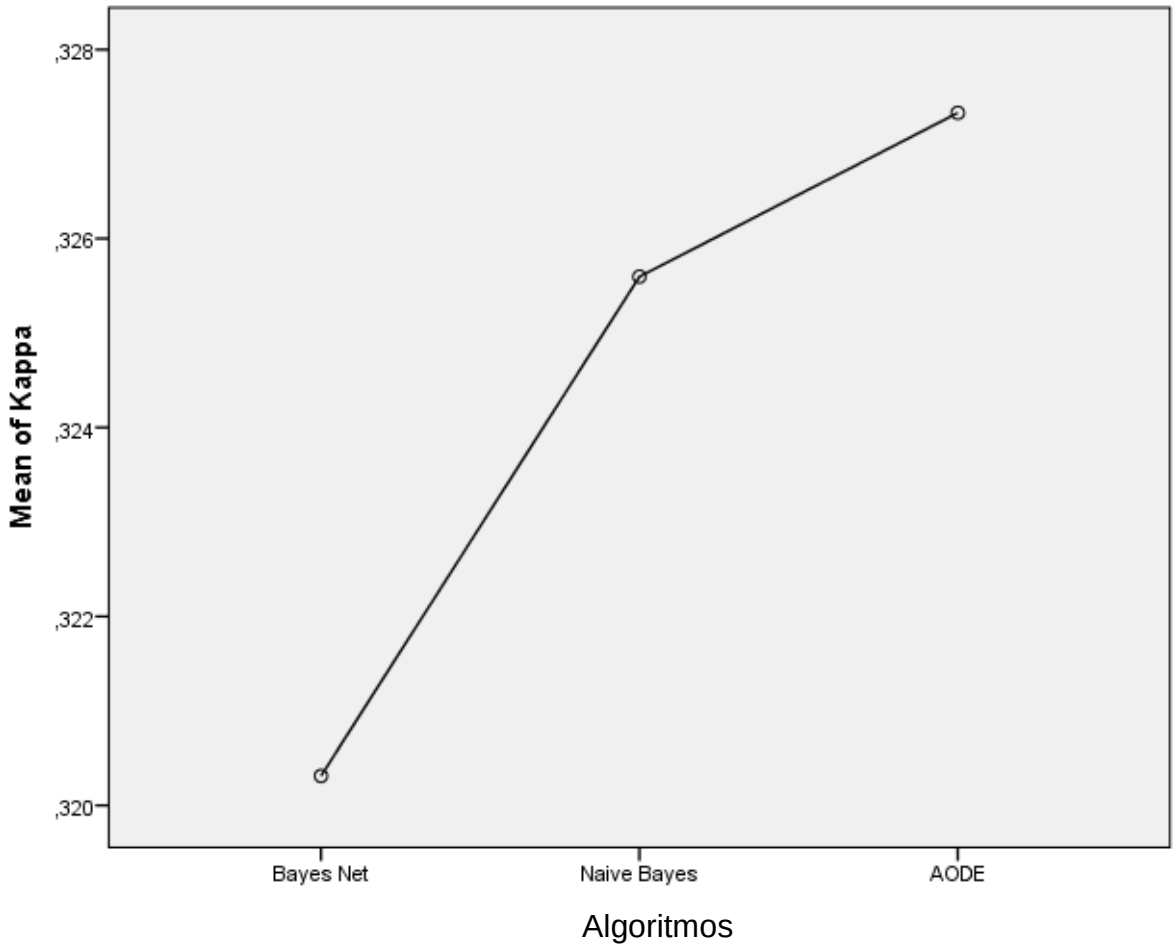
Figura 17 - Box plot das instâncias classificadas incorretamente




Fonte: Dados da pesquisa

A Figura 17 apresenta mediana de instâncias classificadas incorretamente, que foi de 15,93 (±26,66), variando de 0,32, a 26,66 no algoritmo Bayes Net. No algoritmo Naive Bayes a mediana foi de 16,00 (±26,61) com variação de 0,43 a 58,79. No AODE a mediana foi de 10,35 (±23,08), oscilando de 0,16 a 57,19.

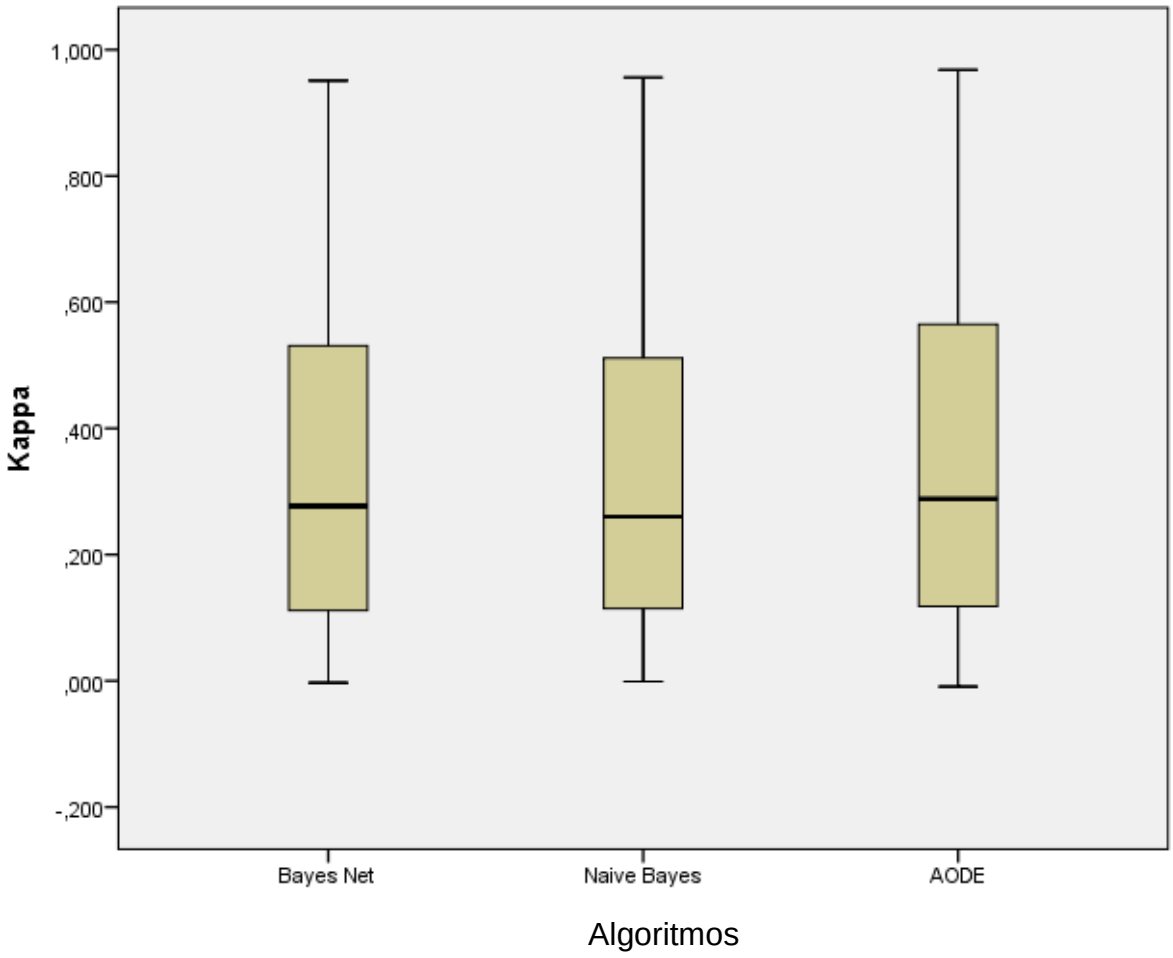
Figura 18 - Média de estatística Kappa




Fonte: Dados da pesquisa

Conforme ilustra a Figura 18 e a Figura 19, a média da estatística Kappa foi de 0,33 (±0,29), no algoritmo Bayes Net; 0,31 (±0,28) no Naive Bayes; e 0,34 (±0,29) no AODE.

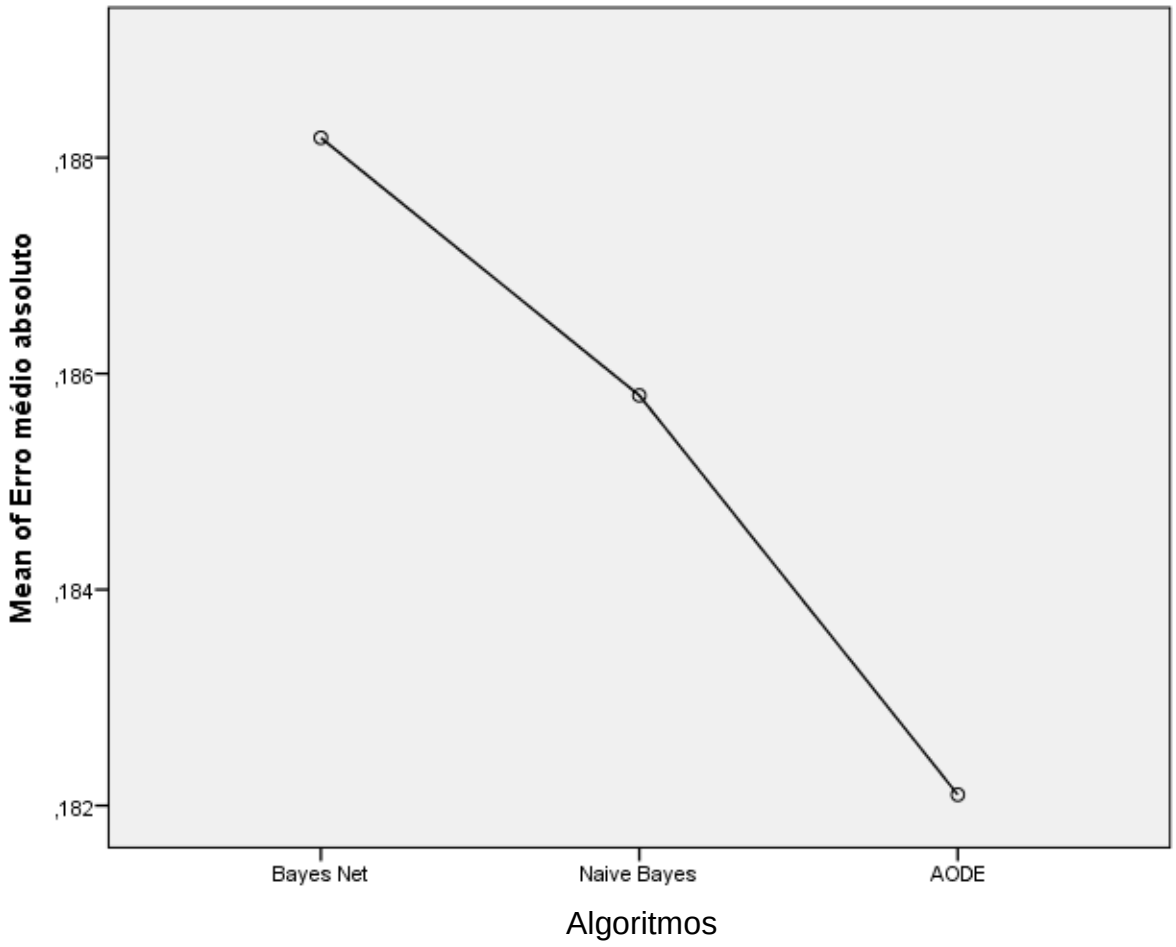
Figura 19 - Box plot da estatística Kappa




Fonte: Dados da Pesquisa

A Figura 19 apresenta a mediana da estatística Kappa, que foi de 0,28 (±0,43), variando de 0,00, a 0,95 no algoritmo Bayes Net. No algoritmo Naive Bayes a mediana foi de 0,26 (±0,40) com variação de 0,00 a 0,96. No AODE a mediana foi de 0,29 (±0,45), oscilando de -0,01 a 0,97.

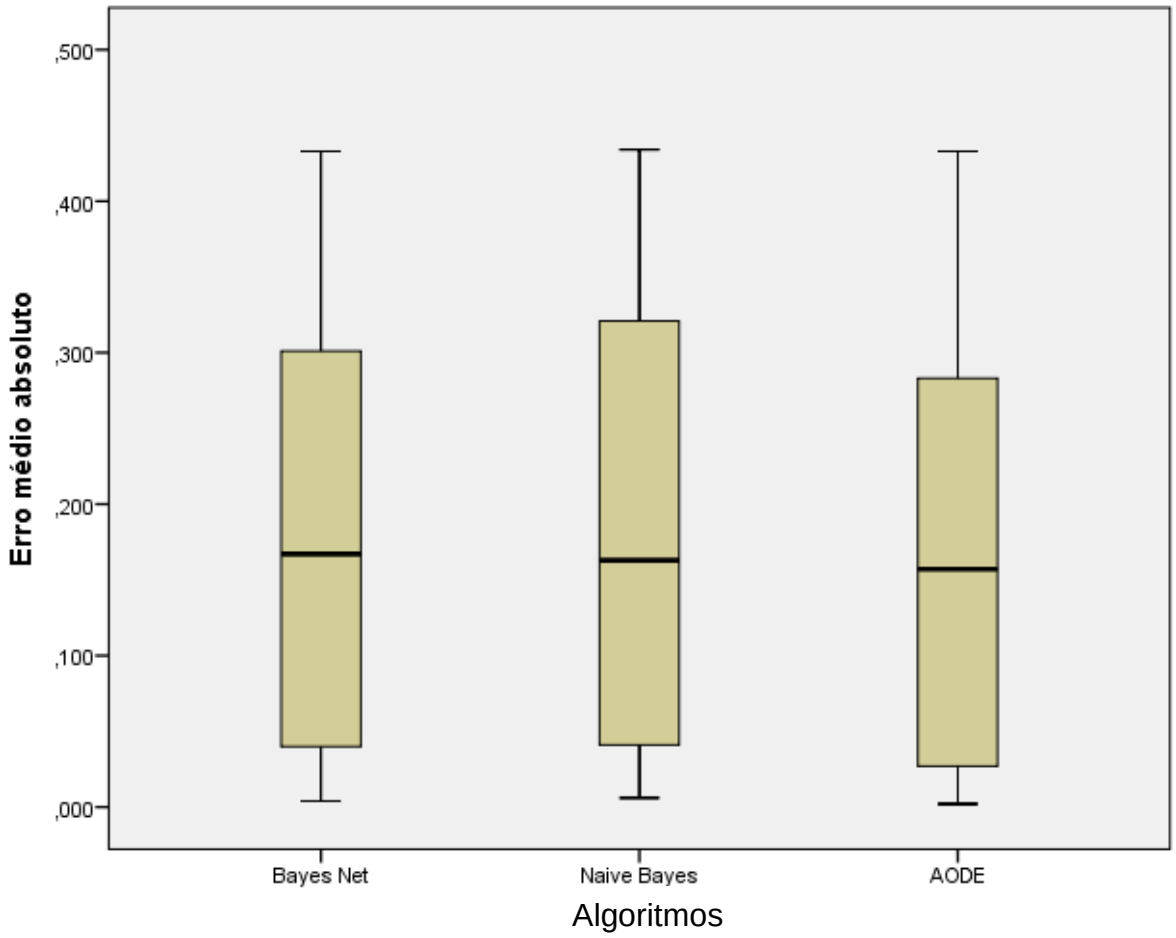
Figura 20 – Média de erro médio absoluto




Fonte: Dados da pesquisa

Conforme ilustra a Figura 20 e a Figura 21, a média de erro médio absoluto foi de 0,19 (±0,14), no algoritmo Bayes Net; 0,19 (±0,14) no Naive Bayes; e 0,18 (±0,14) no AODE.

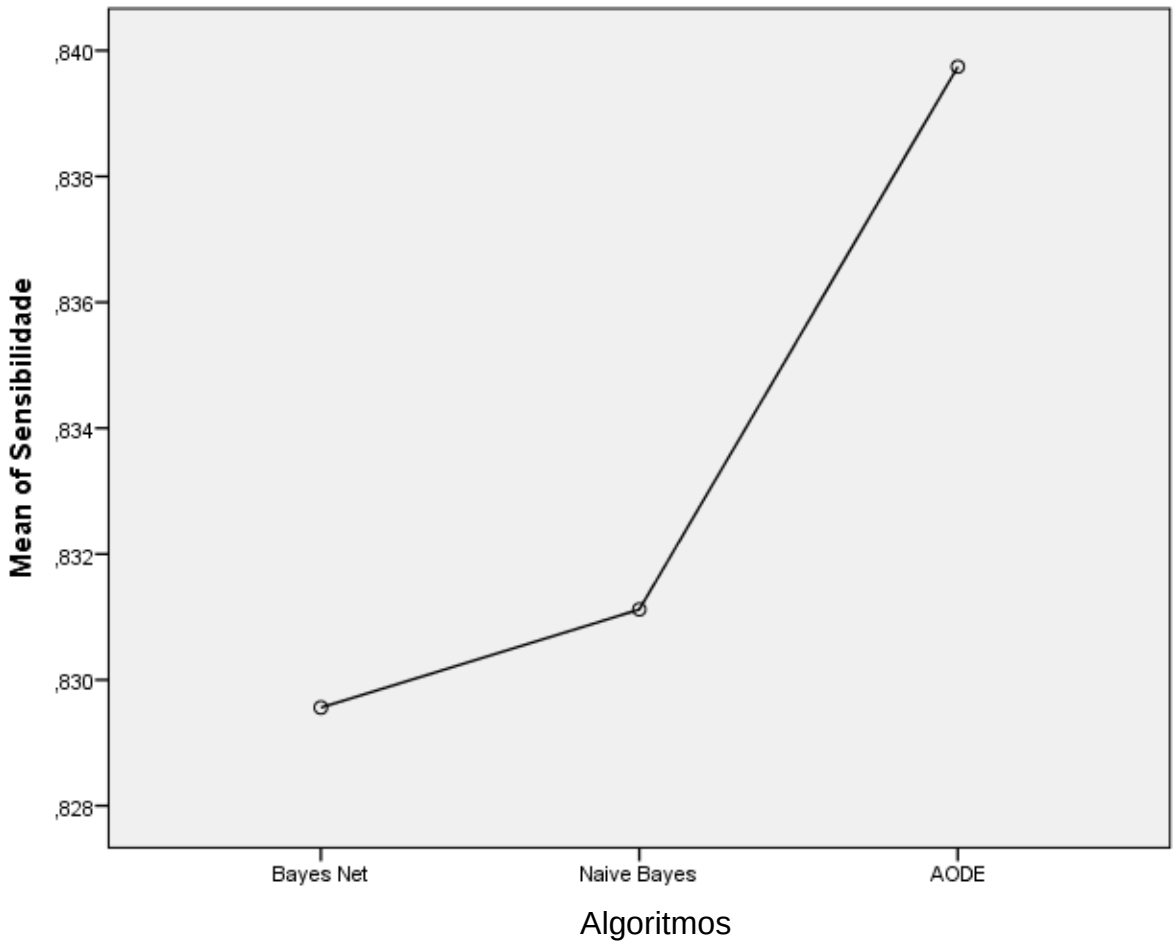
Figura 21 - Box plot do erro médio Absoluto




Fonte: Dados da pesquisa

A Figura 21 apresenta a mediana do erro médio absoluto, que foi de 0,17 (±0,27), variando de 0,00 a 0,43 no algoritmo Bayes Net. No algoritmo Naive Bayes a mediana foi de 0,16 (±0,29) com variação de 0,01 a 0,43. No AODE a mediana foi de 0,16 (±0,26), oscilando de 0,00 a 0,43.

Figura 22- Média de sensibilidade

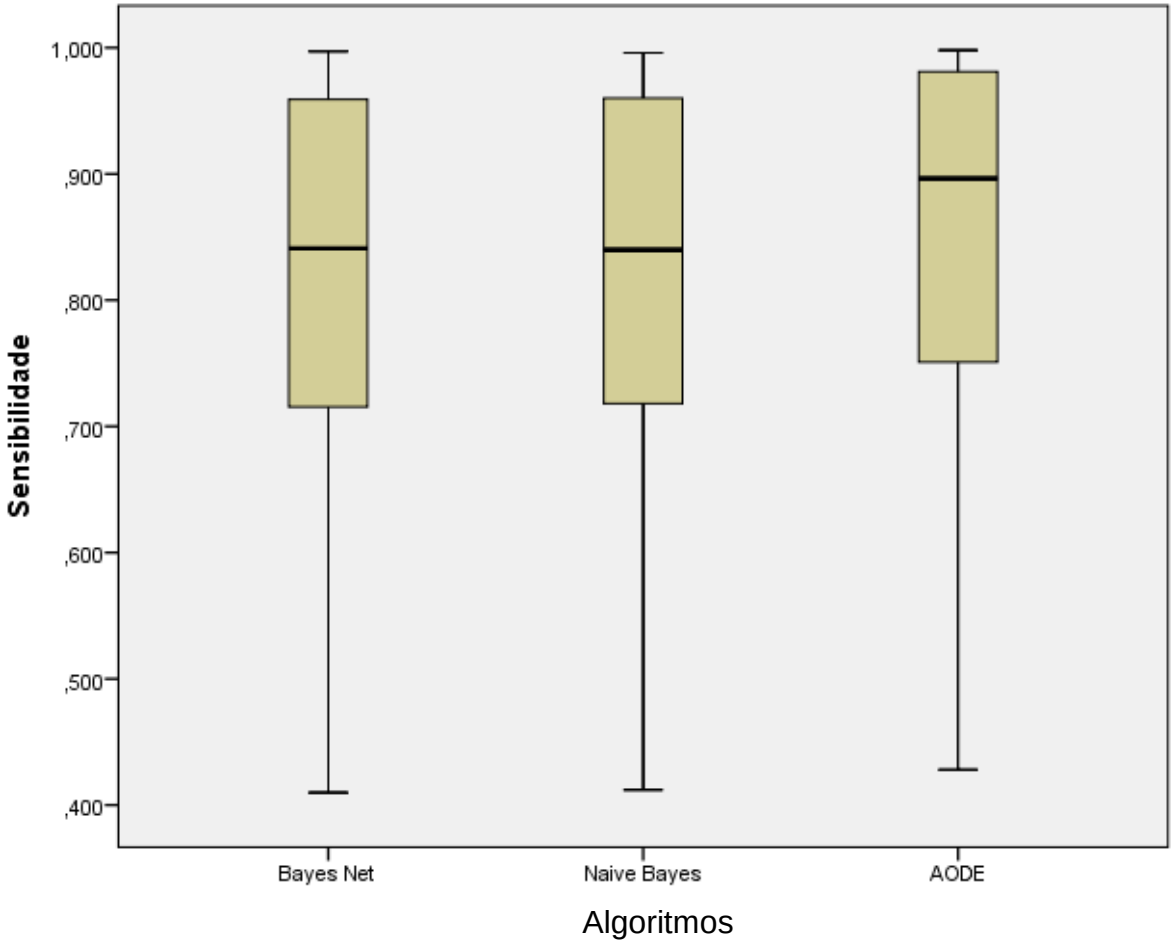



Fonte: Dados da pesquisa

Conforme ilustram a Figura 22 e a Figura 23, a média de sensibilidade foi de 0,82 (±0,15), no algoritmo Bayes Net; 0,83 (±0,14) no Naive Bayes; e 0,84 (±0,14) no AODE.

.

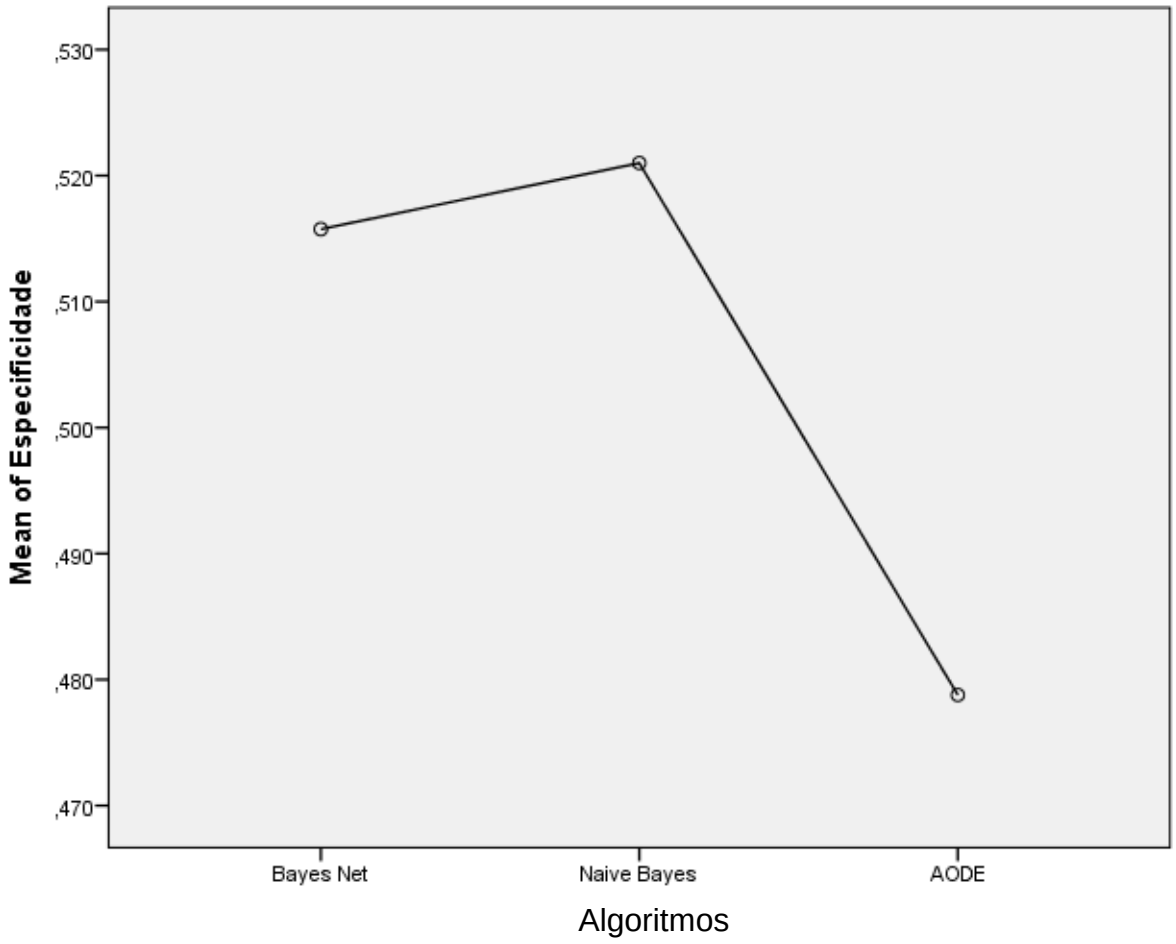
Figura 23 - Box plot da sensibilidade




Fonte: Dados da pesquisa

A Figura 23 apresenta a mediana de sensibilidade, que foi de 0,84 (±0,27), variando de 0,41 a 0,80 no algoritmo Bayes Net. No algoritmo Naive Bayes a mediana foi de 0,84 (±0,27). No AODE a mediana foi de 0,90 (±0,23,), oscilando de 0,43 a 0,99.

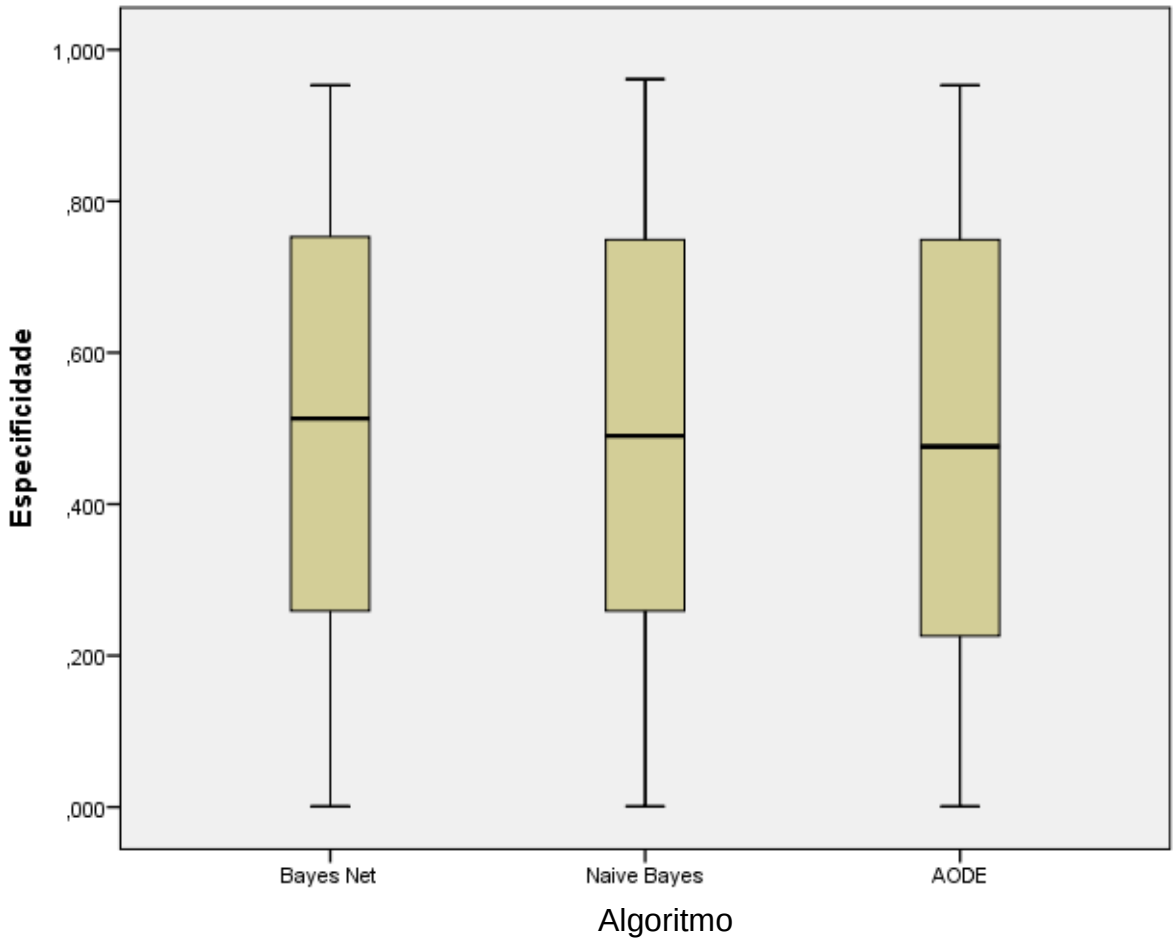
Figura 24 - Média de especificidade




Fonte: Dados da pesquisa.

Conforme ilustra a Figura 24 e a Figura 25, a média de especificidade foi de 0,53(±0,27), no algoritmo Bayes Net; 0,51 (±0,27) no Naive Bayes; e 0,49 (±0,29) no AODE.

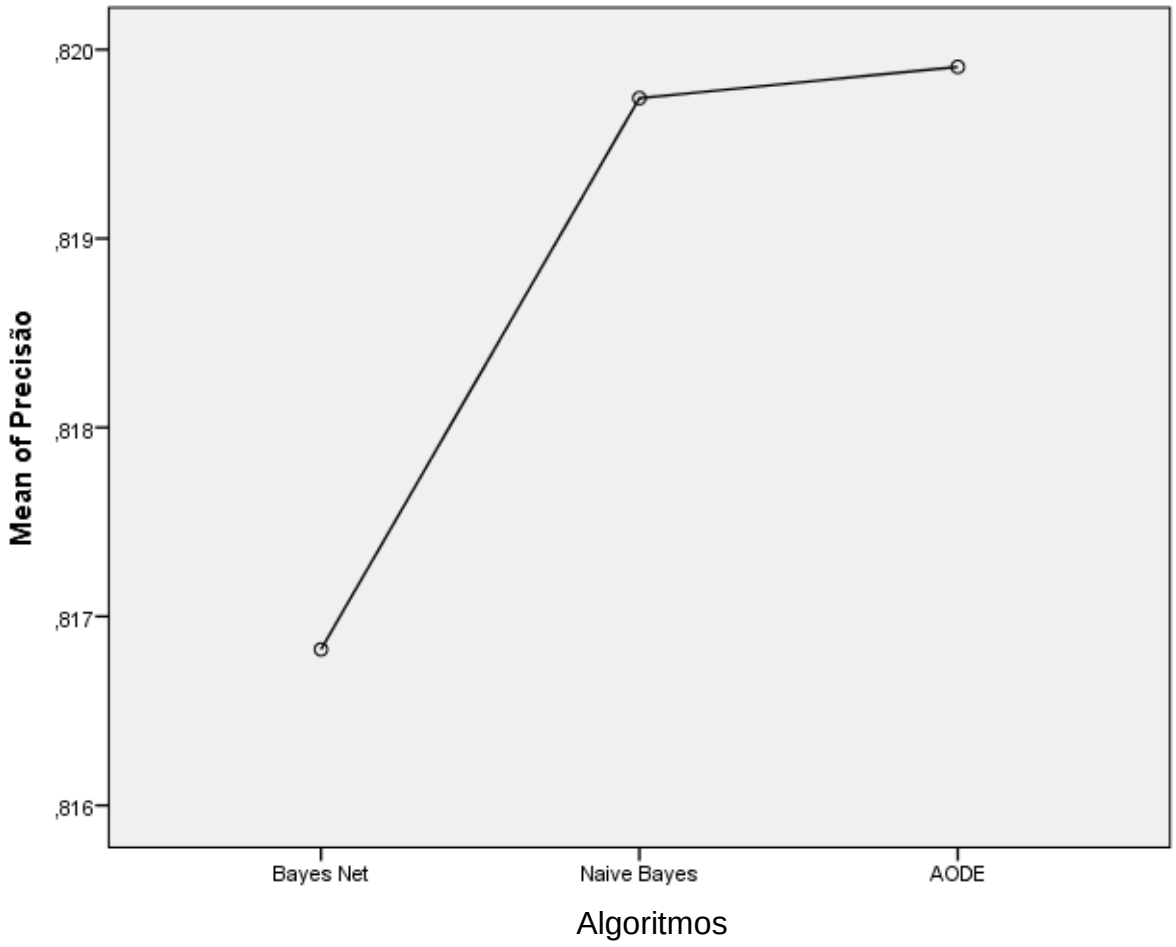
Figura 25 - Box plot da Especificidade




Fonte: Dados da pesquisa

A Figura 25 apresenta a mediana de especificidade, que foi de 0,51 (±0,49), variando de 0,00, a 0,95 no algoritmo Bayes Net. No algoritmo Naive Bayes a mediana foi de 0,49 (±0,49) com variação de 0,00 a 0,96. No AODE a mediana foi de 0,48(±0,52), oscilando de 0,00 a 0,95.

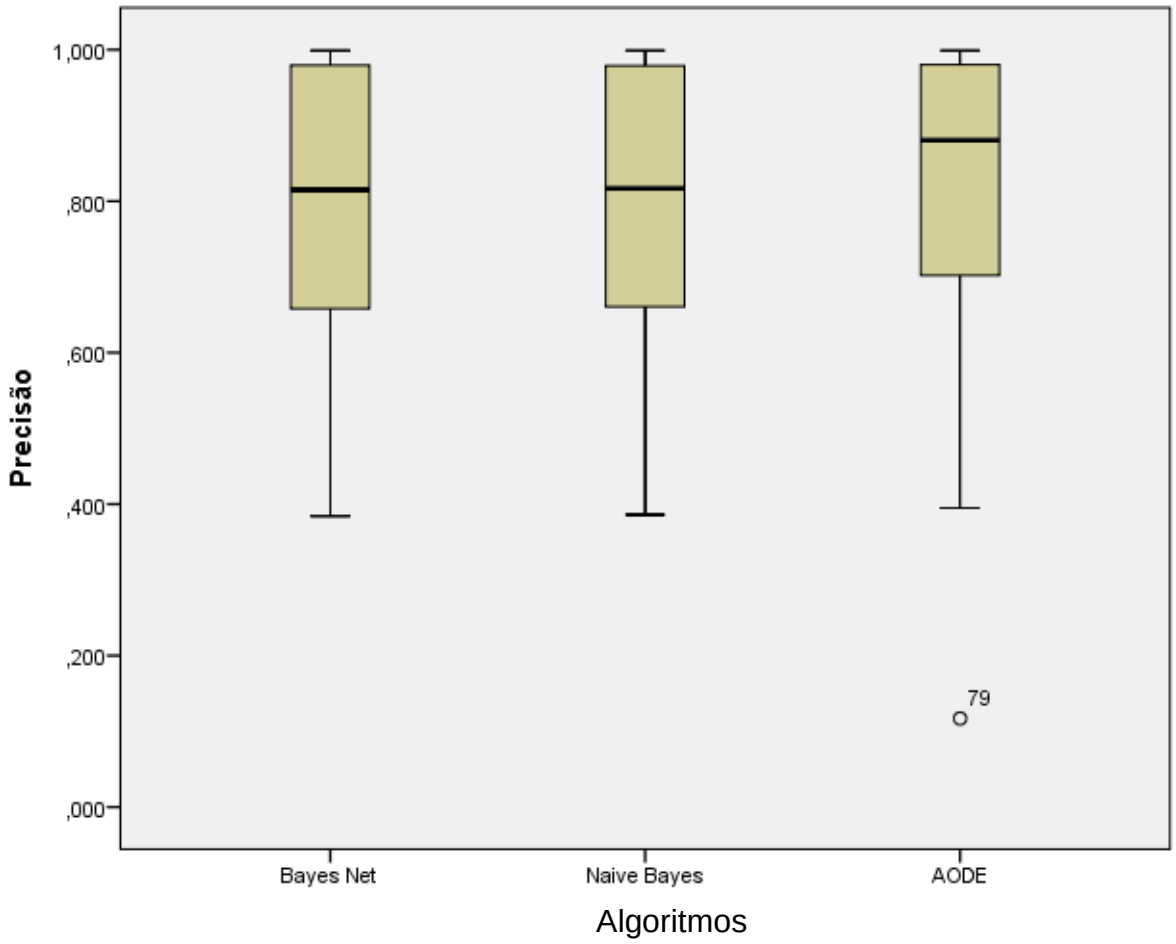
Figura 26 - Média da Precisão




Fonte: Dados da pesquisa

Conforme ilustra a Figura 26 e a Figura 27, a média de precisão foi de 0,81 (±0,16), no algoritmo Bayes Net; 0,82 (±0,16) no Naive Bayes; e 0,83 (±0,17) no AODE.

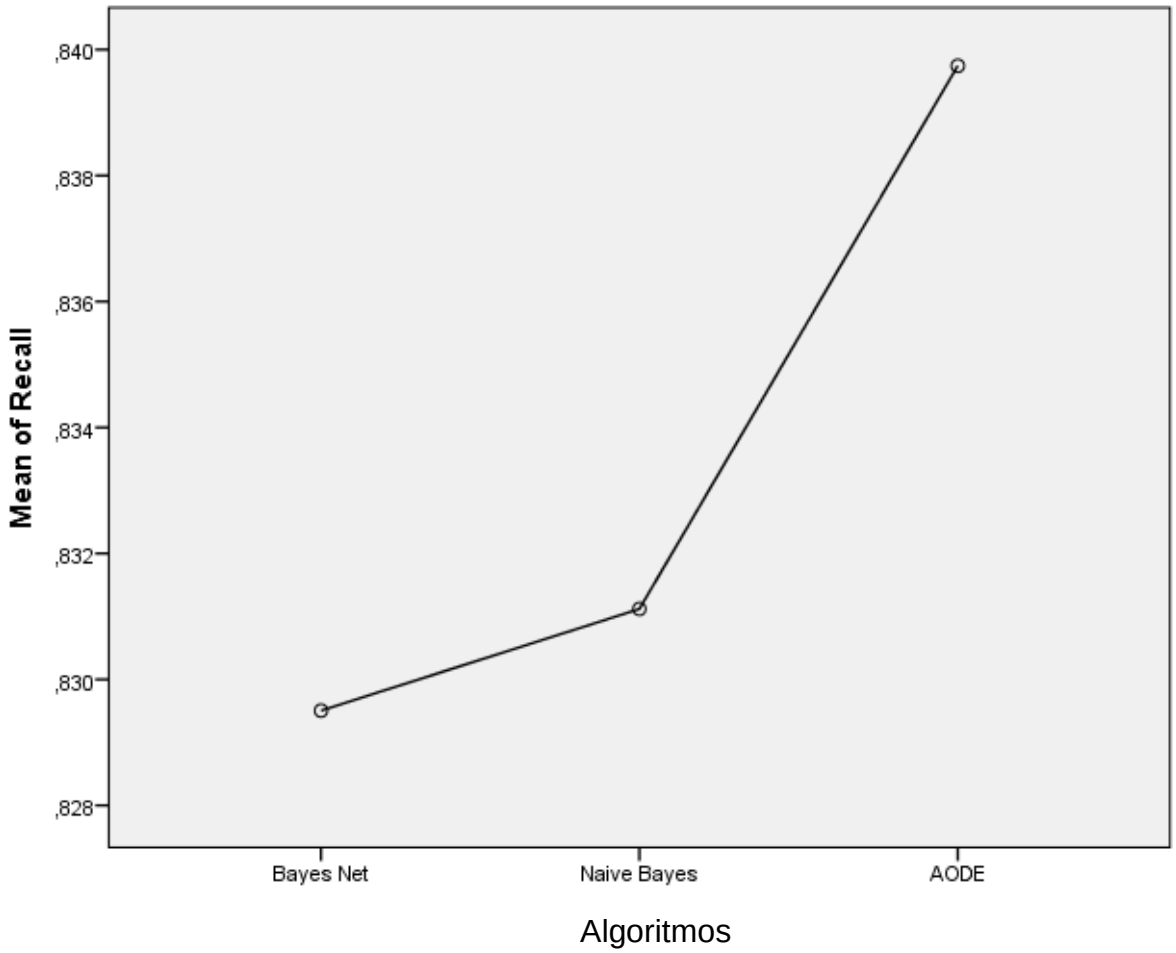
Figura 27 - Box plot da Precisão




Fonte: Dados da pesquisa

A Figura 27 apresenta a mediana de Precisão, que foi de 0,81 (±0,32), variando de 0,38, a 0,99 no algoritmo Bayes Net. No algoritmo Naive Bayes a mediana foi de 0,82 (±0,31) com variação de 0,37 a 0,99. No AODE a mediana foi de 0,88 (±0,28), oscilando de 0,12 a 0,99.

Figura 28 - Média do Recall

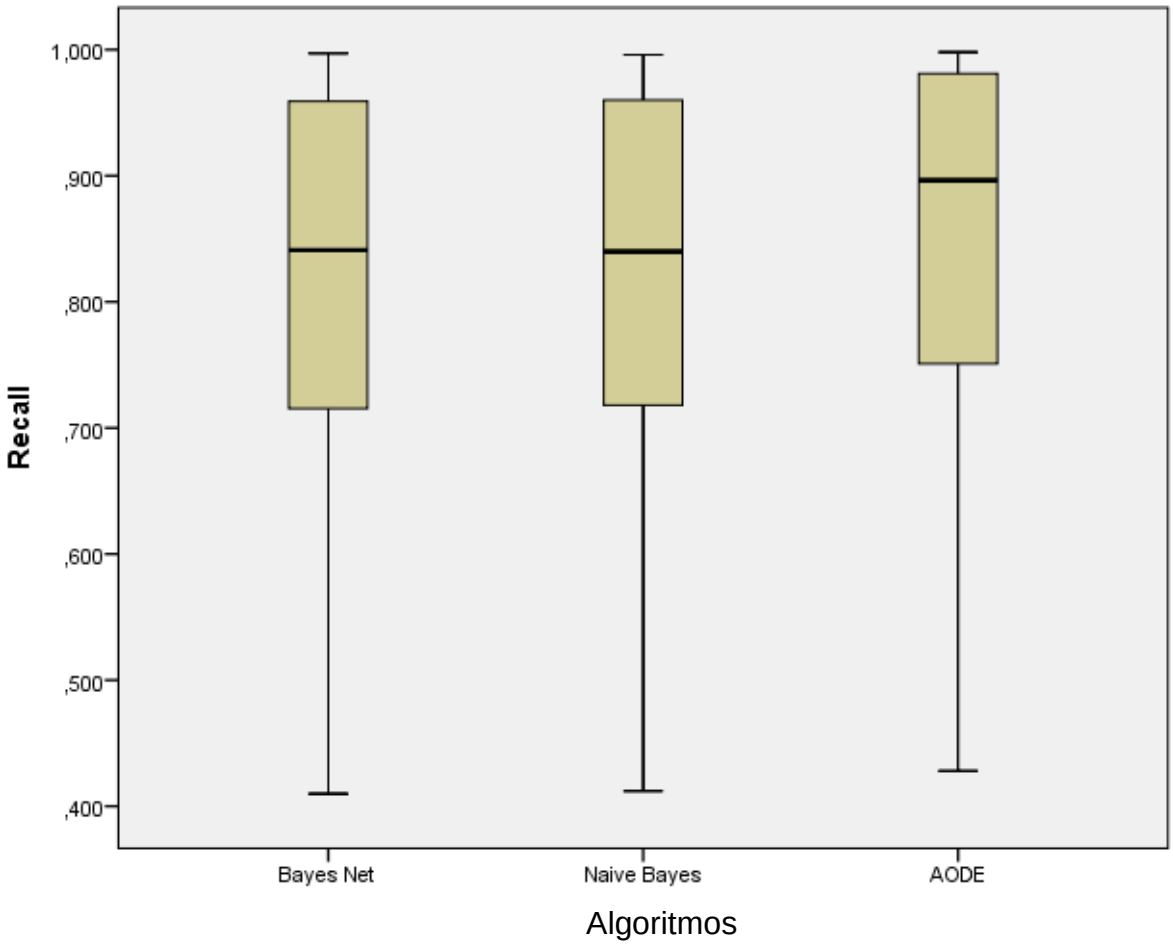



Fonte: Dados da pesquisa

Conforme ilustra a Figura 28 e a Figura 29, a média de recall foi de 0,82 (±0,15), no algoritmo Bayes Net; 0,83 (±0,14) no Naive Bayes; e 0,84 (±0,14) no AODE.

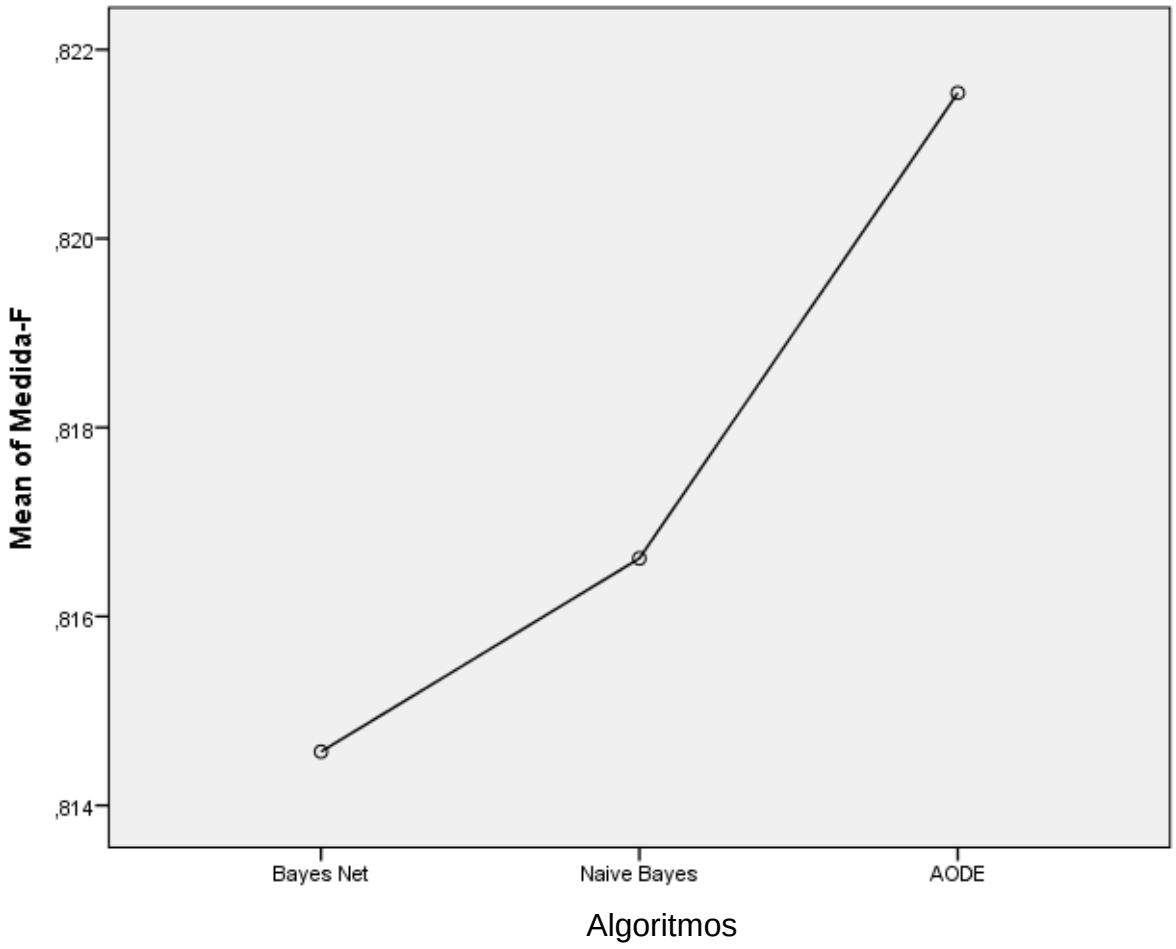
Figura 29 - Box plot do recall

Fonte: Dados da pesquisa




A Figura 29 apresenta a mediana de recall, que foi de 0,84(±0,27), variando de 0,41, a 0,99 no algoritmo Bayes Net. No algoritmo Naive Bayes a mediana foi de 0,84(±0,27) com variação de 0,41 a 0,99. No AODE a mediana foi de 0,90(±0,23), oscilando de 0,42 a 0,99,

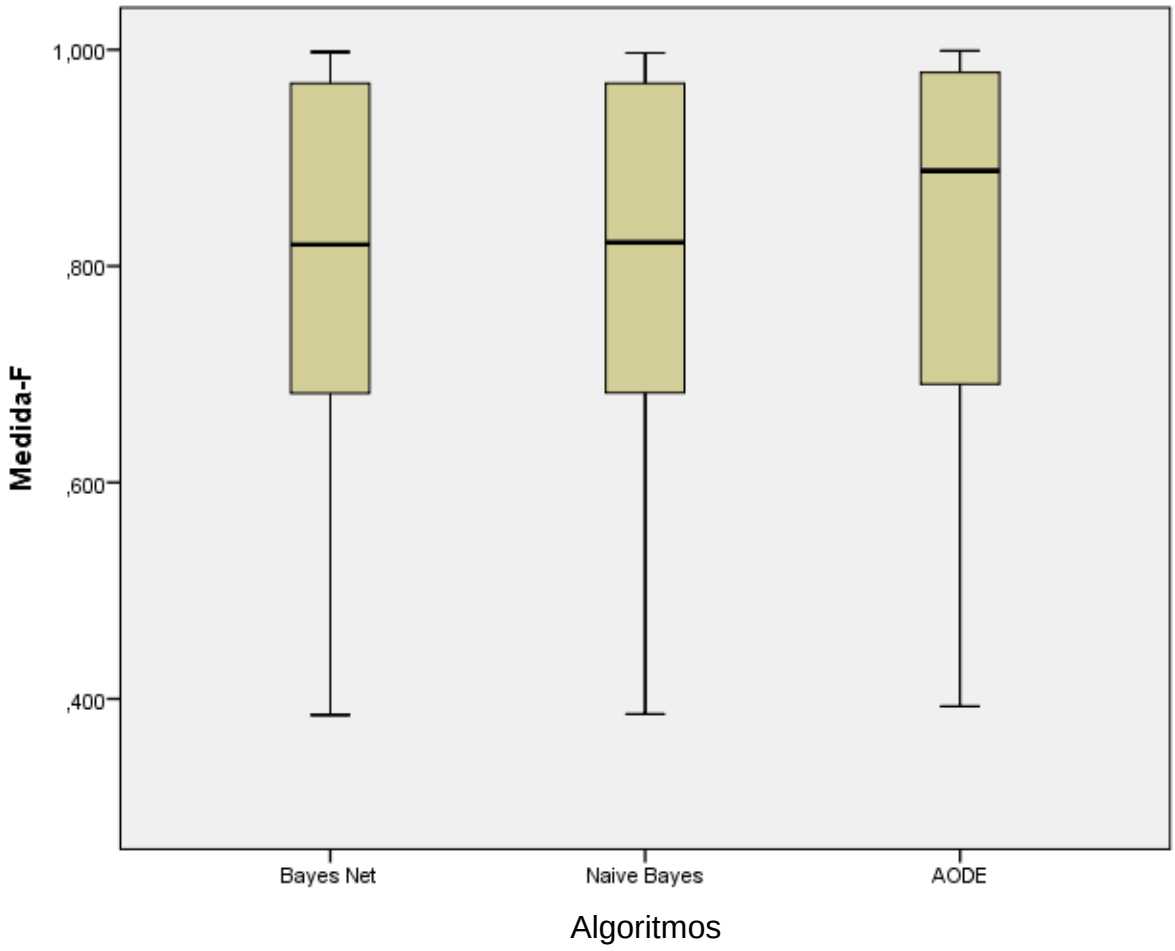
Figura 30 - Média da medida-F




Fonte: Dados da pesquisa

Conforme ilustra a Figura 30 e a Figura 31, a média da medida-F foi de 0,81 (±0,16), no algoritmo Bayes Net; 0,81 (±0,16) no Naive Bayes; e 0,83 (±0,15) no AODE.

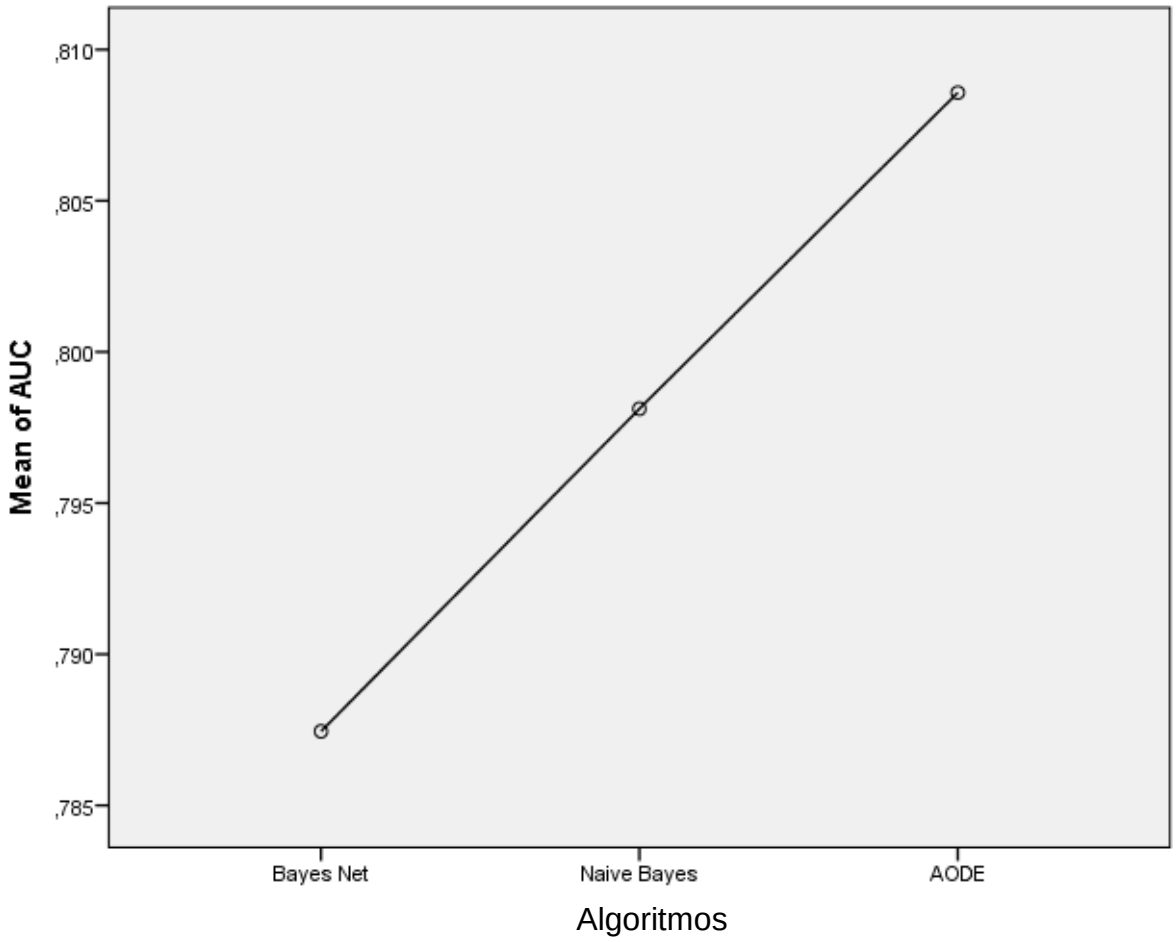
Figura 31 - Blox plot Medida-F




Fonte: Dados da pesquisa

A Figura 31 apresenta a mediana de medida_F, que foi de 0,82 (±0,29), variando de 0,38, a 0,99 no algoritmo Bayes Net. No algoritmo Naive Bayes a mediana foi de 0,82 (±0,27) com variação de 0,37 a 0,99. No AODE a mediana foi de 0,89 (±0,29), oscilando de 0,39 a 0,99.

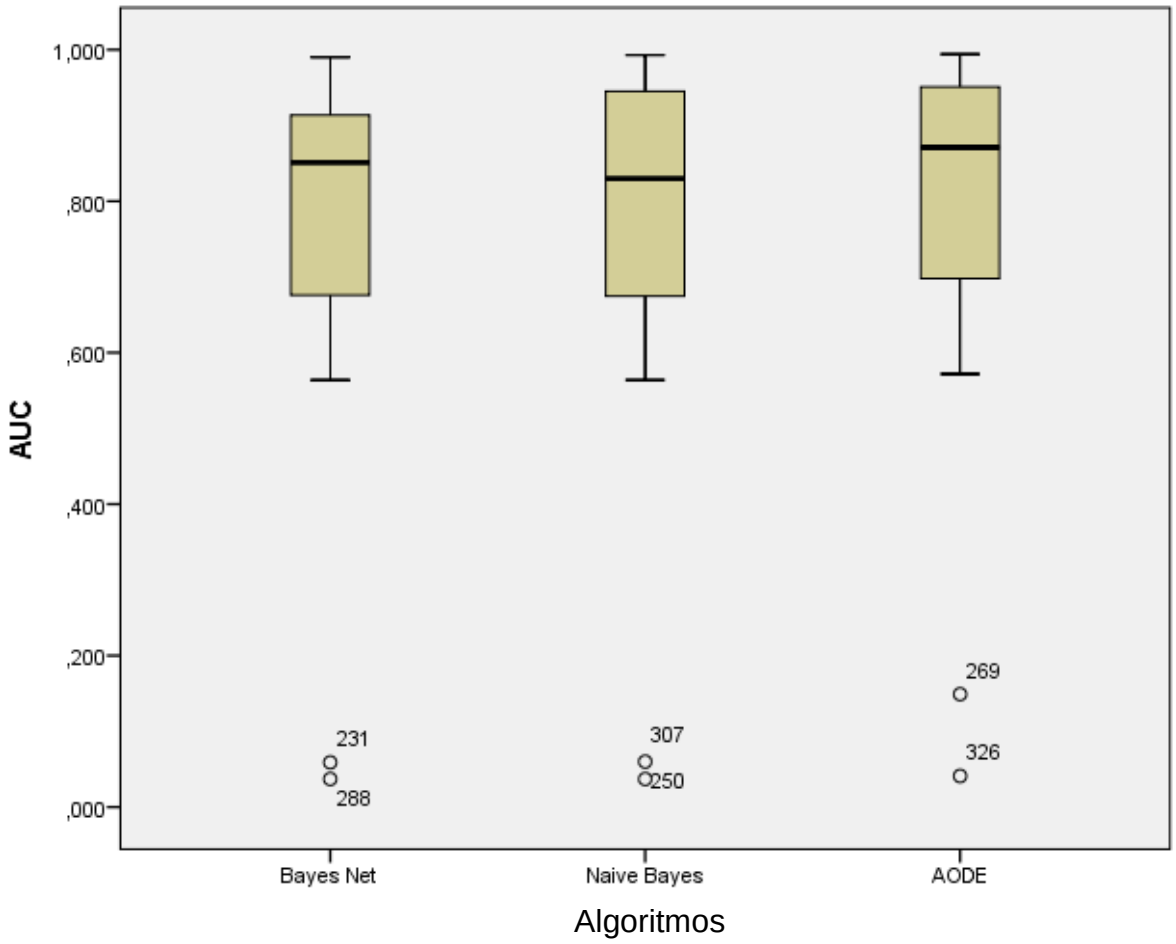
Figura 32 - Média da AUC




Fonte: Dados da pesquisa

Conforme ilustra a Figura 32 e a Figura 33, a média da AUC foi de 0,79 (±0,17), no algoritmo Bayes Net; 0,79 (±0,17) no Naive Bayes; e 0,81 (±0,16) no AODE.

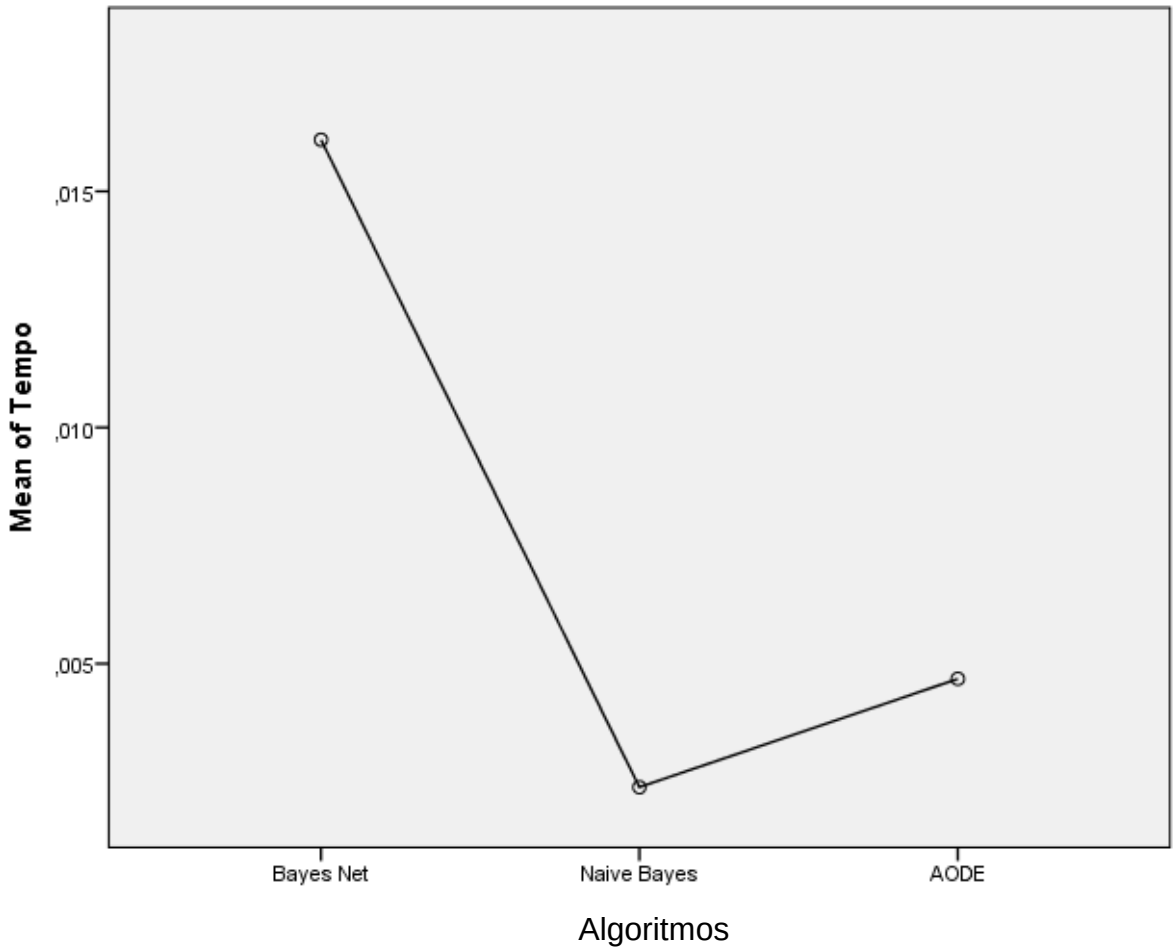
Figura 33 - Blox plot AUC




Fonte: Dados da pesquisa

A Figura 33 apresenta mediana de AUC, que foi de 0,85 (±0,24), variando de 0,04 a 0,99 no algoritmo Bayes Net. No algoritmo Naive Bayes a mediana foi de 0,83 (±0,27) com variação de 0,04 a 0,99. No AODE a mediana foi de 0,87 (±0,26), oscilando de 0,04 a 0,99.

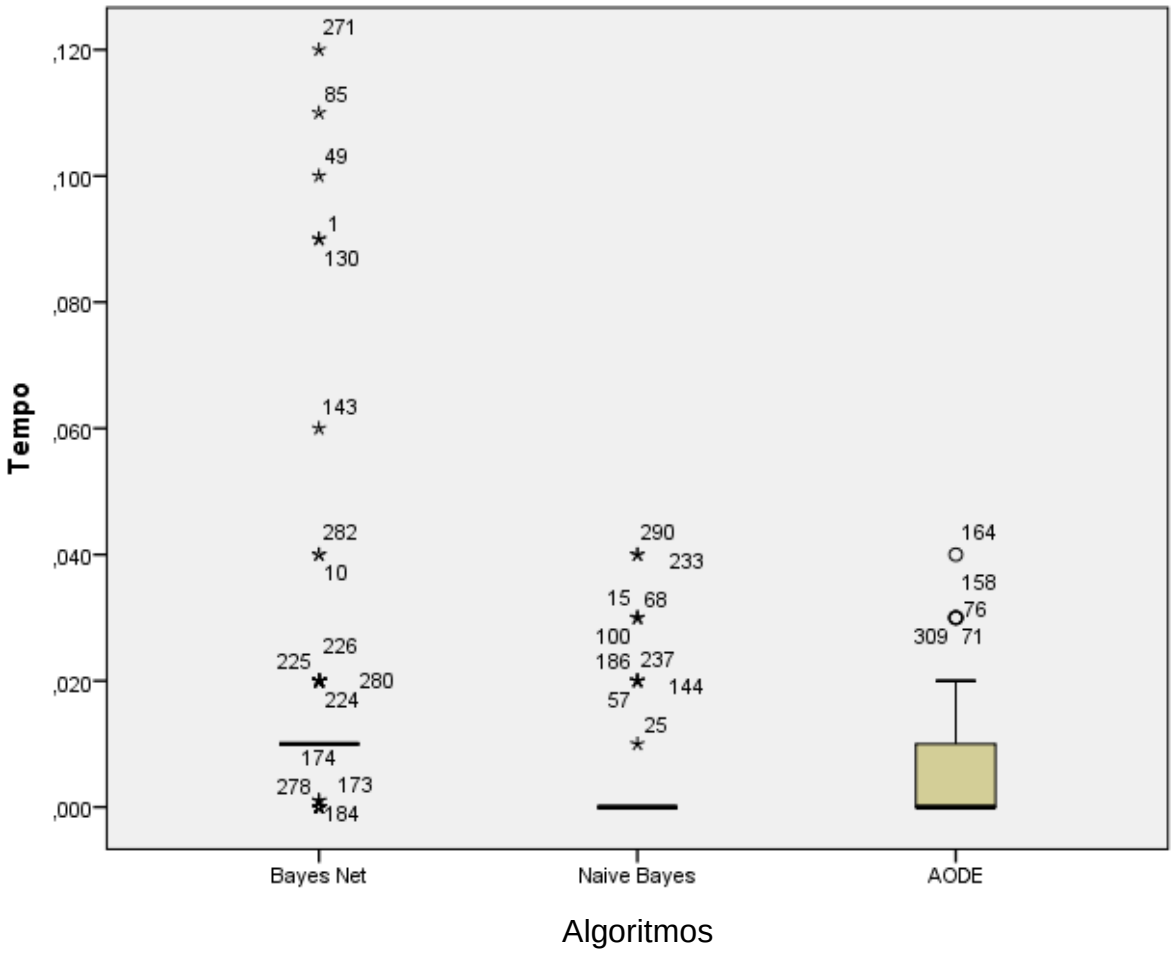
Figura 34 - Média do tempo de execução




Fonte: dados da pesquisa

Conforme ilustra a Figura 34 e a Figura 35, a média do tempo foi de 0,02 (±0,02), no algoritmo Bayes Net; 0,002 (±0,01) no Naive Bayes; e 0,004 (±0,01) no AODE.

Figura 35 - Blox plot do tempo de execução




Fonte: dados da pesquisa

A Figura 35 apresenta a mediana do tempo de execução, que foi de 0,01(±0,00), variando de 0,00, a 0,12 no algoritmo Bayes Net. No algoritmo Naive Bayes a mediana foi de 0,00 (±00) com variação de 0,00 a 0,04. No AODE a mediana foi de 0,00 (±0,01), oscilando de 0,00 a 0,04.

Mediante os resultados apresentados, a tabela 10 ilustra um comparativo das medidas estatísticas analisadas nos experimentos, que revelou o Algoritmo AODE como mais acurado.

Tabela 10 - Comparativo da medidas estatísticas

| Descrição | Algoritmo | Média | IC | Mediana | DP | Min | Max | II |
|---|---|---|---|---|---|---|---|---|
| Instâncias classificadas corretamente | Bayes Net | 82,54 | 79,53-85,54 | 84,06 | 14,76 | 41,04 | 99,67 | 26,66 |
| | Naive Bayes | 82,84 | 79,97-85,72 | 83,99 | 14,42 | 41,20 | 99,57 | 26,61 |
| | AODE | 84,52 | 81,69-87,34 | 89,65 | 14,08 | 42,81 | 99,84 | 23,08 |
| Instâncias classificadas incorretamente | Bayes Net | 17,46 | 14,45-20,47 | 15,93 | 14,76 | 0,32 | 58,95 | 26,66 |
| | Naive Bayes | 17,15 | 14,27-20,03 | 16,00 | 14,42 | 0,43 | 58,79 | 26,61 |
| | AODE | 15,48 | 12,66-18,30 | 10,35 | 14,08 | 0,16 | 57,19 | 23,08 |
| Kappa | Bayes Net | 0,33 | 0,27-0,40 | 0,28 | 0,28 | 0,00 | 0,95 | 0,43 |
| | Naive Bayes | 0,31 | 0,26-0,37 | 0,26 | 0,28 | 0,00 | 0,96 | 0,40 |
| | AODE | 0,34 | 0,28-0,40 | 0,29 | 0,29 | -0,01 | 0,97 | 0,45 |
| Erro médio-absoluto | Bayes Net | 0,19 | 0,16-0,22 | 0,17 | 0,14 | 0,00 | 0,43 | 0,27 |
| | Naive Bayes | 0,19 | 0,16-0,22 | 0,16 | 0,14 | 0,01 | 0,43 | 0,29 |
| | AODE | 0,18 | 0,15-0,21 | 0,16 | 0,14 | 0,00 | 0,43 | 0,26 |
| Sensibilidade | Bayes Net | 0,82 | 0,79-0,85 | 0,84 | 0,15 | 0,41 | 0,80 | 0,27 |
| | Naive Bayes | 0,83 | 0,80-0,86 | 0,84 | 0,14 | 0,41 | 0,99 | 0,27 |
| | AODE | 0,84 | 0,82-0,87 | 0,90 | 0,14 | 0,43 | 0,99 | 0,23 |
| Especificidade | Bayes Net | 0,53 | 0,48-0,59 | 0,51 | 0,27 | 0,00 | 0,95 | 0,49 |
| | Naive Bayes | 0,51 | 0,46-0,57 | 0,49 | 0,27 | 0,00 | 0,96 | 0,49 |
| | AODE | 0,49 | 0,43-0,55 | 0,48 | 0,29 | 0,00 | 0,95 | 0,52 |
| Precisão | Bayes Net | 0,81 | 0,78-0,85 | 0,81 | 0,16 | 0,38 | 0,99 | 0,32 |
| | Naive Bayes | 0,82 | 0,78-0,85 | 0,82 | 0,16 | 0,37 | 0,99 | 0,31 |
| | AODE | 0,83 | 0,79-0,86 | 0,88 | 0,17 | 0,12 | 0,99 | 0,28 |
| Recall | Bayes Net | 0,82 | 0,79-0,85 | 0,84 | 0,15 | 0,41 | 0,99 | 0,27 |
| | Naive Bayes | 0,83 | 0,80-0,86 | 0,84 | 0,14 | 0,41 | 0,99 | 0,27 |
| | AODE | 0,84 | 0,82-0,87 | 0,90 | 0,14 | 0,42 | 0,99 | 0,23 |
| Medida-F | Bayes Net | 0,81 | 0,78-0,84 | 0,82 | 0,16 | 0,38 | 0,99 | 0,29 |
| | Naive Bayes | 0,81 | 0,78-0,84 | 0,82 | 0,16 | 0,37 | 0,99 | 0,27 |
| | AODE | 0,83 | 0,80-0,86 | 0,89 | 0,15 | 0,39 | 0,99 | 0,29 |
| AUC | Bayes Net | 0,79 | 0,75-0,82 | 0,85 | 0,17 | 0,04 | 0,99 | 0,24 |
| | Naive Bayes | 0,79 | 0,76-0,89 | 0,83 | 0,17 | 0,04 | 0,99 | 0,27 |
| | AODE | 0,81 | 0,78-0,85 | 0,87 | 0,16 | 0,04 | 0,99 | 0,26 |
| Tempo | Bayes Net | 0,02 | 0,01-0,02 | 0,01 | 0,02 | 0,00 | 0,12 | 0,00 |
| | Naive Bayes | 0,00 | 0,00-0,00 | 0,00 | 0,01 | 0,00 | 0,04 | 0,00 |
| | AODE | 0,00 | 0,00-0,01 | 0,00 | 0,01 | 0,00 | 0,04 | 0,01 |

IC: Intervalo de confiança / DP: Desvio padrão / II: Intervalo de confiança
Fonte: Dados da pesquisa

# 8 CONCLUSÕES

A mineração de dados é uma técnica amplamente utilizada em várias áreas de conhecimento e busca extrair conhecimento implícito de bases de dados.

Com o desenvolvimento dessa pesquisa pôde-se compreender que o classificador bayesiano usa o teorema de bayes e se destaca pela eficiência, velocidade e acurácia, possibilitando sua utilização no processo de diagnose.

Esta pesquisa possibilitou também compreender os testes diagnósticos, que possibilitam medir a eficiência dos classificadores, podendo-se destacar a sensibilidade, especificidade, entre outros.

Conclui-se também que esta pesquisa atendeu os principais aspectos éticos e legais pertinentes à informática em saúde, pois o projeto foi aprovado pelo Comitê de Ética e Pesquisa em Seres Humanos da Universidade do Extremo Sul Catarinense.

Mediante os estudos realizados, pôde-se compreender o funcionamento de alguns algoritmos de classificação bayesiana, entre eles, o AODE, o NaiveBayes, e o Bayes Net, que foram utilizados na mineração de uma base de dados de osteopenia e osteoporose. Para tal, foi utilizada a shell Weka nos experimentos realizados.

Para realização dos experimentos, foram apreendidos conceitos sobre osteopenia e osteoporose, seus principais sintomas, aspectos epidemiológicos, sendo mais prevalente em pessoas com idade acima dos 50 e mulheres.

Nos experimentos foram utilizados os três algoritmos supracitados, sendo geradas 327 minerações, os quais foram utilizadas no pós-processamento e resultaram na análise da acurácia.

Entre as medidas estatísticas analisadas nos experimentos (instâncias classificadas corretamente e incorretamente, estatística Kappa, Erro médio absoluto, sensibilidade, especificidade, precisão recall, medida-F, e AUC), o algoritmo AODE foi considerado o mais acurado, apesar de um pouco mais lento em relação ao tempo de execução.

Essa pesquisa teve algumas limitações, como a impossibilidade de utilizar as ferramentas Tanagra e Knime pelas poucas medidas de avaliação presentes.

Outra limitação foi a de que o algoritmo AODE só está implementado na versão 3.6.9 do Weka e não possui as medidas de avaliação que a versão 3.7.8 oferece, por isso utilizamos as duas versões no estudo.

## 8.1 TRABALHOS FUTUROS

A partir dos experimentos realizados nessa pesquisa sugere-se como trabalhos futuros: utilizar uma maior quantidade de ferramentas comparando algoritmos Bayesianos com outros algoritmos de classificação.

Minerar a base de dados buscando caracterizar o conhecimento implícito por meio de árvores de decisão, J48, entre outros.

# REFERÊNCIAS


AMED, Abes Mahmed et al. **Atualização terapêutica de Prado, Ramo e Valle:** diagnóstico e tratamento. 24. ed. São Paulo: Artes Médicas, 2012.

AMORIM, Maurício J. V.; BARONE, Dante; MANSUR, André Uebe. Técnicas de Aprendizado de Máquina Aplicadas na Previsão de Evasão Acadêmica. In: XIX SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE 2008), 2008, Fortaleza. **Anais...**. Fortaleza: Sbie, 2008. p. 666 - 674.

ARA-SOUZA, A. L. **Redes Bayesianas**: Uma Introdução Aplicada a Credit Scoring. Projeto de Pesquisa. Iniciação Científica do 19º Simpósio Nacional de Probabilidade e Estatística, 2010.

BAPTISTA, R. S. et al. Desenvolvimento e avaliação de um classificador de padrões para análise do crescimento facial pelo método de maturação vertebral cervical. **Journal of Health Informatics**, v. 3, n. 2, p. 64-68, abr./jun. 2011.

BERRY M.; LINOFF G**. Data Mining Techniques**. New York, John Wiley & Sons, 1997.

BISHOP, N. Characterising and treating osteogenesis imperfecta. **Early Human Development,** v. 86, n. 11, p. 743–746, nov. 2010.

BLAKENEY, W. Stabilization and treatment of Colles; fractures in elderly patients. **Clinical Interventions in Aging**, p. 337, nov. 2010.

BOBBIO, A. et al. E,Improving the analysis of dependable systems by mapping fault tress into Bayesian networks**. Reliability Enginneringand System Safety**, v. 71, p. 249-260, 2001.

CHANDRA, V.; PANDAV, C. S. The importance of non-communicable diseases in developing countries (editorial). **Indian J Community Med**, v.12, p.80-178, 1987.

CHEN, Jianjun et al. The small introns of antisense genes are better explained by selection for rapid transcription than by ‘genomic design’. **A Head Of Print,** United States, p.1-21, 02 set. 2005

CHENG, J.; BELL, D. A.; LIU, W. **Learning Bayesian belief network classifiers:** algorithms and systems. **Lecture Notes In Computer Science**, New York, 2001.

CHICKERING, D. M.; HECKERMAN, D.; MEEK, C. Large-Sample Learning of Bayesian Networks is NP-Hard. **The Journal of Machine Learning Research archive**, v. 5, p. 1287-1330, 2004.

COSTA-PAIVA, Lúcia et al. Prevalência de Osteoporose em Mulheres na Pós-menopausa e. **Revista Brasileiro de Ginecologia e Obstetrícia**, Campinas, v. 25, n. 7, p.507-512, 2007.

CRAIG, J.; PATTERSON, V. Introduction to the practice of telemedicine. **J Tele med Telecare**, United States, v. 11, n.1, 2005.

CUROTTO, C. L.; EBECKEN, N. F. F. **Multi-relational data mining in Microsoft® SQL Server™**. COPPE-UFRJ, 2006.

DEVILLÉ, W. L. et al. Conducting systematic reviews of diagnostic studies: didactic guidelines. **BMC Med Res Methodol**, v. 2, p.9, 2002.

DREISEITL, S. et al. **Comparison of machine learning methods for the diagnosis of pigmented skin lesions.** J Biomed Inform, United States, v. 34, n. 11, Feb. 2001.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. 2. ed. Canada: Wiley-Interscience, 2000.

ESTEVAM JUNIOR, Rafael Hruschka. **Imputaçao bayesiana no contexto de mineração de dados.** 2003. 119 f. Dissertação (Programa de Pós Graduação em Ciências em 2003, Universidade Federal do Rio de janeiro, Rio de Janeiro, 2003.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **American Association For Artificial Intelligence**, p.1-18, 1996.

FLETCHER, R. H.; SUZANNE, W. **Epidemiologia clínica: elementos essenciais.** Artmed. 4ed. Porto Alegre, 2006.

FONSECA, Marcello Porto Alegre**. Classificação Bayesiana de Grandes Massas de Dados em Ambientes Rolap**. 2007. 117 f. Tese (Doutorado). Programa de pós-graduação de engenharia da Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2007.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel Lopes. **Data mining**: uma guia prático: conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005.

HALL, M. et al. The WEKA data mining software: an update. **ACM SIGKDD Explorations Newsletter,** v. 11, n. 1, p. 10, 16 nov. 2009.

HAN, J.; KAMBER, M. **Data Mining Concepts and Techniques**.Morgan Kaufman Publishers, San Francisco, USA, 2006.

HRIPCSAK, G.; ROTHSCHILD, A. S. Agreement, the f-measure, and reliability in information retrieval. **J Am Med Inform Assoc**, v. 12, n. 3, p. 296-298, 2005.

HOLICK, M. F.; CHEN, T. C. Vitamin D deficiency: a worldwide problem with health consequences. **The American Journal of Clinical Nutrition**, p. 07, April 2006.

IP, T. P.; LEUNG, J.; KUNG, A. W. C. Management of osteoporosis in patients hospitalized for hip fractures**. Osteoporosis International**, v. 21, n. S4, p. 605–614, 6 nov. 2010.

JOHNELL, O.; KANIS, J. A. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. **Osteoporosis international: a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA**, v. 17, n. 12, p. 1726–1733, dez. 2006.

KANIS, J. A.; WHO Study Group. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis. Synopsis of a WHO Report. **Osteoporos Int**, v. 4, p. 368-381, 1994.

KNIME. **Knime Data Mining.** Disponível em: <http://www.knime.org/>. Acesso em: 05 nov. 2013.

KONDO, K. Osteoporotic Vertebral Compression Fracturesand Vertebral Augmentation. **Seminars in Interventional Radiology**, v. 25, n. 04, p. 413–424, 15 dez. 2008.

LEEFLANG, M. M. et al. Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. **Ann Intern Med**, v. 149, n. 12, p. 889-897, 2008.

MANARIN, Daiane de Nez. **Aquisição de Conhecimento em Sistemas Especialistas Probabilísticos por Meio da Descoberta de Conhecimento em Base de Dados para Construção de Redes Bayesianas**. 2004. 112 f. Monografia (Bacharel em Ciência da Computação) - Universidade do Extremo Sul Catarinense, Criciúma, 2004.

MASSAD, Eduardo. **Métodos quantitativos em medicina**. Barueri, SP: Manole, 2004.

MELLO, Luis Cesar de. **Descoberta de conhecimento em banco de dados (mineração de dados) e redes bayesianas:** estado da arte. 2001.

MOON, P.; KWON, W.; LEE, Y. Delay-dependent robust stabilization of uncertain state-delayed systems.**International Journal of Control**, p. 1447–1455, 2001.

NEAPOLITAN, R. E. **Learning Bayesian networks**. Upper Saddle River (New Jersey): Pearson Prentice Hall, 2004.

NOLEN, B. M.; LOKSHIN, A. E. Multianalyte assay systems in the differential diagnosis of ovarian cancer. **Expert Opin Med Diagn**, v. 6, n. 2, p. 131-138, 2012.

PANDA, M.; PATRA, M. R. A Comparative Study of Data Mining Algorithms for Network Intrusion Detection. **IEEE Explore**, 2008.

PAPAIOANNOU, A. et al. Diagnosis and management of vertebral fractures in elderly adults.**The American journal of medicine**, 2002.

PETERS, B. S. E.; MARTINI, L. A. Nutritional aspects of the prevention and treatment of osteoporosis. **Arquivos Brasileiros de Endocrinologia & Metabologia,** v. 54, n. 2, p. 179–185, mar. 2010.

POLITI, Jacques. **Implementação de uma Metodologia para Mineração de Dados Aplicada ao Estudo de Núcleos Convectivos**, 2006. 151 f. Dissertação (Mestrado) - Curso de Pós-Graduação em Computação Aplicada, da Universidade de São José dos Campos, São José dos Campos, 2006.

PORTUGAL, Luis Ivo Costa Gomes. **Osteopenia e Osteoporose**: factores modificáveis e não modificáveis. 2012. 87 f. Projeto de Pós Graduação. Curso de Ciências farmacêuticas da Universidade Fernando Pessoa, Lisboa, 2012.

RACHNER, T. D.; KHOSLA, S.; HOFBAUER, L. C. Osteoporosis: now and the future. **Lancet**, v. 377, n. 9773, p. 1276-1287, 2011.

RAISZ, L. G. Pathogenesis of osteoporosis: concepts, conflicts, and prospects. **Journal of Clinical Investigation**, v. 115, n. 12, p. 3318–3325, 1 dez. 2005.

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

ROMÃO, Wesley**. Descoberta de Conhecimento Relevante em Banco de dados sobre Ciência e tecnologia**. 2002. 238 f. Tese de Doutorado. Programa de pós Graduação em Engenharia de Produção. Universidade Federal de Santa Catarina, Florianópolis, 2002.

RUSSELL, S. J.; NORVIG, P. **Inteligência artificial**. Rio de Janeiro: Elsevier, 2004.

SAHAMI, M. The happy searcher: Challenges in the web information retrieval**, Pacific Rim International Conference on Artificial Intelligence**, v. 3157, p. 3-12, 2004.

SANTOS, Edimilson Batista dos. **A Ordenação das Variáveis no Processo de Otimização de Classificação Bayesianos:** Uma Abordagem Evolutiva.2007.114 f. Dissertação (Mestrado) – Curso Ciência da Computação,da Universidade Federal de São Carlos, São Carlos, 2007.

SHABBEER, A. TB-Lineage: An online tool for classification and analysis of strains of Mycobacterium tuberculosis complex**. Infect Genet Evol**, v. 12, n. 4, p. 789-797, Sep. 2012.

SHORTLIFFE, Edward Hance; CIMINO, James J. **Biomedical informatics:** computer applications in health care and biomedicine. 3ed.New York, USA: Springer, 2006.

SIGURDSSON, J. H.; WALLS, L. A.; QUIGLEY, J. L. Bayesian belief nets for managing expert judgement and modelling reliability. **Quality and Reliability Engineering International**, v. 17, n. 3, p. 181–190, maio 2001.

SILVA, Ana Carolina Veiga. **Fatores associados à osteopenia e osteoporose em mulheres submetidas à densitometria óssea**. 2013. 119 f. Monografia de Conclusão de Curso. Curso de Medicina. Universidade do Extremo Sul Catarinense, Criciúma, 2013.

SIMS, N. A.; GOOI, J. H. Bone remodeling: Multiple cellular interactions required for coupling of bone formation and resorption. **Seminars in Cell & Developmental Biology**, v. 19, n. 5, p. 444–451, out. 2008.

SZEJNFELD, Vera Lúcia et al. Conhecimento dos Médicos Clínicos do Brasil sobre as Estratégias de Prevenção e Tratamento da Osteoporose. **Revista Brasileira Reumatologia**, São Paulo, p.251-257, 01 jul. 2007.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining:** Mineração de dados. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009.

TANAGRA. **Tanagra data mining.** Disponível em: <http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html>. Acesso em: 05 nov. 2013.

TENÓRIO, J. M. et al. Artificial intelligence techniques applied to the development of a decision-support system for diagnosing celiac disease. **International Journal of Medical Informatics**, v. 80, n. 11, p. 793-802, nov. 2011.

USDHHS. Bone Health and Osteoporosis: a Report of the Surgeon General.Washinton D.C**. National Library of Medicina Cataloging in Publication**, p.1-28, 2004.

WEBB, G. I.; BOUGHTON, J.; WANG, Z. Not so naive bayes: Aggregating one-dependence estimators. **Machine Learning**, v. 58, p. 5-24, 2005.

WEKA. **Weka.** Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 05 nov. 2013.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A**. Data mining practical machine learning tools and techniques.** 3. ed. Burlington: Morgan Kaufmann, 2011.

YANG, W.; WANG, S. A process mining framework for the detection of healthcare fraud and abuse.**Expert Systems with Applications**. v. 31, p.56-68, 2006.

ZAMORA, J. et al. Meta-DiSc: a software for meta-analysis of test accuracy data. **BMC Med Res Methodol**, v. 6, p. 31, 2006.

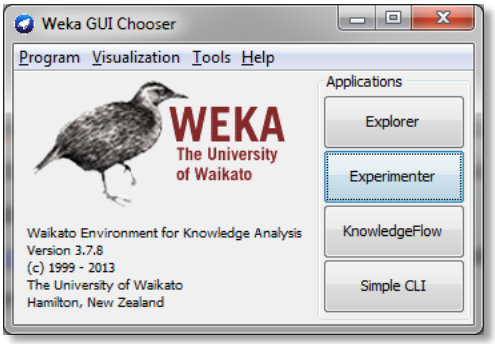
# APÊNDICES

# **APÊNDICE A** – Tutorial para mineração de dados utilizando o algoritmo NaiveBayes na ferramenta Weka

O Weka é uma ferramenta gratuita para classificação de dados entre outros métodos. Para iniciarmos você precisa baixar o software que está localizado no repositório da Universidade de Waikato através do seguinte link: www.cs.waikato.ac.nz/ml/weka/ e efetuar a instalação em seu computador (Figura 36).

Após a instalação e inicialização aparecerá na interface da Figura 36.

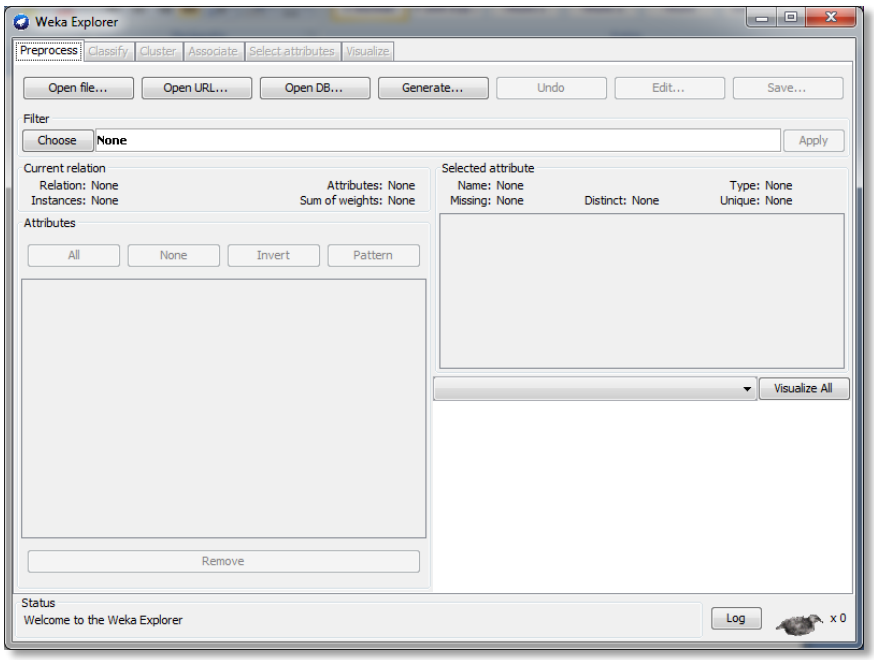
Figura 36 - Tela inicial do Weka




Fonte: Weka (2013)

Na tela inicial o usuário deverá abrir o modo de exploração clicando no botão "*Explorer*". A tela demonstrada na Figura 37 será apresentada.

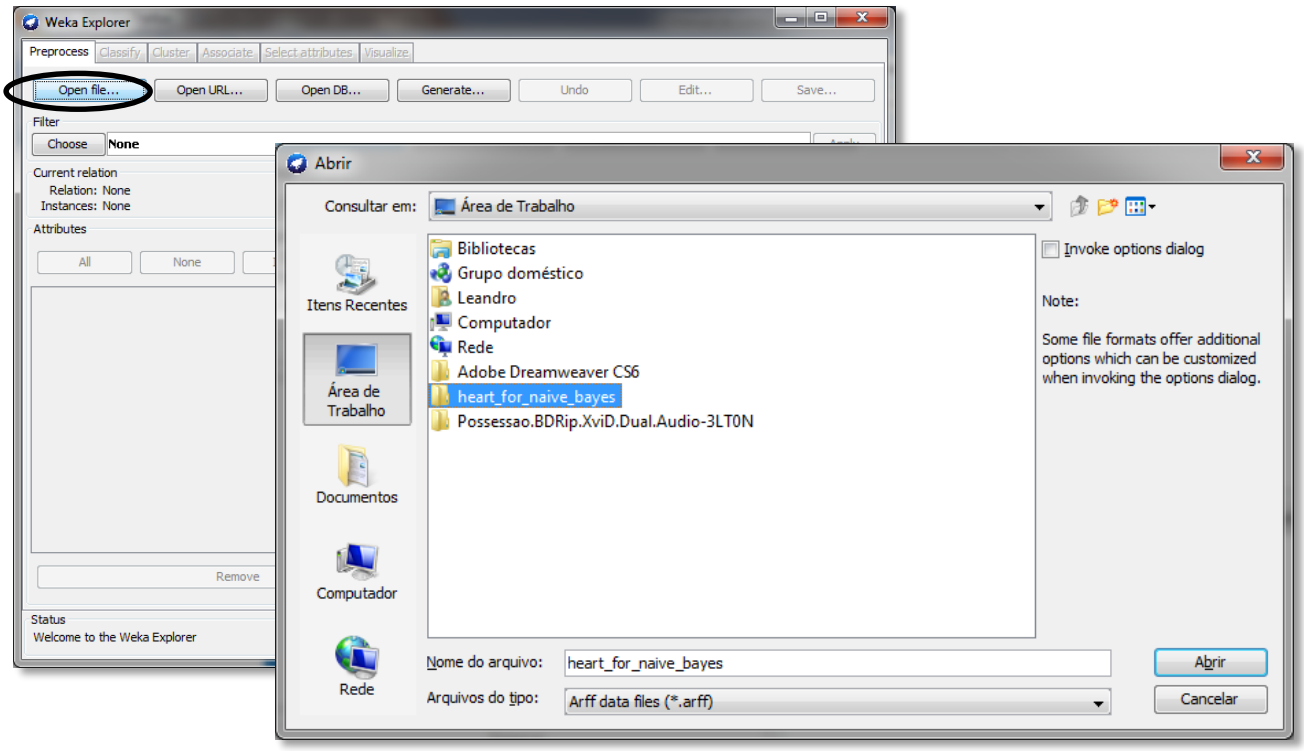
Figura 37 - Tela inicial do modo Explorer




Fonte: Weka (2013)

Ao abrir a tela indicada na Figura 37, o usuário deve clicar em *"Open file"* e informar o diretório da base desejada, como ilustra a Figura 38. Deve-se selecionar um arquivo *"arff"* e clicar em "Abrir". Neste exemplo utilizamos a base "osteoporose.arff".

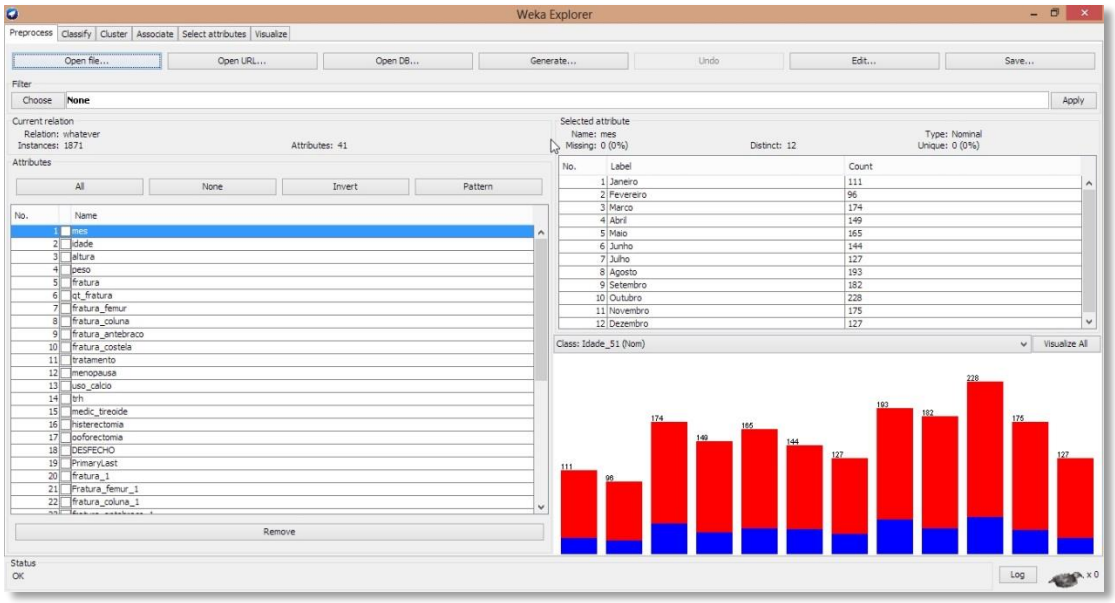
Figura 38 - Seleção da base de dados




Fonte: Weka (2013)

Verifica-se que a base foi carregada corretamente conforme ilustra a Figura 39.

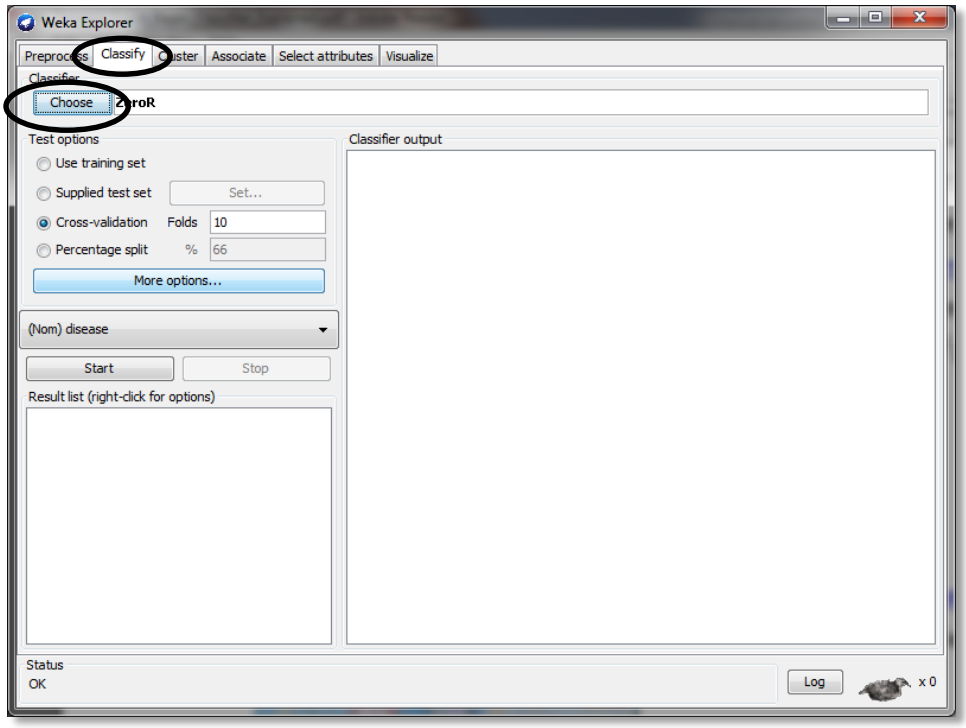
Figura 39 - Base de dados carregada com sucesso




Fonte: Weka (2013)

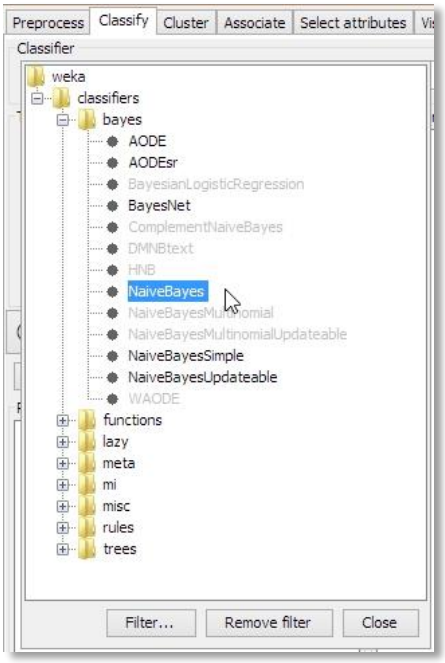
Depois de carregada a base de dados, deve-se escolher o método de classificação. Seleciona-se a aba *"Classify"* e escolhe-se o método NaiveBayes ao clicar no botão *"Choose"*, conforme ilustra a Figura 40 e a Figura 41.

Figura 40 - Aba "Classify"




Fonte: Weka (2013)

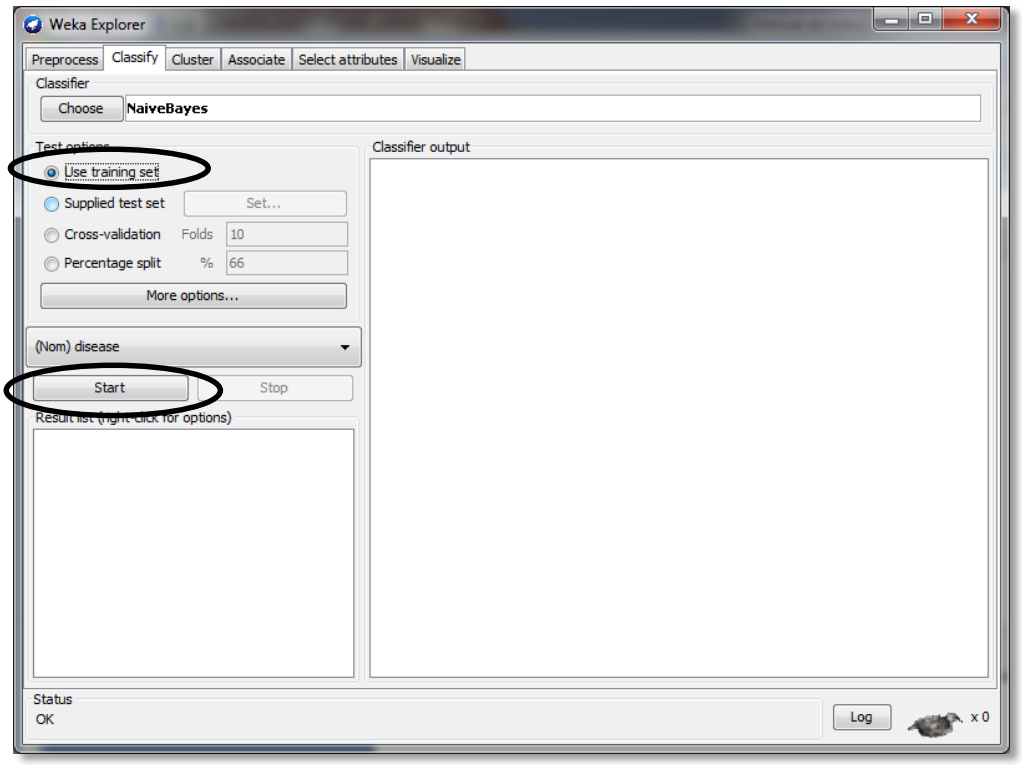
Figura 41 - Seleção do método NaiveBayes




Fonte: Weka (2013)

Depois de escolhido o algoritmo de classificação deve-se clicar em *"Use training set"* e após clicar em *"Start"*, que dará início ao pré-processamento da base, como apresenta a Figura 42.

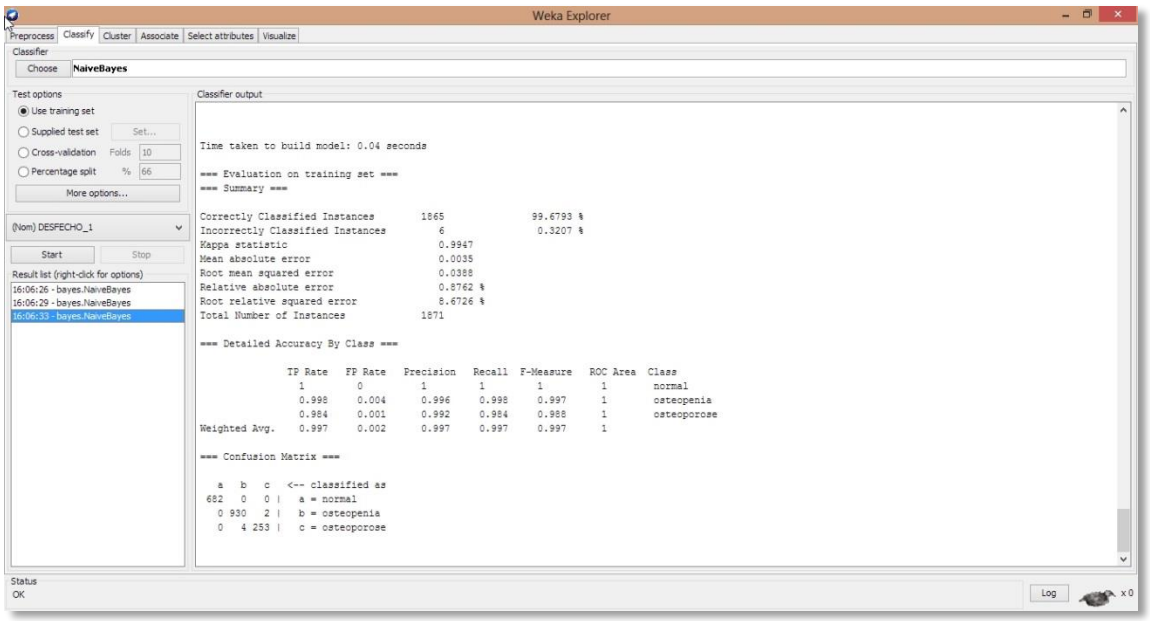
Figura 42 - Tela de pré-processamento




Fonte: Weka (2013)

Após o início do processamento é apresentada a tela de resultados ao lado e o histórico dos últimos processamentos (Figura 43).

Figura 43 - Final do processamento e apresentação dos resultados

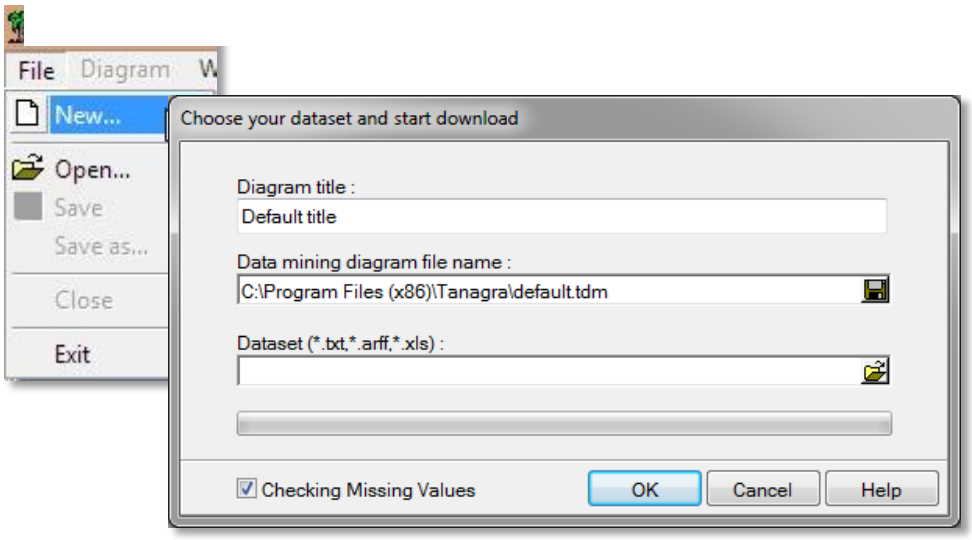



Fonte: Weka (2013)

**APÊNDICE B** – Tutorial para mineração de dados utilizando o algoritmo NaiveBayes na ferramenta Tanagra

Para usar a classificação no Tanagra 1.4.48 primeiramente deve-se criar um novo arquivo. Para isso clica-se em *"File"* e logo em seguida *"New"*. Após abrirá a janela de seleção da base (Figura 44).

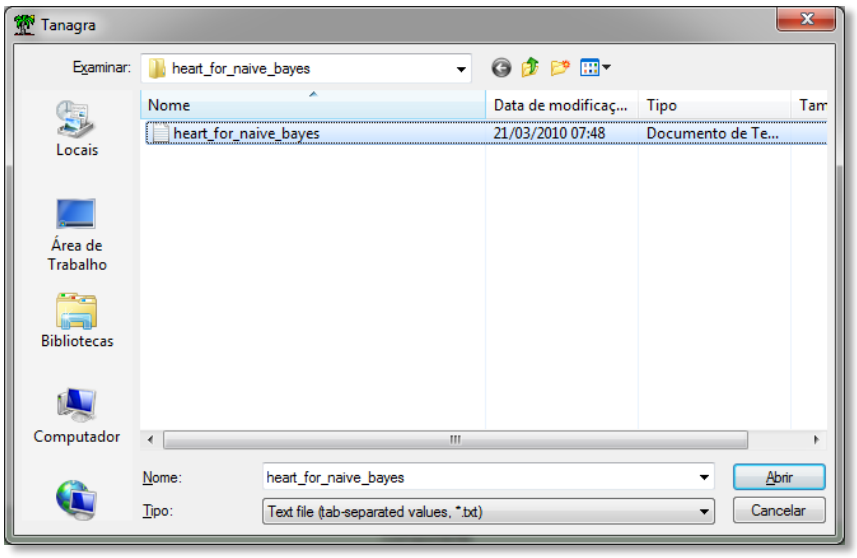
Figura 44 - Criação de um novo arquivo




Fonte: Tanagra (2013)

Nomeia-se a base, informa-se o diretório onde está localizada a base de dados e confirma-se clicando em "OK", como demonstra a Figura 45.

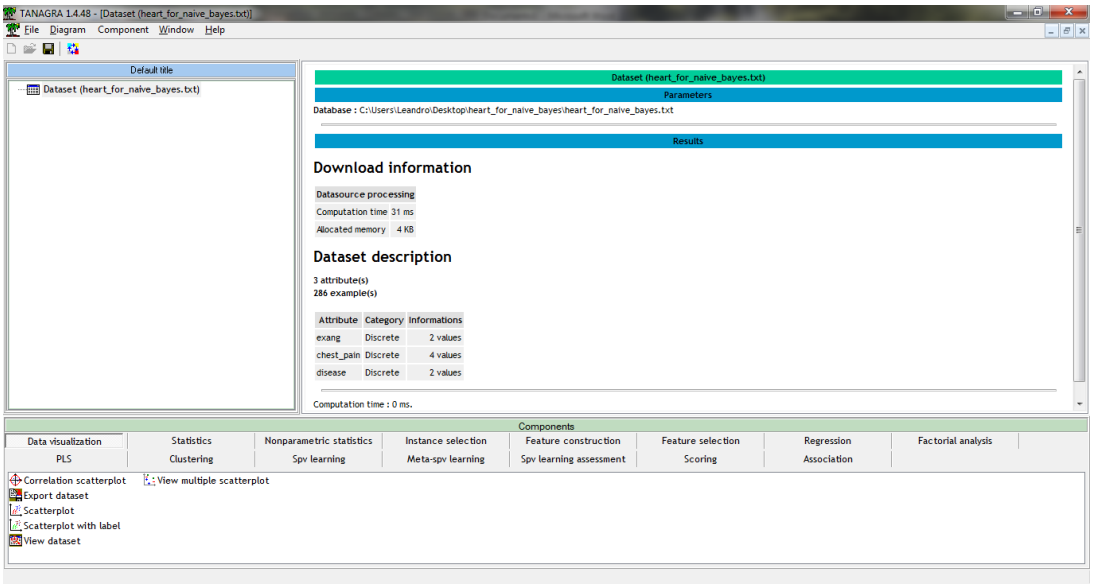
Figura 45 - Seleção da base




Fonte: Tanagra (2013)

A Figura 46 ilustra a tela pós-carregamento da base de dados. O painel esquerdo mostra a base que selecionamos e o painel direito fornece as informações dessa base como: tempo computacional, locação de memória e seus atributos.

Figura 46 – Base carregada com sucesso

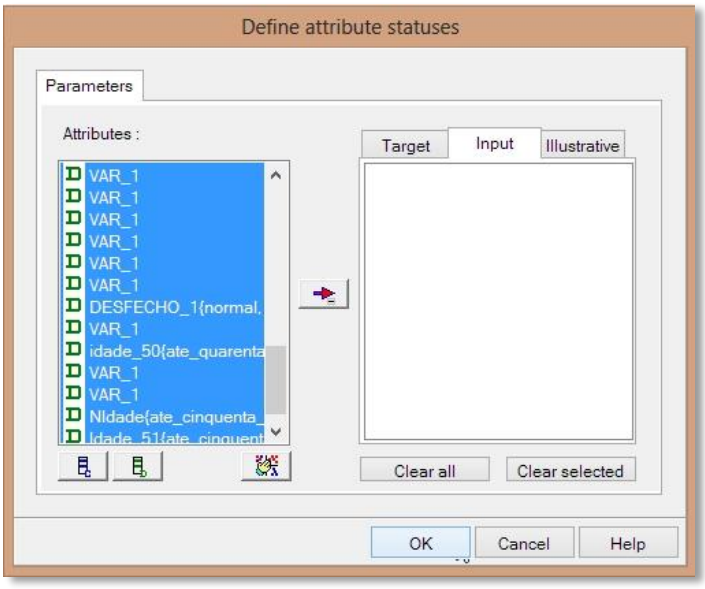



Fonte: Tanagra (2013)

Para fazer o método de classificação, precisamos definir o status da base clicando no botão  que está localizado no submenu  .

Selecionamos todos os atributos (ou os atributos que se deseja analisar) clicando no botão  e após clicamos em  e confirmamos clicando em “OK” (Figura 47).

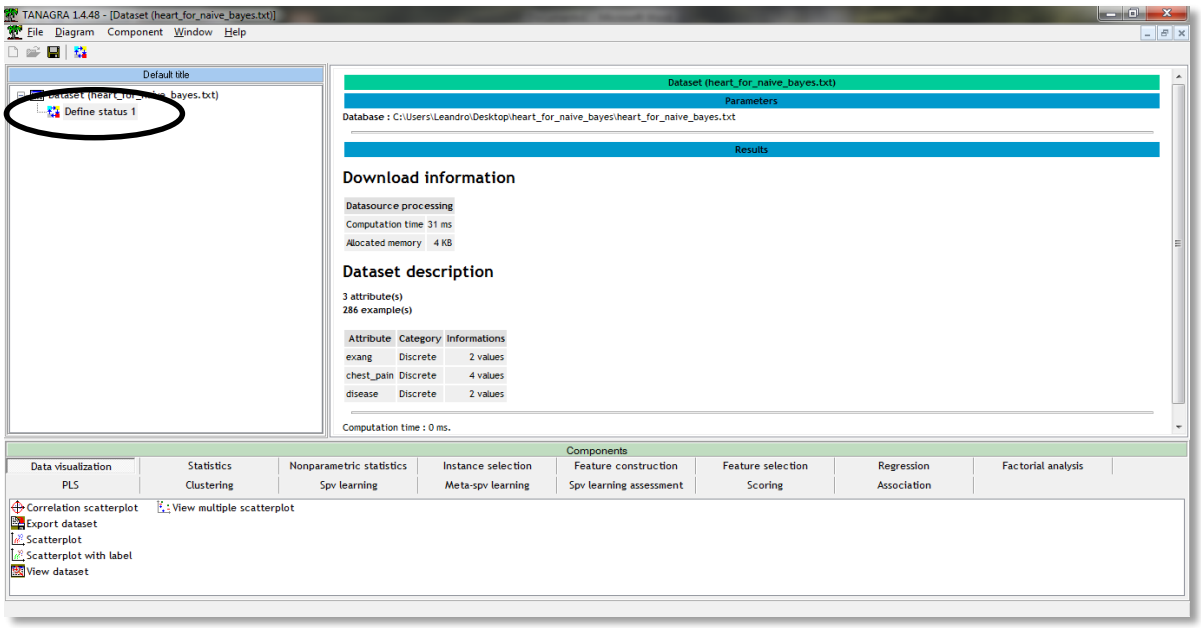
Figura 47 - Tela de seleção dos atributos




Fonte: Tanagra (2013)

Após definir o status, deve-se clicar com o botão direito do mouse em cima do status definido e clicar em *“Execute”*. O ícone que define o status ficará acesso como vemos na Figura 48.

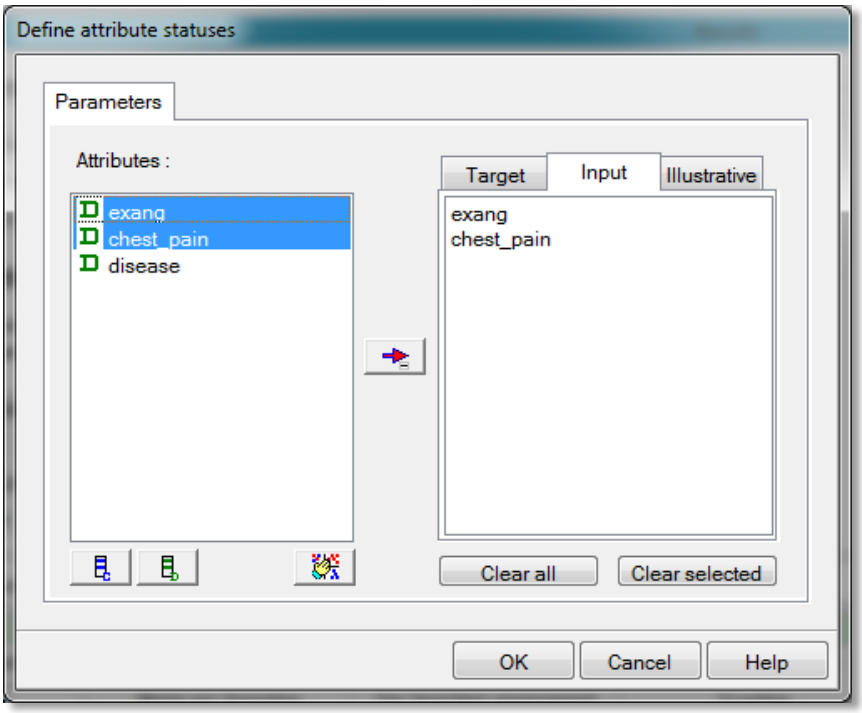
Figura 48 - Definição de status




Fonte: Tanagra (2013)

Após a execução, deve-se configurar os parâmetros como na Figura 49. Onde na aba *"Target"* escolhe–se a saída da classificação.

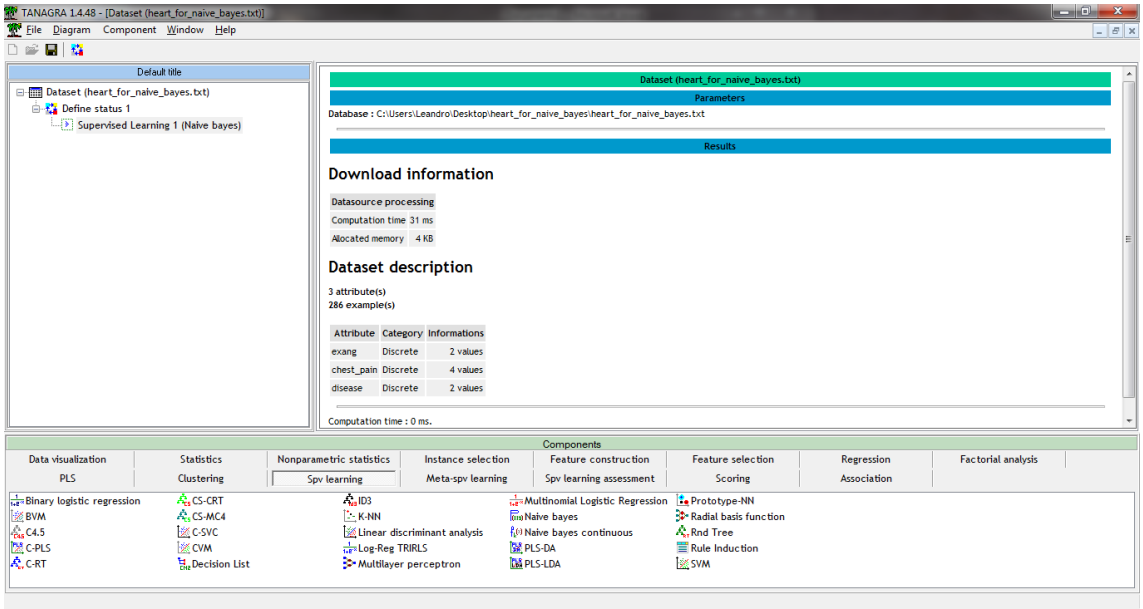
Figura 49 - Seleção dos atributos




Fonte: Tanagra (2013)

Já na aba *"Input"* escolhe-se os atributos de entrada e confirma-se em "OK". Para selecionar o método Naive Bayes clica-se na aba *"Spv Learning"* no sub menu inferior, seleciona-se o classificador Naive Bayes e arrasta-se o ícone até o *"Define status 1"*, como pode-se observar na Figura 50.

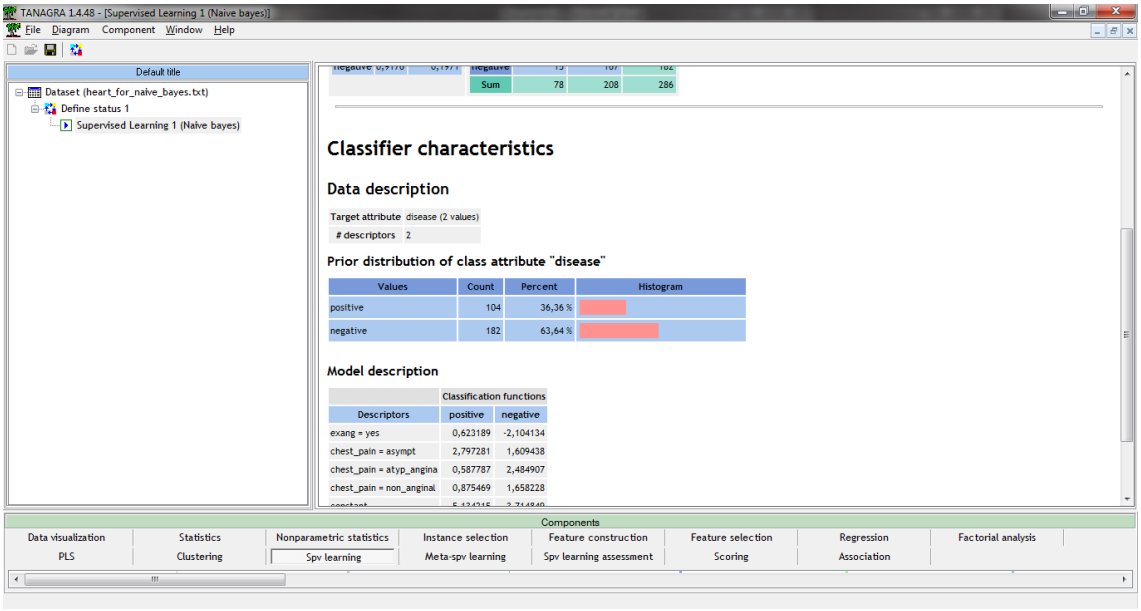
Figura 50 - Seleção do método NaiveBayes




Fonte: Tanagra (2013)

Após clica-se com o botão direito do mouse no módulo e seleciona-se *"Execute"*. Os resultados da classificação aparecem ao lado, como ilustra a Figura 51.

Figura 51 - Execução e resultados



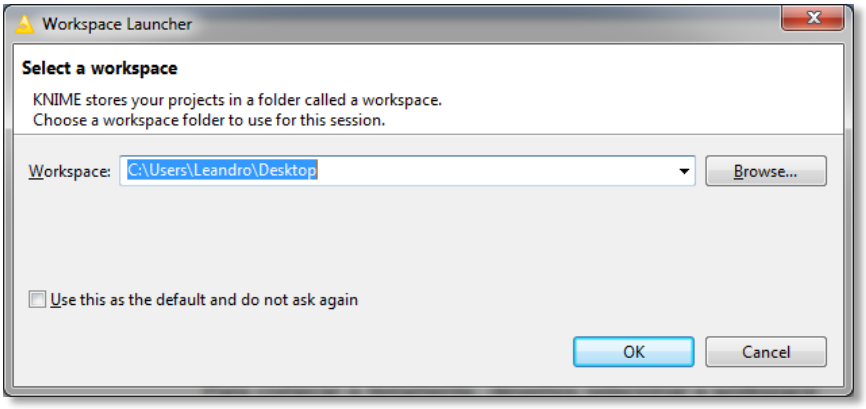

Fonte: Tanagra (2013)

**APÊNDICE C –** Tutorial para mineração de dados utilizando o algoritmo NaiveBayes Learner na ferramenta KNIME

O Knime é um software de classificação, onde é possível classificar uma base de dados, com algoritmos bayesianos, a ferramenta é freeware, e a mesma pode ser baixada no link http://www.knime.org/node/412.

Para inicializar a ferramenta devemos selecionar o workspace que neste exemplo é área de trabalho, como ilustrado na Figura 52.

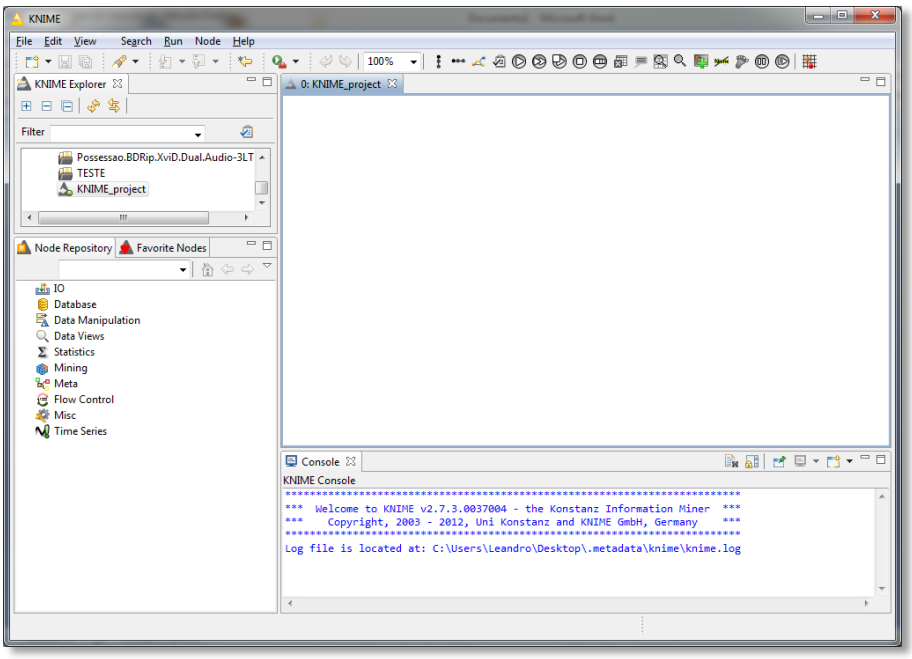
Figura 52 – Definição do workspace




Fonte: Knime (2013)

Depois de selecionar o local onde seus projetos serão salvos, clica-se em “OK”, e aparecerá a tela inicial do programa, como na Figura 53.

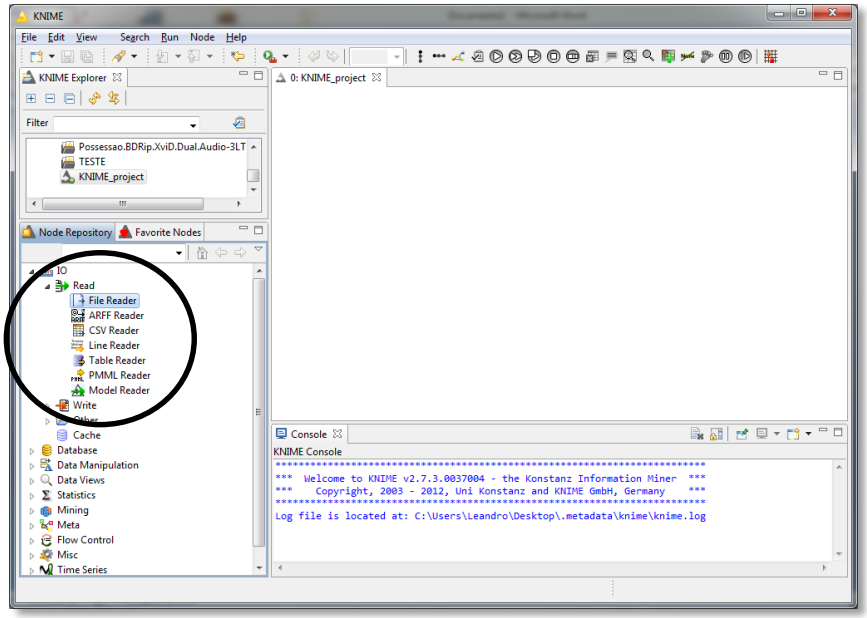
Figura 53 – Tela inicial




Fonte: Knime (2013)

Navega-se no painel lateral expandindo as opções, então deve-se selecionar a opção *"File Reader"* que está contida dentro de *IO* > *Reader* (Figura 54).

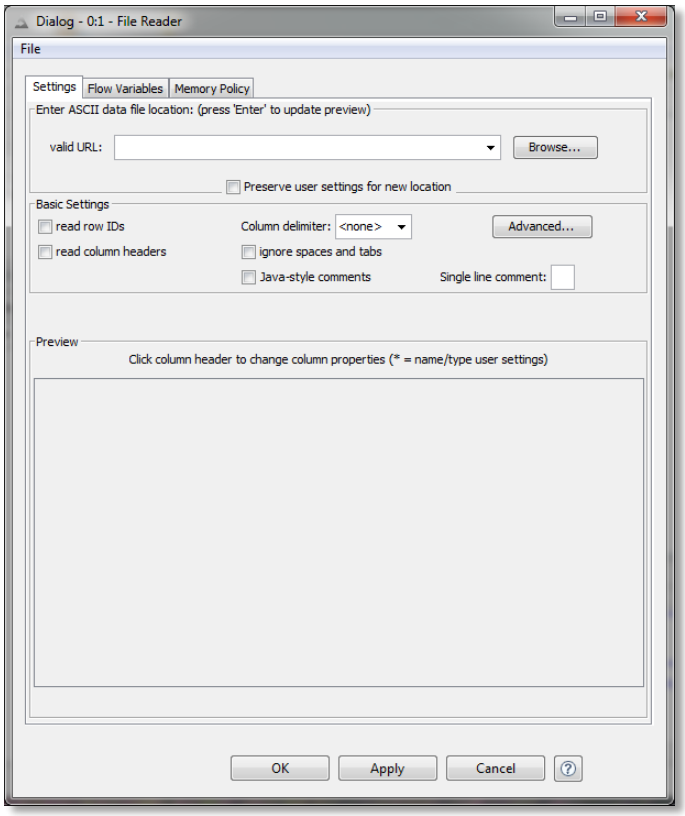
Figura 54 – Seleção de arquivo




Fonte: Knime (2013)

Deve-se escolher a base de dados clicando com o botão direito do mouse sobre o ícone *"File Reader"* e selecionar a opção *"Configure"* que abrirá a tela ilustrada na Figura 55.

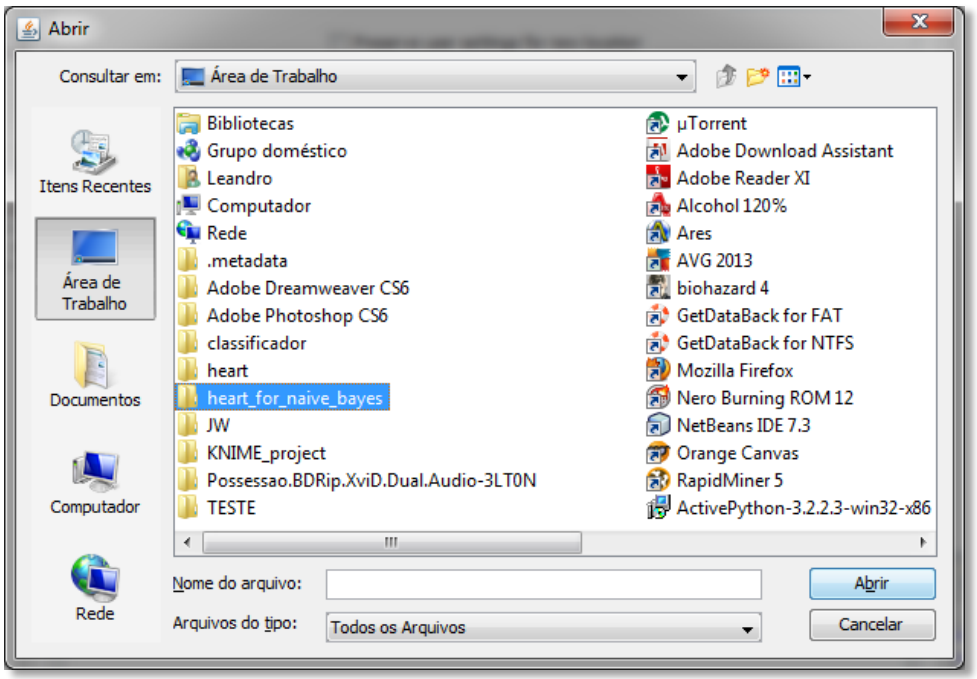
Figura 55 - Seleção da base




Fonte: Knime (2013)

No campo *"valid Url"* devemos procurar a base desejada clicando em browse, que abrirá a tela da Figura 56.

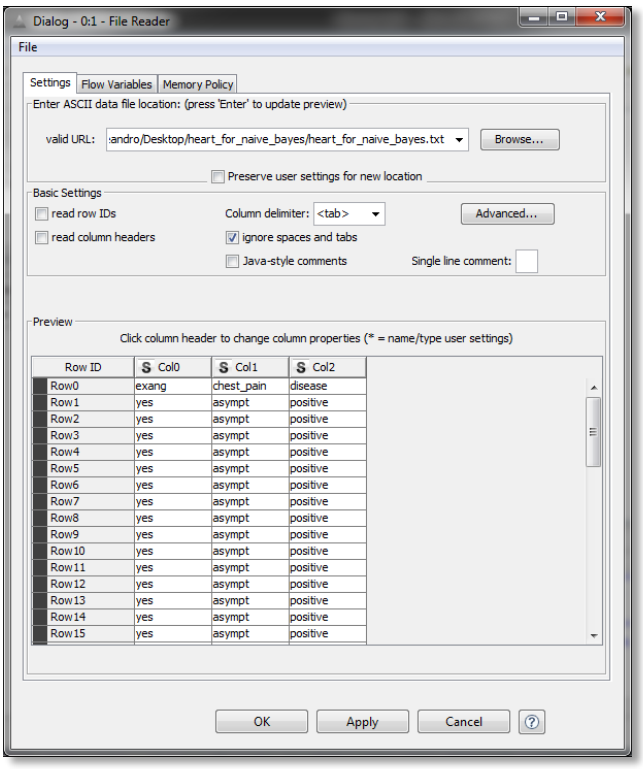
Figura 56 - Selecionando o diretório da base




Fonte: Knime (2013)

Depois de selecionada e aberta a base, concluímos a importação da mesma clicando em “OK”, obtendo assim a tela ilustrada na Figura 57.

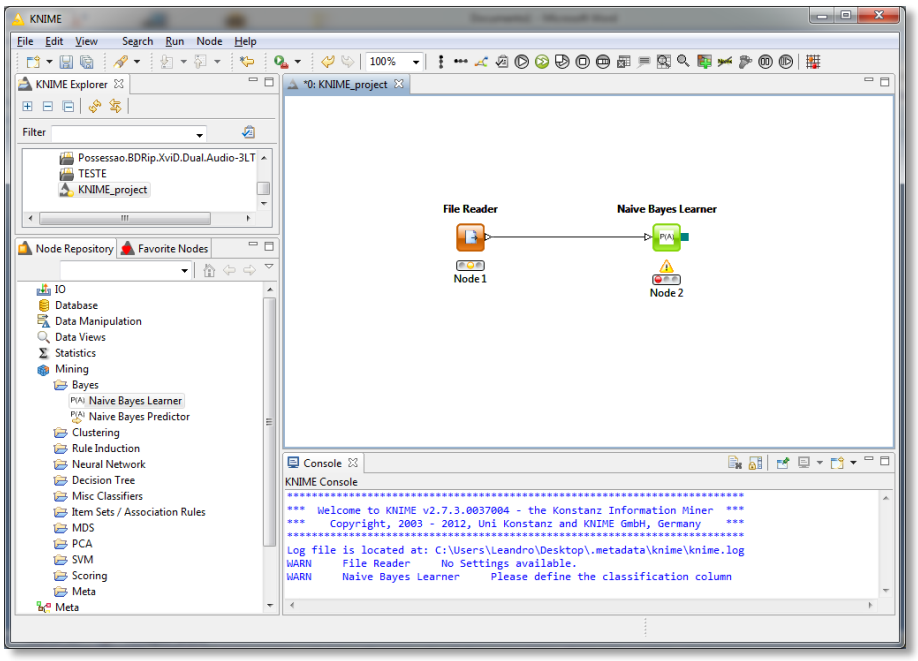
Figura 57 - Conclusão do processo de importação




Fonte: Knime (2013)

Logo depois de fazer os passos acima, escolhe-se o método a ser aplicado navegando pelo painel até Naive Bayes Learner que está contido dentro do menu *Mining* > *Bayes*, ilustrado na Figura 58.

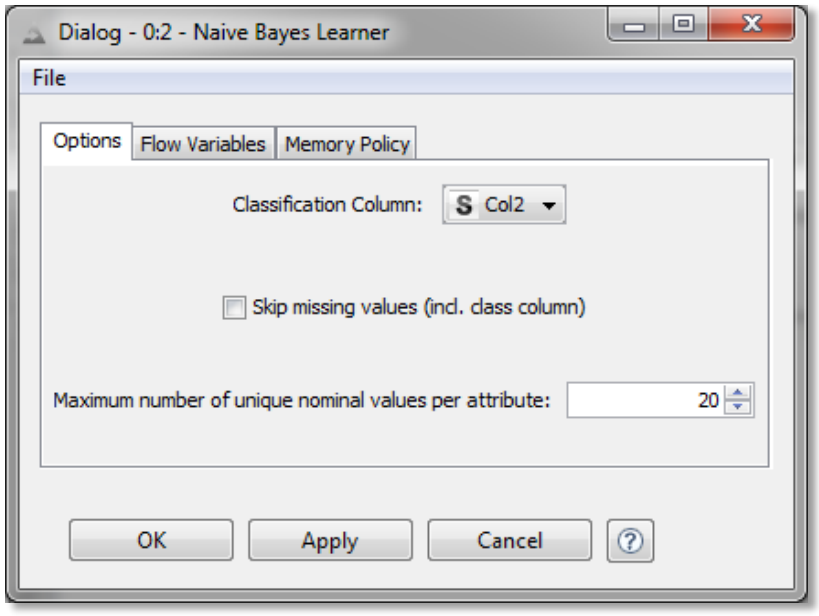
Figura 58 - Aplicação do método de classificação




Fonte: Knime (2013)

Depois disso clica-se em *"Naive Bayes Learner"* com o botão direito do mouse sendo selecionada a opção *"Configuration"*, e escolhe-se qual será a saída em *"Classification Column"* (Figura 59).

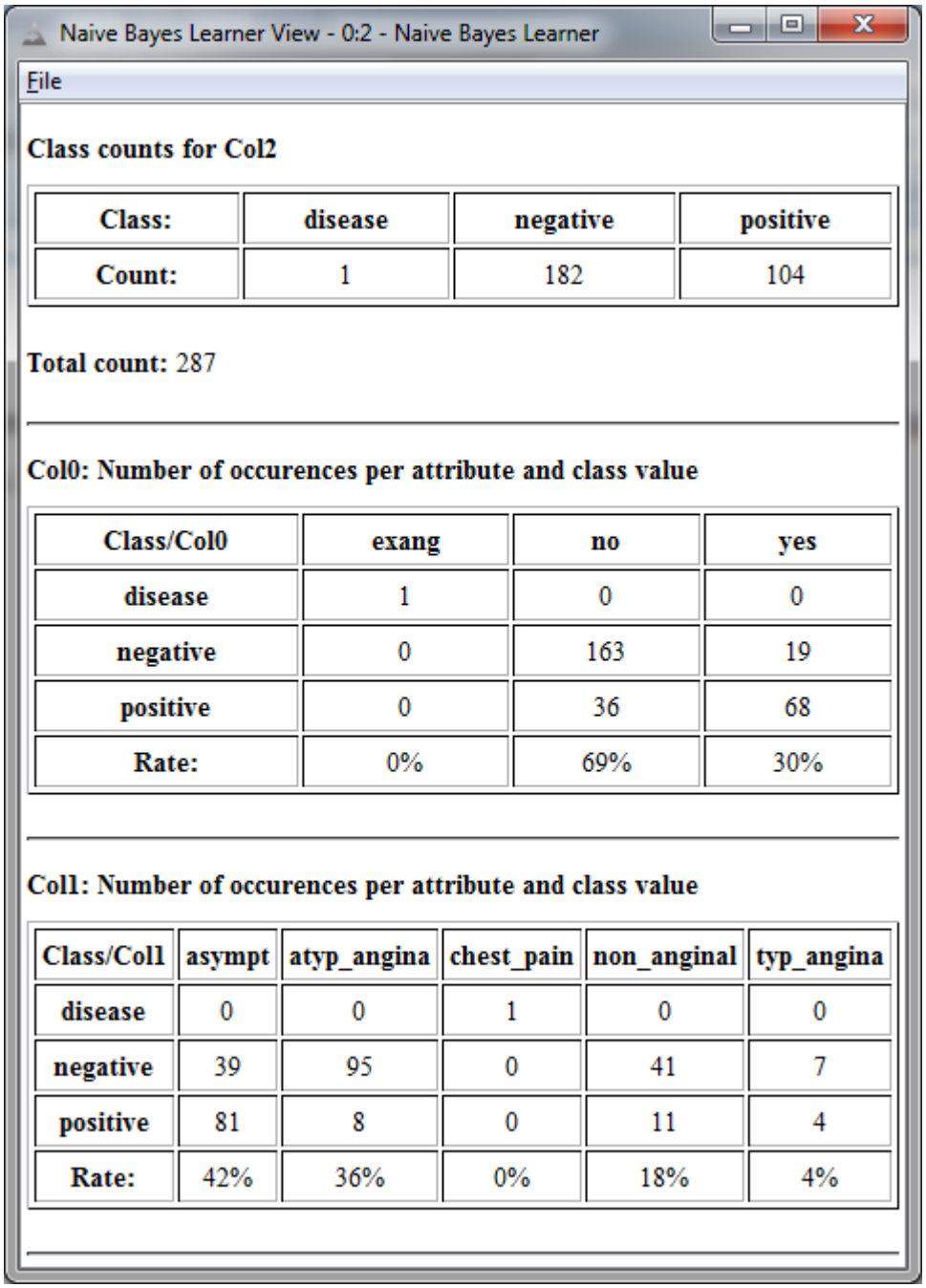
Figura 59 – Configuração do Naive Bayes




Fonte: Knime (2013)

Depois disso clica-se em *“Aply”* e em “OK”. Para finalizar clica-se com o botão direito do mouse e seleciona-se a opção *“Execute and open views”*, que executará e apresentará os resultados, conforme ilustra a Figura 60.

Figura 60 - Execução e apresentação dos resultados




Fonte: Knime (2013)

**APÊNDICE D –** Artigo

# Acurácia dos classificadores bayesianos Naive Bayes, Bayes Net e AODE no apoio ao diagnóstico de osteoporose e osteopenia.

**Leandro Luiz Mazzuchello[1], Everson Bigaton[2], Lucas da Silva Carlessi[3], Ramon Venson[4],Paulo João Martins[5],Priscyla Waleska Targilo de Azevedo Simões[6].**

[1] Curso de Ciência da Computação – Universidade do Sul Catarinense (UNESC) – Criciúma – Santa Catarina.

{leandromaz1988,ramon.venson,pri,pjm,lucascarlessi}@gmail.com

***Resumo*:** *Buscando tratar a incerteza inerente ao diagnóstico biomédico, Nesse contexto, esta pesquisa buscou avaliar a acurácia dos classificadores bayesianos. Os experimentos foram realizados na shell Weka versões 3.6.9 e 3.7.8. Entre as medidas estatísticas analisadas nos experimentos (instâncias classificadas corretamente e incorretamente, estatística Kappa, Erro médio absoluto, sensibilidade, especificidade, precisão recall, medida-F, e AUC), o algoritmo AODE foi considerado o mais acurado, apesar de um pouco mais lento em relação ao tempo de execução.*

***Palavras-chave:*** *Inteligência Computacional. Informática em Saúde. Mineração de Dados. Classificação Bayesiana. Osteoporose.*

***Abstract:*** *Seeking to address the inherent biomedical diagnosis, in this context, uncertainty, this study sought to evaluate the accuracy of Bayesian classifiers. The experiments were performed in Weka shell versions 3.6.9 and 3.7.8. Among the statistical measures analyzed in the experiments (instances classified correctly and incorrectly, Kappa statistic, mean absolute error, sensitivity, specificity, recall accuracy, F-measure, and AUC), the AODE algorithm was considered the most accurate, though a little more slow relative to the runtime*.

***Keywords****: Computational Intelligence. Informatics in Health Data Mining. Bayesian classification. Osteoporosis*.

## 1 Introdução

Ao longo dos anos a Tecnologia da Informação vem contribuindo gradativamente para a saúde, com prioridade à melhoria do acesso aos recursos, fato que foi um dos principais passos para que a saúde e a computação caminhassem juntas, buscando oferecer informações e diagnósticos eficientes por meio do uso das ferramentas tecnológicas (CRAIG; PATTERSON, 2005, tradução nossa).Segundo Shortliffe e Cimino (2006, tradução nossa), dentre as principais áreas de atuação da Informática em Saúde tem-se: os sistemas de informação em saúde; os sistemas de registro eletrônico em saúde; a telessaúde; os sistemas de apoio à decisão; a Mineração de Dados (MD), entre outras. Buscando tratar a incerteza inerente ao diagnóstico biomédico, os classificadores bayesianos são baseados em modelos estatísticos tendo o diferencial em relação aos classificadores clássicos, em determinar a classe a que pertence determinado registro tendo-se como base a probabilidade de um elemento pertencer a determinada classe. Assim, o diagnóstico de doenças como a osteoporose e osteopenia pode se tornar mais rápido e preciso, e particularmente, em que o estado do paciente piora a cada dia agravando o quadro clínico (CHANDRA; PANDAV, 1987, tradução nossa; CHEN et al., 2005, tradução nossa).Como medida de avaliação do conhecimento gerado por um classificador bayesiano tem-se a acurácia que indica a capacidade de um sistema em oferecer um diagnóstico próximo ao real, sendo calculada considerando a sensibilidade e a especificidade. A sensibilidade indica o poder de um algoritmo ou software em classificar como doentes aqueles que realmente são doentes; em contrapartida, a especificidade representa o poder de um aplicativo em sugerir como não doentes, os indivíduos que realmente não apresentam determinada doença. Logo, um software com a alta sensibilidade pode ter baixa especificidade, gerando falsos-positivos levando, por exemplo, indivíduos que não possuem determinada doença a realizar um tratamento desnecessário, o que resultaria em um custo alto e o uso de métodos invasivos de tratamento. Assim, é ideal que um sistema para ser usado na prática clínica apresente acurácia alta (próxima de 100%), resultante de altas taxas de sensibilidade e especificidade combinadas (FLETCHER; SUZANNE, 2006, tradução nossa).

Nesse contexto, pretende-se nesta pesquisa avaliar a acurácia de alguns classificadores bayesianos no apoio ao diagnóstico de osteoporose e osteopenia.

## 2 Metodologia

Este trabalho teve como objetivo avaliar a acurácia de alguns classificadores bayesianos no apoio ao diagnóstico de osteopenia e osteoporose e para isso utilizamos a descoberta de conhecimento em base de dados, tal como sua principal etapa a mineração de dados.

O pré-processamento foi iniciado com a definição do banco de dados que foi submetido ao processo de descoberta de conhecimento em base de dados, sendo realizada a seleção de atributos, limpeza dos dados, binarização e transformações de variáveis. O pré-processamento é fundamental pois aperfeiçoa a etapa central da DCBD que é a mineração de dados (WITTEN; FRANK; HALL, 2011, tradução nossa),

Tendo em vista que a base de dados utilizada nos experimentos foi oriunda de um estudo transversal realizado como Trabalho de Conclusão do Curso de Medicina da Universidade do Extemo Sul Catarinense (SILVA, 2013) com 1871 mulheres submetidas ao exame de densitometria óssea num serviço especializado do sul de Santa Catarina no período de janeiro de 2012 a dezembro de 2012, e aprovado pelo Comitê de ética em Pesquisa da Universidade do Extremo Sul Catarinense sob o protocolo 82939/2012.

Na etapa seguinte foi escolhido os atributos onde a quantidade de variáveis que constituíram o banco de dados foi reduzida de 49 para 42, sendo que as mais associadas ao diagnóstico de osteopenia e osteoporose foram selecionadas (SILVA, 2013). Tal seleção foi feita considerando o objetivo do estudo de Silva (2013) que buscou pesquisar associações multifatoriais a osteoporose e osteopenia entre as variáveis selecionadas.

## 2.4 Limpeza e transformação dos dados

Uma das etapas de suma importância da técnica de descoberta do conhecimento em base de dados é a limpeza e transformação dos dados onde, os registros com informações inconsistentes foram retificados ou excluídos, além disso, verificou-se que algumas variáveis (associadas ao diagnóstico) já estavam dicotomizadas, assim, não havendo a necessidade da binarização.

No estudo de Silva (2013) foram dicotomizadas as variáveis com mais de duas respostas, assim optamos por manter a binarização destas variáveis. Para finalizar esta etapa foi realizada a transformação das variáveis, onde as possíveis respostas de cada atributo foram convertidas de nominal para numérico.

## 2.5 Transformação das Variáveis

Na etapa de transformação das variáveis, campos que continham espaços foram substituídos underline, e campos vazios foram substituídos por pontos de interrogação. Tal funcionalidade foi necessário em virtude das ferramentas utilizadas necessitando de tal ajuste.

## 2.6 Modelo ARFF

Após finalizada a etapa de pré-processamento, a base de dados (.csv) foi exportada para o formato.arff, utilizado na ferramenta onde foram realizados os experimentos de acurácia (Weka). Tal formato (.arff) é dividido em dois cabeçalho principais, @whatever onde são definidas as variáveis e seus atributos, e @data onde são tabulados os dados, sendo que cada informação disponibilizada no @data deve estar de acordo com sua variável correspondente no cabeçalho @whatever.

## 2.7 Mineração de dados

A Mineração de Dados foi iniciada pela seleção das variáveis, seguindo-se da execução dos algoritmos, etapa essa realizada na shell Weka versões 3.6.9 e 3.7.8, pois o algoritmo AODE só está implementado na versão 3.6.9, e a versão 3.7.8 é mais completa nas medidas de avaliação para os demais algoritmos se comparada com a 3.6.9.
Para a mineração de dados utilizou–se apenas as variáveis abordadas no estudo de Silva (2013) fortemente associadas com a osteoporose e osteopenia.Para se minerar a base optou-se

por usar os algoritmos bayesianos disponibilizados na ferramenta Weka: o Naive Bayes, Bayes Net, e AODE, por terem uma acuracidade maior que os demais algoritmos apresentados nas ferramentas.

## 2.8 Descrição dos experimentos

Considerando o estudo prévio que originou a base de dados de nossa pesquisa (SILVA, 2013), foram realizados 327 experimentos: com as variáveis não dicotomizadas e com as dicotomizadas. Os experimentos com as variáveis dicotomizadas consideraram as variáveis apresentadas, que não foram transformadas pela especialista do domínio de aplicação de 2 categorias (apesar de algumas variáveis já possuírem *a priori* duas categorias originais).

Consideraram as variáveis transformadas pela especialista do domínio de aplicação em 2 categorias, quando necessário (apesar de algumas variáveis já possuírem *a priori* duas categorias originais). Tal transformação se deu originalmente no estudo de Silva (2013) pois, além da análise descritiva das variáveis do banco de dados, a pesquisadora realizou também uma análise multivariada por meio da regressão logística, com o objetivo de revelar as variáveis mais fortemente associadas à osteopenia e osteoporose.

Tabela 1 - Características do banco de dados

| **Nome dos atributos** | **Atributos do banco** | **Descrição** | **Numeração atributos** |
|---|---|---|---|
| Numero de identificação* | Protocolo | Numero de identificação | - |
| Prontuário* | Código_URC | Número do prontuário | - |
| Mês* | Mês | Sem descrição | -- |
| Tempo da menopausa* | tempo_menop | tempo da menopausa | - |
| Coluna lombar* | CL_DMO | DMO coluna lombar | - |
| T score da coluna lombar* | CL_T_score | T-score coluna lombar: | - |
| Percentual coluna lombar* | CL_T_percentual | % T-score coluna lombar: % T-score | - |
| Z score coluna lombar* | CL_Z_score | Z-score coluna lombar: | - |
| Percentual da coluna lombar* | CL_Z_percentual | % Z-score DO coluna lombar | - |
| Dmo* | CF_DMO | DMO colo femoral | - |
| T score do colo femural* | CF_T_score | T-score colo femoral | - |
| Percentual T score do colo femural* | CF_T_percentual | % T-score colo femoral: | - |
| Z score do colo femural* | CF_Z_score | Z-score colo femoral | - |
| Percentual do Z-score do colo femural* | CF_Z_percentual | % Z-score colo femoral: | - |
| Total DMO de fêmur* | Femur_total_DMO | DMO fêmur total | - |
| Total T-score Fêmur* | Femur_total_T_score | T-score fêmur total | - |

| | | | |
|---|---|---|---|
| Percentual total de T score de fêmur* | Fêmur_total_T_percentual | % T-score fêmur total | - |
| Total de Z score de fêmur* | Fêmur_total_Z_score | Z-score fêmur total: Z-score | - |
| Percentual de Z score fêmur total* | Fêmur_total_Z_percentual’ | % Z-score fêmur total | - |
| Transformações* | TRANSFORMACOES | SEM DESCRIÇAO | - |
| Índice de massa corporal contínuo* | IMC_cont | IMC continuo | - |
| Índice de massa corporal qualitativa* | IMC_qualitativa | classificação do IMC | - |
| Diagnostico+ | DESFECHO_1 | normal osteopenia e osteoporose | 1 normal, 2-osteopenia e osteoporose |
| Idade até 49* | idade_50 | Idade até 49 | |
| Diagnóstico de osteoporose* | desfecho_osteoporose | Osteoporose | - |
| Diagnóstico de osteopenia* | desfecho_osteopenia | Osteopenia | 1-Osteopenia, 2-normal |
| Percentual do grupos de idades* | NIdade | Percentual do grupo de idades | - |
| Idade até 51* | Idade_51 | Sem descrição | - |
| Fratura+ | fratura_1 | fratura dicotômica | - |

Fonte : dados pesquisa

*- variáveis não dicotomizadas

+ - variáveis dicotomizadas

Tabela 2 - Características do banco de dados

| **Nome dos atributos** | **Atributos do banco** | **Descrição** | **Numeração atributos** |
|---|---|---|---|
| Fratura do fêmur+ | Fratura_femur_1 | FEMUR_ dicotômica | - |
| Fratura de coluna+ | Fratura_coluna_1 | coluna dicotômica | - |
| Fratura do antebraço+ | Fratura_antebraço_1 | antebraço dicotômica | - |
| Fratura da costela+ | Fratura_costela_1 | costela dicotômica | - |
| Menopausa+ | Menopausa_1 | Menopausa dicotômica | - |
| Uso de cálcio+ | Uso_calcio1 | Cálcio dicotômico | - |
| Tratamento hormonal+ | Trh_1 | TRH dicotômico | - |
| Ooforectomia+ | Ooforectomia_1 | ooforectomia dicotômica | - |
| Histerectomia+ | Histerectomia_1 | histerectomia dicotômica | - |
| Idade dicotomizada+ | Idade_1 | idade_dicotomica_media | - |
| IMC dicotomizado+ | IMC_dict | IMC dicotomizado até 25 e 25 ou mais | - |
| Medicamento tireoide+ | Medic_tireode_1 | medicamento para tireoide | 1-sim, 2 não |
| Diagnostico dicotomizado+ | Desfecho_2 | desfecho dicotômico | |
| Idade * | Idade | Idade dos indivíduos por faixa etária | - |
| Altura* | Altura | Altura dos individuo por faixa etária | - |
| Peso* | Peso | Media de peso por faixa etária | - |
| Fratura* | Fratura | Contém fratura | 1-Sim,2-não |
| Quantidade de fratura* | Qt_fratura | Quantidades de fraturas prévias | - |
| Fratura de fêmur* | Fratura_ fêmur | Fratura prévia em colo do fêmur | 1-Sim,2-não |
| Fratura de coluna* | Fratura_coluna | Fratura prévia na coluna | 1-Sim,2-não |
| Fratura de antebraço* | Fratura_antebraço | Fratura prévia em antebraço | 1-Sim,2-não |
| Fratura_costela* | Fratura_costela | Fratura prévia em costela | 1-Sim,2-não |
| Tratamento* | tratamento | Uso de medicamentos | 1-Sim,2-não |

Fonte :dados da pesquisa

*- variáveis não dicotomizada

+ - variáveis dicotomizadas

# 3 Conclusão

Tabela 3 - Comparativo da medidas estatísticas

| Descrição | Algoritmo | Média | IC | Mediana | DP | Min | Max | II |
|---|---|---|---|---|---|---|---|---|
| **Instâncias classificadas corretamente** | Bayes Net | 82,54 | 79,53-85,54 | 84,06 | 14,76 | 41,04 | 99,67 | 26,66 |
| | Naive Bayes | 82,84 | 79,97-85,72 | 83,99 | 14,42 | 41,20 | 99,57 | 26,61 |
| | AODE | 84,52 | 81,69-87,34 | 89,65 | 14,08 | 42,81 | 99,84 | 23,08 |
| **Instâncias classificadas incorretamente** | Bayes Net | 17,46 | 14,45-20,47 | 15,93 | 14,76 | 0,32 | 58,95 | 26,66 |
| | Naive Bayes | 17,15 | 14,27-20,03 | 16,00 | 14,42 | 0,43 | 58,79 | 26,61 |
| | AODE | 15,48 | 12,66-18,30 | 10,35 | 14,08 | 0,16 | 57,19 | 23,08 |
| **Kappa** | Bayes Net | 0,33 | 0,27-0,40 | 0,28 | 0,28 | 0,00 | 0,95 | 0,43 |
| | Naive Bayes | 0,31 | 0,26-0,37 | 0,26 | 0,28 | 0,00 | 0,96 | 0,40 |
| | AODE | 0,34 | 0,28-0,40 | 0,29 | 0,29 | -0,01 | 0,97 | 0,45 |
| **Erro médio-absoluto** | Bayes Net | 0,19 | 0,16-0,22 | 0,17 | 0,14 | 0,00 | 0,43 | 0,27 |
| | Naive Bayes | 0,19 | 0,16-0,22 | 0,16 | 0,14 | 0,01 | 0,43 | 0,29 |
| | AODE | 0,18 | 0,15-0,21 | 0,16 | 0,14 | 0,00 | 0,43 | 0,26 |
| **Sensibilidade** | Bayes Net | 0,82 | 0,79-0,85 | 0,84 | 0,15 | 0,41 | 0,80 | 0,27 |
| | Naive Bayes | 0,83 | 0,80-0,86 | 0,84 | 0,14 | - | 0,99 | 0,27 |
| | AODE | 0,84 | 0,82-0,87 | 0,90 | 0,14 | 0,43 | 0,99 | 0,23 |
| **Especificidade** | Bayes Net | 0,53 | 0,48-0,59 | 0,51 | 0,27 | 0,00 | 0,95 | 0,49 |
| | Naive Bayes | 0,51 | 0,46-0,57 | 0,49 | 0,27 | 0,00 | 0,96 | 0,49 |
| | AODE | 0,49 | 0,43-0,55 | 0,48 | 0,29 | 0,00 | 0,95 | 0,52 |
| **Precisão** | Bayes Net | 0,81 | 0,78-0,85 | 0,81 | 0,16 | 0,38 | 0,99 | 0,32 |
| | Naive Bayes | 0,82 | 0,78-0,85 | 0,82 | 0,16 | 0,37 | 0,99 | 0,31 |
| | AODE | 0,83 | 0,79-0,86 | 0,88 | 0,17 | 0,12 | 0,99 | 0,28 |
| **Recall** | Bayes Net | 0,82 | 0,79-0,85 | 0,84 | 0,15 | 0,41 | 0,99 | 0,27 |
| | Naive Bayes | 0,83 | 0,80-0,86 | 0,84 | 0,14 | 0,41 | 0,99 | 0,27 |
| | AODE | 0,84 | 0,82-0,87 | 0,90 | 0,14 | 0,42 | 0,99 | 0,23 |
| **Medida-F** | Bayes Net | 0,81 | 0,78-0,84 | 0,82 | 0,16 | 0,38 | 0,99 | 0,29 |
| | Naive Bayes | 0,81 | 0,78-0,84 | 0,82 | 0,16 | 0,37 | 0,99 | 0,27 |
| | AODE | 0,83 | 0,80-0,86 | 0,89 | 0,15 | 0,39 | 0,99 | 0,29 |
| **AUC** | Bayes Net | 0,79 | 0,75-0,82 | 0,85 | 0,17 | 0,04 | 0,99 | 0,24 |
| | Naive Bayes | 0,79 | 0,76-0,89 | 0,83 | 0,17 | 0,04 | 0,99 | 0,27 |
| | AODE | 0,81 | 0,78-0,85 | 0,87 | 0,16 | 0,04 | 0,99 | 0,26 |
| **Tempo** | Bayes Net | 0,02 | 0,01-0,02 | 0,01 | 0,02 | 0,00 | 0,12 | 0,00 |
| | Naive Bayes | 0,00 | 0,00-0,00 | 0,00 | 0,01 | 0,00 | 0,04 | 0,00 |
| | AODE | 0,00 | 0,00-0,01 | 0,00 | 0,01 | 0,00 | 0,04 | 0,01 |

IC: Intervalo de confiança / DP: Desvio padrão / II: Intervalo de confiança
Fonte: Dados da pesquisa

A mineração de dados é uma técnica amplamente utilizada em várias áreas de conhecimento e busca extrair conhecimento implícito de bases de dados.

Com o desenvolvimento dessa pesquisa pôde-se compreender que o classificador bayesiano usa o teorema de bayes e se destaca pela eficiência, velocidade e acurácia, possibilitando sua utilização no processo de diagnose.

Esta pesquisa possibilitou também compreender os testes diagnósticos, que possibilitam medir a eficiência dos classificadores, podendo-se destacar a sensibilidade, especificidade, entre outros.

Conclui-se também que esta pesquisa atendeu os principais aspectos éticos e legais pertinentes à informática em saúde, pois o projeto foi aprovado pelo Comitê de Ética e Pesquisa em Seres Humanos da Universidade do Extremo Sul Catarinense.

Mediante os estudos realizados, pôde-se compreender o funcionamento de alguns algoritmos de classificação bayesiana, entre eles, o AODE, o NaiveBayes, e o Bayes Net, que foram utilizados na mineração de uma base de dados de osteopenia e osteoporose. Para tal, foi utilizada a shell Weka nos experimentos realizados.

Para realização dos experimentos, foram apreendidos conceitos sobre osteopenia e osteoporose, seus principais sintomas, aspectos epidemiológicos, sendo mais prevalente em pessoas com idade acima dos 50 e mulheres.

Nos experimentos foram utilizados os três algoritmos supracitados, sendo geradas 327 minerações, os quais foram utilizadas no pós-processamento e resultaram na análise da acurácia.

Entre as medidas estatísticas analisadas nos experimentos (instâncias classificadas corretamente e incorretamente, estatística Kappa, Erro médio absoluto, sensibilidade, especificidade, precisão recall, medida-F, e AUC), o algoritmo AODE foi considerado o mais acurado, apesar de um pouco mais lento em relação ao tempo de execução.

Essa pesquisa teve algumas limitações, como a impossibilidade de utilizar as ferramentas Tanagra e Knime pelas poucas medidas de avaliação presentes. Outra limitação foi a de que o algoritmo AODE só está implementado na versão 3.6.9 do Weka e não possui as medidas de avaliação que a versão 3.7.8 oferece, por isso utilizamos as duas versões no estudo.

## Referencias

Craig, J. and Patterson, V. (2005) Introduction to the practice of telemedicine. J Tele med Telecare, United States, v. 11, n.1,

Shortliffe, Edward Hance and Cimino, James J. (2006) Biomedical informatics: computer applications in health care and biomedicine. 3ed.New York, USA: Springer.

Chandra, V. and Pandav, C. S. (1987) The importance of non-communicable diseases in developing countries (editorial). Indian J Community Med, v.12, p.80-178,

Chen, Jianjun et al.(1987)The small introns of antisense genes are better explained by selection for rapid transcription than by 'genomic design'. A Head Of Print, United States, p.1-21, 02 set.

Fletcher, R. H.and Suzanne, W.( 2006) Epidemiologia clínica: elementos essenciais. Artmed. 4ed. Porto Alegre,

Witten, Ian H., Frank, Eibe and Hall, Mark A.( 2011) Data mining practical machine learning tools and techniques. 3. ed. Burlington: Morgan Kaufmann.

Silva, Ana Carolina Veiga.(2013) Fatores associados à osteopenia e osteoporose em mulheres submetidas à densitometria óssea. 2013. 119 f. Monografia de Conclusão de Curso. Curso de Medicina. Universidade do Extremo Sul Catarinense, Criciúma.

# ANEXOS

**ANEXO A –** Aprovação do projeto no Comitê de Ética e Pesquisa em Seres Humanos

**Universidade do Extremo Sul Catarinense UNESC**
**Comitê de Ética em Pesquisa- CEP**

**Resolução**

Comitê de Ética em Pesquisa, reconhecido pela Comissão Nacional de Ética em Pesquisa (CONEP)/Ministério da Saúde analisou o projeto abaixo.

**Projeto: 153/2009**

**Pesquisador:**
Maria Inês da Rosa
Bruno Silva

**Título: "Prevalência e fatores associados ao tabagismo entre universitários no sul do Brasil".**

Este projeto foi Aprovado em seus aspectos éticos e metodológicos, de acordo com as Diretrizes e Normas Internacionais e Nacionais. Toda e qualquer alteração do Projeto deverá ser comunicado ao CEP. Os membros do CEP não participaram do processo de avaliação dos projetos onde constam como pesquisadores

Criciúma, 09 de setembro de 2009.

*Mágada T. Schwalm*

Coordenadora do CEP