Merisandra Côrtes de Mattos Garcia

AVALIAÇÃO DE MÉTODOS DE DATA MINING E REGRESSÃO LOGÍSTICA APLICADOS NA ANÁLISE DE TRAUMATISMO CRANIOENCEFÁLICO GRAVE

Tese submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Santa Catarina para obtenção do Grau de Doutor em Engenharia Elétrica.

Orientador: Prof. Dr. Fernando Mendes de Azevedo
Coorientador: Evandro Tostes Martins, MSc.

Florianópolis
2015

Ficha de identificação da obra elaborada pelo autor
através do Programa de Geração Automática da Biblioteca Universitária
da UFSC.

Garcia, Merisandra Côrtes de Mattos
Avaliação de Métodos de Data Mining e Regressão Logística Aplicados na Análise de Traumatismo Cranioencefálico Grave / Merisandra Côrtes de Mattos Garcia ; orientador, Fernando Mendes de Azevedo ; coorientador, Evandro Tostes Martins. - Florianópolis, SC, 2015.
182 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia Elétrica.

Inclui referências

1. Engenharia Elétrica. 2. Data Mining. 3. Classificação. 4. Regressão Logística Binária. 5. Traumatismo Cranioencefálico Grave. I. Azevedo, Fernando Mendes de. II. Martins, Evandro Tostes. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia Elétrica. IV. Título.

Merisandra Côrtes de Mattos Garcia

AVALIAÇÃO DE MÉTODOS DE *DATA MINING* E REGRESSÃO LOGÍSTICA APLICADOS NA ANÁLISE DE TRAUMATISMO CRANIOENCEFÁLICO GRAVE

Esta Tese foi julgada adequada para obtenção do Título de "Doutor", e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica.

Florianópolis, 24 de março de 2015.

Prof. Carlos Galup Montoro, Dr.
Coordenador do Curso

Banca Examinadora:

Prof. Fernando Mendes de Azevedo, Dr.
Orientador
Universidade Federal de Santa Catarina

Prof. Julio Cesar Nievola, Dr.
Pontifícia Universidade Católica do Paraná

Prof.ª Anita Maria da Rocha Fernandes, Dr.ª
Universidade do Vale do Itajaí

Prof.ª Silvia Modesto Nassar, Dr.ª
Universidade Federal de Santa Catarina

Prof.ª Daniela Ota Hisayasu Suzuki, Dr.ª
Universidade Federal de Santa Catarina

Prof. Jefferson Luiz Brum Marques, PhD.
Universidade Federal de Santa Catarina

Dedico aos meus amados pais, Tereza Lenir e Libanio, e ao meu esposo Rodrigo.

AGRADECIMENTOS

No momento de escrita destes agradecimentos revivi os mesmos sentimentos de quando finalizei o Mestrado, parece que foi ontem, mas já se passaram 14 anos, no entanto, a mesma emoção me invade agora. Quando chegamos neste momento é porque conseguimos, os longos anos de trabalho e estudo no Doutorado estão sendo finalizados, encerra-se um ciclo na vida e começa-se outro.

No encerramento deste ciclo quero agradecer aos presentes em toda a minha vida, como também aos que conheci ao longo dessa caminhada.

Agradeço primeiramente a Deus, pois quando tudo parecia impossível me deu serenidade para seguir em frente e vencer as dificuldades.

Aos meus pais, Tereza Lenir e Libanio, por sua sabedoria em educar-me e pelas lições de vida, esperança e espiritualidade. Obrigada por acreditarem sempre na educação e que o conhecimento é capaz de transformações, por terem compartilhado das minhas angústias e conquistas.

A minha amada irmã Maristane, sempre atenta a me ouvir e ajudar, ao meu cunhado Paulo que é como um irmão, e aos meus sobrinhos Gabriel e Rafael, a vocês obrigada pelo carinho, incentivo e apoio.

Ao meu amado esposo Rodrigo, companheiro de todas as horas, que durante a elaboração desta tese, sempre ficou ao meu lado, dando todo o suporte necessário. Obrigada por ter me feito companhia e cafés ao longo das madrugadas e finais de semana em que eu trabalhava na tese, você nunca foi dormir antes de mim, ficava sentado ao meu lado, me assistindo trabalhar. Ao Noah que também foi meu companheiro nessas madrugadas, ele que é capaz de me entender sem eu dizer uma única palavra.

À Universidade do Extremo Sul Catarinense (UNESC), instituição na qual trabalho há 14 anos, muito obrigada pela confiança, apoio, incentivo e pela liberação para eu cursar o Doutorado, em especial a Diretora da Unidade Acadêmica de Ciências, Engenharias e Tecnologias Profa. Angela Costa Piccinini, aos Coordenadores do Curso de Ciência da Computação e amigos Rogério Antônio Casagrande e Ana Claudia Garcia Barbosa.

À secretária do Curso de Ciência da Computação e amiga Margarete Dagostim, que juntamente com os professores Rogério e Ana Claudia me ajudaram neste percurso e entenderam as minhas ausências.

Ao Fábio Bif Goularte, Ana Claudia e Rogério que me deram o apoio necessário, substituindo-me na disciplina de Trabalho de Conclusão de Curso I, enquanto eu cursava as disciplinas do Doutorado.

Aos meus colegas professores do Curso de Ciência da Computação da UNESC pelo apoio e incentivo.

Aos meus alunos da 7ª, 8ª e 9ª fase do curso de Ciência da Computação da UNESC que durante os anos do Doutorado, souberam entender as minhas ausências e ansiedades, vocês são um aprendizado constante na minha carreira de professora.

Aos meus alunos de iniciação científica, José Márcio Cassettari Júnior, Maicon Bastos Palhano, Gabriel Felippe, Ruano Marques Pereira e Pedro Arns Júnior, obrigada pelo apoio durante esta tese, formamos uma equipe.

Aos meus alunos de extensão universitária, Diego Buz Fernandes, Tiago Rodrigo da Silva, Nádia Soraida Mateus Pessoa, Allan Januário Ramos, Tiago Aleff da Silva e Leandro Justin Vieira, obrigada pela dedicação e apoio.

Aos meus alunos da terceira idade, do projeto de extensão universitária "Informática para a Melhor Idade", como tem sido gratificante e enriquecedora essa experiência, muito obrigada pelo carinho.

À Universidade Federal de Santa Catarina, ao Programa de Pós Graduação em Engenharia Elétrica e ao Instituto de Engenharia Biomédica por esta oportunidade de aprendizado.

Ao Prof. Dr. Fernando Mendes de Azevedo pela disponibilidade em ter aceitado me orientar, pela confiança, compartilhamento do seu conhecimento e orientações durante esta pesquisa.

Ao Prof. Dr. Roger Walz por me apresentar os dados de Traumatismo Cranioencefálico Grave e ao coorientador desta pesquisa, o médico Evandro Tostes Martins pelo apoio e disponibilização dos dados para a realização desta tese e por ter me possibilitado dar continuidade a sua pesquisa.

Aos professores Dr. Júlio Cesar Nievola, Dra. Daniela Ota Hisayasy Suzuki e PhD. Jefferson Luiz Brum Marques pelas sugestões e comentários na qualificação e defesa.

À Profa. Dra. Anita Maria da Rocha Fernandes que me acompanha desde a graduação, uma das principais incentivadoras da minha vida acadêmica, muito obrigada pela amizade e direcionamentos no decorrer desta pesquisa.

À Profa. Dra. Silvia Modesto Nassar que me acompanha desde o Mestrado, muito obrigada pela amizade, acolhimento, olhar carinhoso e pelas contribuições que enriquecem esta pesquisa.

Aos colegas do Instituto de Engenharia Biomédica e do Laboratório de Informática em Saúde, em especial a Christine Fredel Boos, William Alberto Cruz Castañeda e Cristhian Heck pela amizade e companheirismo nesses anos de Doutorado.

Ele faria da queda um passo de dança, do medo uma escada, do sono uma ponte, da procura um encontro.

(Fernando Sabino, 1956)

RESUMO

O traumatismo cranioencefálico é um problema de saúde pública constituindo-se em uma das principais causas de morbidade e mortalidade no Brasil e no mundo. A análise das relações entre as suas consequências tem despertado interesse em pesquisas na área, a fim de se identificar os indicadores que auxiliam no seu prognóstico, buscando-se evitar o óbito. Estes modelos são tradicionalmente gerados por meio da regressão logística que tem se constituído em uma técnica padrão para análise dos dados em saúde. No entanto, os modelos prognósticos em traumatismo cranioencefálico, como o grave que é o foco desta pesquisa, não conseguem acurácia elevada para a predição do óbito por meio da regressão logística. Sabendo-se disso, avanços em termos da acuracidade da predição podem auxiliar no prognóstico e conduta das pessoas acometidas por traumatismo cranioencefálico do tipo grave. A descoberta de conhecimento em bases de dados por meio da etapa de *data mining* e da integração de técnicas de diferentes áreas como inteligência computacional, reconhecimento de padrões, aprendizado de máquina, estatística e banco de dados, constitui-se em uma alternativa para identificar as relações nestes conjuntos de dados. Considerando-se isto, esta pesquisa consiste na avaliação comparativa de diferentes métodos de *data mining*, a fim de se analisar os modelos gerados e compará-los com o de regressão logística, em uma mesma população de estudo. Nesta pesquisa, se objetiva identificar padrões válidos, avaliando se os métodos de *data mining* empregados se mostram como uma alternativa à regressão logística, baseando-se em critérios de avaliação como acurácia e robustez, os quais se constituem em medidas de qualidade dos padrões descobertos. Os métodos de *data mining* empregados referem-se a indução de árvores de decisão por meio dos algoritmos C4.5 e *Classification And Regression Trees*; o aprendizado baseado em instâncias pelo algoritmo k-vizinhos mais próximos; as redes neurais artificiais por Funções de Base Radial; os classificadores bayesianos pelos algoritmos Naive Bayes e Redes de Crença Bayesiana e o metaclassificador pelo algoritmo *Adaptive Boosting*. No desenvolvimento foram gerados modelos de prognóstico do óbito em traumatismo cranioencefálico grave por meio dos algoritmos supracitados, como também pela regressão logística binária. Os modelos gerados na etapa de *data mining* foram comparados aplicando-se as medidas de avaliação de desempenho (verdadeiros positivos, verdadeiros negativos, acurácia, sensibilidade e especificidade) e de confiabilidade (coeficiente de concordância kappa e área sob a *Receiver-*

Operating Characteristic Curve). Na comparação entre os modelos de *data mining* elencados com maior poder de discriminação em relação a regressão logística, utilizaram-se as medidas de confiabilidade citadas anteriormente, considerando-se Intervalos de Confiança de 95%. Dentre as análises realizadas, nos modelos gerados para predição do óbito em traumatismo cranioencefálico grave, os classificadores bayesianos destacaram-se apresentando medidas de desempenho significativamente mais representativas. O modelo gerado pelo algoritmo Naive Bayes destacou-se em relação aos demais métodos de *data mining* empregados, bem como quando comparado com o modelo de regressão logística binária, classificando corretamente o óbito em 58,2% (IC95%: 55,6-61,8), a acurácia geral do modelo foi de 80,2% (IC95%: 76,9-85,7), sensibilidade de 72,7% (IC95%: 69,8-75,4), especificidade de 84,2% (IC95%: 81,6-87,5), área sob a *Receiver-Operating Characteristic Curve* de 0,851 (IC95%: 0,832-0,870) e coeficiente de concordância Kappa 0,530 (IC95%: 0,519-0,541). Comparando-se os resultados, o algoritmo Naive Bayes mostrou-se, no conjunto de dados estudado, significativamente mais representativo que o modelo de regressão logística binária e os outros modelos de *data mining*.

Palavras-chave: Inteligência Computacional. *Data Mining*. Classificação. Regressão Logística Binária. Avaliação Comparativa. Traumatismo Cranioencefálico Grave.

ABSTRACT

Traumatic brain injury is a public health problem thus becoming a major cause of morbidity and mortality in Brazil and worldwide. The analysis of relations between its consequences has stimulated researches in the area, in order to identify indicators that help its prognosis, seeking avoid death. These models are traditionally generated by logistic regression that has been constituted as a standard technique for analysis of health data. However, the prognostic models in traumatic brain injury, such as severe which is the focus of this research, can not have a high accuracy for prediction of death by logistic regression. Knowing this, advances in terms of prediction accuracy may aid in prognosis and management of people affected by severe brain injury. The knowledge discovery in databases by data mining step and integration of techniques from different areas such as computational intelligence, pattern recognition, machine learning, statistical and database, constitutes an alternative to identify relationships in the data sets. Considering this, this research consists on the comparative evaluation of different data mining methods in order to analyze the generated models and compare them with logistic regression, in the same study population. In this research, the objective is to identify valid standards, assessing whether the data mining methods used are shown as an alternative to logistic regression, based on evaluation criteria such as accuracy and robustness, which constitute quality measures of the discovered patterns. The data mining methods employed refer to decision tree induction through C4.5 algorithms and Classification And Regression Trees; learning based on instances by k-nearest neighbors algorithm; artificial neural networks Radial Basis Function; Bayesian classifiers by algorithms Naive Bayes and Bayesian Belief Networks and the metaclassificador by Adaptive Boosting algorithm. In the development were generated death of prognostic models in severe traumatic brain injury through the aforesaid algorithms, but also by binary logistic regression. The models in data mining stage were compared applying the performance evaluation measures (true positives, true negatives, accuracy, sensitivity and specificity) and reliability (kappa coefficient and area under the Receiver Operating Characteristic Curve). Comparing the data mining models listed with major discrimination in relation to logistic regression, we used the reliability of measurements mentioned above, considering 95% confidence intervals. Among the analyzes, the generated models for prediction of death in severe traumatic brain injury, the Bayesian classifiers stood out, presenting performance measures significantly

more representative. The model generated by Naive Bayes algorithm stood out in relation to other data mining methods employed, as well as when compared to the binary logistic regression model, correctly classifying the death in 58,2% (CI95%: 55,6-61,8), the overall accuracy of the model was 80,2% (CI95%: 76,9-85,7), sensitivity of 72,7% (CI95%: 69,8-75,4), specificity of 84,2% (CI95%: 81,6-87,5), area under the Receiver Operating Characteristic Curve of 0,851 (CI95%: 0,832-0,870) and Kappa coeficient of agreement 0,530 (CI95%: 0,519-0,541). Comparing the results, the Naive Bayes algorithm proved, in the data set studied, significantly more representative than the model of binary logistic regression and other data mining models.

Keywords: Computacional Intelligence. Data Mining. Classification. Binary Logistic Regression. Comparative Evaluation. Severe Traumatic Brain Injury.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BTCR – Boosted Tree Classifiers and Regression
CART – Classification and Regression Trees
CFS – Correlation based Feature Selection
CHAID – Chi-squared Automatic Interaction Detector
DM – Data Mining
GCS – Escala de Coma de Glasgow
E-CHAID – Exhaustive CHAID
ERG – Escala de Resultados de Glasgow
KDD – Knowledge Discovery in Databases
QUEST – Quick Unbiased Efficient Statistical Tree
MLP – Multi-Layer Perceptrons
OR – Odds Ratio
RBF – Função de Base Radial
RFRC – Random Forest Regression and Classification
RNA – Redes Neurais Artificiais
TCE – Traumatismo Cranioencefálico

SUMÁRIO

1 INTRODUÇÃO

A velocidade e o volume de dados gerados pelas instituições nos mais variados segmentos aumentam gradativamente. Os avanços tecnológicos têm facilitado o seu armazenamento, porém a análise desta quantidade de dados tornou-se complexa para a capacidade humana.

Ferramentas estatísticas, modelos matemáticos, consultas estruturadas são comumente utilizadas para auxiliar na obtenção de informações a partir das bases de dados. No entanto, estas ferramentas possuem limitações que podem comprometer a precisão da informação gerada. Neste contexto surge a descoberta de conhecimento em bases de dados que reúne técnicas de diferentes áreas como inteligência computacional, reconhecimento de padrões, aprendizado de máquina, métodos estatísticos e banco de dados a fim de identificar padrões nos dados (GARCIA; MARTINS; AZEVEDO, 2013a).

Uma das principais etapas da descoberta de conhecimento em bases de dados e que muitas vezes se confunde com esta é o *data mining* (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; GOLDSCHMIDT; PASSOS, 2005; HAN; KAMBER; PEI, 2011). O *data mining* tem sido usado para explorar a informação em busca de conhecimento nas mais diferentes áreas, como por exemplo, na biomédica a fim de prever a taxa de sobrevivência em pacientes com câncer de mama, identificar os preditores das infecções do trato urinário, resultados de traumatismo cranioencefálico (LEE et al, 2011).

O Traumatismo Cranioencefálico é um problema de saúde pública considerado uma das principais causas de morbidade[1] e mortalidade[2] no mundo, acometendo a maioria das vítimas em idade produtiva (SAATMAN et al, 2008; WINN et al, 2011), sendo considerada uma epidemia silenciosa (MILLER, 1986; OLIVEIRA et al, 2012).

A análise das relações entre as consequências do traumatismo cranioencefálico e as condições pré-mórbidas, como também a sua gravidade têm sido alvo de estudos e interesses dos pesquisadores que visam determinar indicadores que possibilitem conhecer precocemente o prognóstico de uma vítima desta enfermidade (DIJKERS, 2004; LIN et al, 2010; VIEIRA et al, 2013).

---

[1] Importante indicador de saúde que designa o conjunto de casos de uma dada doença ou a soma de agravos à saúde que atingem um grupo de indivíduos (MENEZES; SILVA, 2001).

[2] Número de óbitos em determinado período ou população (FERREIRA, 2010), sendo uma importante fonte para avaliar as condições de saúde da população (MENEZES; SILVA, 2001).

O prognóstico refere-se a perspectiva de recuperação de uma doença ou estado patológico (OHNO-MACHADO; RESNIC; MATHNE, 2006), constituindo-se em integrante do suporte ao diagnóstico e do programa terapêutico (HANSEN et al, 2011; LUCAS; ABU-HANNA, 1999).

As pessoas acometidas por Traumatismo Cranioencefálico, como o grave que é o foco desta pesquisa, são atendidas em Unidades de Tratamento Intensivo, tendo-se a Medicina Intensiva como uma especialidade heterogênea, complexa e em evolução, que se volta ao diagnóstico, monitorização e tratamento das pessoas com doença grave a fim de recuperá-las para o seu estado de saúde e de qualidade de vida anteriores (BLANCH et al, 2013), portanto, busca-se evitar o óbito. Esta é uma das áreas da Medicina em que o desenvolvimento de modelos para predição do desfecho final merece mais atenção por parte dos pesquisadores (BUENO et al, 2005; SILVA, 2007).

O interesse nesses modelos, que são importantes para o prognóstico de pessoas com doenças graves, e na Medicina Baseada em Evidências, aumenta a necessidade deste tipo de ferramenta de suporte à decisão médica (ABU-HANNA; LUCAS, 2001; WINDELER, 2000).

Neste contexto, pesquisas voltam-se a elaboração dos modelos prognósticos, destacando-se como mais relevantes aqueles voltados a mortalidade, ao tempo de internamento na Unidade de Terapia Intensiva (UTI) ou no hospital e a qualidade de vida pós-UTI (PADILHA et al, 2009; SILVA, 2007; SILVA et al, 2014). A escolha dentre estes é resultante da importância atribuída pelos clínicos e o quão são mensuráveis (ROSENBERG, 2002).

Os modelos prognósticos de doenças graves visam estimar um determinado desfecho, como por exemplo, a probabilidade de mortalidade, estando entre os mais utilizados e testados na Medicina (OHNO-MACHADO; RESNIC; MATHNE, 2006).

A preocupação com o prognóstico também é presente nos familiares das pessoas que sofreram lesões cranioencefálicas, porém os métodos atuais para a determinação deste prognóstico são imperfeitos em função da heterogeneidade dos dados dos pacientes, variedade das causas dos traumas e outros fatores como idade e prevalência de doenças sistêmicas. Portanto, predizer o desfecho de um Traumatismo Cranioencefálico é um processo complexo e cognitivo (SUT; SIMSEK, 2011).

Os modelos prognósticos baseiam-se em um conjunto de variáveis preditivas e equação modeladora, tradicionalmente obtida por meio da regressão logística (SILVA, 2007; THEODORAKI et al, 2010).

Especialmente naqueles modelos que visam a predição de desfechos dicotômicos, como é o caso da mortalidade.

A regressão logística apresenta como vantagem a fácil interpretação, disponibilização de *Odds Ratio*[3] e respectivos intervalos de confiança, bem como indica as variáveis fundamentais para o cálculo do desfecho de interesse (HOSMER; LEMESHOW, 2000). Dentre as desvantagens desta técnica tem-se:

a) o conhecimento *a priori* do desenvolvedor como fundamental para a seleção das variáveis com maior probabilidade de serem preditivas para o desfecho de interesse (BUCHMAN et al, 1994);
b) pode não modelar adequadamente as relações complexas, não lineares e interdependentes pertinentes aos sistemas biológicos (ROSENBERG, 2002);
c) presença de amostras de tamanho pequeno podem gerar modelos de desempenho fraco (SILVA, 2007);
d) não são ideais para trabalhar com dados biológicos complexos, por vezes multidimensionais e armazenados em grandes repositórios de dados, tornando demorado o processo de análise (THEODORAKI et al, 2010).

A complexidade dos dados biológicos impõe desafios para a integração e descoberta de conhecimento potencialmente útil relativo ao prognóstico das doenças, como a aplicação de técnicas de aprendizado de máquina para a identificação de padrões, modelos de predição e classificação a partir de bases de dados clínicas.

Diferentemente dos métodos estatísticos habituais que se baseiam em hipótese e teste, bem como a estrutura matemática dos modelos é fornecida, na abordagem de algumas técnicas de aprendizado de máquina a estrutura é apreendida automaticamente (LUCAS, 2004; TAN; STEINBACH; KUMAR, 2009).

No que se refere a modelos estatísticos em Traumatismo Cranioencefálico tem-se várias iniciativas que empregam a regressão logística, como por exemplo as pesquisas de Bernal et al (2013), Cardozo Junior e Silva (2014), Duncan et al (2011), Guerra et al (2010), Lingsma et al (2010), Martins et al (2009), Murray et al (2007), Mushkudiani et al (2007) e Tjahjadi et al (2013).

[3] Medida de associação entre um indivíduo exposto ter a doença, comparado com a do não exposto (LATORRE, 2004).

Aplicações de *data mining* em Traumatismo Cranioencefálico são abordadas nas pesquisas de Andrews et al (2002), Chesney et al (2009), Dolce et al (2008), Mondello et al (2012), Pang et al (2007), Penny e Chesney (2006), Raeesi et al (2014), Sut e Simsek (2011), Theodoraki et al (2010), entre outros.

Considerando-se a problemática e as motivações expostas anteriormente, esta pesquisa se propõe a avaliar a aplicação de diferentes métodos de *data mining* e compará-los com os resultados da regressão logística aplicada ao prognóstico de Traumatismo Cranioencefálico Grave, a fim de identificar um modelo para predição do desfecho óbito. Para isso, são empregados diferentes métodos para a classificação em *data mining* e como critérios de avaliação são considerados a acurácia[4] e robustez[5].

A classificação é uma das mais importantes e populares tarefas do *data mining*, consistindo na identificação de características comuns entre os registros de uma base de dados associando-as a um atributo específico designado como classe (HAN; KAMBER; PEI, 2011).

Nesta pesquisa aplicam-se diferentes métodos para a classificação de dados em *data mining*, empregando-se sete algoritmos. Destes tem-se a indução de árvores de decisão pelos algoritmos C4.5 e Classification And Regression Trees (CART); o aprendizado baseada em instâncias pelo algoritmo k-Nearest Neighbors (kNN); as redes neurais artificiais por Funções de Base Radial (RBF); os classificadores bayesianos pelos algoritmos Naive Bayes e Redes de Crença Bayesiana (Bayes Net); e o metaclassificador pelo algoritmo Adaptive Boosting (AdaBoost).

Os modelos para Traumatismo Cranioencefálico Grave gerados pelos algoritmos de *data mining* supracitados são comparados entre si e com a regressão logística, que consiste em uma das formas mais empregadas na análise de dados em saúde, a fim de se identificar o que apresenta maior acurácia e robustez.

Os dados utilizados referem-se a registros de pessoas com Traumatismo Cranioencefálico Grave que foram admitidas na Unidade de Terapia Intensiva do Hospital Governador Celso Ramos, em Florianópolis-SC, no período de Janeiro de 1994 a Dezembro de 2003, totalizando 748 registros. Esta base de dados foi coletada e aplicada no

---

[4] Medida de precisão ou de validade que possibilita avaliar o quanto os resultados de uma aferição correspondem ao estado verdadeiro do fenômeno aferido (FLETCHER; FLETCHER; FLETCHER, 2014).

[5] Habilidade de uma técnica estatística de desempenhar razoavelmente bem, mesmo quando as suposições estatísticas inerentes forem violadas (HAIR et al, 2009).

estudo de Martins et al (2009), que desenvolveu um modelo de regressão logística para a predição do óbito. Considerando-se isso, os resultados oriundos desta pesquisa pela aplicação dos modelos de *data mining* são comparados aos de Martins et al (2009), pois este, conforme levantamento bibliográfico das publicações dos estudos científicos brasileiros, concebeu a maior amostra de dados referente ao Traumatismo Cranioencefálico Grave no Brasil.

Mediante a análise das pesquisas realizadas na área de Traumatismo Cranioencefálico, em especial as voltadas ao prognóstico, pode-se observar as ferramentas estatísticas e computacionais adotadas, bem como identificar as possibilidades de desenvolvimento desta pesquisa, principalmente no que se refere aos métodos e algoritmos de *data mining* empregados e a avaliação dos resultados.

Além disso, outra motivação para o desenvolvimento desta pesquisa é que o modelo concebido não consegue acurácia significativa para o prognóstico do desfecho óbito. Sabendo-se disso qualquer avanço em termos de acuracidade e robustez da predição pode gerar ganhos para o prognóstico e a conduta do doente. Daí o interesse em avaliar diferentes modelos de *data mining* e compará-los com a regressão logística, a fim de demonstrar que os métodos de aprendizado de máquina aplicados na análise desses dados podem apresentar melhor desempenho, constituindo-se em uma alternativa as ferramentas mais tradicionais de análise de dados.

Também outro fator motivador consistiu na afirmação realizada no estudo de Theodoraki et al (2010), que devido ao fato de não haver um consenso quanto a um método ideal para o prognóstico em Traumatismo Cranioencefálico, é interessante explorar métodos diferentes.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Avaliar se métodos de *data mining* mostram-se como uma alternativa à regressão logística binária em dados de Traumatismo Cranioencefálico Grave.

1.1.2 Objetivos Específicos

Os objetivos específicos desta pesquisa consistem em:

a) aplicar métodos para descoberta de conhecimento em dados de Traumatismo Cranioencefálico Grave;
b) avaliar por meio de medidas de qualidade em *data mining* os padrões descobertos;
c) comparar o desempenho dos métodos de *data mining* aplicados ao conjunto de dados;
d) comparar os resultados da regressão logística e do *data mining* para a predição do óbito em Traumatismo Cranioencefálico Grave.

1.2 HIPÓTESES

As hipóteses desta pesquisa são:
a) hipótese 1: existe diferença entre as medidas de qualidade apresentadas para os métodos de *data mining* e regressão logística em dados de Traumatismo Cranioencefálico Grave;
b) hipótese 2: existe diferença entre os algoritmos de classificação em *data mining* para a predição do óbito em dados referente a Traumatismo Cranioencefálico Grave.

1.2 JUSTIFICATIVA

Em aplicações médicas é prática usual o desenvolvimento de um modelo de regressão logística usando o conjunto de dados completo, isto ocasiona uma falsa elevação na acurácia da predição dos modelos (PANG et al, 2007; PENNY; CHESNEY, 2006). Além disso, na área de Traumatismo Cranioencefálico vários estudos abordando modelos de prognóstico foram publicados, porém muitas vezes incorporam diferentes métodos e também populações de estudo, inviabilizando a comparação desses modelos.

Devido a isso, esta pesquisa propõe a avaliação de diferentes métodos de *data mining* aplicados ao Traumatismo Cranioencefálico Grave a fim de compará-los com os resultados provenientes de um modelo de regressão logística elaborado com a mesma população de estudo, no caso de Florianópolis-SC, a fim de conceber o prognóstico para o desfecho óbito.

O prognóstico é a previsão de um evento antes da sua possível ocorrência, como por exemplo, o óbito que é um importante resultado, pois é o fim clínico do processo de cuidar (SARABANDO, 2010), buscando-se evitá-lo, a fim de que a pessoa recupere o seu estado de saúde anterior ao traumatismo cranioencefálico.

Os modelos prognósticos no geral têm uma especificidade elevada, ou seja, são capazes de prever o não óbito, mas possuem uma menor sensibilidade, apresentando uma menor acurácia na predição do óbito (HERRIDGE, 2003; SILVA, 2007). Portanto, faz sentido explorar as possibilidades de opções de modelos para otimizar o acerto para predição do óbito. A escolha do desfecho óbito para esta pesquisa, também ocorreu em função de em parte do desenvolvimento da tese se estar reproduzindo um estudo secundário, no que se refere ao modelo de regressão logística concebido por Martins et al (2009).

Segundo Schuster (1992) e Silva (2007), um modelo prognóstico, de uma forma geral, deverá possuir uma boa capacidade discriminativa se o objetivo é identificar os que não sobrevivem. A discriminação consiste na capacidade do modelo distinguir se uma pessoa pode ter o desfecho óbito ou não óbito, baseando-se nas estimativas de mortalidade intra-hospitalar. O poder discriminativo é fornecido pela área sob a curva Receiver-Operating Characteristic (ROC).

O modelo prognóstico baseado em regressão logística para a população de estudo apresenta, segundo Martins et al (2009), acurácia em torno de 55% para a predição do óbito. Considerando-se isso e conforme Dolce et al (2008), a pesquisa se justifica pois alguns estudos têm mostrado que o processo de *data mining* parece ser mais eficiente que as estatísticas convencionais. Assim, pretende-se demonstrar que métodos de *data mining* apresentam acurácia e robustez superiores, podendo-se constituir em melhores ferramentas para o prognóstico do Traumatismo Cranioencefálico Grave. Com isso, pode-se auxiliar no planejamento de futuras estratégias de reabilitação.

No âmbito da análise estatística tem-se um processo conduzido pelo usuário, que se baseia em grande parte pela confirmação de um conjunto de hipóteses que é impulsionado por outras pré-definidas. Enquanto, o aprendizado de máquina é utilizado para gerar as hipóteses, constituindo-se em uma atividade exploratória e baseada em menos pressupostos, sendo impulsionado pelos dados (CHOWRIAPPA; DUA; TODOROV, 2014).

A escolha pelo *data mining* deve-se ao fato que segundo a literatura, técnicas oriundas do aprendizado de máquina, podem superar alguns dos problemas presentes à análise de regressão logística (CLERMONT et al, 2001; DOLCE et al, 2008).

O processo de *data mining*, por meio de métodos específicos de generalização, tem como objetivo a identificação do conhecimento em bases de dados, podendo facilitar a tomada de decisões pela predição da ocorrência de padrões e relações entre os dados (KANTARDZIC, 2011).

Além disso, é útil para derivar modelos de prognóstico médico, podendo contribuir para aumentar a disponibilidade de dados médicos coletados pelo uso sistemático de sistemas de informação hospitalares, clínicos e laboratoriais (THEODORAKI et al, 2010).

A classificação é uma das tarefas cognitivas humanas mais realizadas no auxílio à compreensão do ambiente em que se vive. Esse contexto faz também da classificação uma das tarefas mais utilizadas do *data mining* (TAN; STEINBACH; KUMAR, 2009).

Segundo Dolce et al (2008) a classificação é adequada para ser aplicada na investigação de sinais neurológicos e em diferentes condições clínicas, como naquelas em que se tem escassez de sinais clínicos, como nos casos de estados vegetativos ou de mínima consciência.

A classificação é responsável por aprender uma função alvo que é conhecida como o modelo de classificação. Este modelo pode ser útil para o propósito de uma modelagem preditiva. Assim, quando o conjunto de atributos de um registro desconhecido é submetido a este modelo preditivo, ele é capaz de atribuí-lo automaticamente a um rótulo de classe.

Nesta pesquisa optou-se pelo uso de modelos clássicos em *data mining*, pois de acordo com Wu et al (2008) e Wu e Kumar (2009), respectivamente, no artigo e livro intitulado "*Top 10 algorithms in data mining*", estes algoritmos são identificados como alguns dos mais influentes e que têm sido amplamente utilizados pela comunidade de *data mining*, tendo sido avaliados e listados como os melhores algoritmos de *data mining* atualmente disponíveis. Especialmente, o C4.5, kNN, CART, Naive Bayes e AdaBoost. Isto se justifica inclusive na realização da comparação entre os modelos de *data mining* gerados por estes algoritmos com os de regressão logística binária.

Além disto, muitos destes algoritmos são empregados em *data mining* em especial no contexto médico e biomédico (ESFANDIARI et al, 2014; PARAMASIVAN et al, 2014), permitindo que se realizem predições em Medicina que eram consideradas impossíveis há alguns anos atrás (PARAMASIVAM et al, 2014).

A aplicação de *data mining* fornece novos conhecimentos biomédicos e de cuidados à saúde que podem ser utilizados para apoiar a tomada de decisão clínica como, por exemplo, no processo de diagnóstico, escolha de opções de tratamento e prognóstico, bem como a tomada de decisão administrativa na área da saúde (BELLAZZI; ZUPAN, 2008; YOO et al, 2012).

O *data mining* pode contribuir efetivamente para o desenvolvimento de modelos de predição clinicamente úteis em função de pelo menos três aspectos inter-relacionados: abordagem abrangente para análise de dados que envolve a aplicação de métodos e conhecimentos oriundos de diferentes áreas científicas; capacidade explicativa dos modelos e de usar o conhecimento do domínio no processo de análise dos dados (BELLAZZI; ZUPAN, 2008).

As contribuições dessa pesquisa compreendem:

a) apresentar um modelo com melhor capacidade de discriminação para a partir dos métodos de *data mining* e regressão logística predizer o óbito em Traumatismo Cranioencefálico Grave em uma mesma população de estudo;
b) demonstrar que essas medidas de desempenho geradas a partir dos modelos de *data mining* se constituem em uma ativa e importante área de pesquisa em *data mining* no grupo estudado, pois a avaliação, de acordo com Guillet e Hamilton (2010), é uma das maiores dificuldades no processo de descoberta de conhecimento em bases de dados;
c) utilizar uma combinação de algoritmos de aprendizagem que não foram aplicados em outros trabalhos científicos para o prognóstico na área de Traumatismo Cranioencefálico Grave.

1.3 ESTRUTURA DO TRABALHO

Esta tese é formada por oito capítulos, tendo-se primeiramente a introdução que apresenta o tema proposto, as motivações, os objetivos pretendidos, as hipóteses delineadas para o estudo e a justificativa da realização da pesquisa e escolhas adotadas.

No capítulo 2 contextualiza-se sobre o processo de descoberta de conhecimento em bases de dados, em especial da etapa de *data mining*, fornecendo-se uma visão acerca dos seus conceitos, tarefas, métodos, áreas de aplicação e uma abordagem do *data mining* em relação as técnicas tradicionais de análises de dados como a regressão logística.

As tarefas, métodos e algoritmos de *data mining* aplicados nesta pesquisa são apresentados no capítulo 3, enquanto as medidas de qualidade são descritas no capítulo 4, conforme os modelos empregados que compreendem os classificadores.

No capítulo 5 aborda-se o domínio de aplicação da pesquisa, no caso o traumatismo cranioencefálico, bem como o levantamento do estado da arte.

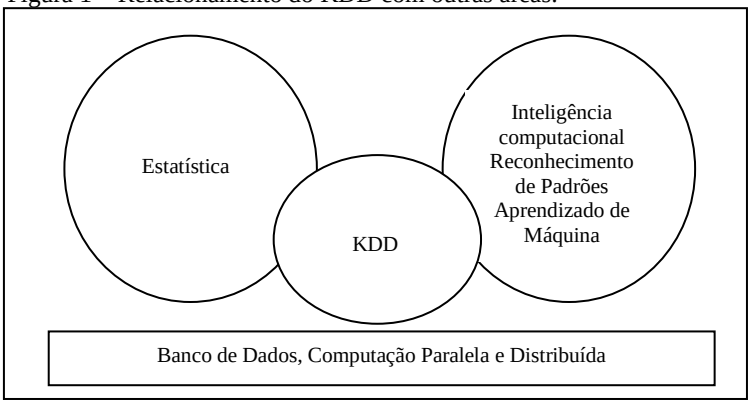
Os materiais e métodos empregados na realização da pesquisa são descritos no capítulo 6, enquanto os resultados obtidos com a pesquisa no que se refere ao modelo de regressão logística, aos modelos de *data mining*, a comparação entre os modelos gerados com estas duas abordagens e a discussão dos resultados são apresentados no capítulo 7. Finalmente, no capítulo 8 tem-se a conclusão.

2 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

A Descoberta de Conhecimento em Bases de Dados, do inglês Knowledge Discovery in Databases (KDD), segundo Fayyad, Piatetsky-Shapiro e Smyth (1996) é um processo automático de identificação de características e relacionamentos nos dados, os quais são transformados em conhecimento útil e compreensível.

O KDD originou-se basicamente de ideias da estatística como a amostragem, estimativa e testes de hipóteses, e de métodos de busca, inteligência computacional, reconhecimento de padrões e aprendizado de máquina, valendo-se também das áreas de otimização, computação evolutiva, processamento de sinais, visualização e recuperação de informações, banco de dados, computação paralela e distribuída (figura 1).

Figura 1 – Relacionamento do KDD com outras áreas.



Fonte: Adaptado de Tan, Steinbach e Kumar (2009).

O KDD é considerado interativo, pois exige a atuação de um especialista na área de aplicação da base de dados, e iterativo, pois se executam repetições de parte ou de todo o processo de KDD a fim de atingir os resultados desejados e de aprimorá-los (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; GOLDSCHMIDT; PASSOS, 2005).

O objetivo do KDD é identificar nos dados padrões válidos, novos, potencialmente úteis e compreensíveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; TURBAN et al, 2008). O

método de KDD também deve apresentar as seguintes características: ser acurado, genérico e facilmente modificado (STEINER et al, 2006).

Os padrões descobertos devem ser válidos em novos dados com algum grau de certeza; ocasionarem algum benefício ao usuário ou tarefa, o que se refere a sua utilidade; e finalmente devem ser compreensíveis depois de algum pós-processamento. Portanto, para avaliar os padrões extraídos devem-se definir medidas de qualidade, como por exemplo, a acurácia da predição em dados novos. Já os conceitos de novo e compreensível são mais subjetivos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Nesta pesquisa, se está focando na identificação de padrões válidos, por isso se realiza uma avaliação comparativa entre diferentes métodos de KDD, a fim de avaliar os padrões encontrados por meio de critérios de avaliação como a acurácia e robustez, os quais se constituem em medidas de qualidade dos padrões descobertos.

As metas do KDD são definidas conforme a intenção de uso, podendo ser de verificação e descoberta. Na verificação o sistema é limitado a constatar a hipótese do usuário. Com a descoberta, o sistema autonomamente encontra novos padrões (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

O processo de KDD (figura 2) envolve algumas etapas essenciais as quais, de acordo com Han, Kamber e Pei (2011), basicamente são o pré-processamento, *data mining* e pós-processamento.

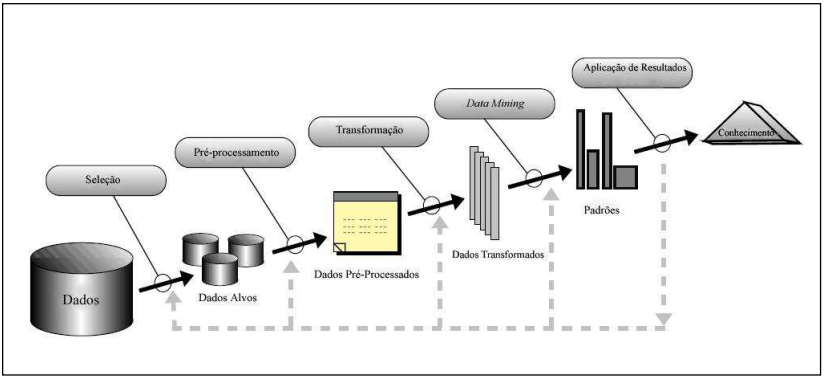
O pré-processamento é a preparação dos dados de entrada, convertendo-os em um formato apropriado para as análises a serem realizadas. Atividades compreendidas nesta fase incluem a seleção dos dados e o resumo das variáveis que serão utilizadas no KDD, a fusão de dados de múltiplas fontes, a limpeza e remoção de ruídos, a adequação de valores que estão fora de contexto. Esta fase é uma das mais trabalhosas e demoradas no processo de KDD, pois a qualidade dos dados interfere no conhecimento a ser extraído pelo *data mining* (TAN; STEINBACH; KUMAR, 2009).

O *data mining* é a principal etapa do KDD, sendo responsável pela busca, no conjunto de dados, dos padrões que podem originar conhecimento útil (DASU; JOHNSON, 2003; HAND, MANILLA, SMYTH, 2001). Para isso, aplicam-se determinados algoritmos para a extração desses padrões na base de dados.

O pós-processamento envolve a visualização, análise e interpretação do conhecimento descoberto na fase de *data mining*. Nesta etapa empregam-se medidas estatísticas ou métodos de teste de hipóteses para desconsiderar resultados que porventura não sejam

legítimos (TAN; STEINBACH; KUMAR, 2009), validando-se o conhecimento gerado pela análise de um especialista do domínio de aplicação e por medidas de qualidade (SASSI, 2006).

Figura 2 – Etapas do processo de KDD.



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

Apesar de ser dependente das outras etapas, o *data mining* é a mais importante, pois explora e analisa grandes conjuntos de dados, extraindo conhecimento por meio de relações e padrões que podem auxiliar na resolução de problemas e tomada de decisão de um determinado domínio de aplicação (SIVANANDAM; SUMATTI, 2006; WITTEN; FRANK, 2005). Considerando-se isto, esta etapa do processo de KDD será abordada mais detalhadamente.

2.1 DATA MINING

O *data mining* integra métodos tradicionais de análise de dados com algoritmos sofisticados a fim de processar conjuntos de dados e identificar relações entre estes, sendo aplicado para a análise de novos tipos de dados, como também dos mais antigos por meio de uma metodologia diferente (TAN; STEINBACH; KUMAR, 2009).

Esta etapa do processo de KDD é composta por várias tarefas, as quais são empregadas conforme o padrão dos dados que se deseja, sendo classificadas como descritivas e preditivas (KANTARDZIC, 2011). As tarefas descritivas são exploratórias e caracterizam os dados de acordo com as semelhanças e diferenças de padrões existentes entre eles, como por exemplo, a associação, agrupamento e detecção de anomalias (HAN; KAMBER; PEI, 2011; TAN; STEINBACH; KUMAR, 2009).

As tarefas preditivas consistem em prever o valor de um atributo com base nos demais, tendo-se como exemplo a classificação, regressão e previsão por séries temporais (KANTARDZIC, 2011).

A associação busca por itens que tendem a ocorrer juntos em uma base de dados, sendo os padrões descobertos representados por meio de regras de implicação ou subconjuntos de características. O objetivo desta análise é extrair os padrões mais interessantes de uma forma eficiente, já que o espaço de busca é exponencial. A análise de associação pode ser aplicada, por exemplo, para a descoberta de genes que têm funções associadas, doenças que ocorrem em conjunto, entre outros (TAN; STEINBACH; KUMAR, 2009).

O agrupamento separa um conjunto de dados em grupos (*clusters*). A divisão ocorre de acordo com as semelhanças entre os dados, assim registros do mesmo *cluster* compartilham similaridades que os diferenciam dos demais grupos (GAN; MA; WU, 2007). Utilização em bases de dados geográficas para identificação de regiões do solo que possuem uso similar é uma das aplicações desta tarefa (HAN; KAMBER; PEI, 2011).

A detecção de anomalias identifica observações com características diferentes dos demais registros, as quais são denominadas de anomalias ou fatores estranhos. Nesta tarefa descobre-se o que realmente é uma anomalia, evitando-se rotular errado um objeto como anômalo. As aplicações da detecção de anomalia voltam-se a detecção de fraude, padrões incomuns de doenças, entre outros (TAN; STEINBACH; KUMAR, 2009).

A classificação constrói uma função que mapeia um registro a fim de categorizá-lo em classes pré-definidas, sendo empregada para variáveis alvo discretas (KANTARDZIC, 2011). Pode ser usada para prever se um internauta fará uma compra em uma livraria online, identificar se um paciente possui uma doença baseando-se nos exames médicos, entre outros (TAN; STEINBACH; KUMAR, 2009).

A tarefa de regressão em *data mining* é semelhante a classificação, diferenciando-se desta por ser usada para variáveis alvo contínuas (valores numéricos). Ela mapeia matematicamente em valores reais os registros de uma base de dados, a partir dos valores de entrada a fim de prever os de saída. Alguns exemplos de aplicações são: estimar as chances de um paciente sobreviver, a demanda de um produto novo no mercado, entre outras (KANTARDZIC, 2011).

A previsão por séries temporais baseia-se no comportamento passado dos dados a fim de verificar e estimar valores futuros. Exemplificando pode-se analisar o comportamento da bolsa de valores e

prever o quanto uma ação irá variar no próximo mês (BERRY; LINOFF, 2004).

As tarefas de *data mining* extraem diferentes tipos de conhecimento, portanto a escolha de qual utilizar deve estar de acordo com o que se deseja extrair da base de dados (HAN; KAMBER; PEI, 2011). Alguns problemas podem ser solucionados com algumas tarefas em específico, enquanto outros precisam de uma utilização conjunta das tarefas, a fim de auxiliar e melhorar o desempenho do KDD (BERRY; LINOFF, 2004).

Após a escolha da tarefa, deve-se optar pelo método de acordo com a representação dos padrões que se deseja. Alguns dos métodos empregados são:

a) lógica *fuzzy*: na lógica clássica os elementos pertencem ou não a um determinado conjunto, já na lógica *fuzzy* os elementos possuem graus de pertinência a cada conjunto existente, podendo pertencer ou não, ou pertencer parcialmente a um ou mais conjuntos, possibilitando raciocínios imprecisos (COX, 2005);
b) redes neurais artificiais: são modelos matemáticos que se assemelham as estruturas neurais biológicas, com o objetivo de simular o mecanismo de processamento do cérebro humano. Possuem capacidade computacional adquirida por meio de aprendizado e generalização (HAYKIN, 2001; REZENDE, 2005);
c) algoritmos genéticos: métodos computacionais adaptativos que se baseiam nos processos genéticos de entidades biológicas e podem ser usados na resolução de problemas de busca e otimização (SIVANANDAM; SUMATTI, 2006);
d) árvores de decisão: recursivamente, por meio de regras simples de decisão, um determinado conjunto de dados é dividido em pequenos grupos (BERRY; LINOFF, 2004; GOLDSCHMIDT; PASSOS, 2005);
e) métodos estatísticos: baseiam-se em teorias da estatística, fornecendo modelos e técnicas tradicionais para análise e interpretação dos dados, como por exemplo, as redes bayesianas, análise discriminante, análise exploratória de dados, entre outras (GOLDSCHMIDT; PASSOS, 2005).

Devido a capacidade de descobrir conhecimento, a fim de compreender as bases de dados, existem muitas áreas em que o *data mining* pode ser aplicado, como por exemplo:

a) engenharia elétrica: na previsão futura de carga do sistema elétrico, pois quanto maior a precisão das estimativas das cargas máxima e mínima em determinados períodos, tem-se uma maior economia para as companhias geradoras e distribuidoras de energia. Aplicando-se técnicas de *data mining* analisam-se dados históricos, gerando-se previsões inclusive por hora (WITTEN; FRANK; HALL, 2011);
b) medicina: descoberta de relações entre doenças e as características sociais da região onde se vive e dos hábitos pessoais, fornecendo-se assim conhecimento para estudos de epidemiologia, análise de diagnóstico, tratamento e prognóstico , auxiliando-se no entendimento das doenças e dos tratamentos (WITTEN; FRANK, 2005);
c) marketing e vendas: atualmente, as empresas podem ter muitas informações sobre seus clientes. Utilizando *data mining* é possível gerar um perfil sobre necessidades e gostos, e assim oferecer a seus consumidores novos produtos, descontos e ofertas com o objetivo de garantir fidelidade (OLSON; DELEN, 2008). Em supermercados, por exemplo, por meio da análise das vendas realizadas podem-se distribuir os produtos nas gôndolas de forma a incentivar o consumo e compras conjuntas (WITTEN; FRANK, 2005);
d) web: motores de busca na internet, como por exemplo, o Google e o Bing, valem-se da aplicação de técnicas de *data mining* nos conteúdos pesquisados pelos usuários a fim de indicar anúncios que possam interessá-los, pois só recebem dos anunciantes quando os usuários entram nesses links (WITTEN; FRANK; HALL, 2011);
e) biomédica: o crescimento da pesquisa biomédica e biotecnologia tem ocasionado um aumento nos dados oriundos de estudos farmacêuticos, de terapias para o câncer, do genoma, buscando-se a identificação de padrões por meio de métodos de *data mining* para a análise desses dados (WANG et al, 2005).

Mais informações sobre tarefas de *data mining* e suas aplicações na área biomédica podem ser encontradas em Esfandiari et al (2014) e Yoo et al (2012) que realizam uma revisão de literatura na área.

2.1.1 Data Mining e Técnicas Tradicionais de Análise de Dados

Os avanços na computação, no final do Século XX, modificaram a capacidade de armazenamento e processamento dos dados, bem como a sua forma de análise, surgindo novos métodos associados ao volume dos dados e a possibilidade de modelar sistemas complexos. Essa abordagem denominada *data mining* consiste na exploração dos dados em busca de uma teoria, enquanto as técnicas tradicionais de análise são uma teoria em busca de dados confirmatórios (PEREIRA, 2001).

Conforme Tan, Steinbach e Kumar (2009) o *data mining* se utiliza de outras áreas de análise de dados, não as substituindo, mas valendo-se de suas pesquisas para agrupamento, classificação e detecção de anomalias.

A estatística é utilizada pela maioria dos algoritmos na construção dos modelos, bem como para validá-los por meio dos testes estatísticos e para avaliar os algoritmos de aprendizagem (WITTEN; FRANK, 2005).

No entanto, inicialmente pensava-se que o *data mining* não necessitaria de analistas estatísticos para a construção de modelos preditivos, porém eles são necessários para a avaliação dos modelos e validação da plausibilidade das predições (BERSON; SMITH; THEARLING, 2000).

O *data mining* diferencia-se das técnicas tradicionais de análise de dados, principalmente na forma como se dá a exploração das relações entre os dados. As técnicas tradicionais, estatísticas, valem-se da verificação, construindo-se hipóteses e as comprovando ou refutando. Esse método depende do levantamento de hipóteses interessantes e da manipulação da complexidade dos atributos, o que por vezes torna-se difícil. Enquanto, no *data mining* o próprio processo gera as hipóteses, garantindo maior qualidade, rapidez e integridade aos resultados (BARBOSA; MACHADO, 2007).

O desenvolvimento do *data mining* deu-se em função da dificuldade prática das técnicas tradicionais de análise, quando aplicadas nos novos conjuntos de dados, devido a (TAN; STEINBACH; KUMAR, 2009):

a) escalabilidade: a maior capacidade de processamento e armazenamento exigem algoritmos de *data mining* escaláveis, capazes de trabalhar com volumosos conjuntos de dados, empregando-se para isso métodos de busca exponencial e algoritmos paralelos e distribuídos;
b) alta dimensionalidade: atualmente as bases de dados apresentam muitos atributos o que lhes confere essa alta

dimensionalidade, porém as técnicas tradicionais de análise de dados foram criadas voltadas a dados de baixa dimensionalidade. A complexidade computacional desses algoritmos tradicionais também aumenta em relação a dimensionalidade;

c) dados complexos e heterogêneos: métodos tradicionais lidam com bases de dados que possuem os mesmos tipos de atributos, por exemplo, contínuos ou categóricos. Porém, nos últimos anos têm-se dados mais complexos precisando-se de técnicas que trabalhem com atributos heterogêneos;
d) propriedade e distribuição de dados: algumas vezes os dados a serem analisados estão distribuídos geograficamente ou pertencem a diferentes organizações, necessitando-se de algoritmos distribuídos de *data mining*;
e) análises não tradicionais: a estatística tradicional baseia-se em hipótese e teste, porém muitas vezes as tarefas de análise de dados envolvem o levantamento e a avaliação de várias hipóteses. Assim, a fim de se automatizar esta tarefa tem-se o *data mining*.

Dentre as técnicas tradicionais de análise de dados, na área biomédica, campo de aplicação desta pesquisa, uma das mais empregadas é a regressão logística que tem obtido sucesso na análise de estudos epidemiológicos (ABREU et al, 2008; PEREIRA, 2001; WALPOLE et al, 2009).

2.1.2 Regressão Logística

A regressão logística é empregada para prever e explicar uma variável categórica binária, sendo a abordagem mais popular para modelar este tipo de resposta, devido a sua robustez e facilidade de interpretação (HAIR et al, 2009), apresentando como distribuição básica a Binomial ou Bernoulli (WALPOLE et al, 2009).

A regressão logística apresenta como vantagem ser menos afetada quando as suposições básicas não são satisfeitas, em particular a normalidade das variáveis. Acomoda também variáveis não-métricas por meio da codificação em variáveis dicotômicas, sendo adequada para prever apenas uma medida dependente de dois grupos (HAIR et al, 2009).

Na análise de regressão logística a variável dependente, *Y*, que é a de interesse se apresenta dicotômica. Tem-se como objetivo descrever *Y*, equação (1), como uma função matemática que é conhecida pelas outras

variáveis qualitativas ou quantitativas do problema (MASSAD et al, 2004):

$$X_1, X_2, \dots, X_k \ (Y = f(X_i), i = 1,2, \dots, k) \quad (1)$$

O modelo de regressão logística é escrito em termos de probabilidade, portanto, dados os regressores *x*, a função logística é representada pela equação (2), *f(x)* é chamada de preditor linear (MASSAD et al, 2004; WALPOLE et al, 2009).

$$p(x) = \frac{1}{1 + e^{-f(x)}} \quad (2)$$

Na regressão logística emprega-se outra forma de inferência que é derivada do uso da razão de chances, a qual determina como os *sucessos das chances = (p/1-p)* aumenta a medida que acontecem modificações nos valores do regressor, equação (3) (DEVORE, 2006; WALPOLE et al, 2009).

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x} \quad (3)$$

A regressão logística para ser ajustada aos dados amostrais exige que os parâmetros β_0 e β_1 sejam estimados. Para isso emprega-se o método da máxima verossimilhança que fornece aos parâmetros os valores que maximizam a probabilidade de se obter o conjunto de dados existentes, tornando-o mais verossímel (DEVORE, 2006; MASSAD et al, 2004). Mais informações acerca de regressão logística podem ser encontradas em Field (2009) e Hair et al (2009).

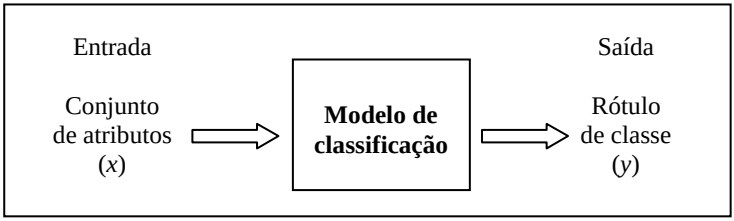
3 TAREFAS, MÉTODOS E ALGORITMOS DE *DATA MINING*

Dentre as várias tarefas, métodos e algoritmos de *data mining* existentes, neste capítulo são abordados somente os que serão utilizados no decorrer desta pesquisa.

3.1 CLASSIFICAÇÃO

A classificação é uma tarefa que consiste na organização de objetos em uma categoria pré-definida, sendo a aprendizagem de uma função alvo *f*, também chamada de modelo de classificação, que mapeia cada conjunto de atributos *x* a um único rótulo *y*, denominado de classe ou objeto de saída (figura 3) (TAN; STEINBACH; KUMAR, 2009).

Figura 3 – Tarefa de classificação.



Fonte: Tan, Steinbach e Kumar (2009).

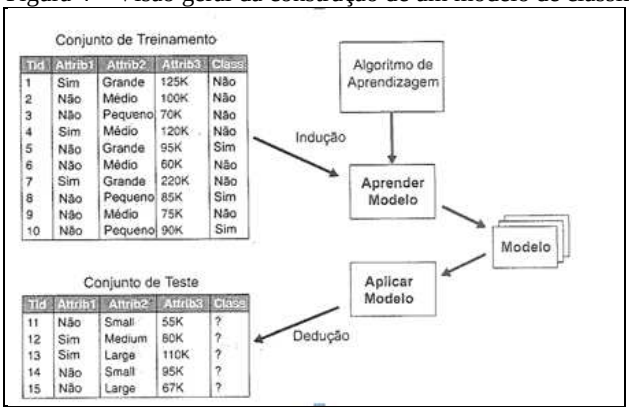
Esta tarefa, de acordo com Tang e Maclennan (2005), é uma das mais populares em *data mining*, sendo aplicada em uma diversidade de problemas. Um classificador encontra propriedades comuns entre um conjunto de registros pertencentes a uma base de dados e os classifica em diferentes classes conforme um modelo. A tarefa é de aprendizado supervisionado, pois os registros têm suas classes pré-definidas a partir do grupo inicialmente analisado (TANIAR, 2008).

Os dados de entrada da classificação são um conjunto de registros, formado por valores de atributos inerentes ao domínio do problema, caracterizado por uma dupla (*x*,*y*), em que *x* é o conjunto de atributos e *y* o atributo especial denominado de classe, atributo alvo ou de categorização (BERRY; LINOFF, 2004). Na classificação o conjunto de atributos, pode ser formado por variáveis discretas e contínuas, no entanto, a classe deve ser um atributo discreto. Isso diferencia a classificação de outra tarefa do *data mining* denominada de regressão cujo *y* é contínuo (TAN; STEINBACH; KUMAR, 2009).

A classificação constrói modelos, partindo do conjunto de dados de entrada, por meio de métodos que incluem classificadores de árvores de decisão, baseados em regras, redes neurais artificiais, máquinas de vetor de suporte e classificadores Bayesianos. Cada um desses métodos emprega um algoritmo de aprendizagem a fim de identificar um modelo mais adequado para o relacionamento entre o conjunto de atributos e o rótulo da classe. O modelo originado pelo algoritmo de aprendizagem deve ser bem adaptado aos dados de entrada e prever corretamente as classes de registros desconhecidos. Portanto, os modelos de classificação devem apresentar boa capacidade de generalização (TAN; STEINBACH; KUMAR, 2009).

A base de dados a ser classificada tem seus conjuntos de atributos que podem ser divididos em dois grupos, denominados de treinamento e teste, respectivamente. Os dados de treinamento são os utilizados na fase de aprendizagem para construir o modelo de classificação, enquanto os de teste são empregados na avaliação do modelo gerado (RUSSEL; NORVIG, 2004). Na figura 4 tem-se uma abordagem geral empregada na resolução de problemas de classificação.

Figura 4 – Visão geral da construção de um modelo de classificação.



Fonte: Tan, Steinbach e Kumar (2009).

A avaliação do desempenho de um modelo de classificação é baseada na contagem dos registros de testes que foram classificados de forma correta e incorreta pelo modelo. Portanto, o conjunto de teste é empregado como uma estimativa da qualidade do classificador, valendo-se dos seguintes critérios (HAN; KAMBER; PEI, 2011):

a) acurácia da predição: refere-se a capacidade do modelo de classificar registros desconhecidos;
b) custo computacional: tempo de processamento do modelo;
c) robustez: capacidade de tomar decisões corretas quando tem-se dados com ruídos ou incompletos;
d) escalabilidade: é capaz de gerar modelos de classificação quando se tem grandes bases de dados.

Nesta pesquisa a qualidade dos classificadores gerados é analisada por meio dos critérios de acurácia e robustez.

Diversos métodos são aplicáveis à tarefa de classificação, como por exemplo, a Indução de Árvores de Decisão, o Aprendizado Baseado em Instâncias, as Redes Neurais Artificiais, os Metaclassificadores e os Classificadores Bayesianos, os quais serão empregados nesta pesquisa.

3.1.1 Indução de Árvores de Decisão

A indução de árvores de decisão é uma metodologia de aprendizado supervisionado muito utilizada para a geração de um modelo classificador em problemas de predição (KANTARDZIC, 2011). Além disso, é uma das formas mais simples e bem sucedidas de algoritmos de aprendizagem (RUSSEL; NORVIG, 2004), apresentando uma acurácia tipicamente alta (HAN; KAMBER; PEI, 2011).

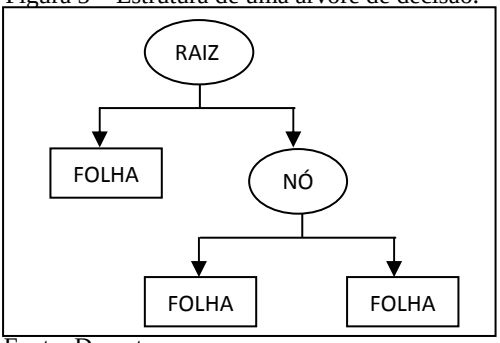
As árvores de decisão têm como entrada um objeto descrito por um conjunto de atributos e fornecem uma decisão, no caso o objeto de saída previsto, que deve estar de acordo com a entrada. Para isso, uma árvore de decisão executa uma sequência de testes, tendo-se cada nó da árvore correspondendo a um teste do valor de uma das propriedades, e as ramificações a partir deste nó são identificadas com os valores possíveis do teste. Cada nó folha na árvore denota o valor a ser retornado se a folha for alcançada (RUSSEL; NORVIG, 2004).

Assim, na estrutura de árvore (figura 5) as ramificações são: o nó que representa o teste feito ao valor de um atributo; as ramificações que são o resultado do teste do nó e a folha que se refere a distribuição das classes (HAN; KAMBER; PEI, 2011).

Na construção de uma árvore de decisão adota-se o particionamento recursivo de dados, dividindo-se o conjunto de treinamento em duas ou mais partições conforme o conjunto de valores de cada atributo, isso se repete até que todos ou pelo menos a maioria dos exemplos em cada uma das partições pertença a um rótulo de classe (GOLDSCHMIDT; PASSOS, 2005). Cada atributo é avaliado por uma medida, como por exemplo, o ganho de informação que determina qual

deles é capaz de separadamente classificar melhor um conjunto de dados (YE, 2003).

Figura 5 – Estrutura de uma árvore de decisão.



Fonte: Do autor.

As árvores construídas podem se apresentar instáveis em função da presença de atributos irrelevantes e de ruídos nos dados de treinamento; a fim de evitar este tipo de problema tem-se uma técnica denominada de poda que por meio de métodos estatísticos remove da árvore aqueles nós que não são relevantes, com isso a classificação torna-se mais rápida e melhor, pois aumenta a capacidade de generalização da árvore (HAN; KAMBER; PEI, 2011). Esta técnica pode ser realizada de duas maneiras, por meio da pré-poda e da pós-poda.

A pré-poda ocorre durante a construção da árvore no momento de avaliação da realização de uma partição em um nó. Consiste em não dividi-lo, transformando-o em folha. Para isso, utiliza-se de um critério de parada, como por exemplo, a ausência de diferença significativa antes ou depois da divisão do nó (HAN; KAMBER; PEI, 2011; KANTARDZIC, 2011).

A pós-poda ocorre após a árvore estar construída e consiste na remoção de nós que não são relevantes para a solução do problema, transformando-os em folha. A decisão de se realizar a poda de um nó é tomada mediante os cálculos das taxas de erro esperadas para a poda e não poda da subárvore de um nó, juntamente com a importância da ramificação. Caso a taxa fique acima do esperado não se realiza a poda, caso contrário, elimina-se a subárvore (HAN; KAMBER; PEI, 2011).

Conforme Tan, Steinbach e Kumar (2009) a pós-poda geralmente tem melhores resultados que a pré-poda, pois toma decisões baseadas

em uma árvore totalmente construída, enquanto a pré-poda pode ser prejudicada pela finalização antecipada da árvore.

Os resultados gerados pelas árvores de decisão, do nó raiz até as folhas, podem ser visualizados por meio de regras de classificação. As condições de teste formam os antecedentes da regra, enquanto que o rótulo da classe na folha da árvore constitui o consequente.

As regras de classificação auxiliam no entendimento do modelo construído, pois mesmo realizando-se a poda das árvores elas podem apresentar um tamanho grande, o que pode tornar complexa a sua compreensão (KANTARDZIC, 2011).

A indução de árvores de decisão ocorre por meio de algoritmos que particionam, recursivamente, o conjunto de treinamento em subconjuntos. Dentre estes algoritmos, tem-se o classificador de aprendizado de máquina C4.5 e o CART, os quais são utilizados nesta pesquisa.

3.1.1.1 Algoritmo C4.5

O C4.5 é um algoritmo de árvores de decisão, desenvolvido por John Ross Quinlan, que executa a simplificação da árvore de decisão, ou seja, a poda, excluindo as regras que não são significativas para a precisão da classificação. De acordo com Dimitoglou, Adams e Jim (2012) o C4.5 encontra hipóteses de alta precisão, porém tem maior custo computacional que o seu antecessor, o ID3[6], em termos de tempo e espaço de busca.

O algoritmo C4.5 trabalha com atributos discretos e contínuos, possui como objetivo gerar um modelo classificador, sendo executado em duas etapas: construção e simplificação da árvore de decisão.

3.1.1.1.1 Construção da Árvore de Decisão

Na construção da árvore separam-se os dados em duas ou mais partições, para isso empregam-se restrições sobre os conjuntos de valores de cada atributo. Esse processo é feito recursivamente até que os dados do conjunto de treinamento pertençam a uma determinada classe. Estes dados são armazenados na estrutura de árvore, a qual é construída em largura, portanto, todos os nós pertencentes a um nível da árvore

[6] Algoritmo também desenvolvido por John Ross Quinlan,trabalha com atributos nominais e constrói a árvore de decisão, porém não realiza a fase de simplificação (poda) (LUGER, 2004).

devem ser processados para então se iniciar a construção do próximo nível (GOLDSCHMIDT; PASSOS, 2005).

A construção da árvore é realizada por meio da avaliação dos pontos de separação para os casos pertencentes a cada nó e identificação daquele que os separa melhor, posteriormente, aplica-se este critério que foi identificado para criar as partições da árvore.

O processo de atribuição dos pontos de separação depende do domínio de cada atributo, podendo ser numérico ou categórico. No caso do atributo ser numérico, a partição possuirá duas ramificações, sendo o nó identificado como menor ou igual ao ponto de separação escolhido, ou maior que este valor. No entanto, se o atributo for categórico, a divisão terá uma ramificação para cada um dos valores do atributo (QUINLAN, 1993).

O C4.5, como a maioria dos algoritmos de indução de árvores de decisão, é baseado no algoritmo de Hunt[7], pois emprega uma estratégia que cresce uma árvore considerando uma série de decisões localmente ótimas sobre qual atributo usar para particionar os dados (TAN; STEINBACH; KUMAR, 2009).

Supondo-se que D_t seja o conjunto de registros de treinamento que estão associados a um determinado nó t e $y = \{y1, y2, ..., yc\}$ sejam as classes. Uma definição recursiva do algoritmo de Hunt é (TAN; STEINBACH; KUMAR, 2009):

a) se todos os registros em D_t pertencerem a mesma classe y_t, então t constitui-se em uma folha rotulada como y_t;
b) se D_t possuir registros que pertençam a mais de uma classe, emprega-se uma condição de teste de atributo para particionar os registros em subconjuntos menores. Um nó filho é criado para cada condição de teste e os registros D_t são distribuídos para os filhos baseando-se nos resultados. Aplica-se o algoritmo recursivamente para cada nó filho criado.

Conforme Quinlan (1993), desenvolvedor do algoritmo C4.5, os pontos de particionamento para cada nó da árvore são calculados pelas medidas de ganho de informação. A seguir, a equação (4), apresenta o cálculo da entropia do conjunto de dados completo, S, sobre os rótulos de classe, cuja unidade de medida utilizada para os valores de informação é *bits*.

---

[7] Algoritmo que constitui a base dos algoritmos de indução de árvores de decisão, como ID3, C4.5e CART (TAN; STEINBACH; KUMAR, 2009).

$$Info(S) = -\sum_{j=1}^{k}\left(\left(\frac{freq(C_j,S)}{|S|}\right).\log_2\left(\frac{freq(C_j,S)}{|S|}\right)\right) \quad (4)$$

Tem-se que: S é qualquer conjunto de amostras, podendo representar a base completa (no caso do nó raiz) ou partições da base de dados; $freq(C_j,S)$ é o número de vezes que a classe C_j acontece em S; $|S|$ é o número de amostras do conjunto S e k o número de classes possíveis.

Na sequência, considerando-se um dos atributos do conjunto de dados S, deve-se dividi-lo em subconjuntos T_1, T_2, T_3...T_n, que representam os possíveis valores de um atributo de teste X. A informação esperada pode ser encontrada como a soma ponderada das entropias nos subconjuntos, equação (5), obtendo-se finalmente o ganho total para o atributo aplicando-se a equação (6).

$$Info_x(T) = \sum_{i=1}^{n}\left(\left(\frac{|T_i|}{|T|}\right).info(T_i)\right) \quad (5)$$

Tem-se que T é a quantidade de ocorrências na partição que está sendo analisada e T_i a quantidade de ocorrências da classe i contida no conjunto T.

$$Gain(X) = Info(T) - Info_x(T) \quad (6)$$

O ganho total, equação (6), mede a informação que se obtém por meio do particionamento T de acordo com o atributo de teste X. Este critério seleciona o atributo de teste X que maximiza o ganho, $Gain(X)$, selecionando-se, portanto, o atributo com maior ganho de informação que será a primeira divisão na construção da árvore. Após a divisão inicial, cada nó filho tem várias amostras da base de dados e todo o processo de seleção de testes e otimização é repetido para cada nó filho (KANTARDZIC, 2011).

O critério de ganho de informação tem apresentado bons resultados quando as árvores de decisão construídas são compactas, porém ele tem uma deficiência grave que é dar preferência a atributos que possuem muitos valores possíveis e por vezes são irrelevantes, como, por exemplo, o código identificador. A solução encontrada no

C4.5 foi adicionar um parâmetro especificado na equação (7) a fim de se ter uma medida adicional chamada de razão do ganho de informação (KANTARDZIC, 2011), definida na equação (8).

$$Split\ info\ (X) = -\sum_{i=1}^{n}\left(\left(\frac{|T_i|}{|T|}\right).\log_2\left(\frac{|T_i|}{|T|}\right)\right) \quad (7)$$

A equação (7) representa a informação gerada pela divisão do conjunto T em n subconjuntos T_i, podendo-se agora definir uma nova medida de ganho:

$$Gain\ ration\ (X) = \frac{gain(X)}{split\ info(X)} \quad (8)$$

A razão do ganho, equação (8), expressa a porcentagem da informação gerada pela divisão que aparenta ser mais útil para a classificação. Após calcular a razão do ganho para todos os atributos do conjunto de dados, aquele que apresentar o maior valor será a raiz da árvore. Neste momento a base de dados é particionada e o processo é repetido para cada novo nó gerado (HAN; KAMBER; PEI, 2011).

Quinlan (1993) demonstrou que a razão do ganho é uma medida superior ao ganho de informação, empregada pelo algoritmo ID3, gerando árvores mais precisas e menos complexas. No entanto, ele salienta que em bases de dados com um número menor de registros a serem classificados pode-se gerar informações incorretas e árvores muito reduzidas quando se emprega a razão do ganho. Portanto, as implementações do C4.5 devem disponibilizar ao usuário os dois métodos, razão do ganho e ganho de informação, a fim de que se possa analisar o que proporciona os melhores resultados conforme a base de dados empregada.

3.1.1.1.2 Simplificação da Árvore de Decisão

A segunda etapa da execução do algoritmo C4.5 consiste na simplificação da árvore de decisão, pelo método de pós-poda, reduzindo algumas subárvores a folhas. Para isso, avalia a importância das regras geradas pela árvore, assim as que não acrescentam conhecimento são podadas, originando-se uma árvore com melhor classificação (HAN; KAMBER; PEI, 2011).

O algoritmo C4.5 emprega o cálculo denominado de poda pessimista, em que calcula-se para cada nó da árvore uma estimativa do limite superior de confiança, U_{cf} , que é calculado usando-se as tabelas estatísticas de distribuição binomial. O parâmetro U_{cf} é uma função de $|T_i|$ e E (taxa de erro encontrada na folha) para um determinado nó. O C4.5 usa o intervalo de confiança padrão de 25% e compara o $U_{25\%}(|T_i|/E)$ para um dado nó T_i com a confiança ponderada de suas folhas, tendo-se como o peso o número total de casos para cada folha. Se o erro predito de um nó raiz de uma subárvore é menor que a soma ponderada de $U_{25\%}$ para as folhas (erro predito para a subárvore), então a subárvore pode ser substituída pelo nó raiz que passa a ser uma folha de uma árvore podada (KANTARDZIC, 2011; QUINLAN, 1993).

3.1.1.2 Algoritmo CART

O algoritmo CART foi desenvolvido em 1984 por Leo Breiman, Jerome Friedman, Richard Oslen e Charles Stone. As árvores de decisão geradas pelo CART são binárias, sendo percorridas da raiz as folhas respondendo a questões do tipo sim ou não (LAROSE, 2005), possui grande capacidade de pesquisa entre os dados, prevendo por meio da classificação o tratamento de variáveis dependentes discretas. O CART possui duas metodologias para a construção da estrutura de árvore, a *Tree Structured Classifiers* e a *Tree Structured Regression*.

Nesta pesquisa o interesse é na *Tree Structured Classifiers*, visto que o CART está sendo usado para a tarefa de classificação, portanto a representação do conteúdo das suas folhas é a classe.

No crescimento de uma árvore para classificação, o CART possui os seguintes elementos: um conjunto Q de perguntas binárias em que $X \in A, A \subset X$, sendo X o conjunto das medidas para a divisão do nó; critério de divisão $\Phi(s,t)$ que é avaliado para a divisão *s* de qualquer nó *t*; critério de parada; regra para atribuir uma classe as folhas da árvore (BREIMAN et al, 1984).

A divisão de cada nó *t* é realizado pelo conjunto Q de perguntas binárias, quando a resposta é sim se segue para o nó esquerdo e nos casos em que é não, dirige-se para o nó direito. Esta divisão de cada nó da árvore é efetuada conforme alguns critérios que buscam pelo melhor ponto para dividi-lo, ou seja, aquele que consegue particionar em subconjuntos homogêneos (BREIMAN et al, 1984).

Considerando-se $\Phi(s,t)$ como uma medida do melhor candidato à divisão *s* de um nó *t*, tem-se a equação (9). Logo, a divisão ideal é aquela que maximiza a medida $\Phi(s,t)$em todas as divisões do nó *t*.

$$\Phi(s|t) = 2P_L P_R \sum_{j=1}^{\#classes} |P(j|t_L) - P(j|t_R)| \quad (9)$$

Sendo:

a) t_L= nó filho esquerdo do nó t;

b) t_L= nó filho direito do nó t;

c) $P_L = \frac{número\ de\ registros\ em\ t_L}{número\ de\ registros\ do\ conjunto\ de\ treinamento}$;

d) $P_R = \frac{número\ de\ registros\ em\ t_R}{número\ de\ registros\ do\ conjunto\ de\ treinamento}$;

e) $P(j|t_L) = \frac{número\ de\ registros\ da\ classe\ j\ em\ t_L}{número\ de\ registros\ em\ t}$;

f) $P(j|t_R) = \frac{número\ de\ registros\ da\ classe\ j\ em\ t_R}{número\ de\ registros\ em\ t}$.

A seleção da melhor divisão dos dados é realizada pelo CART considerando três critérios, os quais são a Entropia, Critério de Gini e Critério de Twoing. Nesta pesquisa aborda-se o Critério de Gini, pois este é o utilizado pela ferramenta Weka, que foi empregada para a realização do *data mining*, na implementação do algoritmo CART (SimpleCart).

O critério ou índice Gini, criado em 1912, é aplicado pelo algoritmo CART para mensurar a impureza de um nó, verificando a heterogeneidade dos dados. Consiste na probabilidade condicional do erro, dado um conjunto de treinamento, selecionado aleatoriamente, que é particionado em um nó *t*, tendo cada classe *j* uma probabilidade *P(j|t)*. O valor de Gini igual a zero caracteriza o nó como puro, enquanto próximo de um o nó é considerado impuro, pois aumenta o número de classes uniformemente distribuídas neste. Empregando-se o critério de Gini, equação (10), tende-se a isolar em um ramo da árvore aqueles registros que representam a classe mais frequente (BREIMAN et al, 1984).

$$G = 1 - \sum_{j=1}^{c} p^2(j|t) \quad (10)$$

Sendo *p(j|t)* a probabilidade *a priori* da classe *j* se formar no nó *t*.

Após particionar o nó, se não houver ganho em dividi-lo novamente, associa-se uma determinada classe à folha gerada.

Na árvore um nó é considerado como folha mediante a determinação de algum critério de parada, como por exemplo, o atributo selecionado como ponto de divisão possui índice de Gini igual a zero; o somatório das probabilidades de um nó é zero; ou a quantidade de casos do nó é inferior aquele predeterminado como mínimo.

Posteriormente, deve-se associar uma classe a este nó que pode ser escolhida baseando-se, por exemplo, na atribuição da classe mais provável, objetivando-se minimizar a taxa de erro do classificador (equação (11)).

$$\max_j(p_j) = \max_j \frac{n_j}{n} \tag{11}$$

Sendo:

a) j o número de classes de 1...j;
b) n o número total de exemplos na folha;
c) n_j o número de exemplos da classe j na folha.

Após a finalização do crescimento da árvore, aplica-se a poda a fim de evitar o seu tamanho excessivo. No algoritmo CART aplica-se a poda por minimização do custo-complexidade.

A poda pela minimização do custo-complexidade compreende as seguintes etapas: desenvolvimento de uma árvore de decisão inicial grande com um erro estimado baixo, realização da poda a fim de originar árvores de dimensão menor, identificar a melhor árvore criada por meio da estimação do erro (BREIMAN et al, 1984).

Considerando-se a árvore de decisão inicial gerada, T_{max}, que deve ser grande a fim de se realizar a poda, $\tilde{T}$ o conjunto de nós terminais para uma subárvore T de qualquer T_{max}, $|\tilde{T}|$ o número de nós terminais de T e $\alpha \geq 0$ o parâmetro de complexidade, a medida de custo-complexidade é definida pela equação (12).

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{12}$$

O custo associado à taxa de má classificação da árvore T é $R(T) = \sum_{t\in\tilde{T}} \emptyset(t)$, sendo $\emptyset(t)$ uma medida de heterogeneidade calculada em um nó $t \in T$. Aumentando o valor de α a partir de zero, tem-se uma sequência de árvores de tamanho decrescente. Depois de

aplicada a poda e gerada uma sequência de subárvores, deve-se selecionar a que será o classificador final.

O classificador final é aquele que possui melhor capacidade de classificação conforme a estimativa de erro encontrada por meio da técnica de cross validation. Logo, a árvore escolhida será a que minimiza o valor do erro calculado.

3.1.2 Aprendizado Baseado em Instâncias

O aprendizado baseado em instâncias, também conhecido como *lazy* (preguiçoso) e aprendizado por memorização, é a forma mais simples de aprendizagem. Este tipo de classificador é considerado livre de modelo, visto que não generaliza um modelo derivado dos dados para predizer as classes (TAN; STEINBACH; KUMAR, 2009).

Este método de aprendizado emprega as instâncias de entrada a fim de atribuir o valor ou classe para as novas que forem apresentadas, para isso utiliza uma medida de distância que define qual membro do conjunto de treinamento está mais próximo da instância que se deseja classificar (WITTEN; FRANK; HALL, 2011).

O aprendizado baseado em instâncias é vantajoso quando a função objetivo é muito complexa para ser generalizada, sendo possível a sua definição por meio de funções de aproximações locais de complexidade menor (MITCHELL; BLUM, 1997). No entanto, em grandes bases de dados a classificação pode ter um custo alto, pois calculam-se os valores de proximidade individualmente entre os exemplos de teste e de treinamento (TAN; STEINBACH; KUMAR, 2009).

Dentre os algoritmos deste método, o mais tradicional é o algoritmo k-vizinhos mais próximos, kNN.

3.1.2.1 Algoritmo k-Vizinhos mais Próximos

O algoritmo kNN é um dos principais métodos de classificação de instâncias, sendo simples e não necessitando de treinamento para ser aplicado (GOLDSCHMIDT; PASSOS, 2005).

Este algoritmo representa cada instância como um ponto de dado em um espaço de *n* dimensões, em que *n* é o número de atributos. Dada uma instância de teste, calcula-se a sua proximidade com os demais pontos de dados do conjunto de treinamento, usando a medida de distância. Os *k* vizinhos mais próximos de uma determinada instância *z*

são os *k* pontos que estão mais perto de *z* (TAN; STEINBACH; KUMAR, 2009).

A medida de distância entre os pontos mais empregada com o kNN e que foi utilizada nesta pesquisa é a distância euclidiana (equação 13). Sendo $X=(x_1, x_2,...,x_n)$ e $Y= (y_1, y_2,..., y_n)$ dois pontos no espaço $\Re^n$, a distância euclidiana é:

$$dist_e(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (13)$$

Portanto, no algoritmo kNN a distância entre cada instância de teste e todas as de treinamento são calculadas, como por exemplo pela equação (13), a fim de se determinar a lista dos vizinhos mais próximos. Evidentemente que podem ser aplicadas outras medidas de distância. Este algoritmo, assim que obtém esta lista, classifica a nova instância baseando-se na classe majoritária entre os k pontos mais próximos (equação 14), sendo *v* o rótulo da classe, y_i é o rótulo de classe para um dos vizinhos mais próximos e *I(.)* a função indicadora que retorna 1 se o argumento for verdadeiro e 0 caso contrário (TAN; STEINBACH; KUMAR, 2009).

$$y' = \underset{v}{argmax} \sum_{(x_i,y_i)\epsilon D_z} I(v = y_i) \quad (14)$$

Nesta abordagem cada vizinho possui a mesma influência na classificação, sendo a classe predita pelo algoritmo kNN dependente do valor de k, sendo uma forma de reduzir o seu impacto a determinação de pesos referentes a influência de cada vizinho mais próximo.

3.1.3 Redes Neurais Artificiais

As Redes Neurais Artificiais (RNA) são modelos matemáticos que se inspiram nas estruturas neurais biológicas e que apresentam capacidade computacional por meio de aprendizado e generalização (HAYKIN, 2001).

O aprendizado é uma das principais características de uma RNA, pois possibilita que ela melhore o seu desempenho ajustando os pesos sinápticos de forma iterativa, conforme a sua resposta ao ambiente. A

generalização de uma RNA refere-se a capacidade de fornecer respostas coerentes para dados que não lhes foram apresentados durante a fase de treinamento (REZENDE, 2005).

Os modelos de RNA consistem de uma série de nós, neurônios artificiais, que representam uma unidade de processamento e são unidos por meio de ligações direcionais que especificam uma relação causal entre os nós conectados. O processamento é distribuído pelas camadas de neurônio e paralelo, pois os neurônios dentro das camadas processam as suas entradas de forma simultânea e independente (KANTARDZIC, 2011; LUGER, 2004).

Dentre as tarefas para as quais as RNA são adequadas tem-se a classificação, pois possuem a habilidade de atribuir um padrão de entrada desconhecido a uma classe, sendo adequado à resolução de problemas com pouco conhecimento das relações entre os atributos e as classes (LUGER 2004; REZENDE, 2005; WANG et al, 2005).

Além da capacidade de aprendizado, as RNA também possuem outras características semelhantes as do cérebro humano que são interessantes para as tarefas de *data mining* (HAYKIN, 2001; KANTARDZIC, 2011):

a) busca paralela: nas RNA o conhecimento é distribuído, portanto a busca pela informação se dá de forma paralela e não seqüencial;
b) capacidade de adaptação: adaptam os pesos sinápticos conforme as modificações no ambiente, tornando-se úteis para classificação de padrões;
c) tolerância a falhas: as RNA são tolerantes a falhas, pois a sua performance não diminui significativamente em condições operacionais adversas, como por exemplo, na presença de dados ruidosos ou em falta;
d) generalização: esta capacidade permite a RNA fazer boas classificações de dados novos e incompletos;
e) resposta a evidências: no contexto da classificação de dados uma RNA pode prover informações não somente sobre a classe selecionada para uma determinada amostra, mas também sobre a confiabilidade da decisão tomada. Isso faz com que a rede seja capaz de rejeitar dados ambíguos e melhorar o desempenho da classificação.

O termo RNA se popularizou a partir de 1980, quando surgiram vários tipos como os Perceptrons de Múltiplas Camadas[8] (Multi-Layer Perceptrons (MLP)) e as Redes de Função de Base Radial (Radial Basis Functions (RBF)).

3.1.3.1 Redes de Função de Base Radial

De acordo com Haykin (2001) nas RNA aprender é encontrar uma superfície em um espaço multidimensional que seja capaz de fornecer o melhor ajuste para os dados do conjunto de treinamento e generalizar é empregar esta superfície para interpolar os dados do conjunto de testes.

Considerando-se esta abordagem têm-se as redes de Função de Base Radial (RBF) que são um tipo de RNA que possuem apenas uma camada intermediária que é formada por neurônios artificiais, cada um deles implementa uma função de base radial. A camada de entrada é não linear e a de saída é linear (BORS, 2001).

A rede RBF é um modelo neural multicamadas, com duas camadas, capaz de aprender rapidamente padrões complexos e resolver problemas que não são linearmente separáveis, adaptando-se a mudanças. Em função disso, têm destaque no domínio de RNA, pois apresentam melhores tempos no processo de treinamento e eficiência computacional (HAYKIN, 2001).

Este tipo de RNA tem uma estrutura semelhante a de uma rede MLP com uma camada oculta, exceto que cada nó da camada oculta utiliza uma função, normalmente a gaussiana para realizar a aproximação de funções (ajuste de curva) em um espaço de alta dimensionalidade. A ativação de um neurônio da camada oculta é determinada pela distância entre os vetores de entrada e peso, produzindo uma resposta localizada para o estímulo de entrada (THEODORIDIS; KOUTROUMBAS, 2006).

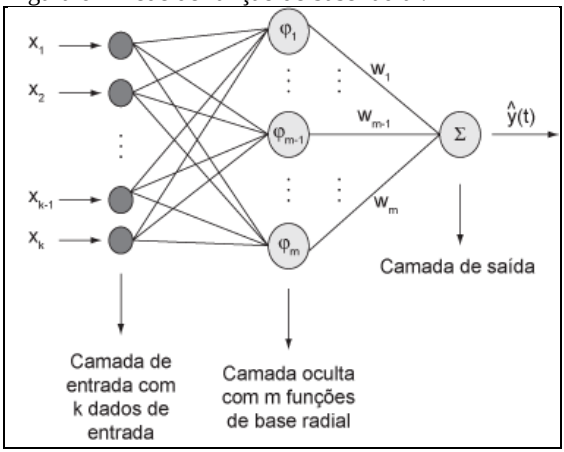
A construção de uma rede RBF envolve três camadas (figura 6) com papéis distintos, as quais são: de entrada, oculta e de saída.

A camada de entrada é formada por nós que conectam a rede ao seu ambiente. A segunda camada, a única oculta da rede, aplica uma transformação não linear do espaço de entrada para o oculto, que na maioria das vezes possui alta dimensionalidade. A distância entre o

[8] As MLP são aquelas redes que se caracterizam pela existência de uma ou mais camadas ocultas entre as camadas de entrada e saída, propagando-se o sinal de entrada para a frente por meio de cada camada da RNA (HAYKIN, 2001).

vetor de entrada e o centro da unidade radial, formado por cada neurônio da camada oculta, é o responsável pela ativação do neurônio. A camada de saída é linear e fornece a resposta da rede ao padrão que foi apresentado à camada de entrada (HAYKIN, 2001).

Figura 6 – Rede de função de base radial.



Fonte: Coelho, Santos e Costa Júnior (2008).

Conforme a figura 6, φ é a função de base radial para o neurônio da camada oculta, sendo que a mais aplicada é a Gaussiana (RUSSEL; NORVIG, 2004). Nos casos em que se tem apenas duas classes, basta um neurônio de saída.

O treinamento das redes RBF é realizado considerando-se a seleção dos centros de cada função de base radial da camada oculta; a determinação do raio da função de base em relação ao centro; o mapeamento do espaço não linear e finalmente o projeto da camada de saída.

Na etapa inicial de processamento do algoritmo, selecionam-se os centros de cada função de base radial da camada oculta de forma aleatória a partir do conjunto de treinamento, devendo este ser representativo do domínio de aplicação (GUERRA, 2006).

Após a inicialização dos centros, calcula-se o raio das RBF que define o espalhamento dos dados representados pela RBF em torno do seu centro (GUERRA, 2006), o qual é encontrado por meio da equação (15), que considera a distância máxima entre os centros dos neurônios da camada oculta, a qual pode ser obtida pela distância euclidiana

(equação (16)) e *H* refere-se a quantidade de neurônios (HAYKIN, 2001).

$$\sigma = \frac{dist_{max}(c_i, c_j)}{\sqrt{2H}}, \forall i \neq j \tag{15}$$

$$dist_{max}(c_i, c_j) = max_{\forall i \neq j}\{\|c_i - c_j\|\} \tag{16}$$

Definido o raio das RBF realiza-se o mapeamento do espaço não linear. Portanto, primeiramente, para cada vetor de entrada *x* que foi apresentado à RNA na iteração *t* calcula-se, pela equação (17), para cada neurônio oculto o seu valor de ativação. A equação (18) consiste na distância euclidiana utilizada pela função da equação (17).

$$u_i(t) = \|x(t) - c_i(t)\|, i = 1, \ldots . H \tag{17}$$

$$u_i(t) = \sqrt{\sum_{i=1}^{H} [x(t) - c_i(t)]^2} \tag{18}$$

Sendo *H* correspondente ao número de funções de base da camada oculta e c_i o vetor dos centros de cada função *i* definido pelos pesos que conectam cada entrada a função de base.

Definidos estes valores pode-se calcular a saída de cada neurônio da camada oculta pela função Gaussiana, equação (19), que consiste na função de base mais utilizada em redes RBF (HAYKIN, 2001; RUSSEL; NORVIG, 2004).

$$\varphi_i(t) = exp\left(-\frac{u_i^2(t)}{2\sigma_i^2}\right) \tag{19}$$

Tendo-se que σ_i é o raio da função de base e o neurônio *i* fornece resposta máxima, ($\varphi_i(t) \approx 1$), para vetores de entrada próximos do seu centro c_i. Assim, segundo Haykin (2001) cada neurônio da camada oculta é ativado sempre que o vetor de entrada estiver próximo do seu centro.

Após a transformação do espaço de entrada não linear em linear, realiza-se o ajuste dos pesos de saída da rede. No caso de problemas de classificação, pode-se optar pela utilização da regra de aprendizagem do

perceptron simples, tendo-se a saída *o* de cada neurônio *k* da camada de saída dada pela regra (20) e $u_k(t)$ definido pela equação (21), tendo-se *H* como a quantidade de funções de base e m_{ki} como os pesos das saídas (HAYKIN, 2001).

$$o_k(t) = \begin{Bmatrix} 1, & u_k(t) \geq 0 \\ 0, & u_k(t) < 0 \end{Bmatrix} \tag{20}$$

$$u_k(t) = \sum_{i=1}^{H} m_{ki}(t)\, \varphi_i(t) \tag{21}$$

Logo, encontrando-se o valor de $u_k(t)$ pode-se definir a saída do neurônio o_k que deve ser *1* para a classe a qual pertence, enquanto as dos outros neurônios será *0*.

A seguir, calcula-se o erro *e*, equação (22), que quando for diferente de zero será utilizado na atualização dos pesos sinápticos empregando-se a equação (23).

$$e_k(t) = d_k(t) - o_k(t) \tag{22}$$

$$m_{ki}(t+1) = m_{ki}(t) + \eta e_k \varphi_i(t) \tag{23}$$

Sendo η a taxa de aprendizagem definida para a rede, $e_k(t)$ o erro na saída, $d_k(t)$ o vetor de saídas desejadas e $o_k(t)$ a saída real da rede.

Após a apresentação de todos os vetores de treinamento à rede, tem-se o final de uma época de treinamento. Deve-se então calcular, para cada época até a convergência do algoritmo, o erro médio quadrático, equação (24), em que *N* é o número total de vetores de treinamento. A convergência ocorre quando o valor obtido para *E* for inferior ao valor máximo permitido ou o algoritmo executar o número de épocas definidas pelo usuário (HAYKIN, 2001).

$$E = \frac{1}{N} \sum_{i=1}^{N} \big(e_k(t)\big)^2 \tag{24}$$

Finalizado o treinamento, na fase de teste do classificador o vetor de entrada será associado a classe que gerar o maior valor de saída para $o_k(t)$ (BISHOP, 1995).

3.1.4 Classificadores Bayesianos

Classificadores bayesianos são aqueles que se valem da estatística para prever as probabilidades de associação a uma determinada classe, como por exemplo, estimar que um determinado registro pertença a uma classe em particular (HAN; KAMBER; PEI, 2011).

Em algumas aplicações, o rótulo da classe de um registro de teste não pode ser previsto com certeza, apesar dos seus atributos serem iguais a alguns do conjunto de treinamento. Nestes casos de incerteza empregam-se estes classificadores, já que combinam o conhecimento prévio que se tem sobre uma hipótese com novas evidências acerca dos dados que serão analisados (TAN; STEINBACH; KUMAR, 2009).

Os classificadores bayesianos baseiam-se no Teorema de Bayes, que foi desenvolvido por Thomas Bayes e publicado, após sua morte, em 1764 em obra intitulada "*An Essay Towards Solving a Problem in the Doctrine of Chances*", porém na época foi ignorado até que Pierre Simon Laplace introduziu o uso da inferência bayesiana em publicação de 1774. Desde então, o teorema de Bayes é bastante aplicado e frequentemente encontrado na literatura de estatística moderna (STIGLER, 1982; WALPOLE et al, 2009). Os métodos bayesianos têm sido aplicados com sucesso em áreas como engenharia, agricultura, ciências biomédicas, entre outras (WALPOLE et al, 2009).

O teorema de Bayes se fundamenta no cálculo da probabilidade *a posteriori*, $P(H|X)$, a partir das probabilidades *a priori*, $P(H)$, e das probabilidades condicionais $P(X|H)$ (DEVORE, 2006).

O teorema de Bayes, também conhecido por lei de Bayes ou regra de Bayes (RUSSEL; NORVIG, 2004), representa uma base teórica para abordagem estatística em problemas de classificação por meio da inferência probabilística. Seja X uma amostra de dados, cujo rótulo de classe é desconhecido e H algumas hipóteses. O teorema de Bayes fornece uma forma de calcular a probabilidade *a posteriori* $P(H|X)$ usando as probabilidades de $P(H)$, $P(X)$ e $P(X|H)$ (KANTARDZIC, 2011). A relação básica é apresentada na equação (25).

$$P(H|X) = \frac{P(H).P(X|H)}{P(X)} \tag{25}$$

Considerando-se que $P(X)$ é desconhecido, esta probabilidade pode ser reescrita pela aplicação da lei ou teorema da probabilidade

total: sejam $H_1, \dots, H_k$ uma coleção de k eventos mutuamente exclusivos e exaustivos com $P(H_i) > 0$ para $i = 1, \dots, k$, então para qualquer outro evento X tem-se a equação (26) (DEVORE, 2006).

$$P(X) = P(H_1).P(X|H_1) + \cdots + P(H_k).P(X|H_k)$$
$$P(X) = \sum_{i=1}^{k} P(H_i).P(X|H_i) \qquad (26)$$

A partir desta expressão, pode-se reescrever o teorema de Bayes substituindo-se $P(X)$, conforme a equação (27).

$$P(H_j|X_j) = \frac{P(H_j).P(X|H_j)}{\sum_{i=1}^{k} P(H_i).P(X|H_i)}, \qquad j = 1, \dots, k \qquad (27)$$

O teorema de Bayes pode parecer pouco aplicável, pois exige três termos (uma probabilidade condicional e duas incondicionais) para calcular uma probabilidade condicional. Porém na prática é útil, porque se tem muitos casos em que se fazem boas estimativas de probabilidade para esses três termos e precisa-se calcular o quarto. Por exemplo, em uma tarefa como o diagnóstico médico, frequentemente tem-se probabilidades condicionais sobre relacionamentos causais e deseja-se chegar a um diagnóstico (RUSSEL; NORVIG, 2004).

Dentre os modelos de aprendizagem de máquina e de classificadores bayesianos um dos mais comumentes utilizados no *data mining* é o Classificador Ingênuo de Bayes ou Naive Bayes (RUSSEL; NORVIG, 2004).

3.1.4.1 Algoritmo Naive Bayes

O algoritmo Naive Bayes é um classificador bayesiano, também chamado de Classificador Ingênuo de Bayes. Essa denominação justifica-se por ser usado nas situações em que se assume a premissa dos atributos serem sempre independentes entre si, o que em muitos casos não ocorre. Na prática, classificadores ingênuos de Bayes podem funcionar bem, mesmo quando a hipótese de independência não é verdadeira (RUSSEL; NORVIG, 2004).

Supondo-se que há um conjunto de treinamento de *m* amostras $S = \{S_1, S_2, ..., S_m\}$sendo cada amostra S_i representada como um vetor de dimensão *n*, $\{x_1, x_2, ..., x_n\}$. Os valores de x_i correspondem,

respectivamente, aos atributos A_1, A_2, ..., A_n. Além disso, tem-se k classes, C_1, C_2, ..., C_k, e cada amostra pertence a uma destas classes. Dada uma amostra de dados adicional X, cuja classe é desconhecida, pode-se prever a sua classe utilizando a maior probabilidade condicional $P(C_i|X)$, sendo i = 1,..., k. Essa é a idéia básica do classificador Naive Bayes, que tem as suas probabilidades calculadas usando o teorema de Bayes, conforme a equação (28) (KANTARDZIC, 2011).

$$P(C_i|X) = \frac{P(C_i).P(X|C_i)}{P(X)} \tag{28}$$

Considerando $P(X)$ constante para todas as classes, basta escolher a classe que maximiza o produto $P(C_i).P(X|C_i)$. As probabilidades *a priori* da classe são calculadas como $P(C_i)$ que corresponde ao número de amostras de treinamento da classe $C_i|m$, em que m é o número total de amostras do conjunto de treinamento, equação (29).

$$P(C_i) = \frac{C_i}{m}, \qquad i = 1, \dots, m \tag{29}$$

Devido ao cálculo de $P(X|C_i)$ ser complexo, em especial para grandes conjuntos de dados, a suposição ingênua de independência condicional entre os atributos é feita. Considerando-se isso, pode-se expressar a $P(X|C_i)$ como um produto, conforme a equação (30).

$$P(X|C_i) = \prod_{t=1}^{n} P(X_t|C_i) \tag{30}$$

Assim, pode-se dizer que um registro é classificado pela equação (31) que calcula a probabilidade *a posteriori* para cada classe C_i, escolhendo-se a classe que maximiza o produto.

$$P(C_i|X) = P(C_i).\prod_{t=1}^{n} P(X_t|C_i) \tag{31}$$

A avaliação da probabilidade condicional $P(X_t|C_i)$ é realizada empregando abordagens diferentes para atributos categóricos ou

contínuos. No caso de um atributo categorizado X_t, a probabilidade condicional $P(X_t|C_i)$ é avaliada conforme a fração do número de instâncias do conjunto de treinamento com classe C_i que recebem como atributo o valor de X_t pelo número total de amostras que pertencem a classe C_i (HAN; KAMBER; PEI, 2011).

No entanto, se o atributo X_t for contínuo têm-se duas opções: pode-se categorizá-lo e proceder a avaliação da forma explicada anteriormente, ou supor uma determinada forma de distribuição de probabilidades para a variável contínua, geralmente a distribuição Gaussiana (equação 32), em que μ_{C_i} e σ_{C_i} são respectivamente a média e o desvio padrão dos valores das amostras de treinamento C_i (HAN; KAMBER; PEI, 2011; RUSSEL; NORVIG, 2004).

$$P(X_t|C_i) = \frac{1}{\sqrt{2\pi\sigma_{C_i}}} exp^{-\frac{\left(x_t-\mu_{c_i}\right)^2}{2{\sigma_{c_i}}^2}} \quad (32)$$

Após a avaliação da probabilidade condicional $P(X_t|C_i)$ os valores encontrados são substituídos na equação (31).

O algoritmo Naive Bayes apresenta problemas com a avaliação de probabilidades *a posteriori* a partir do conjunto de treinamento, pois ele requer que cada probabilidade condicional, $P(X|C_i)$, não seja nula. Porém, nos casos que a probabilidade condicional de classe para um dos atributos for zero o algoritmo pode classificar incorretamente o registro. Portanto, a abordagem de avaliação empregando frações simples pode trazer estes problemas quando se tem poucos dados de treinamento e o número de atributos for grande. Logo, a solução é o emprego da correção de Laplace, equação (33), em que n_c é o número de instâncias com o valor de X_t que pertencem a classe C_i, n é o número total de dados do conjunto de treinamento com classe C_i e t refere-se à quantidade máxima de valores de X_t (HAN; KAMBER; PEI, 2011).

$$P(\mathrm{X_t}|\mathrm{C_i}) = \frac{n_c + 1}{n + t} \quad (33)$$

O Naive Bayes apresenta como características ser um algoritmo robusto na presença de ruídos e de valores faltantes, como também em relação a atributos irrelevantes; possui taxa de erro menor em comparação aos outros classificadores desenvolvidos em *data mining*, porém isso nem sempre é o caso já que a suposição de independência

pode não funcionar bem em alguns domínios (HAN; KAMBER; PEI, 2011; TAN; STEINBACH; KUMAR, 2009).

3.1.4.2 Redes de Crenças Bayesianas

Os classificadores bayesianos simples como o Naive Bayes que possuem a suposição de independência condicional podem ser uma abordagem rígida quando os atributos são um pouco correlacionados. As redes de crença bayesianas, também conhecidas como redes probabilísticas e redes bayesianas, modelam as probabilidades condicionais da classe P(X|Y), codificando a dependência entre as variáveis.

As redes de crenças bayesianas (Bayes Net) possuem dois elementos: um grafo acíclico direcionado que codifica as relações de dependência entre um conjunto de variáveis e uma tabela de probabilidades (HAN; KAMBER; PEI, 2011).

Modelos como este, baseado em rede bayesiana, podem ser usados para aprender as relações causais, auxiliando no entendimento do domínio de aplicação, como também prevendo as consequências de uma determinada intervenção (CLICKERING; HECKERMAN, 1997; WITTEN; FRANK; HALL, 2011).

Nesta abordagem um dos vértices da rede bayesiana é considerado o atributo classe, podendo haver vários atributos classe em uma rede bayesiana. Se tiver um único atributo classe assumindo os valores C_1, ...,C_m, a saída será a distribuição de probabilidade $P[C_1|X]$, ..., $P[C_m|X]$. Caso contrário, se existirem diversos atributos classe, assumindo os valores $C_1^1, \dots, C_1^m$ para a classe 1 e $C_2^1, \dots, C_2^l$ para a classe 2, o algoritmo retorna a distribuição de probabilidade $P[C_1^1|X], \dots, P[C_1^m|X], P[C_2^1|X], \dots, P\left[C_2^l|X\right]$.

A classificação pelas redes de crenças bayesianas é semelhante a pelo algoritmo Naive Bayes, objetivando-se maximizar a probabilidade condicional *a posteriori*, aplicando-se para isso o Teorema de Bayes (equação 28).

Nas redes de crenças bayesianas tem-se uma tabela de probabilidade condicional para cada variável. Por exemplo, para uma variável Y esta tabela especifica a distribuição condicional $P(Y|pais(Y))$. Seja $X=(x_1,...,x_n)$ uma tupla de dados descrita por atributos Y_1, ..., Y_n, tem-se a equação (34) que permite a rede fornecer uma representação completa da distribuição de probabilidade conjunta (HAN; KAMBER; PEI, 2011):

$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P\left(x_i | pais(Y_i)\right) \quad (34)$$

Sendo $P(x_1,\dots,x_n)$ a probabilidade dos valores de X e $P\left(x_i | pais(Y_i)\right)$ os valores correspondentes as entradas na tabela de probabilidade condicional para Y_i.

3.1.5 Metaclassificadores

Os metaclassificadores agregam as previsões de um conjunto de classificadores a partir dos dados de treinamento e executam a classificação recebendo o voto sobre as previsões realizadas por este classificador base.

Estes metaclassificadores podem ser criados de variadas formas, sendo uma delas a manipulação do conjunto de treinamento. Neste caso, desenvolvem-se múltiplos conjuntos pela reamostragem dos dados originais, conforme uma distribuição de amostras que determina o quanto um exemplo pode ser selecionado para treinamento, o que varia de um julgamento para outro. Partindo-se de cada conjunto de treinamento e por meio de um algoritmo de aprendizagem origina-se um classificador. O *boosting* é um exemplo de metaclassificador que manipula seus conjuntos de treinamento (TAN; STEINBACH; KUMAR, 2009).

O *boosting* é iterativo e altera a distribuição de exemplos de treinamento a fim de que os classificadores de base se dediquem aqueles que são mais difíceis de classificar. Nesta abordagem é definido um peso para cada exemplo de treinamento, o qual é empregado pelo classificador base na descoberta de um modelo. Estes pesos são atualizados a cada iteração, aumentando-se os valores daqueles cujos exemplos foram classificados incorretamente (HAN; KAMBER; PEI, 2011).

Ao longo dos anos várias implementações de *boosting* têm sido desenvolvidas, as quais diferem na forma de atualização dos pesos a cada iteração e na combinação das previsões de cada classificador. Dentras estas, tem-se o algoritmo *Adaptive Boosting* que é um dos mais populares segundo Han, Kamber e Pei (2011).

3.1.5.1 Algoritmo *Adaptive Boosting*

O *Adaptive Boosting* (AdaBoost) é um dos algoritmos metaclassificadores mais famosos e utilizados, especialmente pela sua adaptação aos classificadores base, favorecendo os exemplos que foram classificados erroneamente nas classificações anteriores (REIS, 2013; TAN; LI; QIN, 2007; WU et al, 2008).

No AdaBoost considerando-se um conjunto composto por N exemplos de treinamento representado por $\{(x_i, y_i)|j = 1,2, \dots, N\}$, um classificador base C_i depende da sua taxa de erro (equação 35) e tem sua importância definida pela equação (36). Tendo-se que $I(p)=1$ se p for verdadeiro e 0 caso contrário (TAN; STEINBACH; KUMAR, 2009).

$$\epsilon_i = \frac{1}{N}\left[\sum_{j=1}^{N} w_j \, I\left(C_i(x_j) \neq y_i\right)\right] \tag{35}$$

$$\alpha_1 = \frac{1}{2} ln\left(\frac{1-\epsilon_1}{\epsilon_1}\right) \tag{36}$$

O parâmetro α_i apresenta valor positivo se a taxa de erro estiver próxima de zero e valor negativo quando está próxima de um. Ele é usado para atualizar o peso dos exemplos de treinamento, sendo o mecanismo de atualização de pesos do AdaBoost definido pela equação (37), em que $w_i^{(j)}$ corresponde ao peso atribuído para o exemplo (x_i,y_i) na iteração *boosting* de índice j; Z_j é o fator de normalização que assegura que o somatório de $w_i^{(j+1)}$ é igual a um (TAN; STEINBACH; KUMAR, 2009).

$$w_i^{(j+1)} = \frac{w_i^{(j)}}{Z_j} * \begin{cases} exp^{-\alpha_j} se\ C_j(x_i) = y_i \\ exp^{\alpha_j}\ se\ C_j(x_i) \neq y_i \end{cases}, \tag{37}$$

A equação (37) aumenta os pesos de exemplos classificados incorretamente e diminui o das classificações corretas.

Várias versões do AdaBoost foram desenvolvidas, como o AdaBoost.M1 que pode também ser empregado em problemas de classificação com múltiplas classes e não somente binárias.

4 MEDIDAS DE QUALIDADE EM DATA MINING

As relações descobertas pelo *data mining* somente serão relevantes para o usuário se forem consideradas potencialmente úteis, de fácil compreensão e válidas (HAN; KAMBER; PEI, 2011).

A fim de auxiliar nesse processo de determinação da qualidade das relações encontradas pelos algoritmos de *data mining*, os resultados são validados com o objetivo de verificar a solução encontrada, bem como o desempenho do algoritmo, para isso empregam-se medidas de qualidade que consistem em índices estatísticos (GUILLET; HAMILTON, 2010).

Na área de *data mining* não existe um único algoritmo que apresente o melhor desempenho para todos os problemas. Portanto, deve-se compreender os benefícios e as limitações desses algoritmos quando aplicados em conjuntos de dados diferentes, para isso deve-se empregar uma metodologia de avaliação que permita compará-los (DIETTERICH, 1998; TAN; STEINBACH; KUMAR, 2009).

As medidas de qualidade serão abordadas conforme os modelos de *data mining* empregados que compreendem classificadores.

4.1 AVALIAÇÃO DOS CLASSIFICADORES

O grau de relevância das informações adquiridas no processo de *data mining* pode ser observado por meio das avaliações de desempenho do algoritmo classificador, geralmente representada pela taxa de erros resultantes da classificação (WITTEN; FRANK; HALL, 2011).

A análise de desempenho empregada nos classificadores pode ser realizada por meio de uma matriz de confusão (tabela 1), resultante do processo de classificação pelos métodos de testes *holdout* ou *cross-validation* de *k* partes.

No método *holdout* os dados originais são particionados em dois conjuntos, sendo um de treinamento e o outro de teste. O modelo de classificação é induzido a partir do conjunto de treinamento e tem o seu desempenho avaliado no conjunto de teste. Neste método a proporção dos dados reservados para treinamento e para teste fica em 50% para cada conjunto ou 2/3 para treinamento e 1/3 para teste (TAN; STEINBACH; KUMAR, 2009; WITTEN; FRANK; HALL, 2011).

No método *cross-validation*, validação cruzada de *k* partes, segmentam-se os dados em *k* partições de tamanho igual. Durante cada execução, uma das partições é escolhida para teste, enquanto as demais são utilizadas para treinamento. O processo se repete *k* vezes até que

cada partição seja usada para teste uma vez (TAN; STEINBACH; KUMAR, 2009). Esta abordagem foi escolhida em função da vantagem, apontada pelos mesmos autores, de utilizar todos os dados possíveis para treinamento, cobrindo efetivamente o conjunto de dados.

Tabela 1 – Matriz de confusão.

Classe	**Predita C_+**	**Predita C_-**
Verdadeira C_+	***Verdadeiros Positivos***	***Falsos Negativos***
Verdadeira C_-	***Falsos Positivos***	***Verdadeiros Negativos***

Fonte: Adaptado de Goldschmidt e Passos (2005).

As células destacadas em cinza na tabela 1 formam a diagonal principal da matriz de confusão que representa o número de classificações corretas para cada classe. Todos os elementos fora dessa diagonal representam os erros na classificação.

Os dados registrados na matriz de confusão possibilitam identificar os seguintes valores (GUILLET; HAMILTON, 2010):

a) **verdadeiros positivos (VP):** refere-se à quantidade de registros positivos classificados corretamente como tal;
b) **falsos positivos (FP):** corresponde aos registros negativos classificados incorretamente como positivos;
c) **verdadeiros negativos (VN):** registros negativos classificados corretamente como negativos;
d) **falsos negativos (FN):** diz respeito ao número de registros positivos incorretamente classificados como negativos.

As variáveis identificadas por meio da matriz de confusão permitem o cálculo de índices para análise da qualidade de um modelo classificador, tendo-se as seguintes métricas de avaliação de desempenho (GUILLET; HAMILTON, 2010):

a) **sensibilidade:** habilidade do modelo em identificar os registros que pertencem verdadeiramente à classe considerada. Equivale à proporção de verdadeiros positivos, sendo definida pela equação (30);

$$S = \frac{VP}{(VP + FN)} \tag{30}$$

b) **especificidade:** capacidade do modelo em identificar registros que não pertencem à classe considerada. É calculada considerando-se quantidade de negativos verdadeiros, equação (31);

$$E = \frac{VN}{(VN + FP)} \tag{31}$$

c) **acurácia:** grau de exatidão que o modelo apresenta para identificar as classes por meio da relação entre os valores estimado e real, equação (32);

$$A = \frac{VP + VN}{(VP + VN + FP + FN)} \tag{32}$$

d) **erro:** refere-se a taxa de erro da classificação geral do modelo, equação (33);

$$e = 1 - A \tag{33}$$

e) **confiabilidade positiva:** é a capacidade do classificador em identificar corretamente os verdadeiros positivos, equação (34);

$$VPP = \frac{VP}{(VP + FP)} \tag{34}$$

f) **coeficiente de concordância Kappa:** é o coeficiente de avaliação da concordância entre dois ou mais métodos de classificação, baseia-se no número de respostas concordantes entre estes classificadores, sendo calculado pela equação (35), em que P_o é a proporção de acordo observado e P_a é a proporção de acordo devido ao acaso;

$$k = \frac{P_o - P_a}{1 - P_a} \tag{35}$$

O coeficiente Kappa, de acordo com a tabela 2, pode variar entre zero e um, nenhuma e total concordância respectivamente. A maioria dos estatísticos prefere que os

valores kappa sejam maiores que 0,6, sendo ideal aqueles superiores a 0,7. Este índice é considerado como uma medida apropriada da exatidão de um classificador, pois ele representa inteiramente a matriz de confusão, ao invés de incluir apenas os elementos que pertencem a diagonal principal (BRITES, 1996; GORELICK; YEN, 2006; LANDIS; KOCH, 1977; SCHWARTSMANN et al, 2006).

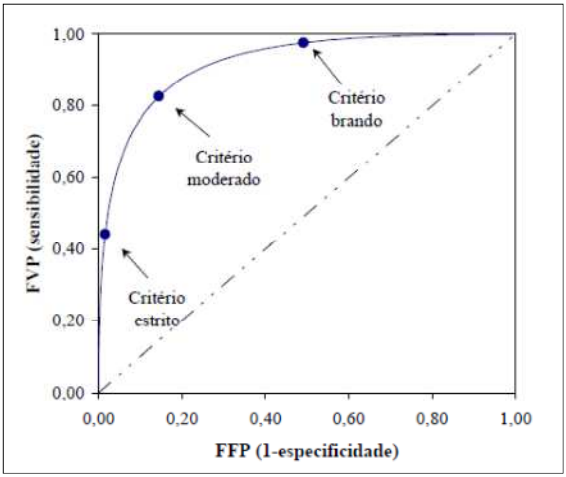
Tabela 2 – Classificação do coeficiente de concordância Kappa.

Índice Kappa	Interpretação
< 0	Ausência de acordo
0 – 0,2	Ruim
0,21 – 0,4	Fraca
0,41 – 0,6	Média
0,61 – 0,8	Boa
0,81 – 1	Excelente

Fonte: Adaptado de Landis e Koch (1977).

g) ***Receiver Operating Characteristic* (ROC) *Curve*:** a curva ROC é um método gráfico para avaliação, organização e seleção, sendo uma ferramenta útil para a avaliação de modelos de classificação, especialmente em domínios em que se tem desproporção entre as classes . A curva ROC consiste em um gráfico da sensibilidade (taxa de verdadeiros positivos) representada no eixo *y* e da taxa de falsos positivos (1-especificidade) representa no eixo *x* (figura 7). Esta medida permite evidenciar os valores em que se tem uma maior otimização da sensibilidade em função da especificidade. Dessa forma, um bom modelo de classificação deve estar localizado o mais próximo possível de um para a sensibilidade e de zero para a taxa de falsos positivos. A área sob a curva ROC fornece outra abordagem para avaliar qual modelo é em média melhor, sendo o ideal quanto mais próximo de 1 (TAN; STEINBACH; KUMAR, 2009).

Figura 7 – Curva ROC.



Fonte: Braga (2000).

As medidas de qualidade em *data mining* auxiliam na identificação do algoritmo mais indicado para resolver um determinado problema, tendo-se algumas bibliografias que trazem uma revisão acerca dos métodos como, por exemplo, em Dietterich (1998); Guillet e Hamilton (2010); Tan, Steinbach e Kumar (2009).

5 TRAUMATISMO CRANIOENCEFÁLICO

Este capítulo, inicialmente, compreende uma revisão do domínio de aplicação da pesquisa, abordando-se o traumatismo cranioencefálico e as suas principais características. Após, são apresentadas algumas das pesquisas desenvolvidas na área formalizando-se os trabalhos correlatos.

5.1 CONCEITOS BÁSICOS

O Traumatismo Cranioencefálico (TCE) consiste em uma agressão ao cérebro, ocasionando lesão anatômica ou comprometimento funcional do crânio, meninges ou encéfalo (DIAMENT; CYPEL, 1996). Esta agressão é causada por uma força física externa que pode diminuir ou alterar o estado de consciência e comprometer as capacidades cognitivas ou motoras (GRAHAM; CARDON, 2008).

A lesão tecidual provoca uma reação inflamatória, aumenta a circulação local e aparece o edema, portanto, necessita de cuidados para limitar os danos. Porém, quando esta lesão é no tecido nervoso intracraniano os problemas são maiores em função da dura-máter e a estrutura óssea não deixarem espaço para o cérebro aumentar de volume. Decorrente disso, tem-se um aumento da pressão intracraniana que acarreta prejuízos das funções vitais causando inconsciência, perda de memória, náuseas, vômitos, convulsões, dores de cabeça e coma (BRUNO; OLDENBURG, 2005).

O TCE ocasiona no sistema nervoso central várias alterações estruturais, fisiológicas e funcionais. Assim, pode acarretar a morte da vítima de TCE como também comprometer as suas capacidades cognitivas, físicas e comportamentais (CÉSPEDES et al, 2001; COLANTONIO et al, 2009).

A lesão encefálica definitiva, estabelecida após o TCE, resulta de mecanismos fisiopatológicos que podem ocorrer imediatamente ao acidente ou levar várias horas e dias para que o cérebro apresente os sinais clássicos de traumatismo. Além disso, as lesões previamente identificadas podem aumentar de tamanho ou surgirem novas nas primeiras 12 a 24 horas após o trauma (OLIVEIRA et al, 2010; TOYAMA et al, 2005).

O TCE é a maior causa de morbidade, mortalidade e limitações neurológicas entre jovens e adultos (ANDRADE et al, 2009; GRAHAM; CARDON, 2008). Este tipo de trauma é também um problema de saúde pública com reflexos socioeconômicos em todo o mundo, pois atinge principalmente indivíduos jovens em idade

produtiva (COLE, 2004; GHAJAR, 2000; IMHOF; LENZLINGER, 2005; SCHETTINO et al, 2006).

5.1.1 Classificação do Traumatismo Cranioencefálico

O TCE é classificado de acordo com o mecanismo da lesão, a gravidade clínica e a avaliação dos danos estruturais (MAAS; STOCCHETTI; BULLOCK, 2008).

Considerando-se o mecanismo da lesão o TCE pode ser fechado ou penetrante. O TCE fechado é aquele proveniente dos acidentes automobilísticos, quedas e agressões, enquanto o TCE penetrante é causado por armas de fogo ou brancas (MAAS; STOCCHETTI; BULLOCK, 2008).

O TCE em relação à gravidade clínica pode ser classificado em leve, moderado ou grave conforme a Escala de Coma de Glasgow, do inglês *Glasgow Coma Scale* (GCS). A escala foi descrita na Revista Lancet, em 1974, pelos pesquisadores da Universidade de Glasgow (Escócia), Teasdale e Jennet, como uma forma prática de avaliar o nível de consciência.

Teasdale e Jennet (1974) salientam que a GCS visa monitorar alterações no nível de consciência nas vítimas de trauma, baseando-se nas funções visuais, verbais e motoras, devendo-se avaliar os resultados mediante índices de pontuação. A correta aplicação desta escala é primordial para que se constitua em um indicador válido da condição clínica do indivíduo.

Desde então, a GCS tem sido uma medida objetiva da gravidade do trauma cranioencefálico, sendo empregada pelos profissionais de saúde para avaliar os achados neurológicos, comparar o efeito de tratamentos e como parâmetro indicador do prognóstico (GABBE; CAMERON; FINCH, 2003). A escala varia de 15 a 3 pontos, compreendo respectivamente, pessoas que abrem os olhos espontaneamente, obedecem a comando e encontram-se orientadas, e aquelas que possuem flacidez muscular, não abrem os olhos ou falam (GHAJAR, 2000).

Os indicadores utilizados na GCS devem ser avaliados de forma independente, pontuando-se conforme a resposta fornecida pelo paciente e os critérios presentes na figura 8. A categoria final da GCS é obtida somando-se os valores de cada indicador que foi avaliado.

O TCE é classificado conforme as seguintes categorias, considerando-se a pontuação na GCS, leve de 14 a 15, moderado de 9 a

13 e grave de 3 a 8 (CARO, 2011; MAAS; STOCCHETTI; BULLOCK, 2008).

Figura 8 – Escala de Coma de Glasgow

Abertura Ocular	
4	Espontânea
3	Ao estímulo verbal
2	Ao estímulo doloroso
1	Ausente

+

Resposta Motora	
6	Obedece comandos
5	Localiza a dor
4	Retirada ao estímulo doloroso
3	Flexão ao estímulo doloroso (postura decorticada)
2	Extensão ao estímulo doloroso (postura descerebrada)
1	Ausente

+

Resposta Verbal	
5	Orientado e conversando
4	Desorientado e conversando
3	Palavras inapropriadas
2	Sons incompreensíveis
1	Ausente

Grave						Moderado					Leve	
3	4	5	6	7	8	9	10	11	12	13	14	15

Fonte: Adaptado de Maas, Stocchetti e Bullock (2008).

Assim, para avaliar se uma pessoa encontra-se em coma, define-se que o escore da GCS menor que 8 é o mais aceito, compreendendo o estado em que não obedece a comandos, não articula palavras e não abre os olhos (GHAJAR, 2000).

A classificação do TCE pela gravidade clínica apresenta algumas limitações, como por exemplo, a variação na sua confiabilidade e a possibilidade de ser usada de forma inconsistente pelos cuidadores nos ambientes de atenção à saúde (STERNBACH, 2000). Também o nível de consciência pode ser afetado em quadros agudos por sedação, paralisias ou intoxicação (BALESTRERI et al, 2004; STOCCHETTI et al, 2004). No entanto, a GCS deve ser considerada como um dos parâmetros de avaliação que é útil na definição do estado da lesão cerebral, sendo reconhecida universalmente por auxiliar na previsão das conseqüências do TCE, formando também a base da tomada de decisão clínica como a necessidade de tomografia computadorizada, intervenção

cirúrgica e medicamentosa (CARO, 2011; MCNETT, 2007; TEASDALE; MURRAY, 2000).

O TCE também pode ser classificado pela avaliação dos danos estruturais por meio de neuroimagem. Este sistema não sofre influência dos fatores, citados anteriormente, sedação, paralisias ou intoxicação (MAAS; STOCCHETTI; BULLOCK, 2008). Desta forma, a classificação do TCE ocorre de acordo com os achados tomográficos (tabela 3), metodologia desenvolvida por Marshall e colaboradores, baseando-se em presença ou ausência de cisternas, lesões de alta ou mista densidade e desvio de linha média (MAAS et al, 2007; MARSHALL et al, 1991).

Tabela 3 – Classificação tomográfica da lesão cerebral difusa.

Categoria	Definição
I	Ausência de lesão intracraniana visível na tomografia
II	Cisternas presentes Desvio de linha média de 0a 5 mm ou lesão densa ou ambos Ausência de lesão de densidade alta ou mista >25ml Inclui fragmentos ósseos e corpos estranhos
III	Cisternas comprimidas ou ausentes Desvio de linha média de 0 a 5 mm Ausência de lesão de densidade alta ou mista >25ml
IV	Desvio da linha média >5mm Ausência de lesão de densidade alta ou mista >25ml

Fonte: Adaptado de Maas et al (2007).

A classificação de Marshall também apresenta algumas limitações, como por exemplo, a não distinção entre o tipo da lesão (hematomas extradurais[9] e subdurais[10]) e a não inserção da hemorragia subaracnóide[11], fatores estes considerados por vários autores como importantes para o prognóstico (CHESNUT et al, 2000; EISENBERG et

[9] Também denominado epidural, consiste em hemorragia entre a dura-máter e a calota óssea do crânio. A pessoa está lúcida e rapidamente fica inconsciente a medida que o hematoma se expande (BRUNO; OLDENBURG, 2012).

[10] Hematoma formado por hemorragia entre o cérebro e a dura-máter, que comprime lentamente o cérebro e provoca o aparecimento de sintomas típicos de TCE após horas ou até mesmo dias (BRUNO; OLDENBURG, 2012).

[11] Caracteriza-se por ruptura e sangramento abrupto no espaço compreendido entre o cérebro e as meninges (TURCATO; PEREIRA; GHIZONI, 2006).

al, 1990; GENNARELLI et al, 1982; ONO et al, 2001). Por outro lado, vários estudos confirmam o valor preditivo de prognóstico.

Conhecidas as diversas formas de classificação do TCE, pode-se dizer de uma forma geral, que este pode ser categorizado em leve, moderado ou grave.

No TCE leve a pessoa não apresenta perda de consciência, porém algumas vezes acontece um período curto de alteração; pode ocorrer síndrome pós traumática que compreende dores de cabeça, tontura, distúrbios de memória ou irritabilidade (WHYTE et al, 2002).

No TCE moderado tem-se de 9 a 13 pontos na GCS seis horas após o TCE, podendo apresentar sintomas semelhantes aos da categoria grave. Este tipo de traumatismo não tem sido estudado de forma tão específica quanto o grave e leve (ANDRADE et al, 2002).

No TCE grave tem-se a consciência comprometida apresentando uma pontuação inferior ou igual a 8 na GCS, podendo também apresentar uma amnésia pós traumática superior a sete dias. Nesta categoria tem-se uma maior incidência de óbito, porém nos casos em que sobrevivem apresentam significativas seqüelas de ordem física, cognitiva ou neurocomportamental (CAMARGO, 2003; WHYTE et al, 2002).

5.1.2 Epidemiologia do Traumatismo Cranioencefálico

O TCE é um dos tipos de traumas que mais acomete a população, ocorrendo geralmente lesões graves que levam a hospitalização. A sua incidência no mundo é estimada em 200 casos para cada 100 mil habitantes (BRUNS JUNIOR; HAUSER, 2003; CARO, 2011).

Este tipo de trauma é uma das principais causas de mortalidade e morbidade no Brasil e no mundo, sendo responsável por aproximadamente 50% dos óbitos associados a eventos traumáticos (BRUNS JUNIOR; HAUSER, 2003; DUTTON et al, 2010; SAATMAN et al, 2008). Além disso, estudos indicam que pacientes com TCE, ao serem comparados com outras vítimas de traumas, possuem um prognóstico pior em termos de mortalidade e morbidade (DUTTON et al, 2010).

As hospitalizações por TCE nos Estados Unidos são de 85 para cada 100 mil habitantes (RUTLAND-BROWN et al, 2006), já na União Européia os dados epidemiológicos indicam 235 internações por 100 mil habitantes (TAGLIAFERRI et al, 2006), valor este semelhante aos encontrados na Austrália (HILLIER; HILLIER; METZER, 1997 apud MAAS; STOCCHETTI; BULLOCK, 2008). No Brasil os estudos

indicam que variam de 36/100 mil habitantes (KOIZUMI et al, 2000) a 106,36/100 mil habitantes (SOARES; SCATENA; GALVÃO, 2008).

Os dados epidemiológicos demonstram que alguns segmentos da população são mais acometidos pelo TCE em função de estarem mais expostos as situações de violência e acidentes. Tanto os estudos internacionais (BRUNS JUNIOR; HAUSER, 2003; MAAS; STOCCHETTI; BULLOCK, 2008; MYBURG et al, 2008; WU et al, 2008) quanto os nacionais (KOIZUMI et al, 2000; MARTINS et al, 2009; MARTINS; SILVA; COUTINHO, 2003; MASINI, 1994; MELO; SILVA; MOREIRA JUNIOR, 2004) indicam que os homens e os jovens, principalmente abaixo dos 35 anos, são mais frequentemente vítimas de TCE. De acordo com Masson (2000) esta incidência ainda se torna maior quando se considera o grau de gravidade do traumatismo.

Estudos nacionais também mostram uma maior ocorrência de TCE na faixa etária abaixo dos 10 anos, por vezes ultrapassando a de adultos jovens (COLLI et al, 1997; KOIZUMI et al, 2000).

Os mecanismos de causa do trauma relacionam-se com as características sócio-econômicas da região de estudo e com a faixa etária da vítima (ROCHA, 2006). Os estudos que utilizam dados de atendimento em serviços de trauma mostram que os acidentes de trânsito são as circunstâncias mais importantes para este agravo no Brasil (MARTINS et al, 2009; MELO; SILVA; MOREIRA JUNIOR, 2004) e no mundo (MYBURG et al, 2008; WU et al, 2008).

A epidemiologia da causa do trauma tem sofrido influencia, nos últimos 20 anos, pelo aumento da idade da população. Notando-se uma diminuição na frequência de acidentes de trânsito e aumento das quedas (NIJBOER et al, 2007). No Brasil, de acordo com as internações hospitalares do Sistema Único de Saúde (BRASIL, 2011), os tipos de causa externa com maior ocorrência são as quedas e os acidentes de trânsito.

No TCE uma das causas importantes que levam ao óbito são os acidentes de trânsito (JOOSSE et al, 2009), o que pode estar relacionado aos indivíduos adultos jovens que são população de risco para este tipo de causa externa (COLANTONIO et al, 2009). No Brasil os acidentes de trânsito também têm apresentado uma maior relação com a mortalidade do que as quedas, provavelmente isso aconteça pelo estado mais grave das vítimas neste tipo de causa (SOUSA, 2009).

Considerando-se a gravidade do TCE, os categorizados como grave estão associados a uma taxa de mortalidade de 30 a 70% (KRAUS; MCARTHUR, 2006; OLIVEIRA; IKUTA; REGNER, 2008). Nos casos de recuperação, tem-se seqüelas neurológicas graves e a

qualidade de vida destas pessoas fica comprometida (FINFER; COHEN, 2001).

5.2 TRABALHOS CORRELATOS

A revisão referente ao TCE pretende apresentar as técnicas empregadas na análise dos dados, bem como aspectos relacionados com a mortalidade no Brasil e no mundo.

Os principais estudos internacionais encontrados na literatura nos últimos 14 anos, de 2000 a 2014, que incluem as variáveis relacionadas com a mortalidade no TCE são apresentados a seguir.

Bruns Junior e Hauser (2003) analisaram estudos epidemiológicos de TCE revisando a metodologia empregada e relatando características da incidência do TCE por idade, sexo, raça, variação geográfica e mortalidade. O objetivo do trabalho consistia na determinação da freqüência de TCE, os grupos de risco e as formas de mortalidade. Para isso, realizaram uma revisão bibliográfica de vários estudos na área, tanto nos Estados Unidos, Austrália e China, como em países da Europa e África. As diferenças nos métodos aplicados nos estudos resultam em estimativas divergentes o que dificulta a comparação entre estes. Considerando as variações, algumas tendências gerais são universais, concluindo-se que o TCE ocorre em frequências mais elevadas nos adultos jovens e nos idosos. Os homens apresentam maior risco de TCE, particularmente durante a adolescência e idade adulta jovem. A mortalidade varia de acordo com a gravidade do TCE, mas é alto em pessoas com ferimentos graves e nos idosos. Além disso, TCE é a maior causa de epilepsia entre os pacientes que sobrevivem.

Perel et al (2006) realizaram uma revisão sistemática dos modelos de prognóstico para TCE, identificando-os, apresentando suas características, investigando a sua qualidade e descrevendo os modelos que foram validados em uma população externa. Os estudos escolhidos foram os que combinam pelo menos duas variáveis para prever qualquer resultado em pessoas com TCE, incluindo-se estudos de acompanhamento clínico e não somente do momento da admissão. Estes trabalhos foram pesquisados no Pubmed e no Embase, selecionando-se aqueles de maior interesse conforme o escopo da pesquisa. Nesta revisão um total de 53 relatórios com 102 modelos foram identificados, sendo que 75% destes modelos incluem menos de 500 pessoas no estudo. A maioria dos modelos, 93%, foram baseados em populações de países de alta renda. Dentre as técnicas estatísticas, a regressão logística foi a mais empregada estando presente em 47% dos estudos. Em relação

a qualidade, menos da metade dos modelos foram validados (38%), dos quais 11% realizaram uma validação externa.

Mushkudiani et al (2008) descrevem as técnicas utilizadas nas pesquisas para a previsão precoce das conseqüências do TCE e identificam aspectos de melhorias. No desenvolvimento da pesquisa revisaram os principais aspectos metodológicos dos estudos (tabela 4), publicados entre 1970 e 2005, que propuseram um modelo de prognóstico baseado em dados de admissão dos pacientes com TCE moderada ou grave. Destes, 31 estudos relevantes foram identificados dos quais 22 relataram menos de 500 casos. Os preditores considerados para o modelo de prognóstico em sua maioria incluíam a idade, a gravidade pela GCS e a reatividade pupilar. A técnica estatística mais comumente empregada nas análises foi a regressão logística, estando presente em 19 trabalhos. Outras técnicas como redes neurais e análise discriminante também foram utilizadas. O desempenho dos modelos estudados foi muitas vezes quantificado pela eficiência (*accuracy)*. A validação do modelo foi abordada em 15 estudos e a validação externa foi realizada em apenas quatro. Eles concluíram que as estratégias de modelagem devem ser melhoradas e a validação externa incluída nas pesquisas.

Tabela 4 – Metodologias empregadas nos modelos de TCE.

Metodologia	Número de modelos (n=31)
Tipos de modelos	
Análise de regressão logística	19 (61%)
Árvore/particionamento recursivo	3 (10%)
Rede neural artificial	1 (3%)
Abordagem bayesiana	5 (16%)
Análise discriminante	3 (10%)
Outros	2 (6%)
Validação interna	
Cross-validation	2 (6%)
Bootstrap	1 (3%)
Holdout	8 (26%)
Validação externa	
Novo conjunto de dados	4 (13%)
Medidas de qualidade	
Acurácia	18 (58%)
Especificidade/sensibilidade	12 (39%)
Curva ROC	4 (13%)
Outros	3 (10%)

Fonte: Adaptado de Mushkudiani et al (2008).

Myburgh et al (2008) realizaram um estudo de TCE para adquirir um perfil detalhado da prevalência, padrões de lesão, estratégias de gerenciamento e o desfecho dos pacientes com TCE admitidos em UTI na Austrália e Nova Zelândia. Dados de 635 pacientes foram utilizados, sendo 74,2% de homens; 61,4% dos TCE foram devido a acidentes de veículos; 24,9% foram as quedas nos pacientes idosos e 57,2% apresentavam TCE grave (GCS < 8); danos cerebrais secundários foram registrados em 28,5% e 34,8% foram submetidos à neurocirurgia antes de serem admitidos na UTI. Em doze meses a mortalidade foi de 26,9% em todos os pacientes e 35,1% em pacientes com TCE grave. Na análise empregaram medidas estatísticas descritivas como média e desvio padrão para os dados normalmente distribuídos, enquanto para os demais calcularam mediana e intervalo interquartil.

Lingsma et al (2010) afirmam que o desfecho de um TCE pode ser muito variável, especialmente nos casos graves. Apesar da associação de muitas variáveis com o resultado as previsões de prognóstico são difíceis de fazer. Neste estudo a análise multivariada identificou a idade, gravidade clínica, anormalidades nos achados tomográficos e variáveis laboratoriais como fatores relevantes a serem incluídos em um modelo de prognóstico. Avanços em modelagem estatística e a disponibilidade de grandes conjuntos de dados têm facilitado o desenvolvimento de modelos de prognóstico que apresentam maior desempenho e generalização.

Coronado et al (2011) descrevem a epidemiologia e as taxas de morte relacionadas ao TCE de 1997 a 2007. Os dados foram analisados a partir do arquivo público que contém dados de certidão de óbito nos Estados Unidos. Neste período identificaram uma média de 53.014 mortes (18,4 por 100 mil habitantes) por TCE. As taxas de morte relacionadas ao TCE: diminui significativamente na faixa etária de 0-44 anos e aumenta acima dos 75 anos; foi três vezes maior no sexo masculino (28,8 por 100 mil habitantes) do que feminino (9,1). As principais causas de morte por TCE foram as armas de fogo (34,8%), veículos (31,4%) e quedas (16,7%). No estudo constataram que a morte por TCE relacionada aos veículos foi maior entre aqueles com idade de 15-24 anos (11,9 por 100 mil habitantes), enquanto os índices por queda foram maiores entre os adultos com idade igual ou maior a 75 anos (29,8 por 100 mil habitantes). No geral, em relação a estudos anteriores, teve-se uma diminuição da taxa global de mortes relacionados com o TCE, bem como para todas as causas de TCE, exceto pelas quedas.

Várias pesquisas têm sido desenvolvidas na área de TCE, como as que compõem o projeto Missão Internacional de Prognóstico e

Análise de Ensaios Clínicos no TCE, do inglês, *International Mission for Prognosis and Analysis of Clinical Trials in Traumatic Brain Injury* (IMPACT). Este projeto consiste em um esforço internacional multidisciplinar para promover programas de investigação clínica na área de TCE buscando: desenvolver e validar modelos de prognóstico para classificação e caracterização do TCE, padronizar a coleta de dados nos estudos, prover recomendações baseadas em evidências para melhorar a sensibilidade e eficiência dos modelos. O projeto foi dividido em duas fases (IMPACT, 2011):

a) IMPACT I: de 2003 a 2006 estabeleceu a base de dados IMPACT e o desenvolvimento de modelos de prognósticos;
b) IMPACT II: de 2007 a 2011 compreende a expansão da base, incluindo dados de estudos recentes, continuação das pesquisas para melhorar o tratamento de TCE.

Uma série de artigos apresentando os resultados do projeto IMPACT têm sido publicados: Butcher et al (2007) descrevem e quantificam a relação entre as causas do TCE e o seu desfecho; Marmarou et al (2007) identificam o valor prognóstico dos componentes da GCS e a reatividade da pupila; Murray et al (2007) determinam por meio de análise univariada e multivariada os fatores prognósticos na admissão após TCE; Mushkudiani et al (2007) demonstram as relações entre as características demográficas em TCE; Steyerberg et al (2008) apresentam o desenvolvimento e a validação internacional de prognósticos baseando-se nos dados de admissão hospitalar; Maas et al (2010) apresentam recomendações para melhorar a concepção e análise de ensaios clínicos em TCE moderada e grave.

Bernal et al (2013) realizaram uma pesquisa cujo objetivo era descrever os fatores prognósticos do traumatismo cranioencefálico grave, a população de estudo foi constituída por 106 pessoas que foram internadas com TCE grave na UTI do Hospital Virgen de La Veja do Complexo Hospitalar Universitário de Salamanca (Espanha). A análise dos dados foi realizada por meio da regressão logística, concluindo-se que a mortalidade foi associada principalmente a um baixo valor de GCS, a hiperglicemia e pupilas midriáticas.

Tjahjadi et al (2013) desenvolveram um modelo preditor para identificar o risco de morte em pessoas acometidas por TCE grave, no estudo foi considerada uma amostra muito pequena, composta por 61 pessoas. O método para geração do modelo foi o de regressão logística, identificando-se como fatores significativos na predição da mortalidade as cisternas comprimidas (classificação tomográfica da lesão cerebral difusa na categoria III – Marshall) e baixa resposta motora.

No caso dos estudos nacionais sobre TCE, em função do número reduzido, empregou-se o critério de revisão das pesquisas nos últimos 23 anos, de 1991 a 2014, apresentando-se os principais.

Masini (1994) desenvolveu uma pesquisa do perfil epidemiológico do TCE no Distrito Federal, utilizando-se de dados do ano de 1991 da Unidade de Politraumatizados e Neurocirurgia do Hospital de Base do Distrito Federal. Realizou análise estatística, identificando a ocorrência de 341 TCE para cada 100 mil habitantes do Distrito Federal e que 60,35% dos casos de óbito por TCE foram causados por acidentes de trânsito. Estes índices quando comparados com os de outras pesquisas internacionais se mostraram elevados.

Colli et al (1997) identificaram as características das vítimas de TCE atendidas no Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, no período de 1990 a 1992. Os dados foram coletados por meio de consulta a Ficha de Atendimento de Urgência e/ou Prontuário. Posteriormente, empregaram análise estatística e os resultados indicaram uma taxa de mortalidade de 6% e que os acidentes com veículos são os principais provocadores de TCE e de óbito. Considerando-se as faixas etárias teve-se o predomínio do sexo masculino em todas as faixas etárias até os 50 anos, acima desta idade a distribuição por sexo foi semelhante. A maioria das ocorrências foram de TCE leve (74,5%), enquanto o grave foi de 7,4%.

Koizumi et al (2000) apresentaram um estudo da morbimortalidade por TCE no município de São Paulo, com dados do ano de 1997, referentes as internações hospitalares e aos óbitos. Estes dados foram obtidos junto ao Ministério da Saúde. Encontraram uma taxa de mortalidade de 10,2% e como fatores relacionados ao óbito a idade, o sexo e os dias de internação, concluindo que no município de São Paulo encontram-se evidências de uma morbimortalidade por TCE alta.

Martins, Silva e Coutinho (2003) realizaram um estudo de TCE grave em Florianópolis no período de 1994 a 2001. Na análise foi comparada a mortalidade em dois períodos: 1994 a 1995 (44%) e 2000 a 2001 (29,7%). Esta queda na mortalidade foi considerada pelos autores como conseqüência do melhor atendimento pré-hospitalar, melhor formação médica, melhora no atendimento na UTI, entre outros.

Melo, Silva e Moreira Junior (2004) descreveram as características de internados com TCE no Hospital Geral do Estado da Bahia (Salvador), no ano de 2001. Os acidentes com meios de transporte foram as principais causas de internamento em vítimas de TCE (40,7%), seguido das agressões físicas com ou sem armas (25,4%) e das quedas

(24%). A análise também comprovou o predomínio das vítimas abaixo de 40 anos e do sexo masculino. A taxa de mortalidade de 22,9% foi considerada elevada conforme a literatura.

Rocha (2006) desenvolveu sua Tese de Doutorado na Faculdade de Medicina da Universidade de São Paulo voltada ao TCE. Realizou a correlação entre dados demográficos, escala de Glasgow e a tomografia computadorizada com a mortalidade em Maceió (Alagoas). Para isso, empregou a correlação de Spearman[12] e regressão logística. A pontuação baixa na GCS apresentou uma correlação linear com a mortalidade, assim, quanto menor a pontuação na GCS, maior a mortalidade. A análise multivariada, por meio da regressão logística, indicou como variáveis preditoras da mortalidade: valores baixos na GCS, presença de anormalidades tomográficas, desvio da linha média e edema difuso.

Martins et al (2009) analisaram a mortalidade em TCE grave na cidade de Florianópolis no período de 1994 a 2003, empregando as análises univariada (teste *t*)[13] e multivariada (regressão logística). No modelo final, por meio da regressão logística, concluiu-se que os fatores relacionados a uma alta mortalidade são: idade maior que 60 anos, presença de hemorragia subaracnóide, GCS de 3 ou 4, pupilas midriáticas e anisocóricas, Marshall Tipo III e ausência de trauma torácico.

Ruy e Rosa (2011) realizaram um estudo para conhecer o perfil epidemiológico dos pacientes com TCE internados na Unidade de Tratamento Intensivo do Hospital São José de Criciúma no período de 2008 a 2009. Trabalharam com a variável dependente TCE e as independentes sexo, cor da pele, tipo de TCE, causa do TCE, seqüelas, desfecho clínico, sinais e sintomas. Na base de dados, composta por 93 registros, realizou-se análise estatística descritiva especificando-se a frequência e porcentagem das variáveis qualitativas, bem como a média e o desvio padrão das quantitativas. Posteriormente, aplicaram-se os testes do *qui-quadrado* e *t* de *Student* para verificar se havia diferença entre as variáveis de acordo com o sexo. Segundo os autores, o estudo está de acordo com a literatura demonstrando o predomínio de TCE em adultos jovens do sexo masculino, em acidentes automobilísticos e uma alta prevalência de óbito.

---

[12] Medida de correlação não-paramétrica que busca a associação entre duas variáveis *x* e *y* (WALPOLE et al, 2009).

[13] Compara dois conjuntos de dados quantitativos por meio dos seus valores médios (BARBETTA; REIS; BORNIA, 2009).

Dantas, Oliveira e Machado Neto (2014) realizaram um perfil epidemiológico do TCE na região Nordeste do Brasil considerando as variáveis morbidade, mortalidade hospitalar, gênero, faixa etária, no ano de 2012. Os dados foram analisados por meio de uma análise estatística descritiva, verificaram que o TCE acomete mais as pessoas do sexo masculino do que feminino, consequentemente também sendo maiores as taxas de mortalidade nos homens. Os acidentes de trânsito são a principal causa do TCE, relacionando-se com a idade ou gênero do paciente.

5.2.1 Data mining e Traumatismo Cranioencefálico

Esta seção dos trabalhos correlatos descreve algumas aplicações de *data mining* para a identificação de padrões em dados de TCE.

Dolce et al (2008) desenvolveram um estudo cujo propósito foi identificar por meio de *data mining*, empregando-se árvores de classificação e regressão, os sinais neurológicos significativos e correlacionados com a previsão do resultado do prognóstico inicial de pacientes em estado vegetativo de etiologia traumática ou não. No estudo analisaram 22 sinais neurológicos, no momento da admissão e depois de 50, 100 e 180 dias, de 333 pacientes em estado vegetativo, dos quais 265 homens e 68 mulheres, gerando árvores de decisão por meio do algoritmo *Classification and Regression Trees* (CART) na ferramenta de *data mining* Weka. O rótulo de classe foi definido como a Escala de Resultados de Glasgow[14] (ERG). As árvores de decisão e o algoritmo CART mostraram-se aplicáveis em neurologia para identificar sinais clínicos importantes também em condições clínicas desfavoráveis, quando se tem escassez de sinais, como é o caso do estado vegetativo. Portanto, o modelo gerado pelo CART mostrou-se adequado para a predição da evolução de pacientes em estado vegetativo. Sinais neurológicos como recuperação dos movimentos espontâneos, *eye tracking* e reflexo óculo-cefálico não observados na admissão ou na fase inicial do acompanhamento clínico e o desaparecimento de automatismo oral quando presentes foram correlacionados positivamente com a ERG 4 ou 5, tendo-se uma predição correta em 89 a 91% dos doentes. Enquanto a alteração do *eye tracking* e o aparecimento de automatismos

[14] Escala utilizada mundialmente para acompanhar a evolução de pacientes com graves lesões cerebrais, como traumatismo cranioencefálico. Emprega cinco notas: 1 (morte), 2 (estado vegetativo persistente), 3 (deficiência grave), 4 (moderada deficiência) e 5 (boa recuperação) (GRZYMATA-BUSSE et al, 2008).

orais são indicativos de mau prognóstico, tendo-se uma ERG 1 ou 2, com uma precisão de 80 a 100%, dependendo do tempo de observação do paciente. No momento da admissão do paciente ou após 50 dias quando reaparece o *eye tracking* e a motilidade espontânea, tem-se uma previsão favorável do resultado do prognóstico em 89 a 90% dos pacientes em estado vegetativo por lesões cerebrais traumáticas, para se tornar irrelevante após os 100 dias. Segundo os autores, o processo de *data mining* parece ser mais eficiente que as estatísticas convencionais, com 90% de precisão no prognóstico após dois meses de admissão do paciente, podendo auxiliar no planejamento de futuras estratégias de reabilitação.

Grzymata-Busse et al (2008) apresentam os resultados da previsão da Escala de Resultados de Glasgow (ERG) para os pacientes acometidos por graves danos cerebrais, como traumatismo cranioencefálico grave. Para isso, utilizam dois métodos de *data mining*: indução de regras pelo LEM2 e geração de redes de crença pelo BeliefSEEKER que são convertidas em um conjunto de regras. O objetivo principal do trabalho foi a comparação dos dois métodos aplicados ao mesmo conjunto de dados descrevendo a ERG para 162 pacientes, utilizando-se 42 atributos, a fim de se dividir os casos em cinco classes correspondentes a pontuação da ERG. Na validação do modelo gerado pelos dois algoritmos o critério de desempenho empregado foi a taxa de erro. Concluíram que o conjunto de regras produzido pelo BeliefSEEKER é muito mais simples do que as produzidas pelo LEM2. No entanto, excluindo-se as regras mais fracas obtidas pelo LEM2 podem-se gerar resultados finais bastante semelhantes apesar das diferenças básicas existentes entre os dois métodos.

Theodoraki et al (2010) construíram modelos de previsão e avaliaram as suas capacidades de prever com precisão a mortalidade em pacientes acometidos por trauma. Esta análise consistiu na comparação das tarefas de *data mining* de agrupamento (*K-means*), classificação (CART, C5.0, CHAID) e associação (*Generalized Rule Induction* - GRI). Os dados foram coletados pela *Hellenic Trauma and Emergency Surgery Society* de trinta hospitais da Grécia. A base de dados é composta por 8.544 registros de pacientes com lesões graves no período de 2005 a 2006. Os resultados obtidos pelos modelos são comparados empregando-se as medidas de especificidade, sensibilidade, valores preditivos positivo e negativo e a curva ROC. Na análise empregaram as ferramentas SPSS e Clementine. Segundo os autores, os algoritmos CHAID e C5.0 proporcionaram um amplo conhecimento da

classificação das lesões, incluindo combinações de características que levam a sobrevida ou ao óbito.

Sut e Simsek (2011) analisaram o desempenho de seis algoritmos de *data mining* a fim de prever a mortalidade em lesões cerebrais. Aplicaram os algoritmos CART, *Chi-squared Automatic Interaction Detector* (CHAID), *Exhaustive* CHAID (E-CHAID), *Quick Unbiased Efficient Statistical Tree* (QUEST), *Random Forest Regression and Classification* (RFRC) e o *Boosted Tree Classifiers and Regression* (BTCR) em um conjunto de dados de 1603 casos de traumatismo cranioencefálico. Antes de aplicar os algoritmos de *data mining*, selecionaram os fatores de risco para o prognóstico da mortalidade por meio da regressão logística. Nesta etapa de seleção dos fatores de risco entre as 19 variáveis de prognóstico, apenas oito foram consideradas influências significativas para a mortalidade (idade, causa do TCE, GCS, pupilas, hemorragia subaracnóide, contusão, hematoma intracerebral e edema cerebral). A avaliação do desempenho foi realizada com base nas medidas de sensibilidade, especificidade, valor preditivo positivo e negativo, acurácia e curva ROC. A previsão da mortalidade foi classificada pelos algoritmos com acurácia variando de 91,1% (CART) a 93% (BTCR). O algoritmo BTCR apresentou a maior taxa de acurácia e uma área significativamente maior sob a curva ROC, tornando-se uma ferramenta potencialmente útil para a predição da mortalidade em traumatismo cranioencefálico.

Raeesi et al (2014) realizaram um estudo acerca da performance de algoritmos de árvore de decisão em *data mining* para predizer as causas do TCE em uma amostra composta por 140 registros. Os algoritmos de árvores de decisão aplicados foram o C5.0, CHAID, QUEST e CART. De acordo com os resultados o algoritmo C5.0 teve a acurácia mais elevada dentre os modelos (81,4%), seguido pelo CART (77,8%).

5.2.2 Data mining e Regressão Logística

Exemplos de pesquisas que realizaram uma análise comparativa dos modelos gerados pela regressão logística e pelos algoritmos de *data mining* são apresentados nesta seção da pesquisa.

Andrews et al (2002) realizaram um estudo a fim de prever a recuperação em pacientes acometidos por traumatismo cranioencefálico, empregando a análise de árvore de decisão e de regressão logística. Para isso, utilizaram dados fisiológicos e da admissão. O objetivo deste estudo foi comparar o resultado da regressão logística com o da árvore

de decisão em um estudo observacional. Os dados referem-se a 124 pacientes adultos com TCE e foram coletados durante a sua estadia na Unidade de Terapia Intensiva. A regressão logística foi empregada para determinar a influência relativa da idade do paciente, categoria da GCS e resposta pupilar na admissão, severidade do traumatismo, entre outros. Na comparação dos resultados bons e ruins, os insultos hipotensivos e a resposta pupilar na admissão foram significativos. No uso das árvores de decisão os autores identificaram que a hipotensão e a baixa pressão de perfusão cerebral são os melhores preditores do óbito, melhorando em 9,2% a acurácia do modelo. A hipotensão foi considerada um preditor significativo para o mau resultado da Escala de Resultados de Glasgow (pontuação 1-3). Os autores concluíram que as árvores de decisão confirmaram alguns dos resultados da regressão logística, mostrando que por meio deste método pode-se obter conhecimento.

Penny e Chesney (2006) realizaram uma comparação de técnicas de análise de dados em lesões traumáticas, empregando métodos de *data mining* e regressão logística para determinar os fatores associados com o óbito. Aplicaram os algoritmos de *data mining* C5.0, CART, rede neural artificial de múltiplas camadas com algoritmo de treinamento *backpropagation* e de regressão logística, bem como a regressão logística convencional realizada no SPSS. A ferramenta de *data mining* utilizada foi Clementine. Os resultados dos modelos gerados foram comparados de acordo com a sua capacidade preditiva. O modelo de regressão logística gerado pelo *data mining* foi mais complexo e incluiu mais variáveis, muitas das quais não eram estatisticamente significativas. No entanto, este modelo foi mais preciso do que o da regressão logística convencional, o que os autores confirmaram pela análise da curva ROC, que foi respectivamente de 0,96 e 0,93. Conforme os autores, dos métodos analisados a rede neural artificial foi a que apresentou melhor precisão na previsão do óbito e a regressão logística convencional foi a que forneceu resultados menos precisos.

Pang et al (2007) desenvolveram um conjunto de modelos que combinam diferentes classes de resultados e fatores de prognóstico empregando metodologias como análise discriminante, regressão logística, árvore de decisão, rede bayesiana e redes neurais artificiais. Os modelos foram desenvolvidos usando dados coletados de 513 pacientes com TCE grave admitidos na *Neurocritical Unit at National Neurosciense Institute of Singapore*, de abril de 1999 a fevereiro de 2003. Os pesquisadores estudaram a correlação entre os fatores prognósticos na admissão do paciente e o desfecho após 6 meses da lesão. Os autores concluíram que dentre os modelos analisados as

árvores de decisão e a regressão logística são mais confiáveis e precisas na previsão do desfecho do TCE, desenvolvendo um modelo de previsão híbrido. Este modelo poderia satisfazer diferentes cenários clínicos encontrados na admissão do paciente.

Chesney et al (2009) analisam dados de traumatismo, dentre os quais o cranioencefálico, em registros de mais de 10 anos do *University Hospital of North Staffordshire* (Reino Unido). O modelo de regressão logística foi utilizado para determinar quais fatores são associados com o óbito durante a internação hospitalar e o algoritmo de *data mining* C5.0 foi aplicado para determinar os fatores associados que podem ser usados para prever a mortalidade. O objetivo do estudo consistiu em identificar os fatores mais importantes que determinam a sobrevida do paciente e comparar a análise de regressão logística com uma abordagem baseada em árvores de decisão. Os dados foram limitados aqueles pacientes com maior gravidade, tendo-se um conjunto de dados com 1.658 registros que foram divididos em 1.111 para o conjunto de treinamento e 547 para teste, tendo-se a mesma proporção de óbitos em ambos os conjuntos. A avaliação da acurácia de ambos os modelos constatou que eles são igualmente precisos quando aplicados ao conjunto de teste (77%), como também em termos de especificidade (regressão logística: 77% e C5.0: 75%) e sensibilidade (regressão logística: 79% e C5.0: 79%). Na análise da curva ROC observou-se que o grau de precisão para o modelo de regressão logística para o conjunto de teste é tão elevado quanto para o de treinamento, confirmando que o modelo é robusto. Os autores concluíram que ambas as técnicas contribuem na determinação dos fatores associados ao óbito, sendo que nenhuma delas superou substancialmente a outra em termos de precisão.

Considerando-se os estudos apresentados neste capítulo da tese, elaborou-se uma tabela com as principais características de cada pesquisa envolvendo *data mining* e lesões traumáticas, inclusive a que constitui esta pesquisa (tabelas 5, 6 e 7).

Tabela 5– Modelos para prognóstico de trauma (*data mining*).

Características	Dolce et al (2008)	Grzymata-Busse et al (2008)	Theodoraki et al (2010)	Sut e Simsek (2011)	Raeesi et al (2014)
N	333	162	8.544	1.603	140
Origem dos dados	S.Anna Institute (Crotone-Itália)	Não Informado (NI)	Hellenic Trauma and Emergency Surgery Society (Grécia)	NI	Khatamolanbya Hospital (Zahdan-Irã)
Período dos dados	1998 a 2006	NI	2005 a 2006	NI	2012-2013
Dados utilizados	Estado vegetativo traumático ou não	TCE grave	Trauma	TCE	TCE
Quantidade de atributos	22	42	10	8	13
Método/ algoritmo	Árvores de Decisão: CART	Indução de regras: LEM2 Redes de crença: BeliefSEEKER	Clusterização (K-MEANS) Classificação (CART, CHAID, C5.0) Associação (GRI)	Árvores de Decisão: CART CHAID E-CHAID RFRC BTCR	Árvores de Decisão: C5.0 CHAID QUEST CART
Medidas de qualidade	Acurácia	Taxa de erro	Sensibilidade Especificidade VPP VPN Curva ROC Acurácia	Sensibilidade Especificidade VPP VPN Curva ROC Acurácia	Acurácia
Validação interna	Cross validation	Cross validation Resubstituição	Cross validation	Cross validation	NI
Ferramenta	Weka	NI	SPSS 17.0 Clementine 12.0	NI	SPSS 16.0 Clementine 12.0
Resultado	Acurácia ~ 90%	BeliefSEEKER erro = 41,98% LEM2 erro = 51,85%	Acurácia C5.0 98,97% CHAID 98,79%	Acurácia CART 91,1% CHAID 91,8% E-CHAID 91,7% QUEST 91,3% RFRC 92,1% BTCR 93%	Acurácia C5.0 81,4% CART 77,8% CHAID NI QUEST NI
Conclusão	DM parece ser mais eficiente	BeliefSEEKER foi melhor e apresentou regras mais simples	CHAID e C5.0 proporcionaram conhecimento	BTCR e RFRC apresentaram melhor desempenho	O C5.0 seguido do CART foram os mais precisos

Fonte: Do autor.

Tabela 6 – Modelos para prognóstico de trauma (*data mining* e regressão logística).

Características	Andrews et al (2002)	Penny e Chesney (2006)	Pang et al (2007)	Chesney et al (2009)
N	124	11.683	513	1.658
Origem dos dados	Regional Head Injury Unit (Edimburgo-Escócia)	University Hospital of North Staffordshire (Reino Unido)	Neurocritical Unit National Neuroscience Institute of Singapore	University Hospital of North Staffordshire (Reino Unido)
Período dos dados	1989 a 1991	1992 a 2003	1999 a 2003	1992 a 1998 e 2001 a 2004
Dados utilizados	TCE	Trauma	TCE grave	Trauma
Quantidade de atributos	19	12	14	21
Método/ algoritmo	Árvores de Decisão: C5.0 Regressão logística (RL)	Classificação (C5.0, CART, RNMLP, RL) RL	Análise de discriminante Árvores de decisão Redes Bayesianas Redes neurais RL	Árvores de decisão : C5.0 RL
Medidas de qualidade	Acurácia	Sensibilidade Especificidade VPP VPN Curva ROC	Acurácia	Sensibilidade Especificidade Acurácia
Validação interna	Cross validation	NI	Cross validation	NI
Ferramenta	See5	SPSS 11.0.1 Clementine 7.0	NI	SPSS Clementine 7.0
Resultado	Acurácia melhorou em 9,2%	Sensibilidade C5.0 65,2% CART 78% RN 96% RL 91,5% Especificidade C5.0 94,3% CART 88% RN 87,7% RL 86,9%	Acurácia variou de 49,79% a 81,49%	Sensibilidade C5.0 79% RL 79% Especificidade C5.0 75% RL 77% Acurácia C5.0 77% RL 77%
Conclusão	Árvores de decisão confirmaram alguns resultados da RL.	As RNMLP apresentaram boa precisão e o modelo de RL foi menos preciso.	Árvores de decisão e RL foram os mais confiáveis e precisos.	Nenhuma técnica superou substancialmente a outra.

Fonte: Do autor.

Tabela 7 – Modelos para prognóstico de TCE grave em Florianópolis.

Características	Martins et al (2009)	Presente Pesquisa
N	748	748
Origem dos dados	Unidade de Terapia Intensiva do Hospital Governador Celso Ramos	Unidade de Terapia Intensiva do Hospital Governador Celso Ramos
Período dos dados	1994 a 2003	1994 a 2003
Dados utilizados	TCE grave	TCE grave
Quantidade de atributos	21	21
Método/ Algoritmo	Regressão logística (RL)	Árvores de decisão: C4.5 e CART Aprendizado baseado em instâncias: kNN Redes neurais: RBF Classificador Bayesiano: Naive Bayes e Bayes Net Metaclassificador: AdaBoost RL
Medidas de qualidade	Acurácia	Sensibilidade Especificidade Acurácia VP VN Curva ROC Coeficiente Kappa
Validação interna	NI	*Cross validation*
Ferramenta	SPSS	SPSS Weka
Resultado	76,9% de predições corretas 87,6% sobrevida 55,6% óbito	Pretende-se identificar um modelo de *data mining* para prognóstico do óbito em TCE grave que apresente maior acurácia e robustez, que o modelo de regressão logística desenvolvido na mesma população de estudo
Conclusão	Fatores relacionados a alta mortalidade são: idade maior que 60 anos, presença de hemorragia subaracnóide, GCS 3 ou 4, pupilas midriáticas e anisocóricas, Marshall Tipo III e ausência de trauma torácico.	Hipótese 1: existe diferença entre as medidas de qualidade apresentadas para os métodos de *data mining* e regressão logística em dados de TCE. Hipótese 2: existe diferença entre os algoritmos de classificação em *data mining* para a predição do óbito referente a TCE grave.

Fonte: Do autor.

Analisando as pesquisas desenvolvidas na área verificou-se que muitos dos resultados são com base em amostras formadas por poucos casos e por vezes com vários atributos, como por exemplo o estudo de Grzymata-Busse et al (2008) o que termina comprometendo a validade do estudo, visto que os dados podem não ser representativos, conforme os critérios estatísticos de poder amostral.

Além disso, os estudos das pesquisas realizadas comparando as técnicas de *data mining* e de regressão logística são inconclusivos para

afirmar que a regressão logística é menos eficiente que o *data mining*. Em termos das medidas de qualidade empregadas para a comparação dos modelos, a maioria dos estudos analisaram somente as de desempenho, não aplicando medidas significativas para a comparação como as de confiabilidade.

Os métodos de *data mining* empregados foram em sua maioria envolvendo árvores de decisão, como é o caso dos trabalhos (tabelas 6 e 7) de Andrews et al (2002), Chesney et al (2009), Dolce et al (2008), Penny e Chesney (2006), Raeesi et al (2014), Sut e Simsek (2011) e Theodoraki et al (2010).

Considerando os motivos expostos anteriormente, bem como a carência de estudos brasileiros nessa área, esta pesquisa consistiu na aplicação de diferentes métodos de *data mining* (árvores de decisão, aprendizado baseado em instâncias, redes neurais artificiais, classificadores bayesianos e metaclassificador) por meio de algoritmos reconhecidos na literatura da área de aprendizado de máquina, sendo muitos deles eleitos por esta comunidade como os dez melhores algoritmos. Também, empregando-se medidas de desempenho e de confiabilidade os modelos gerados são avaliados a fim de se compararem os resultados obtidos, pelos métodos de *data mining* e pela estatística tradicional por meio da regressão logística binária, para a predição do óbito em TCE grave. Buscando-se com isso, confirmar as hipóteses levantadas nesta pesquisa e apresentadas na introdução e na tabela 7.

6 MATERIAIS E MÉTODOS

Os materiais empregados e os métodos, na ordem em que foram executados, para a realização da pesquisa são apresentados neste capítulo.

6.1 DELINEAMENTO DA PESQUISA

Trata-se de uma pesquisa de base tecnológica, sistemática, quantitativa, delineada a partir de conhecimento preexistente, a qual é aplicada na identificação de um método e algoritmo de *data mining* que aperfeiçoe as medidas de desempenho e de confiabilidade de um modelo prognóstico, baseado em uma fonte de dados secundários para a caracterização da prevalência do óbito em Traumatismo Cranioencefálico Grave.

6.2 POPULAÇÃO

Pessoas acometidas por Traumatismo Cranioencefálico Grave que foram admitidas na Unidade de Terapia Intensiva do Hospital Governador Celso Ramos, em Florianópolis-SC, no período de Janeiro de 1994 a Dezembro de 2003.

6.3 BASE DE DADOS

A base de dados secundários refere-se a Traumatismo Cranioencefálico cuja gravidade clínica é classificada como Grave, conforme a Escala de Coma de Glasgow que avalia o nível de consciência da pessoa, segundo as suas funções visuais, verbal e motora. Considerando-se as pontuações na Escala de Coma de Glasgow, o grave tem pontuação de 3 a 8 (CARO, 2011; MAAS; STOCCHETTI; BULLOCK, 2008).

Conforme Martins et al (2009) o Hospital Governador Celso Ramos é um hospital público de referência para o TCE atendendo uma população de aproximadamente um milhão, que compõe a região metropolitana de Florianópolis.

Esta base de dados foi cedida para a realização desta pesquisa pelo médico Evandro Tostes Martins, que desenvolveu junto a colaboradores o estudo “Mortality in Severe Traumatic Brain Injury: A Multivariated Analysis of 748 Brazilian Patients From Florianópolis

City" publicado em 2009 no *The Journal of Trauma Injury, Infection and Critical Care*.

A base de dados é formada por 748 registros, sendo que cada um possui 21 atributos que representam as características relacionadas ao TCE (tabela 8).

Tabela 8 – Base de dados de TCE grave.

Atributos	**Valores**
Sexo	masculino e feminino
Idade	12-30 31-45 46-60 > 60
Glicose	61-110 111-220 221-300 > 300 < 60
Período de atendimento	1994 a 2003
Causa do TCE	acidentes: rodoviário, automóvel, bicicleta; quedas, agressão, outros
Classificação de Marshall	Tipo I, II, III, IV
Hemorragia subaracnóide	sim, não
Trauma associado	sim, não
Tipo de trauma associado	
Face	sim, não
Coluna cervical	sim, não
Coluna tóraco-lombar	sim, não
Tórax	sim, não
Abdominal	sim, não
Membros	sim, não
Outros	sim, não
Escala de coma de Glasgow	3 a 8
Pupilas	isocóricas, mióticas, anisocóricas, midriáticas
Óbito no andar	sim, não
Óbito na UTI	sim, não
Desfecho	óbito, não óbito
Código identificador	1 a 748

Fonte: Garcia, Martins e Azevedo (2013b).

6.4 TAMANHO AMOSTRAL

Mesmo se tratando de uma base de dados secundária, realizou-se o cálculo para o tamanho amostral, buscando verificar se a amostra estudada é representativa para as técnicas empregadas.

Considerando a disponibilidade da base de dados com 748 registros avaliou-se o tamanho da amostra da presente pesquisa, a qual é descrita a seguir.

Na avaliação do tamanho amostral, tomaram-se como base os resultados obtidos sobre este estudo (dados *a posteriori*) em relação aos resultados dos modelos de predição para o desfecho – óbito. Os cálculos foram gerados pelo programa G-Power 3.1.0 (2), onde além da prevalência para o óbito foram utilizados os riscos (OR) estimados para cada variável presente no modelo estimado (NEMES et al, 2009). Sobre estes dados foi assumida uma probabilidade de erro tipo I de 0,05 (nível de significância), uma probabilidade de erro tipo II de 0,20 e uma variação média de 17,0% observada entre as estimativas de risco de cada variável independente de cada modelo isoladamente. O resultado obtido apontou que, para os modelos de Regressão Logistica Binária, bem como, para o Experimento 19 Naive Bayes, o tamanho mínimo de amostra deve ser de 568 casos e para os Experimentos 17 Bayes Net e 10 Naive Bayes o tamanho mínimo de amostra deve ser de 622 pacientes (em cada modelo).

Desta forma, para os modelos com sete variáveis independentes (Regressão Logística Binária e Experimento 19 Naive Bayes) o tamanho de amostra utilizada superou em 180 casos o tamanho mínimo de amostra especificado no cálculo do projeto, apontando que, o poder amostral sobre estes dois modelos alcançou 87,3% (erro tipo 2 = 12,7%). Referente aos modelos onde foram elencadas nove variáveis independentes (experimentos 17 e 10) o tamanho mínimo de amostra foi estimado em 622 casos, ou seja, foram assegurados os parâmetros mínimos de erros, mantendo as estimativas confiáveis (Poder de 80%).

6.5 ESCOPO DA PESQUISA

A pesquisa consiste na avaliação da aplicação de diferentes métodos de *data mining* e a comparação com os resultados da regressão logística aplicada ao prognóstico de Traumatismo Cranioencefálico Grave, a fim de identificar uma melhor técnica para predição do desfecho óbito.

Esta pesquisa foi submetida à apreciação do Comitê de Ética em Pesquisa, sendo aprovada conforme parecer número 931.106.

A abordagem empregada tem seu desenvolvimento dividido em três etapas: definição do problema, modelagem e avaliação dos modelos (figura 9).

O processo de desenvolvimento iniciou-se com a definição do problema em que foi aplicado o *data mining*, observando-se a base de dados bruta no que se refere a estrutura do conjunto de dados em termos dos seus atributos e registros. Esta etapa também foi caracterizada pela definição do especialista no domínio de aplicação que detém o conhecimento sobre o problema, sendo fundamental no processo, pois auxilia na identificação dos objetivos do *data mining*. Os objetivos da aplicação também compõem esta fase e consistem nas características esperadas do modelo de conhecimento.

A modelagem compreende a reelaboração da análise de regressão logística binária, o pré-processamento dos dados e a execução do *data mining*.

A reelaboração da análise de regressão logística binária foi necessária, apesar desta pesquisa se basear em dados secundários do estudo de Martins et al (2009), em que esta técnica já foi empregada no mesmo conjunto de dados. No entanto, nesta pesquisa o modelo elencado da análise de regressão logística foi novamente gerado em função de se buscar mais detalhes para a sua validação. Parâmetros do modelo como Cox & Snell R Square, Nagelkerke R Square, teste de Hosmer-Lemeshow e teste da razão de máxima verossimilhança (*likelihood-ratio test -2LL* ou *-2log*), bem como a equação final para a análise de regressão logística binária foram abordados na reelaboração do modelo, visto que no trabalho de Martins et al (2009) eles não foram explicitados. Considerando que esta pesquisa vai realizar uma comparação da validade de modelos esses parâmetros são fundamentais.

O pré-processamento consistiu na organização e tratamento dos dados para a descoberta de conhecimento em bases de dados, preparando-os para serem submetidos aos algoritmos de *data mining*.

O *data mining* refere-se a etapa de aplicação dos algoritmos a fim de se identificar os padrões presentes nestes dados, sendo que nesta pesquisa aplicaram-se diferentes métodos para a classificação de dados, empregando-se sete algoritmos: indução de árvores de decisão pelos algoritmos C4.5 e CART; aprendizado baseado em instâncias pelo algoritmo kNN; redes neurais artificiais pelo RBF; classificadores bayesianos pelos algoritmos Naive Bayes e Bayes Net; metaclassificador pelo algoritmo AdaBoost.

As árvores de decisão são formas simples de formulação de regras de classificação e como consequência têm sido bastante empregadas como suporte a decisão pelos profissionais de saúde, pois tem se mostrado conveniente para conduzir previsões médicas (WITTEN; FRANK, 2005). As principais vantagens deste método são que produzem resultados levando em consideração as regras mais relevantes, além de possibilitar representações simples do conhecimento, sendo compreensíveis para a maioria das pessoas (KANTARDZIC, 2011). O algoritmo C4.5, empregado nesta pesquisa, além da construção da árvore de decisão implementa a sua simplificação, excluindo aquelas regras que não possuem valores significativos para a precisão do modelo. Assim, reduz a complexidade da árvore o que lhe permite uma classificação mais rápida e eficiente, bem como combate ao problema de *overfitting* (HAN; KAMBER; PEI, 2011). O algoritmo CART realiza uma partição binária recursiva, sendo ajustado mediante estas divisões, a fim de tornar homogêneos os subconjuntos de dados da variável resposta, maximizando a separação das classes (BREIMAN et al, 1984; HAND; MANILLA; SMYTH, 2001). A escolha deste algoritmo ocorreu também em função dos resultados por vezes superiores aos das técnicas estatísticas clássicas. A eficiência dos algoritmos de árvores de decisão, como o C4.5 e o CART, têm sido bem estabelecidas para bases de dados relativamente pequenas (HAN; KAMBER; PEI, 2011).

O aprendizado baseado em instâncias armazena os exemplos históricos (instâncias) na memória e, ao invés de executar uma generalização dos dados fornecidos, emprega uma métrica para encontrar os exemplos passados mais parecidos aos atuais de modo a prever o que acontecerá no futuro com base nos dados anteriores. Apresenta como vantagem em relação aos outros métodos de aprendizado de máquina: a capacidade de adaptar o modelo para dados novos, recuperando da memória exemplos similares que são empregados para predizer a classe (AKINYEMI et al, 2013); nos casos em que a função objetivo é complexa para ser generalizada, ela pode ser definida por meio de funções de aproximações locais com menor complexidade (MITCHELL; BLUM, 1997). O kNN é um dos mais tradicionais e importantes algoritmos deste método, que apesar da sua simplicidade tem se demonstrado robusto na classificação de dados em diversos domínios de conhecimento (BLALOCK, 2003; GUTIÉRREZ, 2010).

As redes neurais artificiais são adotadas nessa pesquisa, pois segundo Olson e Delen (2008) elas possuem capacidade preditiva, produzindo respostas adequadas para dados não conhecidos. Assim,

tornaram-se um método atrativo para tarefas de *data mining* por apresentarem desempenho superior aos modelos convencionais, mostrando-se robusta para a predição (HAYKIN, 2001). As Redes Neurais com Função de Base Radial apresentam como principal vantagem a capacidade de aprender rapidamente padrões complexos e tendências presentes nos dados, ganhando desempenho em relação a outros modelos de redes neurais e apresentando igual poder preditivo (BISHOP, 1995; TAN; STEINBACH; KUMAR, 2009). A arquitetura do RBF se destaca entre os modelos de redes neurais artificiais, pois o processo de treinamento é simples e apresenta eficiência computacional (AZEVEDO; BRASIL; OLIVEIRA, 2000).

Os classificadores bayesianos apresentam excelente acurácia quando aplicados em grandes bases de dados. O classificador bayesiano conhecido como *Naive Bayes* utiliza o Teorema de Bayes como método estatístico e tem como vantagem apresentar um menor tempo de aprendizado preditivo (HAN; KAMBER; PEI, 2011), também tem se mostrado superior a vários outros métodos de classificação quando aplicado especificamente em dados biomédicos (AL-AIDAROOS; BAKAR; OTHMAN, 2012). O *Naive Bayes* é mais preciso quando os atributos são independentes ou fracamente correlacionados, sendo esta última característica presente em muitos atributos de dados referentes a saúde (HICKEY, 2013). O algoritmo Bayes Net prevê a probabilidade de um resultado usando uma estrutura gráfica denominada de rede bayesiana, sendo capaz de aprender diretamente dos dados, usando para isso várias técnicas de busca heurística, com o objetivo de inferir uma rede que melhor represente a distribuição de probabilidade dos dados de treinamento (MITCHELL; BLUM, 1997; NASSIF et al, 2012). Este algoritmo fornece um meio promissor para conceber modelos preditores, sendo aplicado para o desenvolvimento de inferências probabilísticas ou modelos de classificação em diferentes áreas do conhecimento, podendo assim, auxiliar profissionais em processos de tomada de decisão (LI et al, 2010).

O metaclassificador soluciona um problema partindo de saídas individuais propostas por múltiplas soluções alternativas, representa uma das principais linhas de pesquisa em aprendizado de máquina, sendo aplicado em uma variedade de problemas (SILVA; MAIA; FONSECA, 2012). Este método foi concebido a fim de obter melhor desempenho do que classificadores individuais, para isso executa várias vezes o algoritmo base de aprendizagem e combina as hipóteses

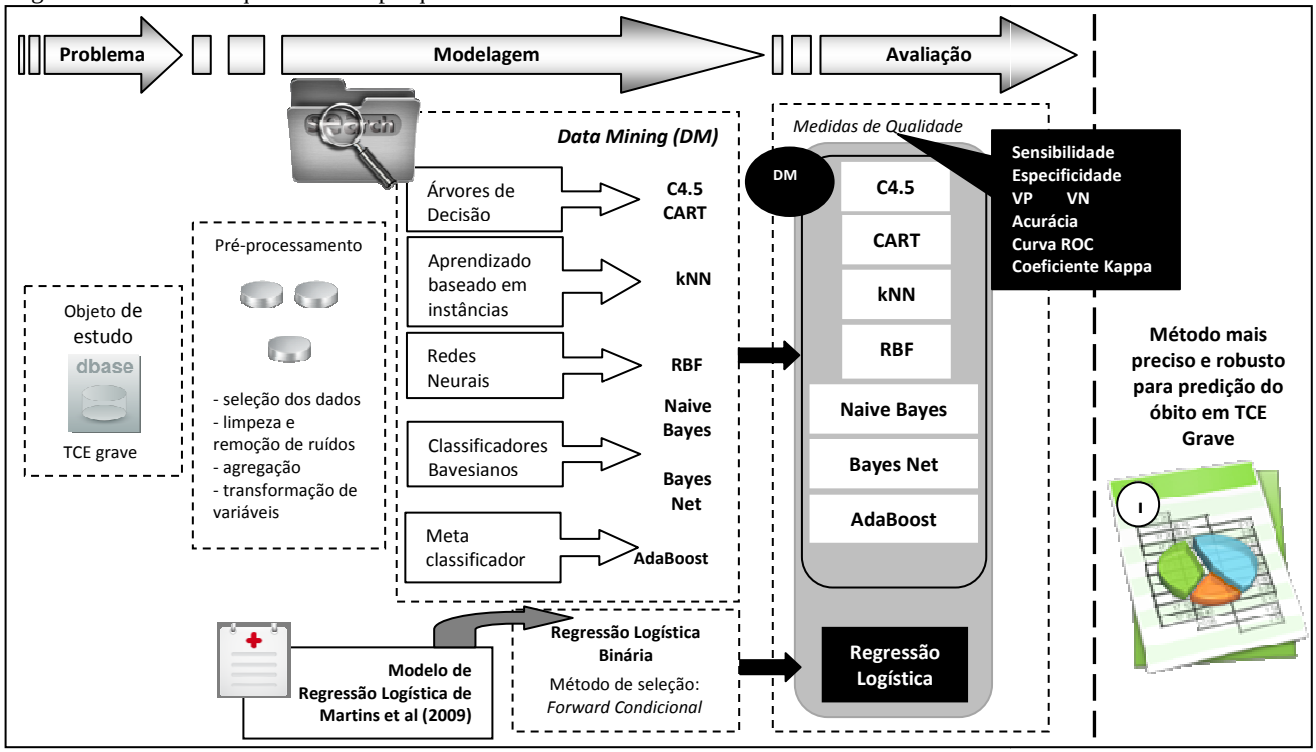
resultantes (OHNO, 2011). O algoritmo AdaBoost é um dos algoritmos mais utilizados dentre os que se baseiam em *boosting*[15], sendo um dos mais importantes métodos de metaclassificadores (FREUND; SCHAPIRE, 1997; TAN; LI; QIN, 2007; WU et al, 2008), isto se deve a adaptação aos classificadores base, sendo o modelo gerado de forma a favorecer as instâncias classificadas erroneamente por classificadores anteriores (REIS, 2013).

A terceira etapa da pesquisa refere-se a avaliação e compreende a análise dos modelos identificados pelos algoritmos supracitados, que são comparados entre si e com o da regressão logística, gerada por Martins et al (2009) e reelaborado por esta pesquisa, empregando-se medidas de qualidade voltadas aos critérios de avaliação de acurácia e robustez. Assim, ao final espera-se demonstrar que um destes algoritmos de *data mining* empregados consiste em uma alternativa válida à regressão logística em dados de TCE grave, identificando-se o método que proporciona um modelo mais preciso e robusto para predição do óbito em TCE grave.

As etapas que compreendem a reelaboração da análise de regressão logística binária, o pré-processamento dos dados, a aplicação do *data mining*, a avaliação dos modelos de *data mining* e regressão logística gerados são abordados em detalhes nas próximas seções deste capítulo.

[15] Algoritmo que cria um metaclassificador selecionando conjunto de dados que posteriormente são unidos por votação pela maioria, a seleção provê dados de treinamento mais informativos para cada classificador gerado (POLIKAR, 2006).

Figura 9 – Desenho esquemático da pesquisa.



Fonte: Do autor.

6.6 ANÁLISE DE REGRESSÃO LOGÍSTICA BINÁRIA

A apresentação dos resultados ocorreu pela estatística descritiva, distribuição absoluta (n) e relativa (%), bem como, pela média e desvio padrão.

O estudo da distribuição de dados das variáveis contínuas ocorreu pelo teste de Kolmogorov-Smirnov, teste não paramétrico que investiga se uma amostra pode ser considerada como proveniente de uma distribuição aproximadamente normal. Baseia-se na função de distribuição acumulada, sendo aplicado para verificar se duas distribuições empíricas são distintas ou se uma delas diferencia-se da considerada ideal (aproximadamente normal) (CALLEGARI-JACQUES, 2003).

A abordagem inicial dos resultados ocorreu pela análise bivariada, com a comparação das variáveis independentes (ou covariáveis) em relação ao óbito. Na análise dos dados categóricos, foi verificada a significância da associação por meio do teste Qui-quadrado de Pearson ($\chi 2$), que estabelece a comparação entre as frequências observadas (reais) e as esperadas, bem como, a análise pelos resíduos ajustados, em que os valores negativos indicam uma frequência real inferior à esperada, enquanto os valores positivos caracterizam uma frequência real superior à esperada. As células que têm resíduos ajustados com valores iguais ou acima de 1,96, em valor absoluto, contribuem significativamente para a relação de dependência entre variáveis comparadas. Destas comparações surgem diferenças, que podem ser grandes ou pequenas; se forem grandes, a H_0 (que pressupõe "bom" ajustamento – distribuição de proporções são semelhantes entre os grupos implicando em independência) deverá ser rejeitada em favor da H_1; se forem pequenas, a H_0 não será rejeitada e as diferenças serão atribuíveis ao acaso. Em tal situação a H_0 (hipótese nula) testa a independência entre as variáveis (CALLEGARI-JACQUES, 2003).

Nas tabelas de contingência em que no mínimo 25% dos valores das células apresentaram frequência esperada menor do que cinco, empregou-se o teste exato de Fischer, sendo que, nos casos em que pelo

menos uma variável apresentasse característica politômica[16] foi utilizada a Simulação de Monte Carlo[17].

Ainda na análise comparativa entre as categorias de desfecho, foi utilizado, como medida de efeito, o *Odds ratio* (OR) bruto, com intervalo de confiança de 95% (IC95%). Assim, calcula-se a probabilidade aproximada de vezes que uma variável pode exercer efeito sobre o desfecho óbito.

Na análise bivariada, para as variáveis contínuas, quando a comparação ocorreu entre dois grupos independentes foi aplicado o teste *t-Student*, o qual tem a função de testar a hipótese nula (H_0) de não-diferença de médias entre dois grupos (óbito e não óbito). Na aplicação deste teste é fundamental avaliar se a distribuição de dados pertenceu a uma população normal, ou se a amostra apresentou dimensão suficientemente grande para se aplicar o Teorema do Limite Central (em geral, $n \geq 30$ para ambos os grupos). Outro fator importante refere-se a verificação da homogeneidade das variâncias, as quais devem ser estatisticamente iguais nos dois grupos (Teste de Levene). A razão desta exigência consiste no teste assumir que as populações originárias destes grupos são iguais em tudo (distribuição, dispersão, entre outros), exceto nos respectivos valores médios (BUSSAB; MORETTIN, 2003). Quando a comparação das variáveis contínuas ocorreu entre três ou mais grupos independentes foi impregada a Análise de Variância (*One Way*) – *Post Hoc Tukey*, onde para garantir a validade da técnica, também, foram testadas a normalidade dos dados (Teste de Kolmogorov-Smirnov) e a homogeneidade de variância (Teste de Levene).

A técnica define como hipótese nula (H_0) a ausência de diferenças entre as médias comparadas, sendo que, o resultado direto do teste aponta se pelo menos uma das médias comparadas difere de forma significativa. A detecção de grupo apresentou maior (ou menor) média estatisticamente significativa foi realizada por meio da análise complementar caracterizada como comparações múltiplas (*Post hoc*). Neste estudo, foi empregado o teste de Tukey, que além de investigar a magnitude das diferenças entre as médias comparadas, é reconhecidamente o mais rigoroso permitindo testar qualquer contraste entre duas médias.

Na investigação das covariáveis (ou variáveis independentes) que poderiam responder pelo desfecho (óbito), foi utilizada a técnica de

[16] Variável com três ou mais categorias.
[17] Método de simulação estocástica mais difundido, operando modelos estatísticos de variáveis descritas por funções probabilísticas (ANDRADE, 2009).

análise multivariada Regressão Logística Binária. Na seleção das variáveis que compuseram o modelo inicial (modelo bruto) adotou-se a estratégia proposta por Victora et al (1997), que utiliza modelos hierarquizados, sendo selecionadas para a análise multivariada todas as variáveis que apresentaram na análise bivariada valores de $p < 0,20$. Assim, evita-se a exclusão de variáveis potencialmente importantes.

A seleção das variáveis com poder de predição sobre o óbito, a partir do modelo saturado, foi realizada pelo método de *Forward condicional*. Desta forma, as variáveis independentes entraram sequencialmente no modelo (*step*) de acordo com o poder discriminatório que elas acrescentam à previsão do desfecho (óbito). A forma de seleção foi baseada no teste de Wald, usado para avaliar a significância dos coeficientes da regressão logística.

A cada novo passo (*step*), gerou-se um novo modelo e o ajuste geral foi avaliado por meio dos seguintes testes: Qui-quadrado da mudança no valor de –2LL (Teste Omnibus) e de Hosmer e Lemeshow. Estes testes permitem analisar, após a inclusão de cada uma das variáveis independentes, se o modelo pode ser considerado capaz de realizar as predições com a acurácia desejada. A estatística de referência L é a função de verossimilhança definida como a probabilidade de obter os resultados da amostra, dadas as estimativas dos parâmetros do modelo logístico. Como essa probabilidade é um valor menor do que um, convencionou-se usar a expressão –2LL (-2 multiplicado pelo logaritmo decimal da probabilidade, do inglês, *likelihood*). O resultado –2LL é uma medida da qualidade de ajuste do modelo estimado aos dados, sendo maior esta qualidade quanto menor for o valor de -2LL. O teste de Hosmer e Lemeshow considera a hipótese estatística de que as classificações em grupo previstas são iguais as observadas, constituindo-se em um teste de ajuste do modelo aos dados (HOSMER; LEMESHOW, 2000).

Também, considerou-se para estabelecer o ajuste dos modelos logísticos as medidas de Cox & Snell e de Nagelkerke (Pseudo R^2). Segundo Hair et al (2009), estas estimativas de ajuste comparam as probabilidades estimadas com aquelas observadas, sendo que valores elevados significam um melhor ajuste do modelo.

Na execução definiu-se em 0,05 a probabilidade de entrada gradual das variáveis ao modelo e de 0,10 para remoção. A significância adotada para o ponto de corte foi de 0,50 para o máximo de 20 iterações. Os níveis de significância inferiores a 0,01 foram considerados significativos com base no critério de Bonferroni.

Os dados empregados nesta pesquisa são secundários e no artigo original foram tratados de forma mais direta, conforme os objetivos dos seus autores Martins et al (2009). No entanto, nesta pesquisa o foco foi a fundamentação da técnica buscando a comparação com os métodos de *data mining.*

A análise dos dados foi realizada no programa Statistical Package for Social Sciences (SPSS) versão 22.0 *trial* para Windows, sendo que, para critérios de decisão estatística adotou-se o nível de significância de 5%.

6.7 PRÉ-PROCESSAMENTO DOS DADOS

O pré-processamento dos dados deve ser aplicado para torná-los mais apropriados ao *data mining*. Segundo Tan, Steinbach e Kumar (2009) divide-se em etapas que compreendem genericamente duas categorias: seleção dos objetos de dados e atributos para a análise ou criação/alteração dos atributos. O objetivo do pré-processamento é melhorar a análise quanto ao tempo, custo e qualidade.

Nesta pesquisa o pré-processamento compreendeu as subfases de entendimento, seleção, limpeza e transformação dos dados.

6.7.1 Entendimento dos Dados

O entendimento consistiu na análise dos dados, composto por 21 variáveis (tabela 8) e 748 registros, fornecidos pelo especialista do domínio de aplicação, a fim de orientar o que era necessário na preparação dos dados para o *data mining*. Os atributos foram analisados em termos do seu significado, tipo de dado, valores assumidos e relevância para os objetivos do *data mining* referentes a predição do óbito em TCE grave. Também se verificou a existência de casos de dados errados, de valores ausentes nos atributos e falta de padronização dos dados, o que ocorreu pouco nesta base. Estas análises para entendimento dos dados foram realizadas utilizando-se a ferramenta SPSS versão 22.0 *trial* para Windows.

A base utilizada referente ao TCE grave apresenta dados com qualidade, visto que já havia sido empregada em pesquisas anteriores, como a de Martins et al (2009). Os dados utilizados nesta pesquisa já se encontravam organizados em uma única tabela, o que facilitou a identificação e agrupamento dos dados relevantes para o *data mining*, realizando-se a seleção dos dados que consistiu na escolha dos atributos a serem considerados.

6.7.2 Seleção dos Dados

Na etapa de seleção dos dados identificaram-se as informações da base de dados, que foram utilizadas na fase de *data mining* do processo de KDD, pois é comum existirem atributos irrelevantes aos objetivos do *data mining*.

Realizou-se uma junção orientada[18]que consistiu na seleção criteriosa dos atributos que podem contribuir na identificação dos padrões nos dados. Esta seleção por junção orientada compreendeu a redução de dados vertical, eliminando-se características irrelevantes e reduzindo o ruído. Considerando-se isso, por meio do módulo de pré-processamento da ferramenta Waikato Environment for Knowledge Analysis (Weka), versão 3.6.11, removeram-se atributos cujos conteúdos não foram considerados relevantes para o problema como, por exemplo, o *código identificador*, *óbito na UTI* e *óbito no andar*. A seleção de atributos relevantes e eliminação daqueles desnecessários, diminui a complexidade do problema e pode auxiliar no desempenho da aprendizagem. A eliminação do atributo *código identificador* ocorreu em função do critério relevância, visto que ele foi considerado inútil ao processo. Enquanto *óbito na UTI* e *óbito no andar*, foram eliminados devido a redundância, visto que o atributo *óbito* já reflete a informação embutida nesses dois atributos. De acordo com Tan, Steinbach e Kumar (2009) a redução de dados verticais pode auxiliar na obtenção de modelos de conhecimento com maior acurácia e concisão, eliminando características irrelevantes e reduzindo o ruído. Após estas eliminações a base de dados passou a ter 18 atributos.

A seleção dos dados pode ser realizada de forma manual ou automática. Nesta pesquisa empregaram-se as duas abordagens.

Na seleção manual o especialista do domínio de aplicação auxiliou na escolha dos atributos que poderiam contribuir para o *data mining*, como também se considerou o entendimento adquirido nesta pesquisa acerca de TCE grave. O método manual é considerado uma das melhores formas de seleção de dados, desde que se conheça o problema de aprendizado e o significado de cada atributo.

---

[18] Modo de junção de dados relevantes em que o especialista em descoberta de conhecimento em bases de dados, juntamente com o especialista do domínio de aplicação escolhem os atributos que podem influenciar no processo (GOLDSCHMIDT; PASSOS, 2005).

Na seleção automática, realizada por meio de algoritmos, empregou-se o método de filtro que é aplicado antes do processo de aprendizado para selecionar o subconjunto de atributos a serem submetidos ao *data mining*.

O método de seleção automática utilizado foi o de filtro supervisionado *CfsSubsetEval*, disponível no módulo de pré-processamento da ferramenta Weka, aplicando-se o método de busca *Best First Search* nas configurações *Forward* e *Backward*.

O *CfsSubsetEval* implementa o *Correlation based Feature Selection* (CFS) que avalia o valor de um subconjunto de atributos considerando a capacidade preditiva de cada característica juntamente com o grau de redundância entre eles, preferindo aqueles que são altamente correlacionados.

O CFS identifica atributos irrelevantes, redundantes e ruídos, selecionando dentre os atributos aqueles que são relevantes. De acordo com Hall (1999) em bases de dados reais o CFS eliminou mais da metade das características, sendo que na maioria dos casos a acurácia dos modelos se igualaram ou foram superiores aos casos em que se usou o conjunto completo dos atributos.

O método de busca heurística *Best First*, também conhecido como melhor escolha, procura otimizar a solução combinando em um único método as vantagens da busca em profundidade e em largura (RICH; KNIGHT; NAIR, 2009). A busca para frente (seleção *forward*) é iniciada sem atributos e os mesmos são adicionados um a um, isoladamente, sendo incorporado o melhor atributo entre os não selecionados baseado no critério de avaliação. Na busca para trás (eliminação *backward*) inicia-se com todo o conjunto de atributos e a cada iteração se vai eliminando o atributo menos importante.

6.7.3 Limpeza dos Dados

A limpeza dos dados compreendeu o tratamento de valores ausentes, realizando-se a eliminação destes no conjunto de dados, por meio da exclusão dos casos que possuíam atributos com valores ausentes. Após esta etapa a base de dados passou a apresentar 728 registros, os quais foram submetidos ao *data mining*.

6.7.4 Transformação dos Dados

A transformação dos dados é aplicada a todos os valores de uma variável e consiste em colocá-los em um formato apropriado, conforme

requerido pelos algoritmos de *data mining*, aplicando-se operações nestes dados na fase de pré-processamento.

Nesta pesquisa empregou-se a conversão de valores nominais para numéricos e a discretização.

A conversão de valores nominais para numéricos ocorreu por meio da aplicação do entendimento que se tem, bem como do especialista do domínio de aplicação para a determinação de uma boa representação. A partir do entendimento do domínio do problema se utilizou o método de remapeamento 1 de *n*, em que para cada valor nominal define-se 1 como a presença de um valor e 0 como a ausência. Neste caso, considera-se que dos *n* valores distintos de um atributo, somente um valor é definido como 1.

A discretização referiu-se a transformação das variáveis contínuas em categorizadas, como por exemplo, idade e glicose, definindo-se a quantidade de categorias e o mapeamento de valores para elas. Primeiramente, ordenaram-se os valores de cada atributo contínuo, os quais foram divididos em *n* intervalos, especificando-se *n-1* pontos de divisão. Após isso, todos os valores de um intervalo foram mapeados para o mesmo valor de categoria.

O método de discretização empregado foi o não supervisionado, realizando-se a partição por meio de duas técnicas: em frequências iguais e pelo método de agrupamento *K-means*, além da inspeção visual dos dados conforme indicado por Tan, Steinbach e Kumar (2009).

Na tabela 9 tem-se todos os atributos que compõem a base de dados após as atividades de pré-processamentos dos dados, os quais serão empregados para a execução do *data mining*, conforme os critérios de seleção dos atributos empregado na pesquisa e expostos anteriormente.

Tabela 9 – Atributos presentes na base de dados

Informações	Nome dos atributos na base de dados
Sexo	sexo
Idade	idade catidade catidade2
Glicose	glicose catglice catglic2
Período de atendimento	biênio
Causa do TCE	causa
Classificação de Marshall	marshall catmarshall
Hemorragia subaracnóide	hsa
Trauma associado	associado
Tipo de trauma associado	
Face	face
Coluna cervical	colcerv
Coluna tóraco-lombar	coltolom
Tórax	tórax
Abdominal	abdômen
Membros	membros
Outros	outros
Escala de coma de Glasgow	gscadm catgsm
Pupilas	pupilas
Desfecho	desfecho

Fonte: Do autor.

6.8 APLICAÇÃO DO *DATA MINING*

Esta etapa compreendeu a aplicação dos métodos e algoritmos de *data mining* anteriormente definidos no escopo desta pesquisa, sendo realizado por meio de sete algoritmos para a classificação de dados, os quais foram: indução de árvores de decisão pelos algoritmos C4.5 e CART; aprendizado baseado em instâncias pelo algoritmo kNN; redes neurais artificiais pelo RBF; classificadores bayesianos pelos algoritmos Naive Bayes (NB) e Bayes Net; metaclassificador pelo algoritmo AdaBoost.

Na execução do *data mining* empregou-se a ferramenta *Waikato Environment for Knowledge Analysis* (Weka), versão 3.6.11, que é distribuída sob os termos da *General Public License* (GNU). A Weka foi desenvolvida na Universidade de Waikato, na Nova Zelândia, sendo utilizada em pesquisas na área de *data mining*. Conforme Witten, Frank e Hall (2011) a Weka implementa em Java diversos algoritmos de *data mining*, permitindo a execução em diferentes plataformas.

A ferramenta Weka é disponibilizada gratuitamente em (http://www.cs.waikato.ac.nz/~ml/weka/).

No desenvolvimento desta pesquisa realizaram-se 19 experimentos, aplicando-se em cada um deles os sete algoritmos citados anteriormente, o método de validação interna adotado para todos foi o *cross validation* de 10 partes. Este método avalia como será a performance do modelo em dados futuros que não foram utilizados no modelo de treinamento, protegendo-se contra estimativas excessivamente otimistas de desempenho do modelo (KUSANO; GLABER, 2014). Nesta pesquisa adotou-se o valor 10 para *k*, conforme indicado pela literatura na área, Kantardzic (2011), Kohavi (1995) e Tan, Steinbach e Kumar (2009), como sendo o ideal. Ainda de acordo com Kohavi (1995) as estimativas são razoavelmente boas com este valor e se reduz a variância.

Os experimentos realizados diferenciaram-se em termos da técnica de seleção de atributos empregada e consequentemente dos respectivos atributos selecionados, salientando-se que o atributo desfecho constitui-se como classe.

Considerando-se o exposto, o delineamento dos 19 experimentos compreendeu (tabela 10):

a) validação interna pelo método *cross validation* de 10 partes;
b) aplicação dos setes algoritmos em todos os experimentos;
c) em 16 experimentos empregou-se a seleção manual dos atributos;
d) em dois experimentos aplicou-se a seleção automática dos atributos por meio do filtro supervisionado CFS pelo método de busca *Best First Search*, na configuração *forward* (experimento 11) e *backward* (experimento 14);
e) no experimento 19 a seleção dos atributos foi realizada por meio da regressão logística.

O algoritmo C4.5, cuja implementação Java na Weka é denominada de J48, foi empregado nesta pesquisa considerando-se o nível de confiança padrão de 25%. O SimpleCart na Weka é uma

implementação do algoritmo CART proposto por Breiman et al em 1984, adotando-se a poda por minimização do custo-complexidade.

O kNN na Weka é o IBk, usando-se nesta pesquisa o *LinnearNNSearch* como algoritmo de busca por força bruta do vizinho mais próximo e a distância euclidiana.

O algoritmo RBF usa a função radial Gaussiana normalizada, sendo os centros das funções de base radial determinados pelo algoritmo *K-means*.

No algoritmo Naive Bayes considerou-se como estimador a distribuição normal. No Bayes Net o algoritmo *SimpleEstimator* foi aplicado para estimar as tabelas de probabilidade da rede bayesiana e o método para pesquisa da estrutura da rede o K2.

Na aplicação do *data mining* pelo metaclassificador heurístico AdaBoost usou-se o método AdaBoost.M1, escolhendo-se como classificador base o algoritmo de árvores de decisão *Decision Stump*, o qual é usualmente aplicado com os algoritmos de *boosting*. O AdaBoost.M1 é uma versão do AdaBoost que possui como diferença a possibilidade dos algoritmos base serem classificadores de multiclasses ao invés de serem binários.

Tabela 10 – Experimentos realizados.

Experimentos	Algoritmos	Seleção dos atributos	Atributos Selecionados
Experimento 1	C4.5, CART, KNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	sexo, catgcs, causa, catmarshall, hsa, pupilas, associado, catidade, catglic2, desfecho
Experimento 2	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	sexo, catgcs, causa, catmarshall, hsa, pupilas, associado, face, colcerv, coltolom, tórax, abdômen, membros, outros, catidade, catglic2, desfecho
Experimento 3	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	catgcs, causa, catmarshall, hsa, pupilas, associado, catidade2, catglic, desfecho
Experimento 4	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	catgcs, causa, catmarshall, hsa, pupilas, associado, catidade2, catglic, face, colcerv, coltolom, tórax, abdômen, membros, outros, desfecho
Experimento 5	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	sexo, causa, catmarshall, hsa, pupilas, associado, catidade, catglic2, gcsadm, desfecho
Experimento 6	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	sexo, causa, catmarshall, hsa, pupilas, associado, face, colcerv, coltolom, tórax, abdômen, membros, outros, catidade, catglic2, gcsadm, desfecho
Experimento 7	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	causa, catmarshall, hsa, pupilas, associado, catidade2, catglice, gcsadm, desfecho
Experimento 8	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	causa, catmarshall, hsa, pupilas, associado, catidade2, catglice, gcsadm, face, colcerv, coltolom, tórax, abdômen, membros, outros, desfecho
Experimento 9	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	marshall, hsa, pupilas, associado, catidade, catglic2, gcsadm, desfecho
Experimento 10	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	marshall, hsa, pupilas, associado, face, colcerv, coltolom, tórax, abdômen, outros, membros, catidade, catglic2, gcsadm, desfecho
Experimento 11	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Automática Filtro supervisionado CFS *Best first search –forward*	sexo, causa, marshall, hsa, pupilas, associado, catidade2, catglice, gcsadm, desfecho
Experimento 12	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	sexo, catgcs, hsa, pupilas, face, tórax, catglice, desfecho
Experimento 13	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	sexo, catgcs, hsa, pupilas, associado, catglice, desfecho
Experimento 14	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Automática Filtro supervisionado CFS *Best first search – backward*	sexo, catgcs, catmarshall, hsa, pupilas, associado, face, tórax, catidade2, biênio, catglice, gcsadm, desfecho
Experimento 15	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	glicose, causa, marshall, hsa, pupilas, associado, catidade2, gcsadm, desfecho
Experimento 16	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	glicose, causa, marshall, hsa, pupilas, associado, face, tórax, abdômen, catidade2, gcsadm, desfecho
Experimento 17	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	sexo, causa, marshall, hsa, pupilas, associado, catidade2, catglic2, gcsadm, desfecho
Experimento 18	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Manual	sexo, causa, marshall, hsa, pupilas, associado, face, colcerv, coltolom, tórax, abdomên, membros, outros, catidade2, catglic2, gcsadm, desfecho
Experimento 19	C4.5, CART, kNN, RBF, NB, BayesNet, AdaBoost.M1	Regressão logística	pupilas, catgcs, Marshall, biênio, hsa, tórax, catidade, desfecho

6.9 AVALIAÇÃO DOS MODELOS DE *DATA MINING* E REGRESSÃO LOGÍSTICA

Os modelos gerados na etapa de *data mining* foram comparados aplicando-se as medidas de avaliação de desempenho e de confiabilidade obtidas para os modelos gerados pelo *data mining*.

As medidas de avaliação de desempenho empregadas foram: verdadeiros positivos, verdadeiros negativos, acurácia, sensibilidade e especificidade. Enquanto as medidas de confiabilidade basearam-se no Coeficiente de Concordância Kappa e na área abaixo da Curva ROC.

Estas medidas de avaliação de desempenho e de confiabilidade foram apresentadas por meio da média aritmética, estimada para cada experimento de *data mining*, como também para cada um dos sete algoritmos empregados. A avaliação das possíveis diferenças significativas entre os experimentos e algoritmos foi realizada por meio da comparação das médias pela técnica de Análise de Variância (*One Way*) – *Post Hoc* Bonferroni.

O critério utilizado para detectar os modelos com maior poder de discriminação foi a área abaixo da Curva ROC, em função de ser o estimador de maior confiabilidade (FLETCHER; FLETCHER; FLETCHER, 2014).

Na análise que envolveu a comparação entre os modelos de *data mining* elencados como apresentando um maior poder de discriminação para o desfecho óbito em TCE grave, assim como em relação a regressão logística, utilizaram-se as seguintes medidas de confiabilidade: Curva ROC e Coeficiente de Concordância Kappa, sendo considerados Intervalos de Confiança de 95% (IC 95%).

A análise dos dados foi realizada no programa Statistical Package for Social Sciences (SPSS) versão 22.0 *trial* para Windows, sendo que, para critérios de decisão estatística adotou-se o nível de significância de 5%.

7 RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados obtidos nesta pesquisa, no que se refere as análises realizadas em dados de TCE grave por meio da regressão logística binária (7.1) e dos métodos de *data mining* empregados (7.2), bem como os resultados da comparação entre os modelos de *data mining* e de regressão logística (7.3). Finalmente, tem-se a discussão dos resultados (7.4).

7.1 ANÁLISE DE REGRESSÃO LOGÍSTICA BINÁRIA

Os resultados apresentados referem-se a uma amostra de 748 observações referentes a uma base de dados secundária (classificada como fonte externa e publicada), ou seja, dados coletados para algum outro tipo de propósito de pesquisa, mas que se adequaram aos propósitos deste estudo (AAKER; KUMAR; DAY, 2001).

As análises apresentadas foram estratificadas segundo o óbito: sim 33,3% (n=249) e não 66,7% (n=499), desta forma a estatística descritiva inicial ocorreu por meio da análise bivariada, uma vez que, envolveu a comparação entre as variáveis independentes e o desfecho, não estabelecendo necessariamente uma relação de causa e efeito entre elas. Sobre estes resultados foram detectadas as variáveis com poder de predição sobre o óbito (tabela 11).

Vale salientar que, no estudo de Martins et al (2009), os resultados referentes a comparação das variáveis independentes em relação ao óbito foram caracterizados como análise univariada (verificando-se a magnitude de associação pelo *Odds Ratio*), definição também utilizada como etapa anterior ao modelo de Regressão Logística Multivariado. No entanto, de acordo com os conceitos básicos de estatística optou-se por manter a definição de análise bivariada, quando duas variáveis são consideradas para obtenção de algum resultado.

Ainda neste tópico, seguem apresentados os resultados para a Análise de Regressão Logística Binária Multivariada, para os modelos saturado e final.

7.1.1 Análise Bivariada

Na tabela 11 caracterizando as variáveis independentes da amostra segundo a classificação óbito e não óbito (análise bivariada) observou-se que, no sexo feminino os casos de óbito foram de 39,3% (n=46), enquanto que no masculino os casos de óbito alcançaram 32,2% (n=203) (p=0,162), implicando em um risco de 1,36 (OR IC95%: 0,91-

2,05) vezes mais chance das mulheres irem a óbito quando comparadas aos homens.

A média de idade foi significativamente mais elevada (p=0,019) no grupo que veio a óbito (36,2±17,6) quando comparado àqueles caracterizados como não óbito (33,1±15,5). Na avaliação das faixas etárias em relação ao óbito, a diferença estatística significativa não se configurou (p=0,184), tendo-se as maiores concentrações de investigados com óbito nas idades de 31 a 45 anos, 33,3% (n=63); de 46 a 60 anos, 34,4% (n=32); e acima de 60 anos, 44,3% (n=31), sendo que sobre esta última o risco mostrou-se significativo, (OR: 1,77; IC95%: 1,06-2,98), indicando que, os investigados com mais de 60 anos apresentaram 1,77 vezes mais chance de irem a óbito quando comparados aqueles com idades de 12 a 30 anos.

Sobre a relação entre óbito e biênio foi detectada associação estatística significativa (p=0,001), de forma que, o grupo que veio a óbito mostrou-se associado aos anos de 1994-1995, 44,9% (n=80), enquanto que no triênio de 2002-2003-2004 a associação ocorreu com o não óbito, 76,7% (n=102). Nos resultados da associação *Linear-By-Linear* (p<0,001) verificou-se que, quanto menor os anos que classificam os biênios, maior a probabilidade de ocorrência de óbito. No que se refere ao risco, os resultados apontaram risco significativo para os biênios de 1996-1997 (OR: 1,78; IC95%: 1,07-2,99) e de 1994-1995 (OR: 2,68; IC95%: 1,63-4,43) em comparação com o triênio 2002-2003-2004.

O nível médio de glicose mostrou-se significativamente mais elevado no grupo que veio a óbito (178,2±79,5), quando comparado ao grupo não óbito (156,8±55,8) (p<0,001). Na análise foram comparadas as faixas dos níveis de glicose em relação ao óbito, a associação significativa se configurou (p=0,003), de forma que, as faixas com os níveis de 201 a 300, 41,2% (n=49), e superiores a 300, 59,3% (n=16) mostraram-se significativamente associadas ao óbito. Sobre a estimativa de risco, este se mostrou significativo para os níveis acima de 300 (IC95%: 3,38; IC95%: 1,39-8,19).

Considerando a causa do TCE, os resultados apontaram significância limítrofe (0,05<p<0,10), sugerindo que, a queda pode estar relacionada ao óbito, 4,8% (n=44). No entanto, os riscos sobre qualquer causa de TCE não se mostraram relevantes, indicando que o óbito independe desta variável para esta amostra.

Na classificação de Marshall, a associação significativa apontou que as classificações para injúria difusa III, 37,8% (n=65); injúria difusa IV, 67,2% (n=39); lesão de massa evacuada, 35,8% (n=86) e lesão de

massa não evacuada, 53,3% (n=16), mostraram-se significativamente associadas ao óbito (p<0,0001). Sobre os riscos estimados para as categorias de classificações de Marshall citadas, estas oscilaram de 3,54 (IC95%: 1,02-12,29) para a lesão de massa evacuada, até 13,06 (IC95%: 3,42-49,43) estimado para a injúria tipo IV, sempre em comparação com a injúria difusa tipo I (categoria de referência).

Sobre a Hemorragia Subaracnóide (HSA), a presença desta característica mostrou-se significativamente associada ao óbito, 40,0% (n=108) (p=0,003), apontando um risco para óbito de 1,64 (IC95%: 1,20-2,24) mais chance que o grupo sem HSA.

Nos resultados referentes a traumas associados, verificou-se que, a ausência deste é que se mostrou associada significativamente ao óbito, 37,6% (n=160), apontando um risco de 1,65 (IC95%: 1,20-2,25) vezes maior para ocorrência de óbito quando comparado ao grupo com presença de traumas associados. Quando a análise foi detalhada em relação aos tipos de traumas associados, observou-se que, o óbito mostrou-se significativamente associado a não ocorrência de traumas na face, 34,8% (n=26) com um risco de 1,69 (IC95%: 1,05-2,70), e ausência de trauma no tórax, sendo que 35,7% (n=217) vieram a óbito (p=0,004) com risco de 1,89 (IC95%: 1,24-2,91). Em relação aos demais tipos de traumas observados neste estudo não ocorreram associações significativas com o óbito, são eles: os traumas no abdômen (p=0,394), cervical (p=0,536), tóraco-lombar (p=0,891), membros (p=0,101), bem como, com outros tipos de trauma (p=0,528).

No que diz respeito a Escala de Coma de Glasgow, foi detectada associação estatística significativa (p<0,0001), de forma que, o grupo com pontuação de 3-4 mostrou-se associado ao óbito, 55,6% (n=153). Ainda, verificou-se que, quanto menor a pontuação Glasgow maior a probabilidade de ocorrência de óbito (*Linear-By-Linear;* 84,806; p<0,0001). Sobre os riscos estimados, verificou-se que, os investigados com pontuações de 3 a 4 (OR: 5,69; IC95%: 3,87- 8,36) e de 5 a 6 (OR: 1,93; IC95%: 1,26-2,94) apresentaram riscos significativos quando comparados aqueles com pontuações de 7 a 8.

Em relação a caracterização das pupilas, a associação significativa (p<0,0001) com o óbito ocorreu com a pupila midriática, 79,5% (n=66), e anisocórica, 37,8% (n=131). Quanto aos riscos para ocorrência de óbito, estes foram estimados em 21,9 (IC95%: 11,32-39,31) para a pupila midriática e de 3,29 (IC95%: 2,23-4,86) anisocórica, quando comparadas a isocórica.

Tabela 11. Caracterização das variáveis demográficas, radiológicas e neurocirúrgicas de investigados com TCE, para o total da amostra e associação pela presença/ausência de óbito.

(continua)

Variáveis	Total amostra [B]	Óbito [A]			*Odds Ratio* (OR) bruto	
		Não (n=499)	Sim (n=249)	p£	OR (IC95%)	p
Sexo						
Masculino	631 (84,4)	428 (67,8)	203 (32,2)	0,162₮	1,0	
Feminino	117 (15,6)	71 (60,7)	46 (39,3)		1,36 (0,91 – 2,05)	0,132
Idade (anos) *						
Média (±DP) test t-S	34,8 (±16,5)	33,1 (±15,5)	36,2 (±17,6)	0,019¶		
12-30	391 (52,7)	270 (61,9)	121 (30,9)		1,0	
31-45	188 (25,3)	125 (66,5)	63 (33,3)	0,184₳	1,13 (0,78-1,64)	0,512
46-60	93 (12,5)	61 (65,6)	32 (34,4)		1,15 (0,71-1,86)	0,563
Superior a 60	70 (9,4)	39 (55,7)	31 (44,3)		1,77 (1,06-2,98)	0,031
Biênio						
2002-2003-2004	133 (17,8)	102 (76,7)	31 (23,3)		1,0	
2000-20001	142 (19,0)	100 (70,4)	42 (29,6)	0,001₳	1,38 (0,8-2,37)	0,246
1998-1999	133 (17,8)	94 (70,7)	39 (29,3)		1,36 (0,78-2,37)	0,272
1996-1997	162 (21,7)	105 (64,8)	57 (35,2)		1,78 (1,07-2,99)	0,036
1994-1995	178 (23,8)	98 (55,1)	80 (44,9)		2,68 (1,63-4,43)	0,001
Glicose*						
Média (±DP) test t-S	193,9 (±65,5)	156,8 (±55,8)	178,2 (±79,5)	<0,001¶		
61-200	586 (79,0)	407 (69,5)	179 (30,5)		1,03 (0,64-1,67)	0,892
201-300	119 (16,0)	70 (58,8)	49 (41,2)	0,003₳	1,73 (0,95-3,15)	0,074
>300	27 (3,6)	11 (40,7)	16 (59,3)		3,38 (1,39-8,19)	0,007
<60	10 (1,3)	6 (60,0)	4 (40,0)		1,55 (0,40-5,91)	0,517
Causa TCE						
Atropelamento	225 (30,1)	143 (63,6)	82 (36,4)		1,0	
Motorista/passageiro	172 (23,0)	123 (71,5)	49 (28,5)		0,69 (0,45-1,07)	0,093
Queda	96 (12,8)	52 (54,2)	44 (45,8)	0,054₳	1,47 (0,91-2,40)	0,126
Motocicleta	182 (24,3)	128 (70,3)	54 (29,7)		0,74 (0,48-1,12)	0,152
Agressão	28 (3,7)	21 (75,0)	7 (25,0)		0,58 (0,24-1,43)	0,244
Bicicleta	24 (3,2)	16 (66,7)	8 (33,3)		0,87 (0,36-2,13)	0,766
Outros	21 (2,8)	16 (76,2)	5 (23,8)		0,55 (0,19-1,54)	0,255
Marshall ***						
Injúria difusa tipo I	22 (2,9)	19 (86,4)	3 (13,6)		1,0	
Injúria difusa tipo II	175 (23,4)	145 (82,9)	30 (17,1)		1,31 (0,36-4,71)	0,683
Injúria tipo III	172 (23,0)	107 (62,2)	65 (37,8)		3,84 (1,10-13,51)	0,035
Injúria tipo IV	58 (7,8)	19 (32,8)	39 (67,2)		13,06 (3,42-49,43)	<0,001
Lesão de massa evacuada	240 (32,1)	154 (64,2)	86 (35,8)	0,0001₳	3,54 (1,02-12,29)	0,052
Lesão de massa não evacuada	30 (4,0)	14 (46,7)	16 (53,3)		7,24 (1,76-29,75)	0,006
Lesão de tronco cerebral	50 (6,7)	41 (82,0)	9 (18,0)		1,39 (0,34-5,73)	0,653
HSA						
Não	481 (64,3)	340 (70,7)	141 (29,3)	0,003₮	1,0	
Sim	267 (35,7)	159 (59,6)	108 (40,4)		1,64 (1,20-2,24)	0,002
Trauma associado						
Não	425 (56,8)	265 (62,4)	160 (37,6)	0,005₮	1,65 (1,20-2,25)	0,002
Sim	323 (43,2)	234 (72,4)	89 (27,6)		1,0	
Tipo de trauma associado						
Face						
Sim	108 (14,4)	82 (75,9)	26 (24,1)	0,037₮	1,0	0,037
Não	640 (85,6)	417 (65,2)	223 (34,8)		1,69 (1,05-2,70)	0,036
Cervical						
Sim	27 (3,6)	20 (74,1)	7 (25,9)	0,536₮	1,0	
Não	721 (96,4)	479 (66,4)	242 (33,6)		1,44 (0,60-3,45)	0,413
Tóraco-lombar						
Sim	7 (0,9)	4 (57,1)	3 (42,9)	0,891₮	1,0	
Não	741 (99,1)	495 (66,8)	246 (32,2)		0,67 (0,47-2,98)	0,594
Tórax						
Sim	141 (18,9)	109 (77,3)	32 (22,7)	0,004₮	1,0	
Não	607 (81,1)	390 (64,3)	217 (35,7)		1,89 (1,24-2,91)	0,003
Abdômen						
Sim	70 (9,4)	43 (61,4)	27 (38,6)	0,394₮	1,29(0,79-2,15)	
Não	678 (90,6)	456 (67,3)	222 (32,7)		1,0	0,395
Membros						
Sim	204 (27,3)	146 (71,6)	58 (28,4)	0,101₮	1,0	
Não	544 (72,7)	353 (64,9)	191 (35,1)		1,36 (0,96-1,94)	0,085
Outros						
Não	740 (98,9)	495 (66,9)	245 (33,1)	0,528₮	1,0	
Sim	8 (1,1)	4 (50,0)	4 (50,0)		2,03 (0,50-8,14)	0,322

Tabela 11. Caracterização das variáveis demográficas, radiológicas e neurocirúrgicas de investigados com TCE, para o total da amostra e associação pela presença/ausência de óbito.

(conclusão)

Variáveis	Total amostra [B]	Óbito [A]			*Odds Ratio* (OR) bruto	
		Não (n=499)	Sim (n=249)	p£	OR (IC95%)	p
Glasgow**						
7-8	311 (41,7)	255 (82,0)	56 (18,0)	<0,0001₳	1,0	
6-5	192 (25,7)	135 (70,3)	57 (29,7)		1,93 (1,26-2,94)	<0,001
4-3	243 (32,6)	108 (44,4)	135 (55,6)		5,69 (3,87-8,36)	<0,0001
Pupilas**						
Isocórica	283 (38,1)	239 (84,5)	44 (15,5)	0,0001₳	1,0	
Anisocórica	347 (46,7)	216 (62,2)	131 (37,8)		3,29 (2,23-4,86)	<0,0001
Midriática	83 (11,2)	17 (20,5)	66 (79,5)		21,9 (11,32-39,31)	<0,00001
Miótica	30 (4,0)	23 (76,7)	7 (23,3)		2,0 (0,79-5,05)	0,274

*Dados ausentes: 6 (0,8%); **Dados ausentes: 2 (0,3%); ***Dado ausente: 1 (0,1%);
****Dados ausente: 5 (0,7%);
A:Percentuais obtidos com base no total de cada categoria do óbito (sim/não); B: Percentuais obtidos com base no total de casos válidos da amostra;
N.A.: Não se aplica;
£: Nível mínimo de significância para a análise bivariada;
₳: Teste Qui quadrado de Pearson; ₮: Teste Qui quadrado de Pearson com correção de continuidade;
¶: Teste t-Student para grupos independentes.

Fonte: Do autor.

Na verificação das inúmeras variáveis investigadas que apresentaram um real potencial para predizer o óbito, foi aplicada ao conjunto de dados a técnica de Análise de Regressão Logística Binária. Técnica esta multivariada que é apropriada para situações nas quais a variável dependente é categórica (dicotômica), assumindo no caso deste estudo: óbito (1) e não óbito (0); sendo que as variáveis independentes podem ser categóricas ou métricas.

O objetivo da regressão logística é gerar uma função matemática, permitindo aos coeficientes associados as variáveis independentes estabelecer a probabilidade de uma observação pertencer a um grupo previamente determinado (óbito), em função do comportamento de um conjunto de variáveis.

Na composição do modelo multivariado inicial (saturado), foram consideradas como variáveis preditoras aquelas que apresentaram nível mínimo de significância igual ou inferior a 0,200 ($p \leq 0,200$) na análise bivariada (HOSMER; LEMESHOW, 2000).

De acordo com os resultados obtidos na análise bivariada (tabela 12), foram excluídas do modelo inicial as variáveis independentes: outros tipos de traumas, trauma cervical, abdominal e tóraco-lombar, por terem apresentado níveis mínimos de significância superiores a 0,200.

No modelo de regressão logística múltiplo saturado (inicial), apresentado pela tabela 12, em relação a capacidade preditiva utilizou-se a estatística do *Omnibus Tests* do Modelo Logístico, que testou a

hipótese de todos os coeficientes da equação logística serem nulos, ou seja, este teste avalia se os coeficientes relacionados as variáveis independentes têm validade estatística ou não. Nesta técnica os coeficientes indicam o quanto aumenta a probabilidade de ocorrência de um evento (óbito) para o aumento de uma unidade na variável independente. O coeficiente pode ser positivo ou negativo. No caso de um coeficiente positivo, quanto maior for o seu valor, maior será o poder preditivo da variável independente sobre a probabilidade de ocorrência de um evento.

De acordo com os resultados a estimativa para o Qui quadrado foi de 253,008 com p<0,0001 (g.l.= 33), ou seja, há evidências de que os coeficientes relacionados as variáveis independentes não são nulos. Desta forma, pode-se acreditar que, as variáveis independentes conseguem responder sobre a ocorrência do óbito.

Na avaliação resumida (sumária) do modelo, têm-se as estimativas referentes aos valores do Cox & Snell R^2 (0,294) e Nagelkerke R^2 (0,408). Estes dois testes são considerados pseudos R^2 e procuram indicar a proporção das variações ocorridas no *log* da razão de chance. Vê-se também o valor calculado para o Teste de Hosmer e Lemeshow que avalia o grau de acurácia do modelo logístico que nesse caso foi de 0,871. Esta estimativa se mostra relevante quando comparada a outros modelos gerados sobre a mesma base de variáveis.

Avaliando o modelo logístico multivariado saturado, pode-se mensurar o seu poder preditivo. O resultado apontou que o percentual de acerto chegou a 58,8%, sendo a especificidade (não óbito classificado corretamente) do modelo de 88,0% e a sensibilidade (óbito classificado corretamente) de 25,5%.

Tabela 12. Validade preditiva do modelo (*Omnibus Tests*) e proporção de classificação correta dos modelos (*Model summary*); proporção de classificação correta dos modelos e estimativas de validade dos coeficientes de regressão (*Hosmer and Lemeshow*) para o modelo saturado.

Model – Omnibus Tests			***Model summary***		
Chi-Square	**Df**	**Sig.**	**-2 Log *likelihood***	**Cox & Snell R *Square***	**Nagelkerke R *Square***
253,008	33	<0,0001	674,218	0,294	0,408
Hosmer and Lemeshow			**Classificação correta**		
			Óbito		**Total modelo**
Chi-square	**Df**	**Sig.**	**Não - VN**	**Sim – VP**	
3,841	8	0,871	427 (88,0)	62 (25,5)	58,8

Fonte: Do autor.

Considerando os resultados detectados no modelo saturado, seguem-se as interpretações com base na força de associação dos fatores de risco elencados para o modelo, com o cálculo de medidas de associação (*odds ratios*) ajustadas simultaneamente para o efeito de múltiplas variáveis de confusão e/ou modificadoras de efeito (tabela 14).

Em relação ao sexo, observou-se que este perdeu poder de associação, pois o sexo feminino passou de um fator de risco para um fator de proteção (OR: 0,85; IC95%: 0,50-1,44), no entanto manteve ausência de significância estatística.

A faixa etária manteve como categoria representativa para predizer o óbito as idades acima de 60 anos (OR: 2,10; IC95%: 1,02-4,33), indicando que, a chance de óbito para investigados com 60 anos ou mais foi 2,10 vezes maior que o grupo com idades de 12 a 30 anos.

O biênio teve aumentada a sua força de associação para predizer o óbito sobre os riscos dos anos de 1996-1997 (OR: 2,11; IC95%: 1,10-4,04) e de 1994-1995 (OR: 3,14: IC95%: 1,64-6,01), em comparação com o triênio 2002-2003-2004.

A classificação para os níveis de glicose perdeu poder de associação ao óbito, pois os riscos estimados para os níveis elevados não mais se mostraram significativas. A ausência de efeito significativo também foi observada para as causas do TCE, classificação de Marshall, trauma associado, traumas na face e nos membros. Portanto, existem fatores modificadores de efeito que estão interferindo sobre a estimativa de risco das variáveis citadas.

Referente a HSA o risco para óbito teve seu efeito aumentado, passando para 1,99 (IC95%: 1,30-3,06) vezes mais chance que o grupo sem HSA.

Na Escala de Coma de Glasgow verificou-se que os riscos, embora significativos, apresentaram um efeito reduzido para a ocorrência de óbito, tanto na pontuação de 5 a 6 (OR: 1,73; IC95%: 1,04-2,85), quanto de 3 a 4 pontos, (OR: 4,06: IC95%: 2,51-6,56). Demonstrando que esta variável deve estar sendo influenciada por outros fatores (ou outras variáveis) da amostra.

Em relação as pupilas, os riscos para ocorrência de óbito mostraram-se reduzidos, em comparação com o modelo bivariado, com estimativa de 11,39 (IC95%: 5,40-4,04) para a pupila midriática e de 2,54 (IC95%: 1,61-4,03) anisocórica, quando comparadas a pupila isocórica.

De acordo com os resultados do modelo inicial (tabela 13), pode-se verificar que, as variáveis que o compuseram estão sendo afetadas por

fatores confundidores ou modificadores de efeito, pois ocorreram diferenças relevantes nos *Odds Ratios* significativos gerados pelo modelo. Desta forma, este modelo (saturado), não se mostrou robusto para predizer a ocorrência do óbito.

Tabela 13. Modelo inicial (saturado) para a Análise de Regressão Logística Múltipla para predizer óbito.

(continua)

Variáveis	Total amostra	Óbito		OR Ajustado	
		Não (n=499)	Sim (n=249)	OR (IC95%)	p
Sexo					
Masculino	631 (84,4)	428 (67,8)	203 (32,2)	1,0	
Feminino	117 (15,6)	71 (60,7)	46 (39,3)	0,85 (0,50-1,44)	0,546
Idade (anos) *					
12-30	391 (52,7)	270 (61,9)	121 (30,9)	1,0	
31-45	188 (25,3)	125 (66,5)	63 (33,3)	1,02 (0,62-1,66)	0,939
46-60	93 (12,5)	61 (65,6)	32 (34,4)	1,34 (0,71-2,53)	0,373
Superior a 60	70 (9,4)	39 (55,7)	31 (44,3)	2,10 (1,02-4,33)	0,045
Biênio					
2002-2003-2004	133 (17,8)	102 (76,7)	31 (23,3)	1,0	
2000-20001	142 (19,0)	100 (70,4)	42 (29,6)	0,93 (0,45-1,91)	0,849
1998-1999	133 (17,8)	94 (70,7)	39 (29,3)	1,01 (0,52-2,17)	0,861
1996-1997	162 (21,7)	105 (64,8)	57 (35,2)	2,11 (1,10-4,039)	0,025
1994-1995	178 (23,8)	98 (55,1)	80 (44,9)	3,14 (1,64-6,01)	0,001
Glicose*					
61-200	586 (79,0)	407 (69,5)	179 (30,5)	0,33 (0,07-1,61)	0,171
201-300	119 (16,0)	70 (58,8)	49 (41,2)	0,37 (0,07-1,93)	0,372
>300	27 (3,6)	11 (40,7)	16 (59,3)	0,39 (0,06-2,50)	0,319
<60	10 (1,3)	6 (60,0)	4 (40,0)	0,22 (0,02-1,26)	0,922
Causa TCE					
Atropelamento	225 (30,1)	143 (63,6)	82 (36,4)	1,0	
Motorista/passageiro	172 (23,0)	123 (71,5)	49 (28,5)	0,71 (0,41-1,26)	0,256
Queda	96 (12,8)	52 (54,2)	44 (45,8)	1,15 (0,61-2,19)	0,569
Motocicleta	182 (24,3)	128 (70,3)	54 (29,7)	0,43 (0,14-1,35)	0,150
Agressão	28 (3,7)	21 (75,0)	7 (25,0)	0,85 (0,48-1,51)	0,588
Bicicleta	24 (3,2)	16 (66,7)	8 (33,3)	0,49 (0,15-1,59)	0,237
Outros	21 (2,8)	16 (76,2)	5 (23,8)	0,47 (0,12-1,91)	0,294
Marshall ***					
Injúria difusa tipo I	22 (2,9)	19 (86,4)	3 (13,6)	1,0	
Injúria difusa tipo II	175 (23,4)	145 (82,9)	30 (17,1)	0,59 (0,14-2,54)	0,485
Injúria tipo III	172 (23,0)	107 (62,2)	65 (37,8)	1,12 (0,26-4,72)	0,875
Injúria tipo IV	58 (7,8)	19 (32,8)	39 (67,2)	3,77 (0,82-17,39)	0,089
Lesão de massa evacuada	240 (32,1)	154 (64,2)	86 (35,8)	0,87 (0,21-3,65)	0,851
Lesão de massa não evacuada	30 (4,0)	14 (46,7)	16 (53,3)	3,82 (0,73-19,79)	0,110
Lesão de tronco cerebral	50 (6,7)	41 (82,0)	9 (18,0)	0,37 (0,07-1,82)	0,222
HSA					
Não	481 (64,3)	340 (70,7)	141 (29,3)	1,0	
Sim	267 (35,7)	159 (59,6)	108 (40,4)	1,99 (1,30-3,06)	0,002
Trauma associado					
Não	425 (56,8)	265 (62,4)	160 (37,6)	0,88 (0,41-1,92)	0,093
Sim	323 (43,2)	234 (72,4)	89 (27,6)	1,0	
Tipo de trauma associado					
Face					
Sim	108 (14,4)	82 (75,9)	26 (24,1)	1,0	
Não	640 (85,6)	417 (65,2)	223 (34,8)	0,75 (0,39-1,43)	0,383
Tórax					
Sim	141 (18,9)	109 (77,3)	32 (22,7)	1,0	
Não	607 (81,1)	390 (64,3)	217 (35,7)	1,99 (0,98-4,04)	0,056

Tabela 13. Modelo inicial (saturado) para a Análise de Regressão Logística Múltipla para predizer óbito.

(conclusão)

Variáveis	Total amostra	Óbito		OR Ajustado	
		Não (n=499)	Sim (n=249)	OR (IC95%)	p
Membros					
Sim	204 (27,3)	146 (71,6)	58 (28,4)	1,0	
Não	544 (72,7)	353 (64,9)	191 (35,1)	0,71 (0,34-1,48)	0,364
Glasgow **					
7-8	311 (41,7)	255 (82,0)	56 (18,0)	1,0	
6-5	192 (25,7)	135 (70,3)	57 (29,7)	1,73 (1,04-2,85)	0,034
4-3	243 (32,6)	108 (44,4)	135 (55,6)	4,06 (2,51-6,56)	<0,0001
Pupilas ****					
Isocórica	283 (38,1)	239 (84,5)	44 (15,5)	1,0	
Anisocórica	347 (46,7)	216 (62,2)	131 (37,8)	2,54 (1,61-4,03)	<0,001
Midriática	83 (11,2)	17 (20,5)	66 (79,5)	11,39 (5,40-24,04)	<0,00001
Miótica	30 (4,0)	23 (76,7)	7 (23,3)	1,50 (0,53-4,23)	0,274

Nota: R^2 de Nalgelkerke =0,207 ; 2LL = 809,447; Prova de Hosmer-Lemeshow p = 0,994;
Ocorreram 20 casos de dados ausentes sobre o conjunto de variáveis: idade [6 (0,8%)],
Glasgow [2 (0,3%)]; Marshall [1 (0,1%)] e Pupilas [5 (0,7%)]; OR: Odds ratio.
Fonte: Do autor.

Considerando ainda como base o modelo inicial, segue-se a implementação da técnica de Análise de Regressão Múltipla buscando o modelo que melhor consegue predizer o desfecho de estudo (óbito).

Esta técnica minimiza o número de variáveis para que o modelo resultante seja facilmente generalizado e estável numericamente. Neste estudo, como forma de seleção dos fatores com potencial real de predição do óbito, foi utilizado o método *forward* que se caracteriza por considerar a variável de maior coeficiente de associação bivariada amostral, observado com a variável resposta (óbito). Conforme Charnet et al (2008) a cada etapa uma variável pode ser incorporada ao modelo, quando esta inclusão não ocorre, o processo é interrompido e as variáveis selecionadas até o momento definem o modelo final.

De acordo com os resultados obtidos, o modelo final (reduzido) foi estabelecido em sete etapas (*step*). Na tabela 14 têm-se as significâncias estatísticas de cada etapa (*Omnibus Tests*), constatando-se na avaliação da validade preditiva do modelo que os coeficientes são significativos a cada etapa, pois o valor da estatística do teste Qui-quadrado calculado aumenta, implicando em nível de significância ainda menor. Desta forma, há evidências de que os coeficientes de regressão (β), relacionados as categorias das variáveis elencadas para regressão logística, estão se mostrando ainda mais relevantes para predizer a presença do óbito, ou seja, os coeficientes apresentaram um maior poder de predição sobre a probabilidade de ocorrência do óbito.

No ajuste geral (*Model summary*), observou-se que, do passo 1 ao 7, quando uma nova variável era inserida, a estatística de máxima verossimilhança (-2log ou 2LL) diminuiu indicando a melhora do modelo, de 809,447 na etapa 1 para 684,047 na etapa 7.

Em contrapartida, os valores do Pseudo R^2 (Cox & Snell e Nagelkerke) aumentaram à medida que os previsores eram adicionados, indicando que a cada etapa de inserção das variáveis o poder de classificação correta da ocorrência de óbito aumentou. Pelas estimativas, tem-se que o Pseudo R^2 de Nagelkerke no passo 1 foi de 20,7%, enquanto que, no passo 7 foi de 39,4%, ou seja, o poder de explicação não chegou a 50%. No entanto, chama-se atenção para o fato de que, embora 39,4% não seja um percentual de explicação muito elevado, observou-se um aumento no poder de explicação do modelo de 90,3% no passo 7 em relação ao passo 1.

A medida de Hosmer e Lemeshow, também de ajuste geral, é o teste que indica se houve diferença estatística significativa entre as classificações observadas e previstas para a presença de óbito, para todos os modelos gerados passo-a-passo (HOSMER; LEMESHOW, 2000). Este teste mede a correspondência dos valores efetivos e previstos para a variável dependente (óbito), tendo-se o melhor ajuste do modelo indicado por uma diferença menor na classificação observada e prevista, bem como, por um valor Qui-quadrado não significante.

Todos os modelos gerados não apresentaram diferenças estatisticamente significativas entre os valores previstos e os observados (tabela 14), o que indica que os modelos estão sendo capazes de produzir estimativas confiáveis para a classificação correta de óbito.

Considerando todas as estimativas para a validade do modelo combinadas, a indicação é que seja aceito o modelo no último passo (*step* 7), como o modelo significativo para predizer a presença do óbito.

Desta forma, para o modelo validado foram elencados como fatores preditores significativos as variáveis: Faixa etária, Biênio, Marshall, HSA, Trauma de tórax, Glasgow e Pupilas. Neste conjunto de variáveis observou-se que o percentual de acerto *a posteriori* chegou a 76,9%, classificando corretamente 55,6% (n=60) dos investigados que foram a óbito (verdadeiros positivos) e 87,6% (n=425) dos casos de não óbito (verdadeiros negativos).

Pelos resultados da tabela 15 (Modelo final), o OR referente a classificação para óbito detectou como potencial fator preditor a variável Glasgow, sendo que os grupos com pontuações 3-4 (OR: 3,97; IC95%: 2,49-6,31) e 5-6 (OR: 1,68; IC95%: 1,03-2,75) apresentaram riscos significativos quando comparados ao com pontuações 7-8.

Tabela 14. Validade preditiva do modelo (*Omnibus Tests*) e proporção de classificação correta dos modelos (*Model summary*), segundo as etapas de seleção (*step*); estimativas de validade dos coeficientes de regressão (Hosmer and Lemeshow) e proporção de classificação correta dos modelos, segundo as etapas de seleção (*step*).

Testes de validade			Etapas – Step						
			1	2	3	4	5	6	7
Model - Omnibus Tests	Qui quadrado		117,779	152,173	193,727	218,404	226,831	234,049	243,179
	G.L.		3	5	11	16	17	18	21
	p		<0,0001	<0,00001	<0,00001	<0,00001	<0,000001	<0,000001	<0,000001
Model summary	-2 LL		809,447[a]	775,053[a]	733,499[b]	708,822[b]	700,395[b]	693,177[b]	684,047[b]
	Cox & Snell		0,149	0,189	0,234	0,259	0,268	0,275	0,284
	Nagelkerke		0,207	0,262	0,324	0,36	0,372	0,382	0,394
Hosmer and Lemeshow	Qui quadrado		0	0,721	5,193	7,645	8,913	6,908	8,499
	G.L.		2	6	8	8	8	8	8
	p		1	0,994	0,737	0,469	0,35	0,547	0,386
Classificação correta	Óbito	Não - VN	468 (96,5)	414 (85,4)	418 (86,2)	439 (90,5)	433 (89,3)	431 (88,9)	425 (87,6)
		Sim - VP	62 (25,5)	122 (50,2)	136 (56,0)	131 (53,9)	134 (55,1)	132 (54,3)	60 (55,6)
	Total modelo		72,8	73,6	76,1	78,3	77,9	77,3	76,9
Variáveis inseridas no modelo a cada etapa			Pupilas	Pupilas Categ	Pupilas Categ Marshall	Pupilas Categ Marshall Biênio	Pupilas Categ Marshall Biênio HSA	Pupilas Categ Marshall Biênio HSA Tórax	Pupilas Categ Marshall Biênio HSA Tórax Faixa etária

a. *Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001;*
b. *Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.*
2LL: Log *likelihood*; Cox & Snell R *Square*; Nagelkerke R *Square*; G.L.: Grau de liberdade;
VN: Verdadeiros Negativos (classificação correta para não óbito);
VP: Verdadeiros Positivos (classificação correta para óbito sim).
Fonte: Do autor.

Conforme a tabela 15, a não ocorrência de trauma no tórax apresentou um risco 2,02 (IC95%: 1,19-3,41) vezes mais elevado de ocorrência de óbito se comparados aqueles que apresentaram este tipo de trauma. Em relação ao HSA, o grupo com a presença desta característica teve um risco 1,86 (IC95%: 1,23-2,81) vezes mais elevado de ocorrência de óbito quando comparados aqueles sem HSA.

As estimativas de risco em relação as pupilas também se mostraram relevantes para a predição, de forma que, os investigados caracterizados como apresentando pupilas midriáticas apresentaram 11,24 (IC95%: 5,42- 23,30) vezes mais chance de ocorrência de óbito quando comparados ao grupo com pupilas isocóricas. Já, aqueles investigados com pupilas anisocóricas o risco foi estimado em 2,65

(IC95%: 1,69- 4,17) e o grupo com pupilas mióticas em 1,47 (IC95%: 0,53-4,07), ambos em comparação ao grupo com pupilas isocórica.

Outras variáveis que compuseram o modelo final apresentaram determinadas categorias relacionadas de forma significativas ao risco, destacando-se o Biênio para os anos de 1994-1995 (OR: 3,17; IC95%: 1,71- 5,88) e 1996-1997 (OR: 2,17; IC95%: 1,16- 4,04) em comparação ao triênio 2002-2003-2004. Como também, os investigados com idade superior a 60 anos (OR: 2,51; IC95%: 1,31-4,84; p=0,006) em comparação aqueles com idades de até 30 anos.

Em relação a classificação Marshall, os resultados apontaram para uma tendência de risco sobre a categoria Injúria tipo IV (OR: 3,63; IC95%: 0,84-15,76; p=0,076), definindo esta variável como a que menos contribui para predizer óbito no modelo.

Tabela 15. Modelo final para a Análise de Regressão Logística para predizer óbito.

Variáveis	Total amostra	Óbito		OR Ajustado	
		Não (n=499)	Sim (n=249)	OR (IC95%)	p
Idade (anos)					
12-30	391 (52,7)	270 (61,9)	121 (30,9)	1,0	
31-45	188 (25,3)	125 (66,5)	63 (33,3)	1,05 (0,66-1,67)	0,839
46-60	93 (12,5)	61 (65,6)	32 (34,4)	1,61 (0,90-2,88)	0,111
Superior a 60	70 (9,4)	39 (55,7)	31 (44,3)	2,51 (1,31-4,84)	0,006
Bienio					
2002-2003-2004	133 (17,8)	102 (76,7)	31 (23,3)	1,0	
2000-20001	142 (19,0)	100 (70,4)	42 (29,6)	1,00 (0,51-1,98)	>0,999
1998-1999	133 (17,8)	94 (70,7)	39 (29,3)	1,15 (0,58-2,25)	0,681
1996-1997	162 (21,7)	105 (64,8)	57 (35,2)	2,17 (1,16-4,04)	0,009
1994-1995	178 (23,8)	98 (55,1)	80 (44,9)	3,17 (1,71-5,88)	<0,0001
Marshall					
Injúria difusa tipo I	22 (2,9)	19 (86,4)	3 (13,6)	1,0	
Injúria difusa tipo II	175 (23,4)	145 (82,9)	30 (17,1)	0,55 (0,13-2,23)	0,410
Injúria tipo III	172 (23,0)	107 (62,2)	65 (37,8)	1,05 (0,26-4,21)	0,922
Injúria tipo IV	58 (7,8)	19 (32,8)	39 (67,2)	3,63 (0,84-15,76)	0,076
Lesão de massa evacuada	240 (32,1)	154 (64,2)	86 (35,8)	0,81 (0,21-3,16)	0,789
Lesão de massa não evacuada	30 (4,0)	14 (46,7)	16 (53,3)	3,18 (0,66-15,26)	0,149
Lesão de tronco cerebral	50 (6,7)	41 (82,0)	9 (18,0)	0,34 (0,07-1,60)	0,173
HSA					
Não	481 (64,3)	340 (70,7)	141 (29,3)	1,0	
Sim	267 (35,7)	159 (59,6)	108 (40,4)	1,86 (1,23-2,81)	0,003
Trauma Tórax					
Sim	141 (18,9)	109 (77,3)	32 (22,7)	1,0	
Não	607 (81,1)	390 (64,3)	217 (35,7)	2,02 (1,19-3,41)	0,009
Glasgow					
7-8	311 (41,7)	255 (82,0)	56 (18,0)	1,0	
6-5	192 (25,7)	135 (70,3)	57 (29,7)	1,68 (1,03-2,75)	0,038
4-3	243 (32,6)	108 (44,4)	135 (55,6)	3,97 (2,49-6,31)	<0,0001
Pupilas					
Isocórica	283 (38,1)	239 (84,5)	44 (15,5)	1,0	
Anisocórica	347 (46,7)	216 (62,2)	131 (37,8)	2,65 (1,69-4,17)	<0,001
Midriática	83 (11,2)	17 (20,5)	66 (79,5)	11,24 (5,42-23,30)	<0,00001
Miótica	30 (4,0)	23 (76,7)	7 (23,3)	1,47 (0,53-4,07)	0,389

Nota: R^2 de Nalgelkerke =0,394; 2LL = 684,047; Prova de Hosmer-Lemeshow p = 0,386; Ocorreram 20 casos de dados ausentes sobre o conjunto de variáveis idade [6 (0,8%)]; Glasgow [2 (0,3%)]; Marshall [1 (0,1%)] e Pupilas [5 (0,7%)]; OR: *Odds ratio.*

Fonte: Do autor.

Considerando o modelo final, montou-se a equação para a regressão logística.

$$Probabilidade\_de\_óbito = 0{,}048 * idade1 + 0{,}475 * idade2 + 0{,}919 * idade3 + 0{,}775 * biênio1 + 0{,}537 * biênio2 + 0{,}668 * biênio3 + 1{,}306 * biênio4 + ((-0{,}589) * Marshall1) + 0{,}069 * Marshall2 + 1{,}335 * Marshall3 + ((-0{,}186) * Marshall4) + 1{,}257 * Marshall5 + ((-1{,}077) * Marshall6) + 0{,}585 * hsa1 + 0{,}730 * traumatórax1 + 0{,}522 * Glasgow1 + 1{,}387 * Glasgow2 + 1{,}222 * pupilas1 + 2{,}377 * pupilas2 + 0{,}967 * pupilas3$$

Tendo-se as variáveis: idade1 correspondendo a idade em anos que varia de 31 a 45, idade2 de 46 a 60, idade3 acima de 60; biênio1 refere-se aos anos de 2000 a 2001, biênio2 de 1998 a 1999, biênio3 de 1996 a 1997 e biênio4 de 1994 a 1995; Marshall1 designa injúria difusa tipo II, Marshall2 injúria difusa tipo III, Marshall3 injúria difusa tipo IV, Marshall4 lesão de massa evacuada, Marshall5 lesão de massa não evacuada e Marshall6 lesão de tronco cerebral; hsa1 corresponde a presença de hemorragia subaracnóide; traumatórax1 significa ausência de trauma torácico; Glasgow1 se refere a Escala de Coma de Glasgow variando de 6 a 5 e Glasgow2 de 4 a 3; pupilas1 são anisocórica; pupilas2 midriática e pupilas3 miótica.

Baseando-se nos resultados apresentados por Martins et al (2009), neste estudo foram obtidas estimativas semelhantes que se mostraram coincidentes em todos os aspectos, sendo que, diferenças relevantes ocorreram na forma de apresentação das análises geradas, adotando-se um enfoque mais detalhado direcionado aos objetivos desta pesquisa.

7.2 ANÁLISE DOS MÉTODOS DE *DATA MINING* EMPREGADOS

A análise referente ao estudo das simulações ocorreu sobre 133 modelos, que consideraram o 19 tipos de experimentos e os sete algoritmos (C4.5, CART, kNN, RBF, Naive Bayes, Bayes Net, AdaBoost.M1), para responder pela ocorrência do desfecho óbito, sendo este considerado o Verdadeiro Positivo (VP) dos modelos.

Também foram analisadas as medidas de desempenho das classificações corretas para óbito (VP), não óbito (VN) e acurácia. Desta forma, não serão abordados os classificadores para os erros das medidas de desempenho, uma vez que se tratam de medidas complementares aos acertos. Calculou-se a estimativa do coeficiente kappa, que fornece a concordância além do que seria esperado pelo acaso; bem como, a curva

ROC que indica a probabilidade de discriminação do modelo gerado para a ocorrência de óbito (índice de exatidão do teste).

Considerando o total de 133 modelos obtidos, inicialmente se buscou caracterizar para o total da amostra o perfil das medidas de desempenho para cada tipo de classificação (experimento e algoritmo) gerada nos modelos. Para tanto, calcularam-se as médias para as medidas de desempenho sobre cada uma das classificações, possibilitando avaliar a ocorrência das melhores estimativas médias na caracterização do óbito.

De acordo com os resultados apresentados pela tabela 16, verificou-se que para os vários tipos de experimentos gerados neste estudo, as médias para as medidas de desempenho apresentaram diferenças significativas para a classificação correta do óbito (VP) ($p<0,001$) e para a acurácia ($p<0,001$). Desta forma, há evidências de que existem experimentos que classificam com maior acurácia a ocorrência de óbito. Neste sentido, mostrou-se significativamente mais elevada a classificação correta (VP) do experimento 19, que classificou em média corretamente 56,5% dos casos de óbito, enquanto que a menor média ocorreu no experimento 13 (33,4%).

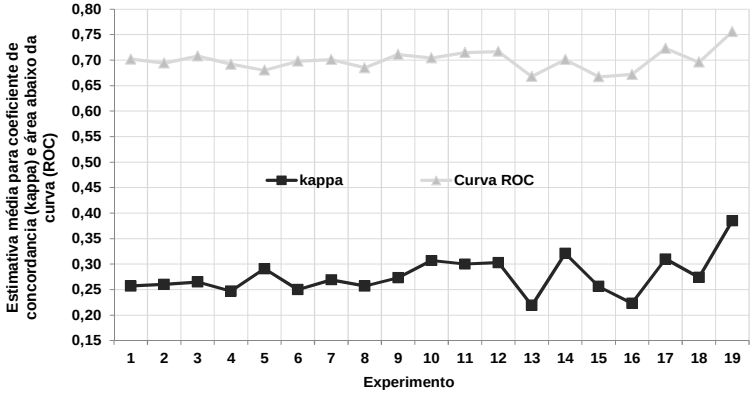
Em relação a acurácia, novamente destaca-se o experimento 19 que, em média, classificou corretamente óbito e não óbito em 77,5% dos modelos, sendo que, o experimento 13 novamente se destacou com a menor acurácia entre os modelos investigados, 65,8%.

Ainda, em relação a classificação correta para ocorrência de óbito, no que refere as medidas que minimizam a presença do acaso, observou-se que a média do coeficiente kappa (figura 10) diferiu significativamente entre os experimentos ($p<0,001$), apontando que, a concordância entre as classificações corretas para o modelo 19 (kappa=0,385) mostrou-se superior a concordância observada nos demais experimentos, sendo que, o maior comprometimento da concordância ocorreu no experimento 13 (kappa=0,219). A diferença significativa também se configurou entre as médias para a Curva ROC ($p<0,001$), tendo-se o maior poder de discriminação do modelo no experimento 19, com estimativa de 0,780, valor este superior as médias dos demais experimentos (figura 10). Destaca-se o fato de que foram considerados como apresentando um menor poder de discriminação os experimentos 13 (0,668), 14 (0,701), 15 (0,667) e 16 (0,672), enquanto que, sobre os demais experimentos a média para a área da curva ROC ficou acima de 0,700 e abaixo de 0,800.

Na informação referente a média para sensibilidade em relação aos diferentes experimentos (tabela 16), mostrou-se significativamente

elevada a estimativa do experimento 19 (74,4%), sendo que, ficaram a cargo dos experimentos 13 e 16 as menores estimativas para sensibilidade, 54,0% e 55,3%, respectivamente. Os resultados para a média da especificidade mostrou-se significativamente mais elevada no experimento 13 (93,7%), sendo menos representativa nos experimentos 2 (87,5%) e 19 (87,5%).

Figura 10- Distribuição das estimativas médias para a área sob a curva ROC e coeficiente de concordância kappa segundo o tipo de experimento.



Fonte: Do autor.

No que se refere aos resultados dos modelos analisados em função dos algoritmos, ocorreu diferença significativa para a classificação correta do óbito ($p<0,01$), indicando que o algoritmo Bayes Net (65,7%), assim como o Naive Bayes (52,9%) foram aqueles que, em média, apresentaram os maiores VP. Também, ocorreu diferença estatística significativa para a classificação correta do não óbito ($p<0,001$), apontando que, o algoritmo AdaBoost.M1 foi o que se destacou de forma representativa (91,4%) para a classificação dos casos VN em relação aos demais algoritmos.

A acurácia mostrou-se significativamente elevada ($p<0,001$) na média alcançada pelo algoritmo Bayes Net, que classificou corretamente – óbito e não óbito – em 76,6% de seus modelos. O algoritmo com menor acurácia média foi o RBF (46,1%).

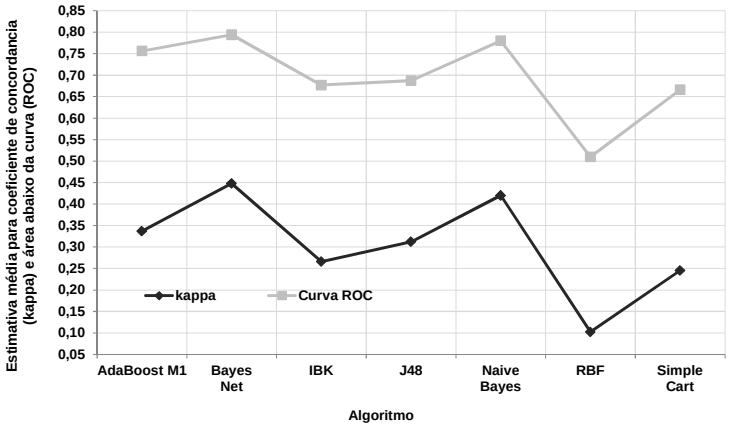
Sobre a concordância além do acaso (figura 11), a diferença significativa ($p<0,001$) apontou que o algoritmo Bayes Net (kappa=0,448) e Naive Bayes (kappa=0,420) apresentaram a maior concordância média quando comparados aos demais algoritmos, sendo

que a menor concordância ficou a cargo do algoritmo RBF (kappa=0,102).

Considerando os resultados para a área da Curva ROC, quando comparadas entre os algoritmos (figura 11), a diferença significativa também se configurou ($p<0,001$). Destacaram-se com a maior média para o poder de discriminação os modelos gerados pelos algoritmos Bayes Net (Área=0,794) e Naive Bayes (Área=0,780), sendo que o menor poder de discriminação foi observado no algoritmo RBF (Área=0,510).

Ainda foram estimadas as medidas de sensibilidade e especificidade, sendo que na primeira as estimativas foram significativamente mais elevadas nos algoritmos Bayes Net (74,3%) e AdaBoost.M1 (64,7%), enquanto para a especificidade os algoritmos AddaBoost.M1 (91,8%) e J48 (C4.5) (89,0%). No que se refere as menores estimativas observadas, tanto para a sensibilidade quanto para a especificidade, o destaque foi o algoritmo RBF, 38,6% e 59,3%, respectivamente.

Figura 11 - Distribuição das estimativas médias para a área sob a curva ROC e coeficiente de concordância kappa segundo o tipo de algoritmo



Fonte: Do autor.

Tabela 16- Escores médios para as medidas de desempenho (VP, VN, acurácia, sensibilidade e especificidade), kappa e Curva ROC estratificadas pelo experimento e algoritmo.

DM	n	Classificação correta (%) Óbito VP	Sobrevida VN	Acurácia	Sensibilidade	Especificidade	Kappa	Curva ROC
Experimento (19)								
1	7	42,4bc	82,4	71,1b	65,0b	88,7c	0,257d	0,702b
2	7	44,3b	84,2	72,9b	63,1c	87,5c	0,260d	0,694b
3	7	43,8b	83,4	72,3b	63,6	88,2c	0,265d	0,708b
4	7	44,9b	84,3	73,3b	62,5c	87,6c	0,247d	0,692b
5	7	40,3c	79,6	68,6c	67,1b	89,6b	0,291d	0,680c
6	7	40,7c	78,9	68,5c	66,7b	88,9b	0,250d	0,698bc
7	7	40,6c	80,1	69,1c	66,8b	89,6b	0,269d	0,701bc
8	7	42,0bc	81,4	70,4bc	65,4b	89,1b	0,257d	0,685b
9	7	42,5bc	80,1	70,0bc	64,9bc	90,0a	0,273b	0,711b
10	7	46,1b	74,3	68,9c	69.8b	90,0a	0,307b	0,704b
11	7	42,9bc	86,7	73,5b	64,5b	90,2a	0,300c	0,715b
12	7	41,6c	84,4	71,7b	65,8b	89,6b	0,303c	0,717b
13	7	33,4d	89,3	65,8c	54,0a	93,7a	0,219d	0,668c
14	7	38,9d	88,6	72,4b	58,4ab	91,9a	0,321b	0,701bc
15	7	35,1d	89,6	71,1b	56,4a	92,9a	0,256d	0,667c
16	7	35,1d	89,5	71,2b	55,3a	92,8a	0,223d	0,672c
17	7	46,4b	75,1	69,5c	69,5b	89,6b	0,310b	0,723b
18	7	42,5b	82,6	71,5b	64,9b	89,9b	0,274b	0,696b
19	7	56,5a	80,2	77,5a	74,4a	87,5c	0,385a	0,756a
p6		<0,001	0,997	<0,001	<0,001	0,009	<0,001	<0,001
Algoritmo (7)								
AdaBoostM1	19	39,5b	91,4b	73,9ab	64,7b	91,8	0,337c	0,756b
Bayes Net	19	65,7a	87,1b	76,6a	74,3a	87,1	0,448a	0,794a
IBK (kNN)	19	40,3b	87,1bc	69,7b	59,7c	84,5	0,266d	0,677b
J48 (C4.5)	19	39,3b	89,1b	72,4b	60,7b	89,0	0,312c	0,687b
Naive Bayes	19	52,9a	85,5bc	75,5a	49,0d	86,4	0,420b	0,780a
RBF	19	22,1d	58,3d	46,1c	38,6b	59,3	0,102e	0,510c
Simple Cart (CART)	19	33,2c	89,2b	70,4b	56,8c	89,2	0,245d	0,666b
p6		<0,001	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001

6: Análise de Variância *One Way – Post Hoc* Bonferroni – médias seguidas de letras iguais na coluna (sobre os resultados de cada variável) não diferem a 5% de significância.

Fonte: Do autor.

Considerando-se os experimentos realizados nesta pesquisa, foram selecionados aqueles que apresentaram as melhores medidas de desempenho, sendo que os principais critérios de seleção foram a Curva ROC e o coeficiente de concordância kappa, pois são medidas de desempenho com elevadas precisão (habilidade do modelo em prever corretamente as classes) e robustez (habilidade do modelo para avaliar ou prever corretamente, utilizando dados ruidosos e com viés).

Conforme consta na tabela 17, dos 19 experimentos desta pesquisa, dez (52,6%) referem-se ao algoritmo Naive Bayes, seis (31,6%) ao AdaBoost.M1; dois (10,5%) ao SimpleCart (CART) e um (5,3%) ao Bayes Net. Portanto, em mais da metade dos experimentos os melhores modelos se caracterizaram por ser do algoritmo Naive Bayes.

De acordo com os resultados, considerando-se em cada modelo uma amostra independente para a verificação da qualidade do ajuste, o modelo com maior robustez e acurácia ocorreu no experimento 19 (Naive bayes). Este experimento foi composto pelas mesmas sete variáveis elencadas pela regressão logística, (pupilas, catgcs, Marshall, biênio, HSA, trauma de tórax e faixa etária), que apresentou as maiores estimativas para a área sob a curva ROC (Área=0,851), assim como para o coeficiente de concordância kappa (0,530). Ainda sobre este modelo, verificou-se 58,2% de classificação correta para óbito, 92,3% de classificação correta para não óbito e uma acurácia de 80,2%.

No que se refere ao modelo mais robusto o destaque foi para o experimento 17 por meio do algoritmo Naive Bayes, em que as variáveis foram definidas pelo especialista de domínio de aplicação, com área sob a curva ROC estimada em 0,808 e concordância kappa de 0,498; classificações corretas para óbito e não óbito de 58,3 e 88,7%, respectivamente; acurácia de 77,4%.

O modelo de menor poder de predição foi o do experimento 8 pelo algoritmo Naive bayes, com área sob a curva de 0,772, coeficiente de concordância kappa de 0,437, acurácia de 75,8% e classificações corretas para óbito e não óbito de 56% e 86,1%, respectivamente.

Tabela 17 - Caracterização do melhor modelo em cada experimento do *data mining*, segundo as medidas de desempenho (VP, VN, acurácia, sensibilidade e especificidade), kappa e Curva ROC.

Modelos CV		Classificação correta						
Experimento	Algoritmo	Óbito VP	Sobrevida VN	Acurácia	Sensibilidade	Especificidade	Kappa	Curva ROC
1	AdaBoostM1	53,6	86,0	75,0	58,0	88,8	0,414	0,787
2	AdaBoostM1	48,8	89,0	75,4	53,2	91,8	0,408	0,793
3	AdaBoostM1	53,6	86,0	75,0	58,0	81,8	0,415	0,795
4	AdaBoostM1	51,2	87,8	75,4	55,7	80,6	0,349	0,802
5	Simple Cart	59,5	81,1	73,8	54,4	83,6	0,409	0,782
6	Simple Cart	59,5	81,1	73,8	63,9	83,9	0,409	0,782
7	Naive Bayes	52,4	87,2	74,4	56,8	82,2	0,419	0,777
8	Naive Bayes	56,0	86,1	75,8	60,4	90,0	0,437	0,772
9	Naive Bayes	57,1	86,6	76,6	61,5	81,4	0,455	0,805
10	Naive Bayes	56,0	87,2	76,6	60,4	82,5	0,452	0,799
11	Naive Bayes	58,3	87,8	77,8	66,7	87,9	0,482	0,807
12	Naive Bayes	59,5	84,8	76,2	63,9	83,6	0,454	0,801
13	AdaBoostM1	44,0	99,3	76,6	48,4	88,2	0,417	0,775
14	AdaBoostM1	44,0	93,3	76,6	50,2	85,7	0,417	0,775
15	Naive Bayes	57,1	83,5	74,6	61,9	86,1	0,418	0,783
16	Naive Bayes	57,6	84,9	75,8	64,7	92,5	0,439	0,776
17	Bayes Net	58,3	88,7	78,9	69,5	87,6	0,498	0,808
18	Naive Bayes	58,3	85,4	76,4	68,7	82,6	0,444	0,802
19	Naive Bayes	58,2	92,3	80,2	72,7	84,2	0,530	0,851

Fonte: Do autor.

No que se refere aos dez melhores modelos (tabela 18) detectados nesta pesquisa, conforme estimativa da área sob a curva ROC, o destaque foi para o experimento 19 pelo algoritmo Naive Bayes, bem como, o experimento 17 pelo algoritmo Bayes Net.

Comparando os dois modelos de maior precisão, em suas estimativas de concordância kappa e área sob a curva ROC, as diferenças significativas se configuraram ($p<0,01$), apontando que o experimento 19 (Naive Bayes – Variáveis: pupilas, catgcs, Marshall, biênio, HSA, trauma de tórax e faixa etária) apresentou medidas de confiabilidade significativamente mais elevadas que o modelo do experimento 17 (Bayes Net – Variáveis: sexo, causa, Marshall, HSA, pupilas, trauma associado, catidade2, catglic2, gcsadm). Portanto, o modelo com maior poder de predição para a ocorrência de óbito foi o do experimento 19 pelo algoritmo Naive Bayes.

No entanto, observou-se que entre o segundo melhor modelo (Experimento 17 – Bayes Net) e os demais modelos elencados na tabela 19, considerando a área sob a curva ROC, as diferenças significativas não se configuraram, apontando que entre estes modelos o poder de discriminação foi semelhante. Ainda, cabe salientar que não se mostraram representativos (alta robustez e elevada confiabilidade) para predizer a ocorrência do óbito os experimentos 1, 2, 5, 6, 7, 8, 13, 14, 15 e 16.

Tabela 18 - Caracterização dos dez melhores modelos, segundo as medidas de desempenho (VP, VN, acurácia, sensibilidade e especificidade), kappa e Curva ROC.

Modelos *data mining*		Classificação correta			Sensibilidade	Especificidade	Kappa	Curva ROC
Experimento	Algoritmo	Óbito VP	Não óbito VN	Acurácia				
19	Naive Bayes	58,2	92,3	80,2	72,7	84,2	0,530	0,851
17	Bayes Net	58,3	87,2	77,2	64,6	79,8	0,498	0,808
10	Naive Bayes	58,3	87,8	77,8	66,7	81,6	0,482	0,807
9	Naive Bayes	57,1	86,6	76,6	61,5	81,4	0,455	0,805
4	AdaBoost.M1	51,2	87,8	75,4	55,7	80,6	0,349	0,802
18	Naive Bayes	58,3	84,8	75,8	66,3	82,7	0,444	0,802
11	Naive Bayes	59,5	84,8	76,2	63,9	83,6	0,454	0,801
19	Bayes Net	56,0	87,2	76,7	63,8	85,4	0,452	0,799
12	Naive Bayes	56,0	87,2	76,6	60,4	82,5	0,452	0,799
3	AdaBoost.M1	53,6	86,0	75,0	58,0	81,8	0,415	0,795

Fonte: Do autor.

7.3 COMPARAÇÃO *DATA MINING* E REGRESSÃO LOGÍSTICA

No que se refere aos modelos para predizer a ocorrência de óbito, pelas técnicas de regressão logística binária e *data mining*, estes foram comparados buscando-se identificar qual apresentou melhor desempenho. Salienta-se que as duas técnicas em questão, além de apresentarem diferenças básicas em suas metodologias e aplicação, também possuem seus resultados influenciados pelas peculiaridades inerentes a amostra utilizada na construção dos modelos.

A seleção dos modelos gerados pelo *data mining* a serem considerados na comparação com a análise de regressão logística binária, levou em consideração as estimativas mais elevadas na avaliação de desempenho sobre a matriz de confusão, a maior área estimada para a Curva ROC e o coeficiente de concordância kappa.

Nos vários modelos gerados pelo *data mining* mostraram-se relevantes aqueles que apresentaram as estimativas mais elevadas para curva ROC, os quais foram oriundos dos experimentos 19 pelo algoritmo Naive Bayes e 17 pelo Bayes Net.

O modelo considerado pela análise de regressão logística elencou sete variáveis (pupilas, catgcs, Marshall, biênio, HSA, trauma de tórax e faixa etária) para predizer a ocorrência de óbito. Enquanto que os modelos gerados pelo *data mining* e selecionados para esta comparação, tiveram: experimento 19 com sete variáveis (Naive Bayes – Variáveis: pupilas, catgcs, Marshall, biênio, HSA, trauma de tórax e faixa etária); experimento 17 dez variáveis (Bayes Net – Variáveis: sexo, causa, marshall, HSA, pupilas, trauma associado, catidade2, catglic2, gcsadm); experimento 10 pelo algoritmo Naive Bayes apresentou as mesmas dez variáveis consideradas no experimento 17.

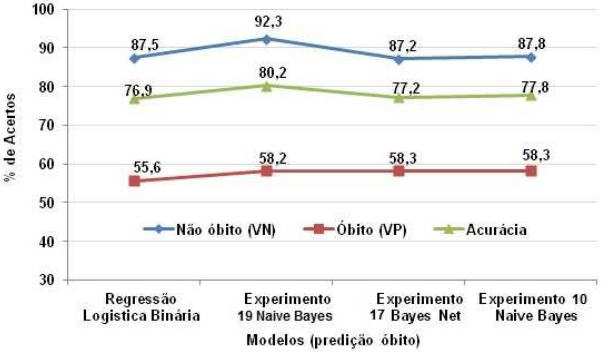
De acordo com os resultados obtidos (tabela 19), verificou-se que a classificação correta para a ocorrência de óbito no modelo de regressão logística foi de 55,6%, enquanto que, para os resultados do *data mining* esta proporção alcançou o máximo de 58,3% nos experimentos 10 e 17. Em relação a classificação correta para a não ocorrência de óbito, no modelo da regressão logística a proporção foi de 87,6% e, esta estimativa mostrou-se semelhante aquelas estimadas pelos modelos dos experimentos 10 e 17. No experimento 19 pelo algoritmo Naive Bayes, com as mesmas variáveis elencadas pela regressão logística, a estimativa para a classificação correta do não óbito alcançou 92,3%, estimativa esta que se mostrou significativamente maior

($p<0,01$) quando comparada aquelas apresentadas pelos demais modelos considerados.

Quanto a acurácia, porcentagem de amostras positivas e negativas classificadas corretamente, o resultado do *data mining* observado no experimento 19 (80,2%) foi significativamente mais elevado ($p<0,05$), quando comparado com as acurácias dos demais modelos (regressão logística: 76,9%; experimento 10: 77,8%; experimento 17: 77,2%).

Desta forma, no que se refere as estimativas das medidas de desempenho verdadeiro negativo ($p<0,01$) e acurácia ($p<0,05$), deve-se acreditar que o modelo de *data mining* do experimento 19 apresentou estimativas mais elevadas quando comparado aos demais modelos (figura 12).

Figura 12 – Medidas de desempenho segundo os modelos.



Fonte: Do autor.

No que se refere as estimativas de sensibilidade e especificidade, observou-se que para a primeira, os resultados para o modelo de *data mining* experimento 19 (72,7%) foi significativamente mais elevado ($p<0,01$) que as estimativas dos demais modelos (regressão logística: 69,4%; experimento 17: 64,6%; experimento 10: 66,7%). Assim, a capacidade do modelo em questão de discriminar dentre os casos de óbito, aqueles que efetivamente vieram a óbito, foi maior quando comparado aos demais modelos.

Considerando as estimativas referente a especificidade, novamente o modelo de *data mining* do experimento 19 (84,2%) apresentou diferença significativa ($p<0,01$), apontando estimativa mais elevada tanto na comparação com a regressão logística (79,9%), quanto com os modelos dos experimentos 17 (79,8%) e 10 (81,6%).

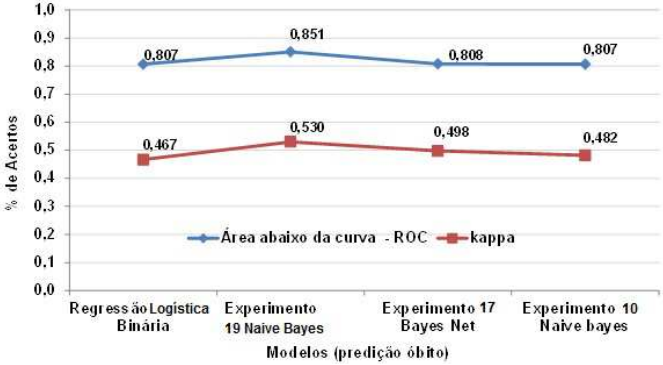
Tabela 19 – Modelos para predição do óbito em TCE Grave.

Testes de validade	Modelos para predizer a ocorrência de óbito			
	Regressão logística binária	*Data mining*		
		Experimento 19 Naive Bayes	Experimento 17 Bayes Net	Experimento 10 Naive bayes
Variáveis elencadas para o modelo final	Pupilas Catgcs Marshall Biênio HSA Tórax Faixa etária	Pupilas Catgcs Marshall Biênio HSA Tórax Faixa etária	Sexo Causa Marshall HSA Pupilas Associado Catidade2 Catglic2 Gcsadm	Sexo Causa Marshall HSA Pupilas Associado Catidade2 Catiglice Gcsadm
Matriz de confusão				
Classificação não óbito - VN	87,6 (84,4-90,5)	92,3 (89,3-95,6)	87,2 (83,9-90,4)	87,8 (83,9-91,4)
Classificação óbito VP	55,6 (52,4-58,7)	58,2 (55,6-61,8)	58,3 (55,1 – 61,3)	58,3 (55,1-61,3)
Acurácia	76,9 (74,2 – 78,9)	80,2 (76,9-85,7)	77,2 (74,3 – 80,4)	77,8 (74,5 – 81,0)
Sensibilidade	69,4 (66,9 – 72,8)	72,7 (69,8-75,4)	64,6 (61,9– 68,0)	66,7 (63,3 – 69,4)
Especificidade	79,9 (76,5 – 83,3)	84,2 (81,6-87,5)	79,8 (75,2 – 83,1)	81,6 (78,6 – 84,8)
Área Curva ROC (IC95)	0,807 (0,783 – 0,831)	0,851 (0,832-0,870)	0,808 (0,784-0,834)	0,807 (0782-0,832)
Kappa	0,467 (0,448 – 0,486)	0,530 (0,519-0,541)	0,498 (0,472– 0,524)	0,482 (0,387 – 0,577)

Fonte: Do autor.

A diferença significativa também se configurou na comparação das medidas de confiabilidade, de forma que para a estimativa referente ao kappa, os resultados apontaram o coeficiente de concordância do modelo de *data mining* do experimento 19 como significativamente maior ($p<0,001$) que os observados nos demais modelos (regressão logística: 0,467; experimento 17: 0,498; experimento 10: 0,482). Esta mesma situação ocorreu na estimativa da área sob a curva ROC, tendo-se o experimento 19 como o mais elevado quando comparado aos modelos de regressão logística (0,807), experimento 17 (0,808), e experimento 10 (0,807) (figura 13).

Figura 13 – Medidas de confiabilidade segundo os modelos.



Fonte: Do autor.

Desta forma, de acordo com as evidências detectadas na análise comparativa entre os modelos de regressão logística e os de *data mining*, conforme apresentados na tabela 19, constata-se nesta pesquisa que o modelo de *data mining* do experimento 19 pelo algoritmo Naive Bayes, composto por sete variáveis, apresentou poder de discriminação (curva ROC), coeficiente de concordância (Kappa), bem como medidas de desempenho significativamente mais representativas. Esta conclusão mostra-se plenamente pertinente, pois estão sendo comparados dois modelos de técnicas diferentes, mas que envolvem exatamente as mesmas variáveis, portanto estão sujeitos a variabilidade semelhante.

7.4 DISCUSSÃO DOS RESULTADOS

O *data mining* tem sido cada vez mais empregado na área biomédica, tendo-se várias pesquisas publicadas em diferentes periódicos internacionais nos últimos anos, auxiliando na identificação de novas hipóteses, na descoberta de relações válidas e principalmente nos modelos prognósticos, a fim de auxiliar a prática clínica.

Métodos de *data mining* têm sido empregados para o diagnóstico e prognóstico de doenças como, por exemplo: Traumatismo Cranioencefálico (RAEESI et al, 2014), Aneurisma (PARAMASIVAM et al, 2014), subtipos de insuficiência cardíaca (AUSTIN et al, 2013), Alzheimer (BRIONES; DINU, 2012), Síndrome da Apnéia do Sono (AL-ANGARI; SAHAKIAN, 2012), Epilepsia (CHAOVALITWONGSE et al, 2011).

No contexto brasileiro, têm-se iniciativas em diferentes áreas de aplicação, no entanto voltado ao traumatismo cranioencefálico e em especial do tipo grave esta é, ao que tudo indica, a primeira pesquisa no Brasil a tratá-lo por meio de métodos de *data mining*.

Nesta pesquisa o modelo para prognóstico do óbito em TCE grave gerado pelo algoritmo Naive Bayes destacou-se em relação aos demais métodos de *data mining* empregados, bem como quando comparado com o modelo de regressão logística desenvolvido por Martins et al (2009) e também projetado nesta pesquisa.

O melhor modelo obtido por meio do Naive Bayes (NB) classificou corretamente o óbito (VP) em 58,2% (IC95%: 55,6-61,8), enquanto a Regressão Logística (Rlog) em 55,6% (IC95%: 52,4-58,7). Na classificação do não óbito (VN): NB 92,3% (IC95%: 89,3-95,6); Rlog 87,6% (IC95%: 84,4-90,05).

A acurácia geral dos modelos foram: NB 80,2% (IC95%: 76,9-85,7); Rlog 76,9% (IC95%: 74,2-78,9). No que se refere a sensibilidade: NB 72,7% (IC95%: 69,8-75,4) e Rlog 69,4% (IC95%: 66,9-72,8), e a especificidade NB 84,2% (IC95%: 81,6-87,5) e Rlog 79,9% (IC95%: 76,5-83,3).

Em relação as medidas de confiabilidade, a área sob a curva ROC para o NB foi de 0,851 (IC95%: 0,832-0,870) e Rlog 0,807 (IC95%: 0,783-0,831), e o coeficiente de concordância kappa para o NB 0,530 (IC95%: 0,519-0,541) e para a Rlog 0,467 (0,448-0,486).

Assim, o modelo gerado por meio do algoritmo Naive Bayes apresentou medidas de desempenho, poder de discriminação e coeficiente de concordância mais representativos do que o modelo de regressão logística binária.

Em relação aos estudos de *data mining* e regressão logística em TCE apresentados nos trabalhos correlatos, considerando-se as pesquisas de Penny e Chesney (2006), Pang et al (2007) e Chesney et al (2009), salienta-se que nos modelos de Penny e Chesney (2006) a rede neural artificial foi a que apresentou melhor precisão na previsão do óbito, sendo a regressão logística binária a que forneceu resultados menos precisos. Conclusão essa referente a regressão logística se corrobora neste estudo, visto que dentre os modelos de *data mining* por meio do classificador bayesiano Naive Bayes teve-se melhores resultados que em relação a regressão logística.

Dentre esses estudos citados anteriormente, Pang et al (2007) foi o que empregou também classificadores bayesianos, verificando-se que a acurácia do modelo desta pesquisa pode ter sido superior aos dele, visto que a acurácia do modelo de Pang et al (2007) variou de 49,79% a

81,49%, porém no seu artigo não especifica dentre os cinco métodos empregados, qual foi exclusivamente a acurácia das redes bayesianas. Pang et al (2007) expõem que as árvores de decisão e a regressão logística em seus modelos mostraram-se mais precisas.

Nesta pesquisa verificou-se que em termos de sensibilidade e especificidade os melhores modelos de *data mining* (experimento 17 e 10) e de regressão logística não se superaram substancialmente, contribuindo igualmente na previsão do óbito, o que corrobora os achados de Chesney et al (2009). No entanto, o experimento 19 de *data mining* mostrou-se mais robusto que os demais, contrariando os resultados de Chesney et al (2009), em que a regressão logística foi o mais robusto.

Considerando-se as pesquisas que descrevem aplicações de *data mining* e TCE, Sut e Simsek (2011) obtiveram em seus modelos para predição da mortalidade acurácia de 91% para o algoritmo CART, enquanto neste estudo o CART se apresentou como melhor modelo somente nos experimentos 5 e 6, com acurácia em ambos de 73,8%, sendo superado pelos demais modelos, como por exemplo, pelos gerados por meio do Naive Bayes, Bayes Net e Adaboost. Além disso, o CART não se encontra neste estudo na caracterização dos dez melhores modelos segundo as medidas de desempenho e confiabilidade.

Também, no estudo de Raeesi et al (2014) dentre os algoritmos empregados o que se destacou foi o C5.0 e o CART com acurácia de 81,4% e 77,8%, respectivamente. Enquanto nesta pesquisa, estes algoritmos foram alguns dos que menos se destacaram em termos de acurácia, bem como em relação as demais medidas de qualidade, apresentando escores médios para a acurácia, Kappa e Curva ROC de 70,4%, 0,245 e 0,666, respectivamente para o CART e de 72,4%, 0,312 e 0,687 para o C4.5 (algoritmo C5.0 é o equivalente ao C4.5 em ferramentas de *data mining* comerciais). Nesta pesquisa na caracterização do melhor modelo em cada experimento, o CART é apresentado como melhor algoritmo nos experimentos 5 e 6, e não aparece na caracterização dos dez melhores modelos. Enquanto, o C4.5 não aparece em nenhuma das caracterizações.

As árvores de decisão foram nesta pesquisa os modelos que apresentaram menor poder de predição, juntamente com o RBF e o kNN.

Alguns estudos comparativos entre diferentes algoritmos de *data mining* como, por exemplo, o de Paramasivam et al (2014) apontam os classificadores bayesianos como bastante precisos em relação a outros métodos de *data mining* como de árvores de decisão e redes neurais

artificiais. Isto se corrobora nesta pesquisa, visto que dentre os sete algoritmos de *data mining* empregados os três melhores modelos foram os de classificadores bayesianos (Naive Bayes e Redes de Crença Bayesiana – Bayes Net). Ainda, além de se destacarem entre os métodos de *data mining*, os classificadores bayesianos em especial o Naive Bayes se destacou, nesta pesquisa e para esta população de estudo, em relação a regressão logística binária, método tradicionalmente empregado nas análises de dados em saúde.

No que se refere as variáveis relacionadas com a mortalidade em TCE Bernal et al (2013) afirmam que a mortalidade associa-se principalmente a GCS, glicemia e pupilas, variáveis estas que também foram identificadas neste estudo, quando realizou-se a seleção dos atributos para o *data mining* de forma automática e manual. Também se confirmaram as variáveis do modelo preditor de Tjahjadi et al (2013) que destaca a classificação de Marshall. Salienta-se, portanto, que as variáveis consideradas nas outras pesquisas como preditoras da mortalidade (BERNAL et al, 2013; MARTINS et al, 2009; MUSHKUDIANI et al, 2008; ROCHA et al 2006; TJAHJADI et al, 2013), também foram elencadas nos melhores modelos desta pesquisa, como é o caso do GCS, pupilas, classificação de Marshall, hemorragia subaracnóide e faixa etária.

A variável biênio elencada no modelo de regressão logística binária de Martins et al (2009) e inclusive nesta pesquisa, apresenta-se como preditora do óbito também nos estudos de Martins, Silva e Coutinho (2003), isto possivelmente foi determinado pelo decréscimo da mortalidade nos anos 2000 em relação a década de 90, período da base de dados constante nos três estudos. Acredita-se que em dados de períodos mais recentes, esta variável não seria elencada como preditora, pois os autores Martins et al (2009) e Martins, Silva e Coutinho (2003) apontam como determinantes para a redução do óbito ao longo dos anos, as melhorias no atendimento pré-hospitalar e na UTI, bem como a formação médica.

Os resultados da pesquisa comprovam as hipóteses, identificando-se diferença entre as medidas de qualidade apresentadas para os métodos de *data mining* e regressão logística em TCE grave, sendo que as medidas de qualidade para os classificadores bayesianos, em especial para o Naive Bayes, apresentou desempenho, poder de discriminação e coeficiente de concordância mais representativo do que o modelo de regressão logística binária. Também se comprovou que existe diferença entre os algoritmos de classificação em *data mining*

para a predição do óbito, destacando-se como melhores, nesta pesquisa, os classificadores bayesianos.

8 CONCLUSÃO

Nesta pesquisa aplicaram-se sete algoritmos variados e recomendados pela literatura da área para a descoberta de conhecimento em dados, no caso de TCE grave, os modelos gerados foram avaliados por medidas de qualidade em *data mining* como de desempenho e de confiabilidade, as quais se complementam na análise comparativa e são fundamentais na identificação de modelos precisos e robustos. Os resultados da regressão logística e do *data mining* para a predição do óbito em TCE grave foram comparados, avaliando-se estes métodos e demonstrando-se por meio das medidas de desempenho e de confiabilidade.

Diferentes métodos para descoberta de conhecimento na base de dados de estudo foram aplicados, como a indução de árvores de decisão, redes neurais artificiais, aprendizado baseado em instâncias, classificadores bayesianos e metaclassificadores. Destacando-se entre estes os classificadores bayesianos que apresentaram os melhores resultados em termos de desempenho e de confiabilidade em aproximadamente 58% dos experimentos realizados.

Os padrões descobertos pela aplicação do *data mining* foram avaliados pelas medidas de qualidade voltadas ao desempenho (verdadeiros positivos, verdadeiros negativos, acurácia, sensibilidade e especificidade) e a confiabilidade (coeficiente de concordância Kappa e área abaixo da curva ROC), a fim de se comparar os métodos de *data mining* aplicados ao conjunto de dados. Concluindo-se que dentre estes critérios a área abaixo da curva ROC foi o estimador de maior confiabilidade na identificação dos modelos de *data mining* com maior poder discriminatório. No caso, o modelo do experimento 19 gerado pelo algoritmo Naive Bayes foi dentre os de *data mining* o que apresentou maior robustez e acurácia, seguido pelo do experimento 17 obtido pelo algoritmo Bayes Net.

Os resultados da regressão logística binária e dos melhores modelos de *data mining* para a predição do óbito em TCE grave foram comparados, identificando-se o que apresentou melhor confiabilidade por meio das medidas de coeficiente de concordância Kappa e curva ROC, considerando-se intervalos de confiança de 95%. O melhor modelo foi o de *data mining*, experimento 19, gerado pelo algoritmo Naive Bayes, concluindo-se que a diferença significativa se configurou nas medidas de confiabilidade, como também nas de sensibilidade, especificidade e acurácia. Dessa forma, neste conjunto de dados o

algoritmo Naive Bayes constitui-se em uma alternativa válida à tradicional regressão logística aplicada na área da saúde.

As hipóteses elencadas foram confirmadas na pesquisa, visto que mediante os resultados obtidos, constatou-se que existe diferença entre as medidas de qualidade apresentadas para os métodos de *data mining* e de regressão logística em dados de TCE grave, como também têm-se diferenças entre os algoritmos de classificação em *data mining* empregados para a predição do óbito em TCE grave.

Dentre os algoritmos de classificação empregados existem diferenças para a predição do óbito em TCE grave, destacando-se entre os dez melhores modelos os algoritmos Naive Bayes, Bayes Net e AdaBoost. Considerando-se todos os experimentos realizados na pesquisa, em mais da metade deles o Naive Bayes foi o algoritmo mais diferenciado visto que conseguiu os melhores resultados em mais da metade dos experimentos (52,6%), seguido pelo AdaBoost (31,6%), CART (10,5%) e Bayes Net (5,3%).

No desenvolvimento da pesquisa constatou-se que existe diferença entre as medidas de qualidade apresentadas para os métodos de *data mining* e regressão logística em dados de TCE grave, sendo as medidas obtidas pelo algoritmo de *data mining* Naive Bayes significativamente mais representativas que as da regressão logística.

As contribuições desta pesquisa consistiram na apresentação de um modelo a partir de um método de *data mining*, para predição do óbito em TCE grave, com melhor capacidade de discriminação; demonstrou-se a metodologia para avaliação dos modelos gerados pelas medidas de desempenho e confiabilidade, as quais se constituem em importante área de pesquisa e possibilitam conclusões referentes ao aspecto computacional, bem como do domínio de aplicação; e a combinação de algoritmos de aprendizagem empregada não foi aplicada em outros trabalhos científicos para o prognóstico na área de TCE grave.

Esta pesquisa explorou as possibilidades de modelos para otimizar a predição do óbito, conseguindo valores de desempenho e de confiabilidade significativamente representativos em relação a um estudo de referência realizado no Brasil na área de TCE grave e com a mesma população de estudo.

Como possibilidade de trabalhos futuros, sugere-se:

a) a ampliação e disponibilização de dados nesta base, inserindo-se atributos, como a pressão intracraniana e biomarcadores, como por exemplo, os marcadores teciduais, de morte e estresse celular. Muitos desses biomarcadores têm sido apontados como promissores para o prognóstico do óbito em

TCE grave. Com isso, valendo-se desses biomarcadores, podem-se identificar modelos de data mining com valores de desempenho e de confiabilidade ainda mais significativos;

b) a realização de validação externa a fim de verificar se é permitida a generalização dos resultados encontrados a outras populações, mostrando-se independentes da amostra, como também para melhoria do conhecimento descoberto;
c) a avaliação da compreensibilidade do melhor modelo de data mining gerado, tanto em termos objetivos quanto subjetivos, visto que a análise de qualidade envolve o desempenho e a confiabilidade, os quais foram avaliados nesta pesquisa, mas também refere-se a questões de compreensibilidade a fim de se viabilizar o entendimento do modelo;
d) a implementação computacional de um comitê de classificadores por meio dos que obtiveram melhores resultados nesta pesquisa.

REFERÊNCIAS

AAKER, D.A.; KUMAR, V.; DAY, G.S. Pesquisa de marketing. São Paulo: Atlas, 2001.

ABREU, M.N.S. et al. Modelos de regressão logística ordinal: aplicação em estudo sobre qualidade de vida. **Cadernos de Saúde Pública**, Rio de Janeiro, v.24, suppl. 4, p. s581-s591, 2008. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-311X2008001600010&lng=en&nrm=iso>. Acesso em: 12 out. 2012.

ABU-HANNA, A.; LUCAS, P.J.F. Prognostic models in medicine AI and statistical approaches. **Special issue of Methods InfMed**, v.40, p. 1-5, 2001. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11310153> . Acesso em: 20 set. 2013.

AKINYEMI, A.E. et al. Instance Based Learning Model for Timing Analysis of Keystrokes to Perform Timing Attacks on the Secure Shell Protocol. **Asian Journal of Computer and Information Systems**, v.1, suppl. 4, p. 114-131, 2013. Disponível em: <http://ajouronline.com/index.php?journal=AJCIS&page=article&op=view&path%5B%5D=577&path%5B%5D=338>. Acesso em: 12 out. 2014.

AL-AIDAROOS, K.M.; BAKAR, A.A.; OTHMAN, Z. Medical data classification with Naive Bayes approach. **Information Technology Journal**, v.11, p. 1166-1174, 2012. Disponível em: <http://scialert.net/qredirect.php?doi=itj.2012.1166.1174&linkid=pdf>. Acesso em: 12 out. 2014.

AL-ANGARI, H.M.; SAHAKIAN, A.V.. Automated recognition of obstructive sleep apnea syndrome using support vector machine classifier. **IEEE Transactions on Information Technology in Biomedicine**, v.16, n.3, p. 463-468, 2012. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/22287247 >. Acesso em: 12 out. 2014.

ANDRADE, A. F. et al. Traumatismo cranioencefálico moderado. In: ASSOCIAÇÃO MÉDICA BRASILEIRA; CONSELHO FEDERAL DE MEDICINA. **Projeto Diretrizes**. 2002, 2 v. Disponível em: < http://www.projetodiretrizes.org.br/projeto_diretrizes/105.pdf >. Acesso em: 02 ago. 2011.

ANDRADE, A.F.; PAIVA; W.S.; AMORIN, R. L. O.; FIGUEIREDO, E.G.; RUSAFA NETO, E.; TEIXEIRA, M. J.. Mecanismos de lesão cerebral no traumatismo cranioencefálico. **Revista da Associação Médica Brasileira**, São Paulo, v.55, n.1, p.75-81, 2009. Disponível em: <http://www.scielo.br/pdf/ramb/v55n1/v55n1a20.pdf>. Acesso em: 27 jul. 2011.

ANDREWS, P. J. D. et al. Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression. **Journal of Neurosurgery**, v.97, n.2, p. 326-336, aug. 2002. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/12186460 >. Acesso em: 03 nov. 2011.

AUSTIN, P. C.; TU, J. V.; HO, J. E.; LEVY, D.; LEE, D. Using methods from the data mining and machine learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. **Journal of Clinical Epidemiology**, v.66, n. 4, p. 398-407, apr. 2013. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/23384592 >. Acesso em: 10 Jun. 2014.

AZEVEDO, F.M.; BRASIL, L. M.; OLIVEIRA, R. C. L. Redes neurais com aplicações em controle em sistemas especialistas. Florianópolis: Bookstore, 2000.

BALESTRERI, M. et al. Predictive value of Glasgow Coma Scale after brain trauma: change in trend over the past ten years. **Journal of Neurology, Neurosurgery & Psychiatry**, v. 75, p. 161-162, 2004. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1757441/pdf/v075p00161.pdf>. Acesso em: 27 jul. 2011.

BARBETTA, P.A.; REIS, M. M.; BORNIA, A. C. **Estatística:** para cursos de engenharia e informática. 2.ed. São Paulo: Atlas, 2009.

BARBOSA, D. C. C.; MACHADO, M. A. Mineração de dados usando o software WizRule em bases de dados de compras de TI. **Revista Eletrônica de Sistemas de Informação**, v.6, n.1, jan-jun. 2007. Disponível em: < http://revistas.facecla.com.br/index.php/reinfo/article/view/184/93>. Acesso em: 20 maio 2011.

BELLAZZI, R.; ZUPAN, B. Predictive data mining in clinical medicine: Current issues and guidelines. **Int J Med Inform**, v.77, suppl. 2, p. 81-97, 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17188928>. Acesso em: 12 132 2014.

BERNAL, E.F.; GIL, F.J.R.; CORRAL, J.C.M.; PRIETO, L.A.M.; ROBLEDO, J.G. Factores prognósticos del traumatismo cranioencefalico grave. Med. Intensiva, v.37, n.5, p.327-332, 2013. Disponível em: < http://www.sciencedirect.com/science/article/pii/S0210569112002069 >. Acesso em: 20 out. 2014.

BERRY, M. J. A.; LINOFF, G. **Data mining techniques:** for marketing, sales and customer relationship management. 2. ed. Indiana: Wiley Publishing, 2004.

BERSON, A.; SMITH, S. J.; THEARLING, K. **Building data mining applications for CRM.** New York: McGraw Hill, 2000.

BISHOP, C. M. **Neural networks for pattern recognition.** New York: Oxford University Press, 1995.

BLALOCK, E.M. A Beginner's Guide to Microarrays. Massashusetts: Kluwer, 2003.

BLANCH, L. et al. The future of intensive care medicine. **Medicina Intensiva**, v.37, n. 2, p. 91-98, 2013. Disponível em:

<http://www.medintensiva.org/en/the-future-intensive-care-medicine/articulo/S0210569112003750/>. Acesso em: 26 set. 2014.

BORS, A. G. Introduction of the Radial Basis Function (RBF) Networks. **Online Symposium for Electronics Engineers (OSEE)**, v.1, DSP algorithms: Multimedia, p. 1-7, feb. 2001. Disponível em: < http://www-users.cs.york.ac.uk/adrian/Papers/Others/OSEE01.pdf>. Acesso em: 17 jul. 2011.

BRAGA, A.C.S. **Curvas ROC: aspectos funcionais e aplicações**. 2000.267 f. Tese (Doutorado em Engenharia de Produção e Sistemas) – Universidade do Minho, Braga, 2000. Disponível em: < https://repositorium.sdum.uminho.pt/retrieve/7/tese_doutACB.pdf>. Acesso em: 05 maio 2015.

BRASIL. Ministério da Saúde. Sistema de Informações Hospitalares do SUS (SIH/SUS). **Morbidade Hospitalar do SUS por Causas Externas.** Disponível em: < http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/eiuf.def>. Acesso em: 15 ago. 2011.

BREIMAN, L.; FRIEDMAN, R. A.; STONE, C. J. Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.

BRIONES, N.; DINU, V. Data mining of high density genomic variatn data for prediction of Alzheimer´s disease risk. **BMC Medical Genetics**, v.13, n.7, 2012. Disponível em: <WWW.biomedcentral.com/1471-2350/13/7>. Acesso em: 14 nov. 2014.

BRITES, R. S. **Verificação de exatidão em classificação de imagens digitais orbitais: efeito de diferentes estratégias de amostragem e avaliação de índices de exatidão.** 1996. 113 f. Tese (Doutorado em Ciência Florestal) – Universidade Federal de Viçosa, Vicosa, 1996. Disponível em: <http://alexandria.cpd.ufv.br:8000/teses/ciencia%20florestal/1996/111128f.pdf>. Acesso em: 14 jun. 2011.

BRUNO, P.; OLDENBURG, C. **Enfermagem em pronto-socorro.** 2.ed. Rio de Janeiro: Senac, 2012.

BRUNS JUNIOR, J.; HAUSER, W. A. The epidemiology of traumatic brain injury: a review. **Epilepsia**, v. 44, p. 2-10, oct. 2003. Disponível em: < http://onlinelibrary.wiley.com/doi/10.1046/j.1528-1157.44.s10.3.x/pdf> . Acesso em: 27 jul. 2011.

BUCHMAN, T.G.; KUBOS, K. L.; SEIDLER, A. J.; SIEGFORTH, M. J. A comparison of statistical and connectionist models for the prediction of chronicity in a surgical intensive unit. **Critical Care Medicine**, v.22, n.5, p.750-762, 1994. Disponível em: < http://journals.lww.com/ccmjournal/Abstract/1994/05000/A_comparison_of_statistical_and_connectionist.8.aspx>. Acesso em: 15 nov. 2010.

BUENO, L.O.; GUIMARÃES, H.P.; LOPES, R.D.; SCHNEIDER, A.P.; LEAL, P.H.R.; SENNA, A.P.R.; JULIANO, Y.; MACHADO, F.R.; AMARAL, J.L.G. Avaliação dos índices prognósticos SOFA e MODS em pacientes após parada cardiorrespiratória em Unidade de Terapia Intensiva. **Revista Brasileira de Terapia Intensiva**, v. 17, n.3, p. 162-164, jul.-set. 2005.Disponível em: < http://www.scielo.br/scielo.php?script=sci_serial&pid=0103-507X&lng=en&nrm=iso>. Acesso em: 22 jul. 2014.

BUSSAB, W.O.; MORETTIN, P. A. Estatística básica. 5.ed. São Paulo: Saraiva, 2003.

BUTCHER, H. et al. Prognostic value of cause of injury in traumatic brain injury: results from the IMPACT study. **Journal of Neurotrauma**, v. 24, n. 2, p. 281-286, feb. 2007. Disponível em: < http://online.liebertpub.com/doi/abs/10.1089/neu.2006.0030>. Acesso em: 14 jul. 2011.

CALLEGARI-JACQUES, S.M. **Bioestatística:** princípios e aplicações. Porto Alegre: Artmed, 2003.

CAMARGO, C. I. A. Traumatismo cranioencefálico. In: TEIXEIRA, E. (Org.). **Terapia Ocupacional na Reabilitação Física.** São Paulo: Roca, 2003. p. 117-125.

CARDOZO JÚNIOR, M.C.L; DA SILVA, R.R. Sepse em pacientes com traumatismo cranioencefálico em unidade de terapia intensiva:

fatores relacionados à maior mortalidade. **Revista Brasileira Terapia Intensiva**, v.26, suppl. 2, p. 148-154, 2014. Disponível em: <http://scielo.br/pdf/rbti/v26n2/0103-507X-rbti-26-02-0148.pdf>. Acesso em: 12 out. 2014.

CARO, D. H. J. Traumatic brain injury care systems: 2020 transformational challenges. **Global Journal of Health Science**, v. 3, n.1, p.19-29, april 2011. Disponível em: < www.ccsenet.org/gjhs>. Acesso em: 27 jul. 2011.

CÉSPEDES, J. M. et al. Factores de pronóstico em los traumatismos craneoencefalicos. **Revista de Neurología**, v. 32, n.4, p. 351-364, 2001. Disponível em: <http://www.portalciencia.net/vdc/pronotce.pdf> Acesso em: 26 jul. 2011.

CHAOVALITWONGSE, W.A.; POTTENGER, R.S.; WANG, S.; FAN, Y.; IASEMIDIS, L.D. Pattern and network-based classification techniques for multichannel medical data signals to improve brain diagnosis. **IEEE Transactions on Systems, Man and Cybernetics**, v.41, n.5, p. 977-988, Sept. 2011. Disponivel em: < http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=5729372&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D5729372>. Acesso em: 04 fev. 2014.

CHARNET, R.; FREIRE, C. A. L.; CHARNET, E.M.R.; BONVINO, H. **Análise de modelos de regressão linear:** com aplicações. 2. ed. Campinas: Unicamp, 2008.

CHESNEY, T. et al. Data mining trauma injury data using C5.0 and logistic regression to determine factors associated with death. **International Journal of Healthcare Technology and Management**, v.10, n. 1/2, p.16-26, 2009. Disponível em: < http://www.inderscience.com/info/inarticle.php?artid=23725>. Acesso em: 22 ago. 2011.

CHESNUT, R.M. et al. Early indicators of prognosis in severe traumatic brain injury. **Journal of Neurotrauma**, v.17, n.6-7, p. 556-627, jun./jul. 2000. Disponível em: < http://online.liebertpub.com/doi/pdfplus/10.1089/neu.2000.17.555> . Acesso em: 14 jul. 2011.

CHOWRIAPPA, P.; DUA, S.; TODOROV, Y. Introduction to machine learning in healthcare informatics. **Machine Learning in Healthcare Informatics**, 2014. Disponível em: < http://link.springer.com/chapter/10.1007/978-3-642-40017-9_1#page-1 >. Acesso em: 12 out. 2014.

CLERMONT, G. et al. Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models. **Critical Care Medicine**, v.29, p. 291-296, 2001. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/11246308 >. Acesso em: 12 out. 2014.

CLICKERING, D.M.; HECKERMAN, D. Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. Techincal report, 1997. Disponível em: < http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.53.6141&rep =rep1&type=pdf>. Acesso em: 03 mar. 2010.

COELHO, L.S.; SANTOS, A.A.P.; COSTA JUNIOR, C.A. Podemos prever a taxa de câmbio brasileira? Evidência empírica utilizando inteligência computacional e modelos econométricos. **Gestão & Produção**, v.15, n.3, São Carlos, Sept./dec. 2008. Disponível em: < http://www.scielo.br/scielo.php?pid=S0104-530X2008000300016&script=sci_arttext >. Acesso em: 26 mar. 2011.

COLANTONIO, A. et al. Trends in hospitalization associated with traumatic brain injury in publicly insured population, 1992-2002. **The Journal of Trauma Injury, Infection, and Critical Care**, v.66, n.1, p. 179-183, 2009. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19131822 >. Acesso em: 26 jul. 2011.

COLE, T. B. Global road safety crisis remedy sought: 1.2 million killed, 50 million injured annually. **The Journal of the American Medical Association**, v. 291, n. 21, p. 2531-2532, jun. 2004. Disponível em: < http://www.jama.ama-assn.org >. Acesso em: 26 jul. 2011.

COLLI, B.O. et al. Características dos pacientes com traumatismo cranioencefálico atendidos no Hospital das Clínicas da Faculdade de

Medicina de Ribeirão Preto. **Arquivos de Neuro-Psiquiatria**, São Paulo, v. 55, n. 1, p. 91-100, 1997. Disponível em: < http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0004-282X1997000100015&lng=en&nrm=iso >. Acesso em: 01 ago. 2011.

CORONADO, V. G. et al. Surveillance for traumatic brain injury – related deaths – United States, 1997-2007. **Centers for Disease Control and Prevention Morbidity and Mortality Weekly Report**, v. 60, n. 5, p. 1-32, may 2011. Disponível em: < http://www.cdc.gov/mmwr/preview/mmwrhtml/ss6005a1.htm>. Acesso em: 14 jul. 2011.

COX, E. et al. **Fuzzy modeling and genetic algorithms for data mining.** San Francisco: Morgan Kaufmann, 2005.

DANTAS, I.E.F.; OLIVEIRA, T.T.; MACHADO NETO, C. D.. Epidemiologia do traumatismo crânio encefálico (TCE) no nordeste no ano de 2012. **REBES**, 2014. Disponível em: < http://www.gvaa.com.br/revista/index.php/REBES/article/viewFile/2573/1985>. Acesso em: 12 out. 2014.

DASU, T.; JOHNSON, T. **Exploratory data mining and data cleaning.** New Jersey: Wiley Publishing, 2003.

DEVORE, J. L. **Probabilidade e estatística:** para engenharia e ciências. São Paulo: Pioneira Thomson Learning, 2006.

DIAMENT, A.; CYPEL, S. **Neurologia infantil.** 3. ed. São Paulo: Atheneu, 1996.

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. **Neural Computation**, v. 10, n. 7, p. 1895-1923, 1998. Disponível em: <http://www.iro.umontreal.ca/~kegl/ift3390/2006_1/Lectures/l06_ApproximateTestsDietterich.pdf>. Acesso em: 12 nov. 2010.

DIJKERS, M.P.; Quality of life after traumatic brain injury: a review of research approaches and findings. **Arch Phys Med Rehabil**, p. 21-35, 2004.

DIMITOGLOU, G.; ADAMS, J. A.; JIM, C. M. Comparison of the C4.5 and Naive Bayes classifier for the prediction of lung cancer survivability. **Journal of Computing**, v. 4, n.8, aug. 2012. Disponível em: < http://www.journalofcomputing.org/volume-4-issue-8-august-2012 >. Acesso em: 26 set. 2012.

DOLCE, G. et al. Clinical signs and early prognosis in vegetative state: a decisional tree, data mining study. **Brain Injury**, v. 22, n. 7-8, p. 617-623, jul. 2008. Disponível em: < http://informahealthcare.com/doi/abs/10.1080/02699050802132503 >. Acesso em: 08 jul. 2011.

DUNCAN, C.C. et al. Evaluation of traumatic brain injury: Brain potentials in diagnosis, function, and prognosis., **International Journal of Psychophysiology** MD, EUA, v.82, p. 24-40, 2011. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21356253 >. Acesso em: 25 out. 2014.

DUTTON, R. P. et al. Trauma mortality in mature trauma systems: are we doing better? An analysis of trauma mortality patterns, 1997-2008. **Journal of Trauma Injury, Infection, and Critical Care**, v. 69, n. 3, p. 620-626, sep. 2010. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/20093983>. Acesso em: 26 jul. 2011.

EISENBERG, H.M. et al. Initial CT findings in 753 patients with severe head injury. A report from the NIH traumatic coma data bank. **Journal of Neurosurgery**, v.73, n.5, p. 688-698, nov. 1990. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/2213158>. Acesso em: 01 ago. 2011.

ESFANDIARI, N. et al. Knowledge Discovery in medicine: current issue and future trend. **Expert Systems with Applications**, v.41, n.9, p. 4434-4463, 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417414000232 >. Acesso em: 12 out. 2014.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n.11, p.27-34, nov. 1996.

Disponível em: < http://dl.acm.org/citation.cfm?doid=240455.240464 >. Acesso em: 01 mar. 2010.

FERREIRA, A.B.H. Mini Dicionário Aurélio da Língua Portuguesa. Curitiba: Positivo, 2010.

FIELD, A. **Descobrindo a estatística usando SPSS**. 2. ed. Porto Alegre: Artmed, 2009.

FINFER, S. R.; COHEN, J. Severe traumatic brain injury. **Resuscitation**, v. 48, n. 1, p. 77-90, jan. 2001. Disponível em: <http://www.sciencedirect.com/science/article/pii/S030095720000321X >. Acesso em: 14 jul. 2011.

FLETCHER, R.H.; FLETCHER, S.W.; FLETCHER, G.S. **Epidemiologia Clínica**: Elementos Essenciais. 5.ed. Porto Alegre: ArtMed, 2014.

FREUND, Y.; SCHAPIRE, R.E. A decision-theoretic generalization of on-line learning and na application to boosting. **Journal of Computer and System Sciences**, v.55, p. 119-139, 1997. Disponível em: <http://www.face-rec.org/algorithms/Boosting-Ensemble/decision-theoretic_generalization.pdf>. Acesso em: 20 out. 2014.

GABBE, B. J.; CAMERON, P. A.; FINCH, C. F. The status of the Glasgow Coma Scale. **Emergency Medicine**, v. 15, n.4, p. 353-360, aug. 2003. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/14631703>. Acesso em: 28 maio 2011.

GAN, G.; MA, C.; WU, J. **Data clustering:** theory, algorithms and applications (ASA-SIAM Series on Statistics and Applied Probability). Philadelphia: SIAM, 2007.

GARCIA, M. C. M.; MARTINS, E. T.; AZEVEDO, F. M. Agrupamento Fuzzy e Regressão Logística Aplicados na Análise de Traumatismo Cranioencefálico Grave. In: CONFERÊNCIA IBERO-AMERICANA DE COMPUTAÇÃO APLICADA 2013/INTERNATIONAL ASSOCIATION FOR DEVELOPMENT OF THE INFORMATION SOCIETY (IADIS), 2013, Porto Alegre/RS.

Anais da Conferência IADIS Ibero-Americana Computação Aplicada 2013. São Leopoldo/RS: IADIS, 2013a. p. 103-110.

_____. Decision tree induction to prediction of prognosis in severe traumatic brain injury of Brazilian patients from Florianopolis city. In: 2013 IEEE 13th INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOENGINEERING (BIBE), 2013, Chania. 13th IEEE International Conference on BioInformatics and BioEngineering, 2013b. p. 1-4.

GENNARELLI, T. A. et al. Influence of the type of intracranial lesion on outcome from severe head injury. **Journal of Neurosurgery**, v. 56, n. 1, p. 26-32, jan. 1982. Disponível em: < http://thejns.org/doi/abs/10.3171/jns.1982.56.1.0026?journalCode=jns >. Acesso em: 03 nov. 2011.

GHAJAR, J. Traumatic brain injury. **The Lancet**, v. 356, p. 923-929, sep. 2000. Disponível em: <http://web.uvic.ca/psyc/skelton/Teaching/General%20Readings/Ghajar%20TBI%202000.pdf >. Acesso em: 14 jul. 2011.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining:** um guia prático. Rio de Janeiro: Elsevier, 2005.

GORELICK, M. H.; YEN, K. The kappa statistic was representative of empirically observed inter-rater agreement for physical findings. **Journal of Clinical Epidemiology**, v. 59, n. 8, p. 859-861, aug. 2006. Disponível em: < http://www.jclinepi.com/article/S0895-4356%2806%2900024-2/abstract>. Acesso em: 02 fev. 2012.

GRAHAM, D.P.; CARDON, A. L. An update on substance use and treatment following traumatic brain injury. **Annals of the New York Academy of Sciences**, v. 1141, p. 148-162, oct. 2008. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18991956>. Acesso em: 28 maio 2011.

GRZYMATA-BUSSE, J. W. et al. Prediction of severe brain damage outcome using two data mining methods. In: CONFERENCE ON HUMAN SYSTEM INTERACTIONS, 2008, Poland. **Anais…**Poland: IEEE, 2008. p. 585-590. Disponível em: <http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4581506&url=

http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4581506>. Acesso em: 9 jun. 2011.

GUERRA, F. A. **Análise de métodos de agrupamento para o treinamento de redes neurais de base radial aplicadas à identificação de sistemas.** 2006. 149 f. Dissertação (Mestrado em Engenharia de Produção e Sistemas) – Pontifícia Universidade Católica do Paraná, Curitiba, 2006. Disponível em: <http://www.produtronica.pucpr.br/publico/ppgeps/conteudo/dissertacoes/pdf/F%C3%A1bioGuerra.pdf>. Acesso em: 21 nov. 2010.

GUERRA, D.S. et al. Fatores associados a hipertensão intracraniana em crianças e adolescentes vítimas de traumatismo crânio-encefálico grave. **J Pediatr**, Rio de Janeiro, v.86, suppl. 1, p. 73-79, 2010. Disponível em: <http://www.jped.com.br/conteudo/10-86-01-73/port.asp>. Acesso em: 20 out. 2014.

GUILLET, F.; HAMILTON, H. J. **Quality measures in data mining.** Chichester: Springer, 2010.

GUTIÉRREZ, V. A. L. **Classificação semi-supervisionada baseada em desacordo por similaridade.** (Mestrado) – Universidade de São Paulo, São Carlos. 2010. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-21062010-142145/en.php>. Acesso em: 14 set. 2014.

HAIR, J.F.; BLACK, B.; BABIN, B.; ANDERSON, R. E.; TATHAM, R. L. Análise Multivariada de Dados. 6.ed. Porto Alegre: Bookman, 2009.

HALL, M.A. **Correlation-based feature selection for machine learning.** 1999. 178p. Thesis (Doctor of Philosophy), Department of Computer Science - University of Waikato, Hamilton, NewZealand, 1999. Disponível em: < https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf>. Acesso em: 16 fev. 2012.

HAN, J.; KAMBER, M.; PEI, J. **Data mining:** concepts and techniques. 3. ed. San Francisco: Morgan Kaufmann, 2011.

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of data mining.** Cambridge: MIT Press, 2001.

HANSEN, E.O. et al. Classificação internacional de funcionalidades de doenças e prognóstico médico em pacientes idosos. **Ver Med**, Minas Gerais, v.21, n. 1, p. 55-60, 2011. Disponível em: <http://rmmg.medicina.ufmg.br/index.php/rmmg/article/viewArticle/342 >. Acesso em: 10 out. 2014.

HAYKIN, S. **Redes neurais:** princípios e práticas. Porto Alegre: Bookman, 2001.

HERRIDGE, M.S. Prognostication and intensive care unit outcome: evolving role of scoring systems. **Clin Chest Med**, v.24, suppl. 4, p. 751-762, 2003. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/14710702 >. Acesso em: 19 nov. 2014.

HICKEY, S. Naive Bayes classification of public health data with greedy feature selection. **Communications of the IMMA**, v.13, suppl. 2, p. 87-98, 2013. Disponível em: < http://connection.ebscohost.com/c/articles/95612798/naive-bayes-classification-public-health-data-greedy-feature-selection>. Acesso em: 15 nov. 2014.

HOSMER, D.W.; LEMESHOW, S. Applied logistic regression. New York: John Wiley & Sons, 2000.

HSIEH, F. Y. Sample size tables for logistic regression. **Statistics in Medicine.** John Wiley, v.8, p. 795-802, 1989.

IMHOF, H. G.; LENZLINGER, P. M. Management of traumatic brain injury. **European Journal of Trauma and Emergency Surgery**, v. 31, n. 4, p. 331-343, aug. 2005. Disponível em: < http://link.springer.com/article/10.1007%2Fs00068-005-2061-5>. Acesso em: 9 jun. 2011.

INTERNACIONAL MISSION FOR PROGNOSIS AND ANALYSIS OF CLINICAL TRIALS IN TRAUMATIC BRAIN INJURY. **Welcome**

to TBI-IMPACT: improving the care for traumatic brain injury. 2011. Disponível em: <http://www.tbi-impact.org/?p=home/news>. Acesso em: 26 set. 2011.

JOOSSE, P. et al. Outcome and prognostic factors of traumatic brain injury: a prospective evaluation in a Jakarta University Hospital. **Journal of Clinical Neuroscience**, v. 16, n. 7, p. 925-928, jul. 2009. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0967586808004165 >. Acesso em: 05 out. 2011.

KANTARDZIC, M. **Data mining:** concepts, models, methods and algorithms. 2. ed. New Jersey: John Wiley & Sons, 2011.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE (IJCAI), 1995. p. 1137-1143. Disponível em: <http://dl.acm.org/citation.cfm?id=1643047 >. Acesso em: 9 nov. 2014.

KOIZUMI, M. S. et al. Morbimortalidade por traumatismo crânio-encefálico no município de São Paulo, 1997. **Arquivos de Neuropsquiatria**, São Paulo, v. 58, n.1, p. 81-89, mar. 2000. Disponível em: < http://www.scielo.br/scielo.php?pid=S0004-282X2000000100013&script=sci_arttext>. Acesso em: 22 nov. 2012.

KRAUS, J. F.; MCARTHUR, D. L. Epidemiology of brain injury. In: EWANS, R. W. (Eds.). **Neurology and Trauma.** 2. ed. New York: Oxford University Press, 2006. p. 3-18.

KUSANO, K.; GLABER, H.C.. Comparison and validation of injury risk classifiers for advanced automated crash norification systems. **Traffic Injury Prevection**, v.15, n.1, p; S126-S133, 2014. Disponível em: <http://dx.doi.org./10.1080/15389588.2014.927577>.Acesso em: 10 dez. 2014.

LANDIS, J. R.; KOCH, G.G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159-174, mar. 1977. Disponível em: < http://www.jstor.org/stable/2529310 >. Acesso em: 02 fev. 2012.

LAROSE, D.T. **Discovering Knowledge in Data**: An Introduction to Data Mining. New Jersey: John Wiley & Sons, 2005.

LATORRE, M.R.D.O. Medidas de risco e regressão logística. In: MASSAD, E. et al (Eds). **Métodos quantitativos em medicina.** Barueri: Manole, 2004. p. 337-350.

LEE, T.T. et al. Application of data mining to the identification of critical factors in patient falls using a web-based reporting system. **International Journal of Medical Informatics**, v. 80, n. 2, p. 141-150, feb. 2011. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21115393>. Acesso em: 08 jul. 2011.

LI, J. et al. Bayes Net classifiers for prediction of renal graft status and survival period. **International Journal of Medicine and Medical Science**, v.1, n. 4, p. 215-221, 2010. Disponível em: <http://www.eng.utoledo.edu/~gserpen/Publications/IJMMS%202010%20Article.pdf>. Acesso em: 12 nov. 2014.

LIN, M.R.; Chiu, W.T.; Chen Y.J.; Wy Y.U.; Huang S.J.; Tsai, M.D.; Longitudinal changes in the health-related quality of life during the first year after traumatic brain injury. **Arch Phys Med Rehabil**, p. 474-480, 2010.

LINGSMA, H. F. et al. Early prognosis in traumatic brain injury: from prophecies to predictions. **The Lancet Neurology**, v. 9, n.5, p. 543-554, may 2010. Disponível em: <http://www.thelancet.com/journals/laneur/article/PIIS1474-4422%2810%2970065-X/abstract>. Acesso em: 14 jul. 2011.

LUCAS, P. Bayesian analysis, pattern analysis, and data mining in health care. **Current opinion in Critical Care**, v.10, n. 5, p. 399-403, 2004. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15385759>. Acesso em: 13 out. 2014.

LUCAS, P.J.F; ABU-HANNA, A. Prognostic methods in medicine. **Artif Intell Med**, v.15, suppl. 2, p. 105-119, 1999. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10082176> . Acesso em: 20 set. 2013.

LUGER, G. F. **Inteligência artificial:** estruturas e estratégias para a solução de problemas complexos. 4. ed. Porto Alegre: Bookmann, 2004.

MAAS, A. I. R. et al. Prognostic value of computerized tomography scan characteristics in traumatic brain injury: results from the IMPACT study. **Journal of Neurotrauma**, v. 24, n. 2, p. 303-314, mar. 2007. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17375995>. Acesso em: 14 jul. 2011.

_____. IMPACT recommendations for improving the design and analysis of clinical trials in moderate to severe traumatic brain injury. **Neurotherapeutics**, v. 7, n.1, p. 127-134, jan. 2010. Disponível em: < http://www.sciencedirect.com/science/article/pii/S1933721309002219 >. Acesso em: 14 jul. 2011.

MAAS, A. I. R.; STOCCHETTI, N.; BULLOCK, R. Moderate and severe traumatic brain injury in adults. **The Lancet Neurology**, v. 7, n. 8, p. 728-741, aug. 2008. Disponível em: <http://www.thelancet.com/journals/laneur/article/PIIS1474-4422%2808%2970164-9/abstract>. Acesso em: 14 jul. 2011.

MARMAROU, A. et al. Prognostic value of the Glasgow coma scale and pupil reactivity in traumatic brain injury assessed pre-hospital and on enrollment: an IMPACT analysis. **Journal of Neurotrauma**, v. 24, n. 2, p. 270-280, feb. 2007. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17375991 >. Acesso em: 14 jul. 2011.

MARSHALL, L. F. et al. A new classification of head injury based on computerized tomography. **Journal of Neurosurgery**, v. 75, n. 1, p. S14-S20, nov. 1991. Disponível em: <http://thejns.org/doi/abs/10.3171/sup.1991.75.1s.0s14?journalCode=sup >. Acesso em: 17 jul. 2011.

MARTINS, E. T. et al. Mortality in severe traumatic brain injury: a multivariated analysis of 748 Brazilian patients from Florianopolis City. **The Journal of Trauma**, v. 67, n. 1, p. 85-90, jul. 2009. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/19590314 >. Acesso em: 19 mar 2011.

MARTINS, E. T.; SILVA, T. S.; COUTINHO, M. Estudo de 596 casos consecutivos de traumatismo craniano grave em Florianópolis – 1994-2001. **Revista Brasileira de Terapia Intensiva**, v. 15, n. 1, p. 15-18, jan./fev. 2003. Disponível em: < http://rbti.org.br/download/artigo_201062917553.pdf>. Acesso em: 26 jul. 2011.

MASINI, M. Perfil epidemiológico do traumatismo crânio-encefálico no Distrito Federal em 1991. **Jornal Brasileiro de Neurocirurgia**, v. 5, n. 2, p. 61-68, maio/ago. 1994. Disponível em: < http://www.abnc.org.br/ed_det.php?edcod=32>. Acesso em: 31 jul. 2011.

MASSAD, E. et al. **Métodos quantitativos em medicina.** Barueri: Manole, 2004.

MASSON, F. Epidemiology of severe cranial injuries. **Annales Françaises d' Anesthèsie et de Rèanimation**, v. 19, n. 4, p. 261-269, apr. 2000. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10836112 >. Acesso em: 26 jul. 2011.

MCNETT, M. A review of the predictive ability of Glasgow coma scale scores in head injured patients. **The Journal of Neuroscience Nursing**, v. 39, n. 2, p. 68-75, apr. 2007. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17477220 >. Acesso em: 31 jul. 2011.

MELO, J. R. T.; SILVA, R. A.; MOREIRA JUNIOR, E. D. Características dos pacientes com trauma cranioencefálico na cidade de Salvador, Bahia, Brasil. **Arquivos de Neuropsquiatria**, São Paulo, v. 62, n. 3-A, p. 711-715, set. 2004. Disponível em: < http://www.scielo.br/pdf/anp/v62n3a/a27v623a.pdf>. Acesso em: 31 jul. 2011.

MENEZES, A.M.B.; SILVA, L.C.C. Noções básicas de epidemiologia. vol. 1. Rio de Janeiro: Revinter, 2001.

MILLER, W. G. The neuropsychology of head injuries. **The neuropsychology handbook**, Behavoural and clinical perspectives, p. 347-375, 1986. New York: Springer.

MITCHELL, T.M.; BLUM, A. Machine Learning. New York: McGraw-Hill, 1997.

MONDELLO, S.; JEROMIN, A.; BUKI, A.; BULLOCK, R.; CZEITER, E.; KOVACS, N. Glial neuronal ratio: a novel index for differentiating injury type in patients with severe traumatic brain injury. **Journal of Neurotrauma**, v. 29, p. 1096-1104, 2012. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3325554/>. Acesso em: 30 set. 2014.

MURRAY, G. D. et al. Multivariable prognostic analysis in traumatic brain injury: results from the IMPACT study. **Journal of Neurotrauma**, v. 24, n. 2, p. 329-337, feb. 2007. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17375997 >. Acesso em: 14 jul. 2011.

MUSHKUDIANI, N. A. et al. Prognostic value of demographic characteristics in traumatic brain injury: results from the IMPACT study. **Journal of Neurotrauma**, v. 24, n. 2, p. 259-269, 2007. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17375990>. Acesso em: 14 jul. 2011.

_____. A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. **Journal of Clinical Epidemiology**, v. 61, n. 4, p. 331-343, april 2008. Disponível em: < http://www.sciencedirect.com/science/article/pii/S0895435607002892 >. Acesso em: 14 jul. 2011.

MYBURG, J. A. et al. Epidemiology and 12-month outcomes from traumatic brain injury in Australia and New Zealand. **The Journal of Trauma Injury, Infection, and Critical Care**, v. 64, n. 4, p. 854-862, april 2008. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18404048 >. Acesso em: 14 jul. 2011.

NASSIF, H. et al. Logical differential prediction Bayes Net improving breast cancer diagnosis for older women. **AMIA Anny Symp**, p. 1330-1339, 2012. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3540455/pdf/amia_2012_symp_1330.pdf>. Acesso em: 12 out. 2014.

NEMES, S.; JONASSON, J.M.; GENELL, A.; STEINECK, G. Bias in odds ratio by logistic regression modeling and saple size. BMC Medical Research Methodology, v.9, n.56, 2009. Disponível em: < http://www.biomedcentral.com/1471-2288/9/56>. Acesso em: 28 maio 2015.

NIJBOER, J. M. M. et al. Two cohorts of severely injured trauma patients, nearly two decades apart: unchanged mortality but improved quality of life despite higher age. **The Journal of Trauma**, v. 63, n. 3, p. 670-675, sep. 2007. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/18073618>. Acesso em: 31 jul. 2011.

OHNO, A. **Detecção de mudanças em problemas de classificação a partir de classificadores sociais** – Programa de Pós-Graduação em Informática**.** 2011. Pontifícia Universidade Católica do Paraná, Curitiba, 2011. Disponível em: http://www.ppgia.pucpr.br/lib/exe/fetch.php?media=dissertacoes:2011arnaldoohno_versaofinal.pdf>. Acesso em: 14 nov. 2014.

OHNO-MACHADO, L.; RESNIC, F.S.; MATHNE, M.E. Prognosis in critical care. **Annual Review of Biomedical Engineering**, v.8, p. 567-599, 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16834567>. Acesso em: 10 out. 2014.

OLIVEIRA, I. B. et al. Traumatismo cranioencefálico: considerações anatomofuncionais e clínicas. **Revista Saúde e Pesquisa**, v. 3, n.1, p. 99-106, jan./abr. 2010. Disponível em: <http://www.cesumar.br/pesquisa/periodicos/index.php/saudpesq/article/viewArticle/1090>. Acesso em: 26 jul. 2011.

OLIVEIRA, C. O.; IKUTA, N.; REGNER, A. Biomarcadores prognósticos no traumatismo crânio-encefálico grave. **Revista**

Brasileira de Terapia Intensiva, São Paulo, v. 20, n. 4, p. 411-421, dec. 2008. Disponível em: < http://www.scielo.br/pdf/rbti/v20n4/v20n4a15.pdf >. Acesso em: 27 jul. 2011.

OLIVEIRA, E.; LAVRADOR, J. P.; SANTO, M. M.; ANTUNES, J. L. Traumatismo Crânio-Encefálico: Abordagem Integrada. **Acta Médica Portuguesa**, v.25, n.3, p. 179-192, 2012. Disponível em: <http://actamedicaportuguesa.com/revista/index.php/amp/article/view/43>. Acesso em: 12 out. 2014.

OLSON, D. L.; DELEN, D. **Advanced data mining techniques.** New York: Springer, 2008.

ONO, J. et al. Outcome prediction in severe head injury: analyses of clinical prognostic factors. **Journal of Clinical Neuroscience**, v. 8, n. 2, p. 120-123, mar. 2001. Disponível em: <http://www.sciencedirect.com/science/article/pii/S096758680090732X>. Acesso em: 31 jul. 2011.

PADILHA, K.G.; SOUSA, R.M.C; SILVA, M.C.M. Disfunções orgânicas de pacientes internados em unidades de terapia intensiva segundo o Logistic Organ Dysfunction System. **Ver Esc Enferm**, p. 1250-1255, 2009.

PANG, B. C. et al. Hybrid outcome prediction model for severe traumatic brain injury. **Journal of Neurotrauma**, v. 24, n.1, p. 136-146, jan. 2007. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17263677>. Acesso em: 01 ago. 2011.

PARAMASIVAM, V.; YEE, T. S.; DHILLON, S. K.; A methodologial review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aortic aneurysm disease. **Biocybernetics and biomedical engineering**, v.34, p. 139-145, 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0208521614000266>. Acesso em: 19 out. 2014.

PENNY, K.; CHESNEY, T. A comparison of data mining methods and logistic regression to determine factors associated with death following

injury. In: ZANI, S. et al (Eds.). **Data analysis, classification and the forward search:** studies in classification, data analysis, and knowledge organization. Heidelberg: Springer, 2006. p. 417-423.

PEREIRA, B. B. Estatística em psiquiatria. **Revista Brasileira de Psiquiatria**, v. 23, n. 3, p. 168-170, set. 2001. Disponível em: < http://www.scielo.br/pdf/rbp/v23n3/a10v23n3.pdf >. Acesso em: 20 out. 2012.

PEREL, P. et al. Systematic review of prognostic models in traumatic brain injury. **BMC Medical Informatics and Decision Making**, v. 6, n. 38, nov. 2006. Disponível em: < http://www.biomedcentral.com/1472-6947/6/38>. Acesso em: 14 jul. 2011.

POLIKAR, R. Ensemble based systems in decision making. **IEEE Circuits and Systems Magazine**, v.6, n.3, p. 21-45, 2006. Disponível em: <http://users.rowan.edu/~polikar/RESEARCH/PUBLICATIONS/csm06.pdf>.Acesso em: 02 ago. 2014.

QUINLAN, J. R. **C 4.5:** programs for machine learning. San Mateo: Morgan Kaufmann, 1993.

RAEESI, A.; EBRAHIMI, S.; NIA, L. I.; ARJI, G.; ASKANI, M. An investigation of data mining techniques of the performance of a decision tree algorithm for predicting causes of traumatic brain injuries in Khatamolanbya Hospital in Zahdan city, 2012 to 2013. **Journal of Health Management & Informatics**, v.1, n.2, april 2014. Disponpivel em: < http://jhmi.sums.ac.ir/index.php/JHMI/article/viewFile/14/9>. Acesso em: 27 nov. 2014.

REIS, W.A.D. Detecção de sinais de trânsito através do método de classificação Adaboost. **UNOPAR Cient.** Exatas Tecnol, v.12, n. 1, p. 27-34, 2013. Disponível em: <http://revistas.unopar.br/index.php/exatas/article/view/1047>. Acesso em: 15 nov. 2014.

REZENDE, S. O. **Sistemas inteligentes:** fundamentos e aplicações. Barueri: Manole, 2005.

RICH, E.; KNIGHT, K.; NAIR, S. B. Artificial Intelligence. 3.ed. India: Tata McGraw-Hill, 2009.

ROCHA, C. M. N. **Traumatismo cranioencefálico:** correlação entre dados demográficos, escala de Glasgow e tomografia computadorizada de crânio com a mortalidade em curto prazo na cidade de Maceió, Alagoas. 2006. 195 f. Tese (Doutorado em Ciências) – Faculdade de Medicina, Universidade de São Paulo, São Paulo, 2006. Disponível em: < http://www.teses.usp.br/teses/disponiveis/5/5151/tde-21062007-145931/pt-br.php>. Acesso em: 31 jul. 2011.

ROSENBERG, A. Recent innovations in intensive care unit risk-prediction models. **Current Opinion in Critical Care,** p. 321-330, 2002. Disponível em: <http://europepmc.org/abstract/MED/12386493>. Acesso em: 15 nov. 2014.

RUSSEL, S.; NORVIG, P. **Inteligência artificial.** 2. ed. Rio de Ja 147 Campus, 2004.

RUTLAND-BROWN, W. et al. Incidence of traumatic brain injury in the United States, 2003. **Journal of Head Trauma Rehabilitation**, v. 21, n. 6, p. 544-548, nov./dec. 2006. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/17122685>. Acesso em: 01 ago. 2011.

RUY, E. L.; ROSA, M. I. Perfil epidemiológico de pacientes com traumatismo crânio encefálico. **Arquivos Catarinenses de Medicina**, v. 40, n. 3, p. 17-20, 2011. Disponível em: < http://www.acm.org.br/revista/pdf/artigos/873.pdf>. Acesso em: 10 nov. 2012.

SAATMAN, K. E. et al. Classification of traumatic brain injury for targeted therapies. **Journal of Neurotrauma**, v. 25, n. 7, p. 719-738, jul. 2008. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2721779/ >. Acesso em: 22 ago. 2011.

SARABANDO, A. C.L. **Um estudo do comportamento de Redes Bayesianas no prognóstico da sobrevivência no cancro da próstata.**

2010. (Mestrado de Informática Médica) – Faculdade de Ciências, Universidade do Porto, 2010.

SASSI, R. J. **Uma arquitetura híbrida para descoberta de conhecimento em bases de dados:** teoria dos rough sets e redes neurais artificiais mapas auto organizáveis. 2006. 169 f. Tese (Doutorado em Sistemas Eletrônicos) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2006. Disponível em: < http://www.teses.usp.br/teses/disponiveis/3/3142/tde-16032007-163930/>. Acesso em: 03 ago. 2011.

SCHETTINO, G. et al. **Paciente crítico:** diagnóstico e tratamento. São Paulo: Manole, 2006.

SCHUSTER, D. P.. Predicting outcome after ICU admission: the art and science of assessing, **Chest**, v.102, n. 6, p. 1861-1870, dec. 1992. Disponível em: < http://journal.publications.chestnet.org/data/Journals/CHEST/21661/1861.pdf >. Acesso em: 14 abr. 2012.

SCHWARTSMANN, C. R. et al. Classificação das fraturas trocantéricas: avaliação da reprodutibilidade da classificação AO. **Revista Brasileira de Ortopedia**, v. 41, n. 7, p. 264-267, jul. 2006. Disponível em: <http://www.scielo.br/scielo.php?script=sci_nlinks&ref=000098&pid=S1413-78522010000400004000010&lng=en >. Acesso em: 02 fev. 2012.

SILVA, A. J. B. M. **Modelos de Inteligência Artificial na Análise da Monitorização de Eventos Clínicos Adversos, Disfunção/Falência de Órgãos e Prognóstico do Doente Crítico.** 2007. 276 p. Tese (Doutorado) – Universidade do Porto, 2007. Disponível em: <repositório-aberto.up.pt/bitstream/.../1/117876_W_4_SIL_001_01_P.pdf>. Acesso em: 20 out. 2014.

SILVA, J.L.C.; MAIA, J.E.E; FONSECA, N.L.S. Identificação de ataques em redes de computadores comitê de classificadores. **XXX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos**, p. 263-276, 2012. Disponível em:

<http://ceresd.facom.ufms.br/sbrc/2012/ST6_1.pdf>. Acesso em: 15 out. 2014.

SILVA, L.M.S.; MARTINS, L.F.; SANTOS, M.C.F.C.; OLIVEIRA, R. M. Índices prognósticos na prática clínica de enfermagem em terapia intensiva: revisão integrativa. **Revista Eletrônica de Enfermagem**, p. 179-90, 2014. Disponível em: <https://www.revistas.ufg.br/index.php/fen/article/view/22830/16457>. Acesso em: 27 nov. 2014.

SIVANANDAM, S. N.; SUMATHI, S. **Introduction to data mining and its applications.** Berlin: Springer, 2006.

SOARES, B. A. C.; SCATENA, J.H.G.; GALVÃO, N. D. Evolução e características da morbidade por acidentes e violências na Grande Cuiabá- Mato Grosso. **Revista Espaço para a Saúde**, v. 9, n. 2, p. 26-38, jun. 2008. Disponível em: < http://www.ccs.uel.br/espacoparasaude/v9n2/Artigo%2059-2008%20_Editado_.pdf >. Acesso em: 12 out. 2012.

SOUSA, R. M. C. Perfil de morbimortalidade relacionado a acidentes e violências no Brasil. In: SOUSA, R. M. C. et al (Org.). **Atuação no trauma:** uma abordagem para a enfermagem. São Paulo: Atheneu, 2009, v. 01, p. 17-28.

STEINER, M.T.A. et al. Abordagem de um problema Médico por meio do Processo de KDD com ênfase á Análise Exploratória dos Dados. **Revista Gestão & Produção**, v.13, n. 2, p. 325-337, 2006. Disponível em: <http://www.scielo.br/pdf/gp/v13n2/31177.pdf>. Acesso em: 20 nov. 2014.

STERNBACH, G. L. The Glasgow coma scale. **Journal of Emergency Medicine**, v. 19, n. 1, p. 67-71, jul. 2000. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10863122>. Acesso em: 29 maio 2011.

STEYERBERG, E. W. et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. **PLoS Medicine**, v. 5, n. 8, p. 1251-1261, aug. 2008. Disponível em: <http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0050165 >. Acesso em: 01 ago. 2011.

STIGLER, S. M. Thomas Bayes´s Bayesian Inference. **Journal of the Royal Statistical Society**, v. 145, n. 2, p. 250-258, 1982. Disponível em: < http://www.jstor.org/stable/2981538>. Acesso em: 12 out. 2012.

STOCCHETTI, N. et al. Inaccurate early assessment of neurological severity in head injury. **Journal of Neurotrauma**, v. 21, n. 9, p. 1131-1140, sep. 2004. Disponível em: < http://online.liebertpub.com/doi/abs/10.1089/neu.2004.21.1131 >. Acesso em: 29 maio 2011.

SUT, N.; SIMSEK, O. Comparison of regression tree data mining methods for prediction of mortality in head injury. **Expert Systems with Applications**, v. 38, n. 12, p. 15534-15539, nov./dec. 2011. Disponível em: < http://www.sciencedirect.com/science/article/pii/S0957417411009018 >. Acesso em: 17 mar. 2012.

TAGLIAFERRI, F. et al. A systematic review of brain injury epidemiology in Europe. **Acta Neurochirurgica**, v. 148, n. 3, p. 255-268, mar. 2006. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/16311842>. Acesso em: 10 dez. 2011.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introdução ao data mining:** mineração de dados. Rio de Janeiro: Ciência Moderna, 2009.

TAN, C.; LI, M.; QIN, X. Study of the feasibility of distinguishing 151 cigarettes of diferentes brands using na Adaboost algorithm and near-infrared spectroscopy. **Anal Bionak Chem**, v.389, n. 2, p. 667-674, 2007. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/17641880>. Acesso em: 25 out. 2014.

TANG, Z. H.; MACLENNAN, J. **Data mining with SQL Server 2005.** Indianapolis: Wiley, 2005.

TANIAR, D. **Data mining and knowledge discovery technologies.** New York: IGI Global, 2008.

TEASDALE, G.; JENNET, B. Assessment of coma and impaired consciousness: a practical scale. **The Lancet**, v. 304, n. 7872, p. 81-84, jul. 1974. Disponível em: < http://www.sciencedirect.com/science/article/pii/S0140673674916390>. Acesso em: 26 set. 2012.

TEASDALE, G. M.; MURRAY, L. Revisiting the Glasgow coma scale and coma score. **Intensive Care Medicine**, v. 26, n. 2, p. 153-154, mar. 2000. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/10784300>. Acesso em: 26 mar. 2011.

TJAHJADI, M. et al. Early mortality predictor of severe traumatic brain injury: A single center study of prognostic variables based on admission characteristics. **The India Journal of Neurotrauma**, v.10, p. 3-8, 2013. Disponível em: <http://www.ijntonline.net/article/S0973-0508(13)00029-4/fulltext.>. Acesso em: 12 out. 2014.

THEODORAKI, E. M. et al. Innovative data mining approaches for outcome prediction of trauma patients. **Journal Biomedical Science and Engineering**, v. 3, p. 791-798, aug. 2010. Disponível em: < www. scirp.org/Journal/PaperInformation.aspx?papaerID=2378 >. Acesso em: 22 ago. 2011.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern recognition.** 3. ed. Orlando: Academic Press, 2006.

TOYAMA, Y. et al. CT for acute stage of closed head injury. **Radiation Medicine**, v. 23, n. 5, p. 309-316, aug. 2005. Disponível em: < http://www.ectsia.org/radac/JC11040602.pdf >. Acesso em: 22 ago. 2011.

TURBAN, E.; LEIDNER, D.; MCLEAN, E.; WETHERE, J. **Tecnologia da informação para Gestão:** transformando os negócios na economia digital. 6.ed. Porto Alegre: Bookman, 2008.

TURCATO, C.; PEREIRA, S. W.; GHIZONI, M. F. Hemorragia subaracnóide. **Arquivos Catarinenses de Medicina**, v. 35, n. 2, p. 78-84, abr./jun. 2006. Disponível em:

< http://www.acm.org.br/revista/pdf/artigos/373.pdf >. Acesso em: 26 set. 2011.

VICTORA, C.G.; HUTTLY, S.R.; FUCHS, S.C.; OLINTO, M.T.A. The role of conceptual frameworks in Epidemiological Analysis: a hierarchical approach. **International Journal of Epidemiology**, v.26, p.224-227, 1997.

VIEIRA, R.C.A. et al. Qualidade de vida das vítimas de Trauma Cranioencefálico seis meses após o trauma. **Rev. Latino-Am. Enfermagem**, v.21, n. 4, p. 1-8, 2013. Disponível em: <http://www.scielo.br/pdf/rlae/v21n4/pt_0104-1169-rlae-21-04-0868.pdf>. Acesso em: 12 out. 2014.

WALPOLE, R. E. et al. **Probabilidade e estatística para engenharia e ciências.** 8. ed. São Paulo: Pearson Prentice Hall, 2009.

WANG, J. T. L. et al. **Data mining in bioinformatics.** London: Springer-Verlag, 2005.

WHYTE, J. et al. Reabilitação do paciente com traumatismo cranioencefálico. In: CURRIE, D.; DELISA, J.; MARTIN, G.**Tratado de medicina de reabilitação.** 3. ed. Barueri: Manole, 2002. p. 1255-1298.

WINDELER, J. Prognosis What does the clinician associate with this notion? **Stat med**, v.19, p. 425-430, 2000. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/10694727>. Acesso em: 20 out. 2014.

WINN, H.R.; BULLOCK, M.; HOVDA, D.; SCHOUTEN, J.; MAAS, A. Youmans Neurological Surgery. **Chapter 323 – Epidemiology Traumatic Brain**, v.4, p. 3270-3275, 2011.

WITTEN, I. H.; FRANK, E. **Data mining practical machine learning tools and techniques.** 2. ed. San Francisco: Morgan Kaufmann, 2005.

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

WU, X.; KUMAR, V. **The top ten algorithms in data mining.** New York: Chapman and Hall, 2009.

WU, Xindong et al. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, n. 1, p.1-37, 2008. Disponível em: < http://www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf>. Acesso em: 15 out. 2010.

WU, Xing et al. Epidemiology of traumatic brain injury in Easter China, 2004: a prospective large case study. **The Journal of Trauma,** v. 64, n. 5, p. 1313-1319, may 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/18469656>. Acesso em: 21 jul. 2011.

YE, N. **The handbook of data mining.** New Jersey: Lawrence Erlbaum Associates, 2003.

YOO, I. et al. Data mining in healthcare and biomedicine: a survey of the literature. **Journal of Medical Systems**, v. 36, n. 4, p.2431-2448, aug. 2012. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/21537851 >. Acesso em: 07 nov. 2012.