A Medical Data Mining Application Covering Patients Over 90 Years Old by Using the CART Algorithm

Firat Ismailoglu

A Dissertation submitted to

the School of Computing Sciences of the University of East Anglia

in partial fulfilment of the requirements for the degree of Master of Science

Abstract

It is certain that the elderly population and life expectancy have been increasing rapidly in the world as well as in the UK for the last few decades and according to several independent organisations, say U.S. Census Bureau, that increase will be permanent. Thus, issues around the elderly people will gain increasing importance. Data mining, nontrivial process of identifying novel patterns in the data, has been used to medical databases for several years and has yield satisfactory and useful results. At this point, we have performed a data mining application by using Oldest Old Dataset which consists of the records of 393 patients over 90 years old. For this application we have taken the advantage of the CART algorithm. In parallel with this medical data mining task, we have discussed the issues around medical data mining in terms of the challenges and potentials it has. Finally, at the end of our data mining application, we have found several patterns about a patient's mortality rate and death risk which have been regarded novel and interesting by our medical expert.

Acknowledgements

I would like to show my gratitude to my supervisor, Professor Vic Rayward-Smith (School of Computing Sciences, University of East Anglia), for his guidance and support throughout my thesis.

I would like to thank the medical expert of this study, Dr Phyo Myint (Norfolk & Norwich University Hospital) who has allowed me to take the advantage of his broad knowledge.

I would also like to thank Dr Patrick Musonda (Norfolk & Norwich University Hospital) for his help.

Last but not least, I would like to thank my father Ali Riza Ismailoglu, my mother Zekiye Ismailoglu and my brother Dr Murat Ismailoglu for their endless and motiveless love.

Contents

Chapter 1

Background and Introduction

1.1 Introduction

Without doubt that one of the most urgent issues of the last few decades has been the regular increase in the elderly population and life expectancy in the world. Indeed, in the year 2000, the number of people over 90 in the world was around 8 million, whereas, within only ten years, that number has increased by 50% and nowadays it has reached about 12 million. Furthermore, according to the U.S Census Bureau, by 2035, the number of people over 90 is projected to be more than 40 million, so it will be five times larger in 25 years time. In contrast, the total world population is projected to increase by about only 30% in that duration (1)**.** This suggests that age distribution will change significantly over the next decades.

The other significant consideration regarding the ageing population is that this is observed in the mostly developed countries and especially in Europe. In fact, in 2006, the wallchart of Population Reference Bureau (which is a non-profit organisation established in 1929 in USA to inform people and raise awareness of them about population health and the environment) showed that nearly all the countries of the world with an elderly population are in Europe (See Figure 1.1).

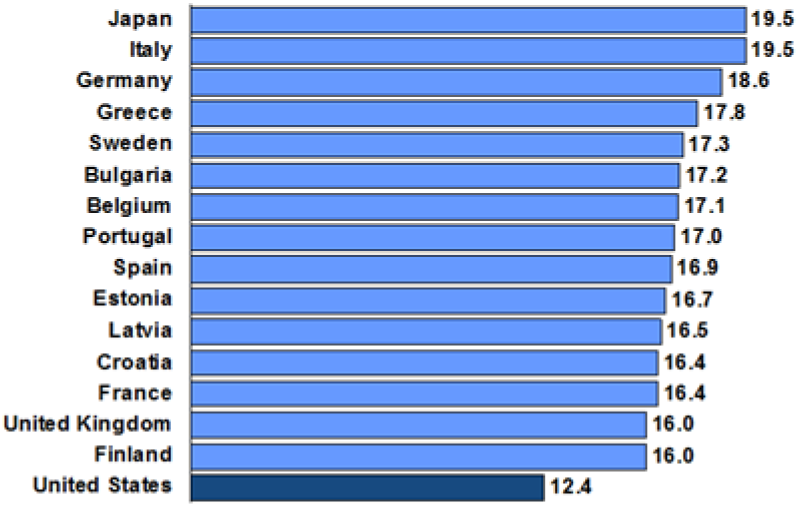


Figure 1.1 Percentage of people over 65 by Country **(2)**

As we see from the figure above, the UK has the fourteenth oldest population in the world. In the year 2008, 16 percent of people were over 65 and 2 percent of them were over 85. Based on The Office for National Statistic (hereafter ONS) estimation, the percentage of people over 85 is projected to grow to 5 percent within the next 25 years. Furthermore, according to ONS, the fastest population increase has been in the number of those aged 85 and over, the "oldest old". The figure below gives more details about the current and possible changes to the elderly population in the UK.

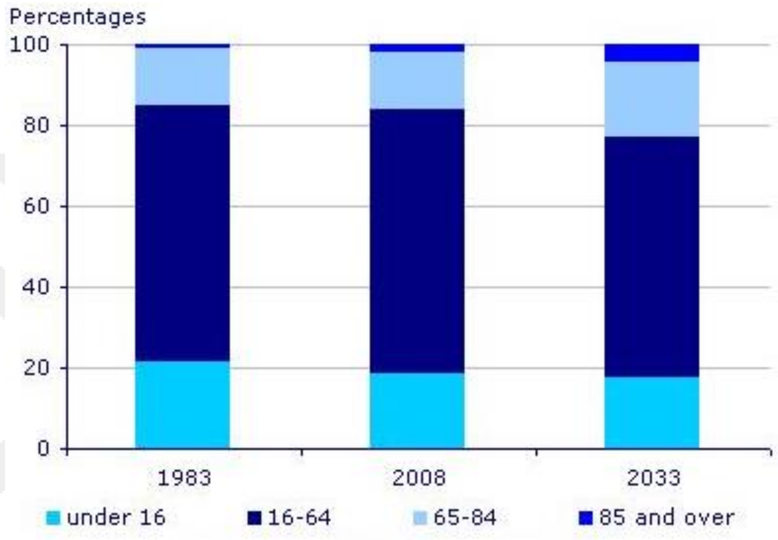


Figure 1.2 Percentage by age, UK, 1983, 2008, 2033 **(3)**

The increase in the number of older people has largely been provided by increases in life expectancy. In the UK, the life expectancy as a whole is assumed to rise from 77.8 years in 2008 to 83.1 years in 2033, while for females, it will increase from 81.9 years to 86.9 years, according to the National Statistics. More specifically, at this point it is useful to look at the National Statistics figures for Norwich and Aberdeen since we will concentrate on these cities later.

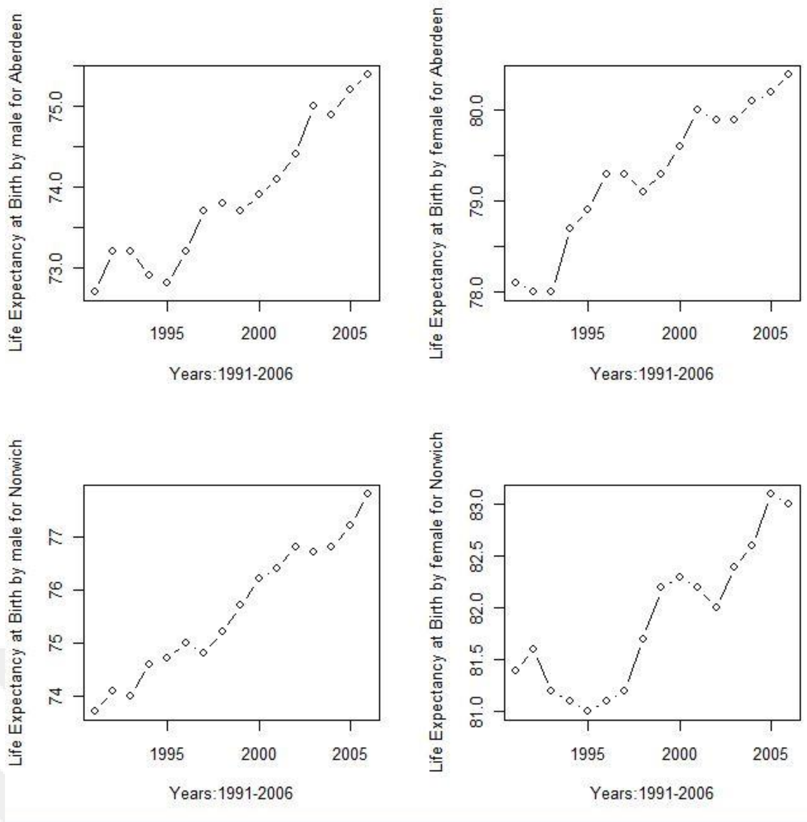


Figure 1.3 Life Expectancy at birth by gender, for Aberdeen and Norwich **(4)**

In the statistical bulletin in 2009, the ONS drew attention o the fact that although there was a clear increase in the life expectancy, healthy life expectancy, which is the expected years of life in good or reasonably good health, and disability-free life expectancy, which is the expected years of life without long-term illness or disability, might not match that increase. However both these figures increased between 1981 and 2006. Another remarkable fact about the healthy life expectancy is that the gap in healthy life expectancy between men and women is smaller than for the total life expectancy. For example, in 2004, the gap was 2.3 years for healthy life expectancy, whereas it was 4.4 years for total life expectancy. See (http://www.bgs.org.uk/index.php?option=com_content&view=article&id=473:lifeexpectancy&catid=89:intelligence&Itemid=208).

All the discussion made so far points out that the old population and issues around the elderly people will gain increasing importance in the very near future for the world as well as for the UK. In this context, the health of elderly people and their health expenditure are important and will gain increasing importance. In addition, when we analyse the NHS expenditure in terms of demography, we see that NHS spending per capita is the most for elderly people (>85) (See Figure 1.4).

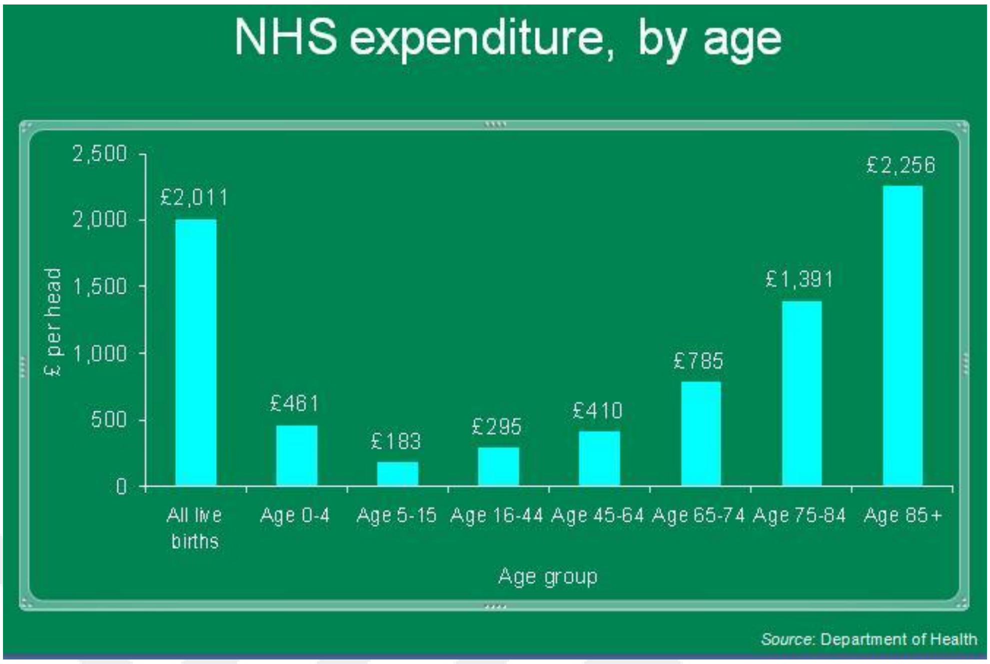


Figure 1.4 NHS expenditure, by age

This is because the elderly tend to have much greater need for health and social services than the young. For instance, nearly two thirds of general and acute hospital beds are used by people over 65 (5), and in 2009, more than £20 billion were allocated to those over 85 in England. As a result of this, even a very small increase in the elderly population would bring a remarkable burden on the UK economy, so any attempt that intends to reduce or at least fix that budget will be financially important.

Another issue regarding the elder population in the UK, which is our main interest in this study, is their living arrangements. In this sense, older people are more likely to live alone than the young. For example, in Scotland, 62% of people aged over 85 are projected to be living alone by 2033. Moreover, that percentage is 77 for women but only 38 for men (6**).** More generally, 3 in 5 women aged 75 and above live alone, whereas slightly less than one third of men in the same age group live alone in the UK (7). To understand better the effects of ageing on the living arrangements, see Figure 1.5.

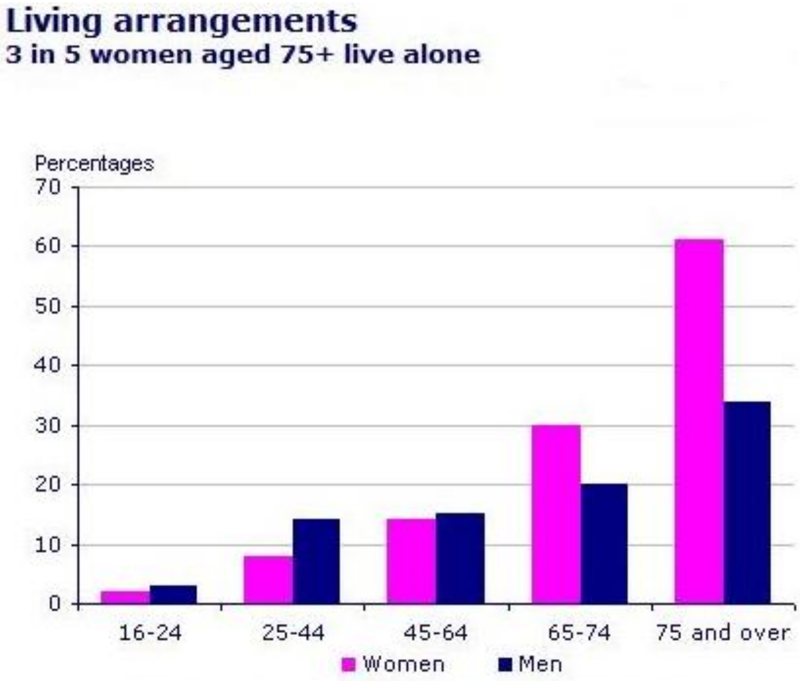


Figure 1.5 People living alone by sex and age, Great Britain, 2007 (8)

There are two underlying reasons explaining the gender differences in the living arrangements. Those are, firstly the difference between total life expectancy of women and men and, secondly, women’s tendency to marry men older than themselves (9).

Despite the fact that more than half of the most aged live alone, these people tend to struggle with chronic illnesses and disabilities. As a matter of fact, almost half of the disabled people are aged 65 or older and they mostly suffer from problems relating to movement and to vision and hearing (10)**.** Further, sensory impairments become increasingly common; since around 80% of those over 60 have a visual impairment and 75% of them have a hearing impairment. Consequently, these disabilities can result in reduction in the ability of older people to look after themselves, so they need personal care (5).

1.2 Statement of Problem

There is a clear regular increase in the most aged (>=90 years) in the UK and many of them suffer from chronic illnesses and disabilities. However, the older people are sometimes seen as a drain on scarce resource within the NHS and funding allocated for the older people is not often seen as a priority. Furthermore, older patients in hospital are generally considered to have poor prognoses and this is perhaps a misconception. Additionally, insufficient research has addressed the epidemiology of the most aged who are admitted to hospitals with an acute illness.

1.3 The aim and The Objectives

The aim of this project is to provide a novel approach to estimate the mortality rate of patients over 90 years old and admitted to the hospital due to various medical causes.

The initial objective of this project is to compute the probability of death which can occur any time in the treatment of the most aged people and which can occur after a week or longer of a patient admission to the hospital. We expect that such index will provide a unique measure that will be determined how urgent and important the patient's situation is. Furthermore, it may yield an opinion about whether the patient requires a critical care or more specifically, it can be used when determining if the patient admitted to intensive care unit (ICU). Secondly, we are expecting to find the most important predictors to classification of death.

1.4 Significance of the Study

This study can be constructive for data mining researches, practitioners who study on gerontology and geriatrics or bioinformatics, people trying to develop new models aiming to optimize hospital resources or working in insurance and risk analysis sectors.

One of the main differences that distinguishes the existing studies in medical data mining and this project is that the vast majority of the previous ones have focused on a particular illness, such as breast cancer, prostate cancer or some heart problems, whereas in this study, we deal with various kind of illness that the most aged suffer from. That is, our prospective findings, for instance probability of death after a long stay in the hospital, are not for certain specified illness. However, we delimit this project with elderly patients.

Finally, it is clear that to be able to predict the risk of death of ones in such an important age group will be a good acquisition in terms of optimising health spending.

Chapter 2

Medical Informatics and Data Mining

2.1 Introduction

The term medical informatics (or health informatics) refers to the scientific discipline that is an intersection of computer science, data analysis and medicine (11). Along with extensive amounts of data gathered in medical database, it has become difficult to extract useful information from those databases for decision support and manual data analysis has become inadequate. In order to satisfy this need, medical informatics has used data mining, machine learning, pattern recognition and visualisation (12). However, in this study we will emphasise data mining applications used in medical databases, since it is our primary interest.

In general, data mining is a step in knowledge discovery that is known as a non-trivial extraction of previously unknown and potentially useful information from data and offers methodological and technical solutions to manage medical data analysis and construction of prediction models (13). Data mining algorithms are able to learn from past instances in clinical data and model non-linear as well as linear relationships between the independent and dependent variables. The resulting model generated by a data mining algorithm can usually provide a high-quality diagnostic opinion to the decision makers. Thus, data mining techniques are widely used in medical research.

In general, medical data mining tasks can be considered under two main groups, which are description and prediction. Finding interesting patterns and clusters is the goal of descriptive data mining in medicine, whereas the predictive one aims to construct a predictive model that makes reliable predictions and helps physicians improve their prognosis and diagnosis. However, the difference between descriptive and predictive data mining is not always clear-cut. That is, while interesting patterns found with descriptive data mining techniques can be used for predictive purposes, a clear predictive model, such as a decision tree, is able to highlight interesting patterns, so has descriptive qualities (14). In contrast, the major difference is that prediction requires the data to include a response variable which can be categorical or numerical, and according to the variable type, the predictive task may be building classification models or regression models, if the type is categorical or numerical, respectively (13). To conclude, in all cases the result of both descriptive and predictive data mining applications should bring previously unseen knowledge to end-users.

The rest of this chapter is organised as follows. Section 2.1 covers the techniques used in medical data mining, and then section 2.2 provides potential of clinical data mining. Finally, section 2.3 covers challenges in medical data mining.

2.2 The Techniques used in medical data mining

Since one of the most important application areas of data mining is health care industry (15), a wide range of data mining techniques which come in various flavours are used in such area (13). However, we are going to cover here some of the most commonly used ones.

2.2.1 The Naive Bayesian Classifier

The first steps on the way to medical data mining have been the Bayesian classifiers. Among the other techniques used in medical data mining, the naive Bayesian classifier is the perhaps the simplest and oldest one which has been used since the early days. Its performance is often not less than the other more sophisticated approaches (16). The naive Bayesian classifier takes advantage of Bayes’ Theorem, which is used for calculating conditional probabilities, to estimate probabilities of individual variable values and then thanks to these probabilities, the data miners are able to classify the new entities. In addition, the naive Bayesian classifier is used as a benchmark algorithm and is often tried before any other advanced method (16).

The naive Bayesian classifier (also called simple Bayes) is based on the assumption that each predictor is conditionally independent of the other. Further, such an assumption allows the data miner to build a fast, highly scalable model. On the other hand, in the case where there is a high dependency between attributes, it is very likely to lead to loss of accuracy. To overcome such a problem, Konenko (1991) advanced the semi naive Bayesian classifier and Pazzani (1996) developed backward sequential elimination and joining algorithms which are able to detect dependencies among attributes (17).

2.2.2 Decision Trees

As the name suggests, a decision tree is a flowchart-like tree structure, where each nonleaf node symbolises a test on an attribute, whereas each leaf node holds a class label and each branch denotes an outcome of the test (18). Decision trees use recursive data partitioning and represents set of a decisions. These decisions generate rules for the classification of a dataset. Moreover, they are able to discover unexpected relationships, use categorical or continuous data and handle missing data. On the other hand, decision trees suffer from finding some weak relationships.

The data mining community has developed tailored decision tree learners for medical databases. For example a decision tree learner ASSISTANT was developed purposely to handle the particular characteristics of medical datasets, binarisation of continuous attributes and to tolerate incompletely specified training examples by Konenko et al in 1986 (12). Additionally, ASSISTANT-I, ASSISTANT—R and ASSISTANT-R2 are reimplementations of ASSISTANT which is for top down induction of decision trees, unlike the other learners, ASSTANT-R2 generates one decision tree for each diagnosis (19) (12).

As a solid example, in 2006, Lu et al. worked on classifying 8,259 elderly patients who were 65 or older with impaired mobility nursing by using decision trees, and then he achieved sensitivity around 69% (20). This study is important in terms of being the first published study which used a data mining method to classify elderly patients with impaired mobility.

According to the pool at KDnuggets (www.kdnuggets.com/polls/2006/data_mining_methods/.htm , April 2006), decision trees are the most frequently used data mining algorithms and it is possible to find several reasons explaining such a situation. For instance, decision tree classifiers do not require any background knowledge, they can deal with high dimensional data, its classifiers have first-rate accuracy and decision tree induction algorithms can be used in many application areas, such as medicine and financial analysis (18).

Finally, the most popular decision trees areID3, C4.5, C5 and CART (21).

2.2.3 Artificial Neural Networks

Basically, artificial neural networks (hereafter ANN) is a non-linear predictive model consisting of an interconnected group of nodes that simulate neurons. According to Bellazzi et al.(2008),ANNs are the most popular artificial intelligence-based data modelling algorithms used in clinical data mining due to their good predictive performance (13). Indeed, Delen stated that since ANNs possess high-speed calculation memory, compression and filtering, they are capability of solving various complex classification and prediction problems (22). For instance, predicting the outcome of liver disease, automated electrocardiographic (ECG) interpretation and image analyses in malignant melanoma and breast cancer (23).

ANNs are more suitable for medical diagnosis, when they are compared with more classical approaches, say, rule based systems or system based on probabilistic methods (24). Nevertheless, ANNs may suffer from difficulties with generalisation and especially with interpretability (23). Furthermore, ANNs often need long training times and tend to suffer from overfitting. Fortunately, by using the cross validation method to determine the appropriate stopping time for the training of ANN, it is possible to avoid from such problem (24).

2.2.4 Support Vector Machines (SVM)

Support Vector Machines (hereafter SVM) is a relatively young supervised learning system and is used to solve problems in nonlinear classification, density estimation, and function estimation. A support vector maps the data to higher-dimensional space using a kernel function, which is a define function that reflects the similarity between an input and the set of support vectors, then finds an optimal separating hyperplane or set of hyperplanes to perform a linear discrimination in that space by minimizing the classification error and maximizing the geometric margin between the classes (25). See Figure 2.1.

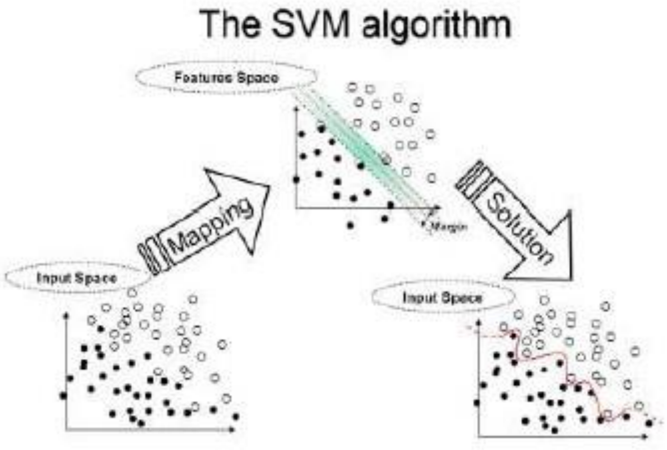


Figure 2.1 The SVM algorithm

SVMs have a vast number of application areas including pattern recognition problems such as handwriting recognition, object recognition, speaker identification, face detection and text categorisation (26). Moreover, SVMs have also been successfully applied to a wide range of biological applications. One of the most popular biomedical applications of support vector machines is automatic classification of microarray gene expression profiles. That is, SVMs are capable of examining the gene expression profile obtained from a tumour sample or from peripheral fluid and arrives at a diagnosis or prognosis (27). However, the most common application of SVMs in medical databases is image classification and this respect, SVMs are claimed to be the best classification algorithm in terms of predictive accuracy (28).

Although SVMs have an outstanding generalisation performance on numerous problems, there are some drawbacks related to SVMs. The major disadvantages of SVMs are their problems with human interpretation. In fact, they struggle with high algorithmic complexity because of the required quadratic programming. In addition, both in training and testing phases, there is a limitation in terms of speed and size.

2.2.5 The k-nearest neighbours (k-NN)

The k-nearest neighbours is a further type of classification that finds a group of k-objects in the training set which are closest to the test object and classifies the test object based on their prevailing class . The closeness mentioned here is defined in terms of distance metric, such as Euclidian, Minkowski and Manhattan. The selected k training objects become the “k nearest neighbours” of the unknown object. Thereafter, the unknown object is assigned to the most common class among its k nearest neighbours. In general, determining k is experimentally, but k is usually a small positive integer and the larger the number of training objects, the larger of the value of k will be (18). Moreover, it should be noted, unlike the other top classification methods, k-NN does not construct a classifier in advance; it is, therefore, suitable for data streams, and it does not train and test a residual classifier first and use it on new samples. Hence k-NN is a preferred choice, especially when simplicity and accuracy are the predominant issues (29). In contrast, a k-NN classifier is inadequate compared to Bayes error and is sensitive to redundant and noisy features. Further, adding more features may increase the error rate for a k-NN classifier (30).

In 2010, Liu et al. developed a fall detection system with the help of a k-NN classifier to detect a fall incident which is the leading cause of the fatal and nonfatal injuries for adults aged 65 and over; the system achieved a correct rate of 84.44% on fall detection and lying down event detection (31). In addition, k-NN has been used in many applications including recognition of handwriting, satellite image and EKG patterns.

2.3 Potential of Medical Data Mining

As is well known, since early 1980’s, along with the rapid increase in data storage capability of computers, almost all types of records have been held in electronic platforms. Concurrently, hospital information systems (HIS) have become able to store large amounts of laboratory examinations, patient records and so on which has led to the more structured and more standardised hospital databases. Hence it is highly expected that data mining methods will find interesting patterns from the medical databases, because reuse of the data stored is significant for medical research and human beings cannot deal with such a huge amount data (32).

Although medical data mining is relatively new area compared with the other data mining applications such as ones in banking or retail industry, it has made adequate progress. . Indeed, according to a poll carried out by kdnuggets.com, in the year 2008, 9.3% of data mining studies were carried out in health care industry (15)**.** Especially, along with the well equipped hospitals becoming widespread, using the diagnostic tools such as ultra- sound and computed tomography (CT) has increased, so the language used in medicine has become more common which has allowed the data miners to get results which are well accepted.

Yet another issue regarding potential of medical data mining is to the use of the classification rules. That is, classification rules extracted from the data are generally useful for medical problems that have been applied particularly in the area of medical diagnosis (33). Furthermore, every single path in decision trees, which is the one of the most popular techniques of data mining, can be regarded as a decision rule (13). Thus, decision trees are able to produce set of rules for a particular medical area, which leads to acquire novel insight for medical databases.

Overall, it is expected that along with the improved technology, more diagnostic tools will be developed to allow doctors to diagnose more precisely. Fortunately, on the other hand, thanks to the increase in the data storage capability of computers, we will be able to store such data in a structured way. Thus, human being will need to simple yet successful data mining tools, such as decision trees, in the future as well, like it has happened in the past.

2.4 Challenges in Medical Data Mining

We have considered the challenges appearing in medical data mining under two main topics.

2.4.1 Issues around Data Quality

2.4.1.1 Heterogeneity

As we have mentioned before, along with good progress in hospital information systems, increase in computer's data storage ability and reduction in computing costs, the abundance of medical data have appeared. However, the data are often in various databases. Indeed, it is common that medical data are held in a paper-based record as an unstructured free text, and medical procedures often employ imaging as a preferred diagnostic tool. Further, in order to monitor a patient's organ, different kinds of imaging techniques like magnetic resonance imaging (MRI), *single photon emission computed tomography* (SPECT), positron emission tomography (PET), and a collection of electrocardiogram (ECG) or electroencephalograph (EEG) signals are used and the other clinical information as well as the practitioner's interpretation often accompany to these images (34).

It is common that the same kind of information is stored in totally different formats or the same operations are performed with different data models due to lack of standardisation (35). Hence, medical databases are often heterogeneous and the process of determining a diagnosis has several components. As a result, many researchers suffer from the challenge of information integration from heterogeneous data sources and this heterogeneity seriously affects carrying out productive research (36). Thus, there is a need to develop an integrated systems for medical databases and developing such a system is not only necessary for the data mining community, but also for clinicians, since the system would allow clinicians to easily review data collected from administration, laboratory, pharmacy and other departments of a hospital for a single patient (37).

2.4.1.2 Personalised Data and Missing Values

Another challenge regarding medical data is that such data often contain a clinician's interpretation; it therefore tends to be subjective. To make matter worse, medicine goes from "one size fits all" to "personalised medicine", where physicians make treatment and diagnostic decisions which are exactly tailored to the individual's patient's characteristics and his or her specific disease. Advances in medical imaging and clinical data repositories (CBRs), gene sequencing, molecular imaging technology can be listed as the reasons of such personalisation. Hence the ratio of interpretation in a medical data is increasing, so data mining systems must be personalised with respect to what is known about each patient (38).

As with other types of databases, medical databases suffer from missing data, which often arises through incomplete data being recorded or human error in recording or transcription (39). Problems with missing data can be minimized by removing variables and/or cases that have a large number of missing values or replacing missing data with their associated statistical descriptors, such as the mean value for a variable. However, the first approach may lead to bias, since cases with a high proportion of missing data may be associated with the outcome of interest, and the latter may introduce bias into the data (40).

2.4.1.3 Difficulties Originated from the Textual Based Content of Medical Data

There are some significant obstacles originating from the textual based content of medical data which can limit the operations with medical data. That is, the conceptual structure of medicine generally consists of word descriptions with quite formal constraints on the vocabulary, so medical data cannot be often put into equations, formulas, equations, and models that moderately reflect the relationships among the data (41). Furthermore, medicine has no comparable formal structure into which a data miner can organize information, using perhaps clustering or regression models; it is therefore poorly characterised mathematically (41).

In medicine, terms, such as disease names or complaints, are entitled according to several criteria, say, anatomic location, frequency (e.g. chronic, acute) or type of the bacteria causing the disease, and that usually leads to a exponentially increase in the type of a disease, diagnosis and so on. For instance, if we take two kinds of ulcer, chronic and acute, and 200 body locations, we get 400 types of ulcers, if we add ten types of bacteria, we have 4000; etc. and this is just for ulcers (42). Thus, it is sometimes difficult to find enough cases necessary for a good data mining application. What's more, in medicine, even elementary concepts have no canonical form, which is a preferred notation that encapsulates all equivalent forms of the same concept, and that result in problems in constructing indexes and statistical tables (41). On the other hand, there are some efforts that aim to ease problems around standardisation of medical language and its terminology; major ones are the UMSL project, SNOMED and the GALEN programme from United States National Library of Medicine, UK National Health Service and the European Community respectively (42).

2.4.2 Ethical, Legal and Social Issues

Considerable progress in networking, storage and monitoring technology in medical informatics has led to excessively large databases. In tandem with this dramatic increase in digital data, concerns about the individual's privacy have emerged globally and these issues have further developed with using World Wide Web (43). As a result, privacy, confidentiality and security of a patient's health information has become more significant, since accessing the information becomes easier through modern electronic communications (44).

Before proceeding, it is useful to explain privacy, confidentiality and security issues, although many times these overlap and have effects on each other, because it is very likely that a data miner dealing with medical databases will encounter these challenges individually. According to Berman (2002), confidentiality is broken when a researcher shares a patient's private information with another unauthorised party, while security is broken when an authorised individual is able to access a patient's record. Additionally, privacy issues reveal when subjects in a research area cause unanticipated intrusions in their personal life. In a medical data mining task, risk of loss of confidentiality, or security, or privacy, or all of them always exists (45).

There has been a debate around when a data mining application violates privacy of human data.Finding a proper answer to this question is not as simple as it seems. For instance, consider a basic medical diagnosis model for public use: a classifier predicting the likelihood of an individual getting a terminal illness. Sharing the results with an insurer could result in erosion of the privacy of an individual, as the classifier uses public information (e.g. age, address, or cause of death of ancestors), which the insurer is presumed not to know (46). However, some data miners put forward the idea that their research should not be classified as a human subject research, since they use patient records, not patients; but according to the regulations made by Department of Health and Human Services, it is accepted that using a patient's record in research means that the patient is considered a participant in the research (45). Furthermore, the Equifax/Harris Consumer Privacy Survey showed that more than 78% of respondents believe that computer technology represents a threat to their personal privacy and that the area of use of computers must be restricted. At the same time, up to 96% of respondents believe that their private information should never be used for other purposes without permission (47).

In order to achieve individual privacy, data must be preprocessed rather than using raw data. In this regard, one of the methods that can overcome privacy violation is to de-identify (anonymise) the data by substituting any explicit identifying information, such as name and address, with some randomised values (48). Nevertheless, as anonymised data could not be verified for correctional or additional data and it is impossible to return to the patient's original data, such techniques become less popular. Hence, instead of anonymised data, the idea of anonymous data referring to the records where patient-identification has been removed has emerged (41).

Another challenge in medical data mining from a legal perspective is data ownership. This is a legally complex issue and still an open question. In legal theory, ownership is determined by who is allowed to sell a particular item of property (41). Although there are laws concerning ownership rights, it is necessary to establish efficient mechanisms to protect the holders' legal possession of the data (48). In this sense, digital watermarking is the most common technology used to ensure security and protection of multimedia data (49). In addition, Bertino (2005) et al. proposed a framework to deal with protecting data ownership and privacy by integrating techniques of binning and digital watermarking (48).

Chapter 3

Case Study

3.1 Understanding the problem domain

3.1.1 Objectives

We will try to learn the knowledge located in the Oldest Old Data Set by using the CART algorithm. More precisely, the possible associations between a patient's input variables and mortality of the patient will be investigated. In order to perform this, we will use two kinds of target variable and try two kinds of approaches. The first target variable indicates if the patient has died in the hospital during his or her treatment, whereas the second one shows if the patient has died after a long stay, say a week or more, in the hospital. Moreover, in the first approach, the original dataset without any change will be used, whereas, in the second approach, we will use the converted dataset including fields which consist of figures that are the absolute difference between the original value and its corresponding column mean. Additionally, we are anticipating revealing the most important variables that explain mortality of the patients over 90 years old.

3.2 Understanding the Data

3.2.1 Data Collection

The medical data used in this study comes from the gerontology department of three hospitals, namely Norwich Norfolk University Hospital (NNUH), Aberdeen Hospital and Woodend Hospital-Aberdeen. The data were collected from 01/11/2008 to 06/06/2009 and it is stored in a Stata file. It consists of clinical records of 393 patients who are over 90 years old and each patient was recorded with 97 inputs including both categorical variables and quantitative variables.

3.2.2 Data Quality

In general, the Oldest Old Data Set is of high quality; but of course suffers from some redundant and erroneous entries which are the typical problems found in medical data.

Hence, there is some unreliability in the data. Moreover, many medications and conditions are frequently referred to by a variety of names, for instance pyrexia, hyperthermia and fever all refer to an identical condition. Therefore the number of case of some conditions is relatively low.

Yet another issue relating to our dataset is the relatively high proportion of missing data. That is, more than 90 percent of some predictors are missing due to various reasons. Such fields are discarded from the data set.

3.3 Data Preparation

This stage is regarded as the most critical step in a data mining process, since the later steps largely depend on it (39). Thus it has been one of the focus points of this study.

Before starting the data mining steps, firstly the data were converted to spreadsheet (Excel File) format to make it suitable for the data mining package, PASW Modeler.

3.3.1 Handling the Outliers

It is very usual to encounter data objects which do not comply with the general behaviour or model of the data in a database. In this regard, they lie close to the limits of the data range or go against the trend of the remaining data. These kinds of data objects are called outliers or anomalous observations (18) (50). On the other hand, as the mechanism that generated the data is unknown, the definition of outliers is generally personal and speculative at best (51).

Since they may represent errors in data entry and can affect the model that will be constructed influentially, it is necessary to identify outliers in the data sets. Furthermore, even in the cases where an outlier is valid and not in error, certain statistical methods can be affected from the presence of outliers and are likely to yield unstable results (50). There are several methods generated to identify outliers for continuous variables, but the major ones are the followings:

- In the first method, an acceptable range of values for each variable is initially determined, and then any value out of the range is regarded as an outlier. Generally, it is accepted that the upper limit of the range is three times and the lower limit of the range is minus three times the standard deviations from the mean for a continuous variable. However, in this method the assumption of there is no interaction among the variables is made.

- The second approach is more sophisticated than the first. That is, it takes the multivariate nature of the data into account and fits a candidate model between a dependent variable and a set of independent variable and then finally deems those showing the large deviation from the fitted model as outliers (51).

- Another technique is to use clustering algorithms to cluster the data into smaller subsets and those containing an extremely small number of observations are identified as outliers.

Among the methods specified above, we have used the first one which is also provided by PASW Modeler's Data Audit node. In this sense, we have labelled the values further than 3 standard deviations as outliers and the values further than 5 standard deviations as extremes. From this perspective, 10 outliers and 4 extremes for the gcs (Glasgow coma score); 9 outliers for crp (C-reactive protein); 7 outliers and 2 extremes for urea; 6 outliers for temp (temperature); 5 outliers and 2 extremes for rr (respiratory rate) and creatinine; 5 outliers and an extreme for sodium; 4 outliers for diasbp (diastolic blood pressure); 4 outliers and an extreme for sato2 (oxygen saturation) and wcc (white blood cell count); 3 outliers for pulse (pulse rate) and hb (haemoglobin) fields have been detected.

PASW Modeler provides four options to handle outliers and extreme values. These include replacing outliers and extreme values with the nearest value that would not be considered problematic or with the null or system-missing value, discarding records with outlying or extreme values for the specified field, and discarding or nullifying extreme values only. However, we have preferred to replace the outliers and extremes with the nearest accepted value, as the data we have is relatively small and may not be able to tolerate discarding the cases where there is an outlier or an extreme value.

For the other dataset including fields which consist of figures that are the absolute difference between the original value and its corresponding column mean, the number of the extremes and the outliers are quite similar, so we have handled them in the same way. Consequently, those values have been replaced with the nearest acceptable value.

3.3.2 Missing Data Handling

Broadly speaking, missing values are values that should have been recorded but are not included in the available data (50). Missing data anomaly is quite common in datasets regardless of the size of the dataset, since the data collection and data storage processes are affected by various reasons (52). Nevertheless, the issue of missing data must be addressed before starting the data mining algorithm, as ignoring this problem is very likely to introduce bias into the models being evaluated and to result in inaccurate data mining conclusions (53). Additionally, when the management of the missing values is the issue, it should be borne in mind that inappropriate treatment of missing data can cause large errors or false results (54). Consequently, handling missing is an important part of the knowledge discovery process and requires maximum care.

The missing values we have are non-string values that have been left blank and have not been defined as “missing” in the both source files, namely the original dataset and the converted dataset. Thus they are recognised as system-missing values and displayed as $null$ by PASW Modeler, the data mining package we have used for this project. In this sense, the fields that have missing values are shown in the table below:

Name of the Field	Number of Missing Value	Missing Value Percentage
temp,meantemp	3	0.8%
crp,meancrp	3	0.8%
rr,meanrr	1	0.3%
pulse,meanpulse	1	0.3%
sysbp,meansysbp	1	0.3%
diasbp,meandiasbp	1	0.3%
sato2,meansato2	1	0.3%

Table 3.1 The number of missing values and their associated percentage in each field

Dealing with missing values can be examined under two main headings:

Removal of missing data: In this case, the records and/or fields with missing values are simply discarded. This approach can be effective and practical if the discarded fields are not crucial to the data analysis and/or the data contain small amount of missing values (34). On the other hand deleting records with missing values may be dangerous, because this approach would lead to a biased and smaller subset of the data, so the model generated may not represent the problem (50).

Imputation of missing data: Imputation methods can be divided into two categories: simple and complex. In the simple method, a missing value can be imputed by a user-defined value, mean or mode (the most probable value) of the attribute, mean of the records that have the same class label. For a complex method, a model is generated with a data mining algorithm, then the results of that model is used to predict the values appeared as missing. The first method is simple and straightforward, but it tends to risk introducing some bias in the model, whereas the second one has a better chance for generating less bias, since it uses the most information available in the present (51). Besides using the most information available, using a statistical imputation to handle missing data has several attractive properties. For example, once the imputations have been generated, the complete data set can be analysed with any data mining algorithm, as the imputation phase and analysis phase are separated. Therefore, the imputation model generated does not have to be true model (55).

Following the discussion above, we have used statistical imputation to handle missing data for the fields suffering from missing values, namely temp, crp, rr, pulse, sysbp, diasbp and sato2 for the original data set and, meantemp, meancrp, meanrr, meanpulse, meansybp, meandiasbp and meansato2 for the converted one. In addition, the statistical method used to fill the missing values here is the CART algorithm which also includes a fully automatic missing-value handling mechanism and it is provided by PASW Modeler.

3.3.3 Random Partitioning

In order to achieve a more robust model and to get a more realistic estimate of how that model would perform with unseen data, it is often necessary to partition the data into three separate subsets, namely training, testing and validation (56). However, if we mainly concentrate on finding the best model that has the highest accuracy rather than considering exactly how well it will do, we might use only training and validation partitions, but applying the model to test data will provide an unbiased estimate of how well it will do with new data. Further, in the cases where the data set is relatively small, as we have, there is always a risk of overfitting and when using the only validation data to assess the models generated and then select the model that does best with the validation data, we may overestimate the accuracy of our model (51) (57).

To draw the samples from the dataset for the subsets mentioned above, random sampling without replacement, which implies that each record is selected at most once, has been used. Therefore every sample from the dataset has the same chance of being selected in the subset. However, as such sampling does not guarantee that all values are represented in the target attributes, namely death and deathlg; we are assuming that the data is reasonably distributed.

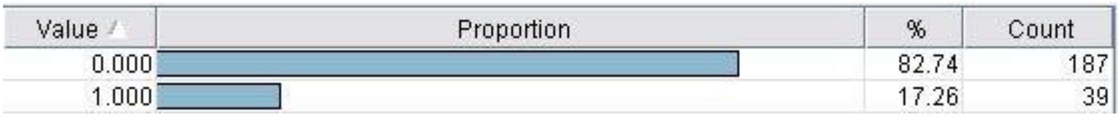
Because different random splits will result in different error estimates, determining the size of subsets is always significant and at the same time difficult. Indeed, according to Rafeet (2003) calculating the rigorous sample size is one of the most complex questions in data mining (51). That is, if the raining set is small, then the resulting model will not be very robust and will exhibit a lack of generalisation. In contrast, if the test is small, then the confidence in the estimated error rate will be low (58). Hence many theoretical formulas aiming to provide a rigorous solution to the optimum sample size problem have been developed, but these studies have been inadequate, as most of such formulas can only make a suggestion about the minimum sample size and generally they deal with quantitative variables (51).

Given the relatively size of the database, we have partitioned the original dataset and the converted dataset into training, testing and validation in a way that the training set, the testing set and the validation set contains 60%, 25% and 15% of the records with random samplings respectively.

3.3.4 Balancing

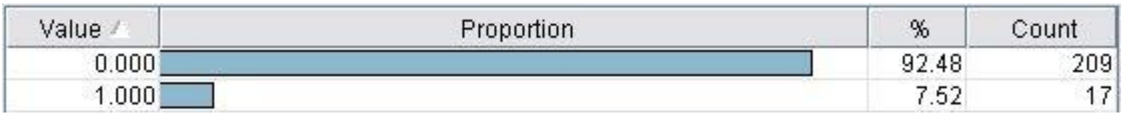
Decision trees are very sensitive to unbalanced datasets which include one class that is more heavily represented than the others, since they are designed to optimise overall accuracy without taking into account the relative distribution of each class (59) (60). Thus a decision tree can be easily misled, if one class dominates the training set and so it will tend to ignore the small classes while concentrating on classifying the large ones accurately (60). Moreover, in such cases, the average classification accuracy (percentage of testing examples correctly recognised by the system) is no longer a secure criterion to evaluate the classification performance. Hence it requires using more complex evaluation criteria (61). Finally, in an extreme case, if the classifier could not see any example for a particular target value, it would be meaningless to expect that the classifier could predict that target value.

For both datasets, we have identical target values with identical distribution. These are death, which shows if the patient has died in the study, and deathlg that indicates if the patient has died after a long stay, i.e. any stay equal to or greater than a week in the hospital. For both fields, 1 stands for yes and 0 stands for no. In this sense, 82.74% of the attribute called deadth in the training set is 0 whereas 17.26% of this attribute is 1 (See Figure 3.1). Similarly, for deathlg attribute in the training set, 92.48 percent is 0 and only 7.52 percent is 1 (See Figure 3.2). As a result, the training set we have requires balancing prior to constructing a model.



Value	Proportion	%	Count
0.000		82.74	187
1.000		17.26	39

Figure 3.1 Distribution of death in the training set



Value	Proportion	%	Count
0.000		92.48	209
1.000		7.52	17

Figure 3.2 Distribution of deathlg in the training set

There are various approaches trying to balance the class distribution in the training set. However, the most popular ones are oversampling the minority class and downsampling the majority class (62). In the first approach, the minority class samples are replicated, while majority class samples are preserved to equate the number of those classes. In the latter case, many majority samples are ignored, while the minority ones are preserved. Meanwhile, Japkowics (2000) revealed that despite their simplicity, oversampling and downsampling are very effective (63).

PASW Modeler provides both methods discussed above thanks to its balance node. However, we have preferred to use the oversampling method, since the number of minority class in the training set was too low for both target fields. Indeed, if the downsampling method had been preferred, the classifier would have learned from 39 cases for death attribute, and 17 cases for deathlg attribute, which would have led to a lack of generalisation ability for the training set (58).

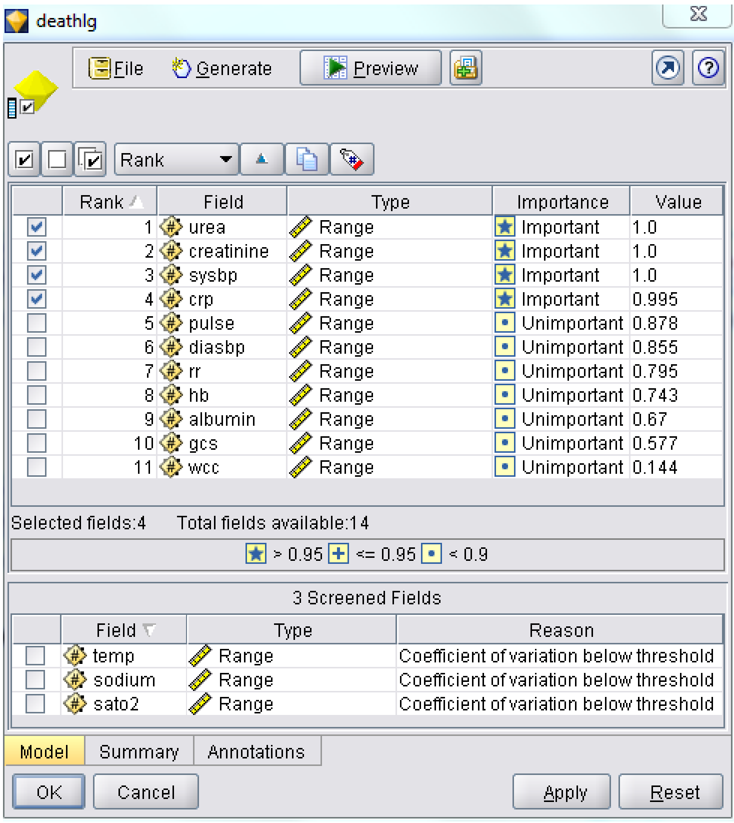
3.3.5 Feature Subset Selection

Generally, datasets contain more information than is needed to build a model which may reduce learning accuracy and/or a model's generalisation ability. To make matters worse, irrelevant features originated from noisy fluctuations in data can mask the underlying natural patterns in the data (64). Feature selection as a preprocessing step addresses this concern. In this step, input variables otherwise called features that are most relevant to predicting a target variable are selected by using various techniques. With an effective feature selection mechanism, the data mining algorithm can learn faster, higher accuracy can be provided, the model will be generated can generalize better from data and using and understanding of the results of the data mining process becomes much more easier (58).

The process of feature selection can be implemented in two ways: manually and/or with an algorithm. Firstly, the domain expert can select a subset of features in the initial data set, based on his or her knowledge about the application domain and the goals of the data mining work (58). In our case, the dataset initially had 97 features ranging from categorical to numerical about a patient, such as age, gender, date of admission, troponin I value and so on. However, the domain expert reduced that number to sixteen in a way that there would be fourteen input variables which are all numerical and two categorical output variables, namely death and deathlg explained in the previous sections.

Secondly, the number of features can be reduced by using an algorithm (e.g. CART) or some automated procedures. Again, in our case, after narrowing the feature space with the help of the domain expert, we have implemented an automatic procedure for the output called deathlg in the original dataset.

PASW Modeler, has ranked the list of the features which were ordered according to their importance. Here, the measure used to rank the fields on a percentage scale defined as 1 minus the p-value (65). According to that measure, the program has found urea, creatinine, sys and crp as the most important features in determining to the target feature, deathlg in the original dataset. Additionally, as is known, the fields that have low variance do not play a major role in predicting the target field. Hence those fields should be discarded from the dataset in order to catch the advantages of the feature subset selection process mentioned at the beginning of this section. At this point, PASW Modeler uses the coefficient of variation (CV), which does not require understanding the context of the mean of the data, to detect the fields suffering from low variance. In our case, distributions with CV < 0.1 were considered low variance and the features, temp, sodium and sato2, have been regarded as a problematic and have been discarded with the other unimportant fields from the explanatory variable set (See Figure 3.3).



	Rank	Field	Type	Importance	Value
☑	1	urea	Range	Important	1.0
☑	2	creatinine	Range	Important	1.0
☑	3	sysbp	Range	Important	1.0
☑	4	crp	Range	Important	0.995
□	5	pulse	Range	Unimportant	0.878
□	6	diasbp	Range	Unimportant	0.855
□	7	rr	Range	Unimportant	0.795
□	8	hb	Range	Unimportant	0.743
□	9	albumin	Range	Unimportant	0.67
□	10	gcs	Range	Unimportant	0.577
□	11	wcc	Range	Unimportant	0.144

Selected fields:4 Total fields available:14

★ > 0.95 + <= 0.95 • < 0.9

3 Screened Fields

	Field	Type	Reason
□	temp	Range	Coefficient of variation below threshold
□	sodium	Range	Coefficient of variation below threshold
□	sato2	Range	Coefficient of variation below threshold

Figure 3.3 The List of the features determining deathlg

For the target field, deathlg, the program has found that urea, creatinine, sysbp and crp take the most important fields in order of priorities under the criteria explained before, so only these fields have been used to construct the decision tree for the output, deathlg.

3.4 Data Mining

There are several data mining methods. However, due to the benefits of the CART algorithm given in 3.4.1.1, here we used the CART algorithm to reach the objectives stated at the beginning of this case study. In this sense, the main characteristics of the CART algorithm, its advantages and disadvantages are discussed below.

3.4.1 Classification and Regression Trees (CART)

Classification and regression trees (hereafter CART) introduced by Breiman et al. (1984) are prediction models constructed by recursively partitioning a data set and fitting a model to each partition (66). Furthermore, the thing that makes the CART algorithm special is that it can deal with both continuous and categorical dependent variables. In this sense, if a dependent variable is continuous, CART produces regression trees, whereas if the variable is categorical; CART produces classification trees. However, in both trees, the major aim of the CART algorithm is to construct a tree that has a relatively small number of branches and intermediate nodes and high predictive power, where entities are correctly classified at the terminal nodes (67). Constructing such a tree can be summarised as follows:

(1) Put all objects to root node.

(2) Split each predictor at all its possible split points (e.g. X= t_1 is a split, where X is a variable, t_1 is a value) (64).

(3) For each split point, split the parent node into two child nodes in a way that the objects with values lower than the split point are sent to the "left"; otherwise the objects are sent to the "right" (See Figure 3.4 for an example).

(4) Select the variable and split point that provides the highest decrease in impurity.

(5) According to the selected split point specified at the step 4, perform the split.

(6) Repeat steps 2 to 5, using each node as a new parent node, until the tree becomes the largest (64).

(7) Prune the tree back by using cross-validation to decide the optimal sized tree.

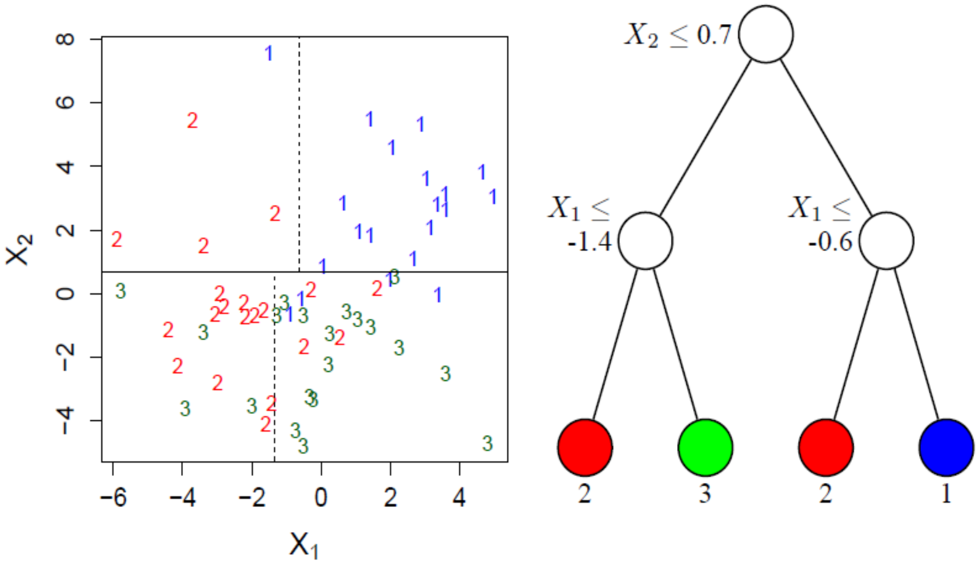


Figure 3.4 Partitions and decision tree structure for a CART classification model (66).

There are two important considerations, when constructing a CART tree. Firstly, impurity measure for a node varies mainly according to the type of a dependent variable. That is, if a tree has a numerical response variable, then the impurity is defined as the total sum of squares of the response variable around the mean of each node and computed as follows

$$\text{impurity} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where, n is the number of objects and $\bar{y}$ is the mean of the related column (68) (64).

However, if a tree has a categorical dependent variable, the impurity can be defined in three ways. Most commonly, the Gini index, proposed by Breiman et al. (1984), is used for computing the impurity, while Twoing and ordered Twoing are other methods used for such a task (69). The Gini index, the generalization of the binomial variance, of a node with n objects k possible classes is defined as

$$\text{Gini} = 1 - \sum_{j=1}^{c} \left(\frac{n_j}{n} \right)^2$$

where n_j is the number of objects from class j present in the node (64).

Another important issue around constructing a CART tree is to decide the size of the tree. The earlier work on decision trees did not allow for pruning and the trees grew until they satisfied a stopping rule specified before. On the other hand Breiman et al. (68) showed that no rule designed to stop tree growth can guarantee that it will uncover all important interactions between explanatory variables (70). Hence a tree is first grown to it's largest size where all terminal nodes are either small (no more than a predefined number of objects) or pure (all objects in the node have identical values of the response value) (64). Thereafter, the tree is pruned back in order to avoid overfitting, because such large trees tend to fit the noise.

In attempt to understand the importance of optimal size in decision trees and the effects of overfitting more precisely, the figure below can assist us. Clearly, after passing the optimal tree size, the error rate on the training data gets close to zero (in fact the error rate becomes zero, if the terminal nodes contain data from a only one single class), whereas the test error rate on new data (what we are interested in for prediction) begins to increase since the algorithm overfits the data and nodes which are only predict noise or random variation in the training data (71).

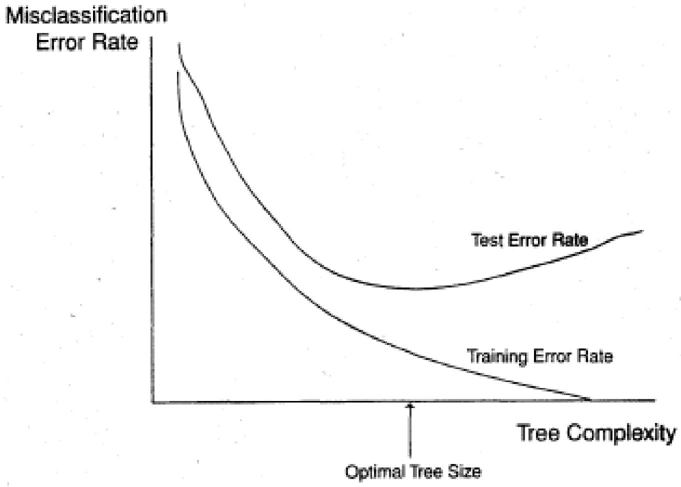


Figure 3.5 Misclassification error rates for both training and testing data as a function of tree complexity

To deal with the problem discussed above, the CART algorithm employs cross-validation procedure. In this regard, cross validation tells us when to stop pruning. That is, firstly Cost-complexity measure is defined as:

$$R\ \alpha\ (T) = R\ (T) + \alpha\ |T|$$

where R(T) is misclassification rate, |T| is the number of terminal nodes in the tree and α is a penalty imposed on each node (70). Moreover, the less α is, the larger the size of the tree is and for each value of α, there is a smallest tree minimising the cost-complexity measure (R α (T)) (64). For example, when α is 0, the tree has reached its largest size (T_0) and increasing in the complexity parameter, α, leads to a nested sequence of trees that decreases in size:

$$\boldsymbol{T_0 \gg T_1 \gg T_2 \gg T_3 \gg \dots \gg T_k \gg \dots}$$

At this point, the optimal tree size is determined by cross validation. In this sense, the data are split up into several, broadly speaking 10, nonoverlapping equal-size pieces. For each piece of the data, a tree is constructed with 90% of the data and the remaining 10% is used for testing the tree. The prediction error, which is misclassification error for classification trees, is matched with the subtrees of the complete dataset using α value, and then the optimal sized tree becomes the one with the lowest cost-complexity measure (64). However, Breiman et al. offered 1 S.E. rule that is to choose the smallest tree whose cross-validation (CV) costs do not exceed the minimum CV costs plus 1 times standard error of the CV costs for the minimum CV costs tree, since CV costs of several subtrees are close to the minimum (72).

3.4.1.1 Advantages and Disadvantages of the CART Algorithm

The CART algorithm has been used successfully for more than two decades and its advantages can be outlined as follows:

- The CART algorithm is easy to interpret.
- There is no strict rule to stop a CART tree’s construction, contrary to other decision trees using stopping rules. This is therefore; the important structure of the data is not overlooked by a stopping rule.

- CART can cope with any data structure, e.g. categorical, numerical and ordinal.
- Unlike the other decision trees using multi-way splits which partition the data rapidly so may fail to discover the data broadly, CART uses binary split. Hence it has ability to detect more rules.
- CART can handle missing values thanks to the surrogate splitters accommodating in it.
- Like the other decision trees, it is easy to interpret a CART’s results.
- Since CART is a nonparametric method, it makes no distributional assumptions of any kind of explanatory and response variables. Hence it has a capability of handling numerical data that are highly skewed or multi-model and categorical predictors having ordinal structure (73).
- CART is computationally faster than other methods, such as neural networks (72).

In contrast, the CART algorithm suffers from many aspects:

- Like the other types of the decision trees, a small change in the value of a dependent variable may result in a large change in the predicted value of the response.
- CART fails to capture strong linear structure. That is, a very large tree may be produced to capture very simple linear relationships (74).
- CART uses only univariate split, so does not use combinations of variables at each node.
- The tree built by the CART algorithm is not often globally optimal, although it is optimal at each split.
- Some important software packages (e.g., SAS) do not include CART, since it is not a standard analysis technique.

Finally the components of the CART algorithm are summarised in the table below:

Task :	Prediction
Model Structure:	Tree
Score Function:	Cross-Validated Loss Function
Search Method:	Greedy Local Search
Data Management Method:	Unspecified

Table 3.2 The components of the CART algorithm

3.4.2 The Application Process of the CART Algorithm

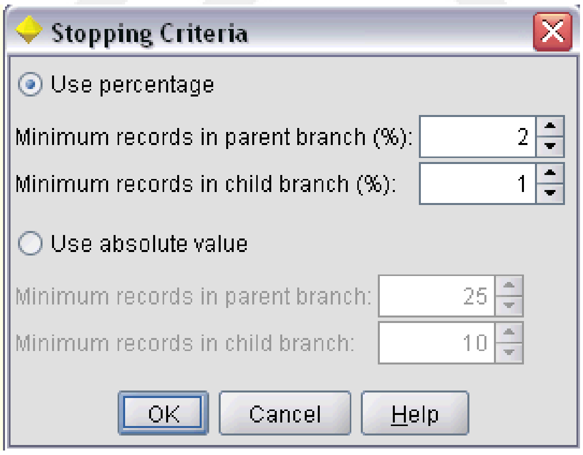
PASW Modeler allows us to use CART under C & R Tree Node. This node, first examines the input fields to decide the best split that is measured by the reduction in an impurity index and then generates two subgroups from the split which is subsequently split into two more subgroups, and so on, until one of the stopping criteria is satisfied (65). Thereafter, the C & R tree is pruned optionally, based on the cost-complexity algorithm mentioned in the previous section to avoid the overfitting.

Along with the data mining process, we have built three different decision trees with the CART algorithm. Hereafter, Model I will refer to the first model derived from the original dataset. For this model, the target variable is death, which indicates if the patient has died in the study. Moreover, for this model, neither balancing nor feature selection mechanisms have been implemented, since they resulted in high decrease in the accuracy of the model. Secondly, Model II will refer to the second model which was again derived from the original dataset. For Model II, the output variable is deathlg, which shows that if the patient has died after a long stay (e.g. one week or more). For this model, we have balanced the distribution of the target variable and used the feature selection based on the results of the feature selection node. Finally, Model III will refer to the third model derived from the converted dataset and implemented the balancing mechanism. In summary, the table below shows those three models with their properties:

Model Name	**Dataset Used**	**Target**	**Algorithm Used**	**Balancing**	**Feat. Selection**
Model I	The Original	death	CART	No	No
Model II	The Original	deathlg	CART	Yes	Yes
Model III	The Converted	death	CART	Yes	No

Table 3.3 The properties of the models

For all the decision trees constructed, the default values of C & R Node have been used. According to these default values, the maximum number of surrogate is 5, the minimum change in the impurity is 0.0001 and its measure for categorical targets is the Gini index. In addition, a minimum record in parent branch is 2% and a minimum record in child branch is 1%. Thus if either a parent node (to be split) includes less than 2% of total records or if a child node (to be created) includes less than 1% of total records, such split will be prevented (See Figure 3.6). Finally, the pruning applied here is based on removing bottom-level splits that do not contribute considerably to the accuracy of the tree (65). See Figure 3.7.



Stopping Criteria

● Use percentage

Minimum records in parent branch (%): 2

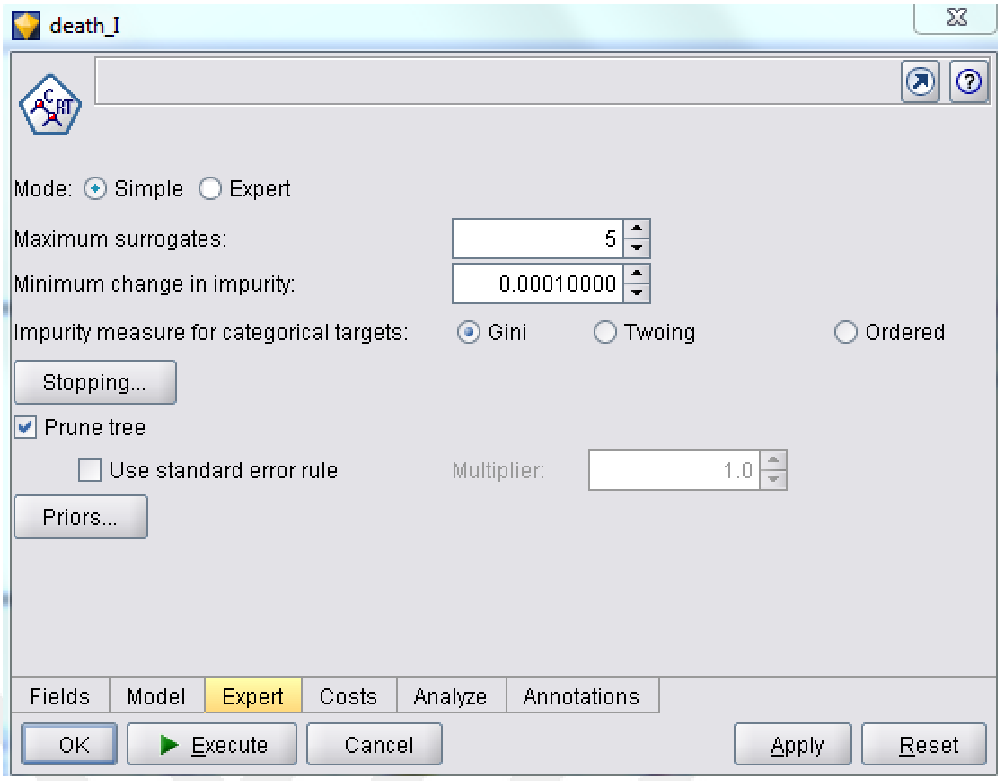
Minimum records in child branch (%): 1

○ Use absolute value

Minimum records in parent branch: 25

Minimum records in child branch: 10

Figure 3.6 Stopping criteria of the decision tree



Mode: ● Simple ○ Expert

Maximum surrogates: 5

Minimum change in impurity: 0.00010000

Impurity measure for categorical targets: ● Gini ○ Twoing ○ Ordered

☑ Prune tree

□ Use standard error rule Multiplier: 1.0

Figure 3.7 The Default values in C & R Node

3.4.3 The Chi-square Test for Goodness of Fit

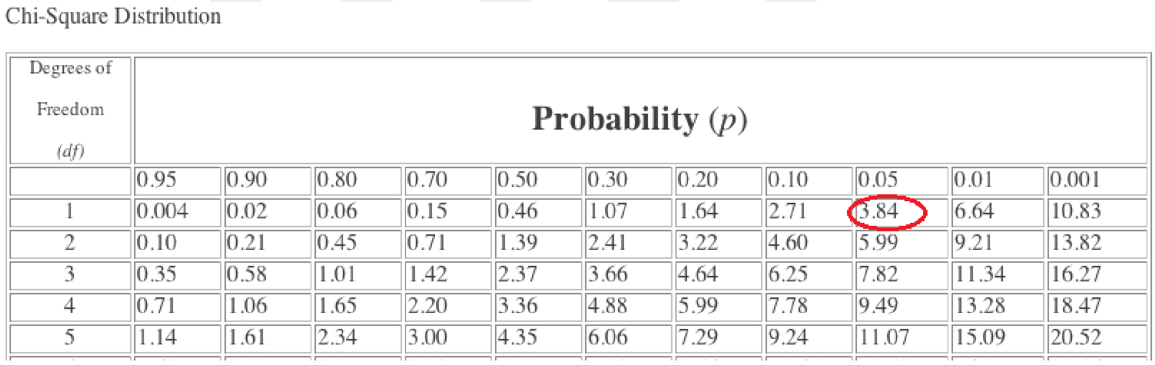
In order to test the goodness of fit between theoretical and experimental data and compare these, Chi-square (χ^2) test is often used (75). Here, observed values are those that we have obtained empirically through the testing dataset (real values)p; theoretical or expected values are those obtained from the models we have generated, namely Model I, Model II and Model III.

In the Chi-square test, first a null hypothesis is stated, and then according to the comparison between the χ^2 value found and the critical value read from the χ^2 distribution table, it is determined whether the null hypothesis is accepted. That is, if the χ^2 value is equal to or greater than the critical value, the null hypothesis is rejected; else it is accepted to be true. The test statistic χ^2 is computed as:

$$\chi^2 = \sum \frac{(E-O)^2}{E}$$

where O is the observed frequency in each category and E is the expected frequency in the corresponding category. To read the critical values from the χ^2 distribution table, there is a need for the level of significance α and the number of degrees of freedom. Choosing level of significance is an arbitrary task, but generally a level of 5% is chosen for no better reason than it is conventional (76). The degrees of freedom (df) refers to the number of values that are free to vary. In other words it refers to the number of independent pieces of information. Generally, the number of degrees of freedom is calculated as the number of rows minus one times the number of columns minus 1 for a *contingency table* (77).

Let the null hypothesis state that there is no significant difference between the expected and the observed frequencies and let the level of significance be 5%. Moreover, since the contingency tables will consist of two rows and two columns for each model, df is calculated as (2-1)(2-1) = 1. In this case, the critical value will be 3.84. See the table below.



Chi-Square Distribution

Degrees of Freedom *(df)*	**Probability** (*p*)										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52

Table 3.4 Chi-Square Distribution

For the three models, there are two categories, 0 standing for no and 1 for yes. χ^2 for each of these models, using the above formula, are found in the following tables.

Table 3.5 Calculation of χ^2 for Model I

Class	$\boldsymbol{O}$	$\boldsymbol{E}$	$(\boldsymbol{O}\text{-}\boldsymbol{E})$	$(\boldsymbol{E}-\boldsymbol{O})^2$	$\dfrac{(\boldsymbol{E}-\boldsymbol{O})^2}{\boldsymbol{E}}$
0 (No)	67	73	-6	36	0.493
1 (Yes)	20	14	6	36	2.571
Total	87	87	0	--	3.064 = χ^2

Table 3.6 Calculation of χ^2 for Model II

Class	$\boldsymbol{O}$	$\boldsymbol{E}$	$(\boldsymbol{O}\text{- }\boldsymbol{E})$	$(\boldsymbol{E}-\boldsymbol{O})^2$	$\frac{(\boldsymbol{E}-\boldsymbol{O})^2}{\boldsymbol{E}}$
0 (No)	81	77	4	16	0.21
1 (Yes)	6	10	-4	16	1.6
Total	87	87	0	--	1.81 = χ^2

Table 3.7 Calculation of χ^2 for Model III

Class	$\boldsymbol{O}$	$\boldsymbol{E}$	$(\boldsymbol{O}\text{- }\boldsymbol{E})$	$(\boldsymbol{E}-\boldsymbol{O})^2$	$\frac{(\boldsymbol{E}-\boldsymbol{O})^2}{\boldsymbol{E}}$
0 (No)	67	63	4	16	0.254
1 (Yes)	20	24	-4	16	0.666
Total	87	87	0	--	0.92 = χ^2

The tables denote that all of χ^2 is less than the critical value, 3.84. Thus the null hypothesis is accepted to be true. As a result, there is no significant difference between the expected and the observed frequencies for all fitted models.

3.4.4 Evaluation of the Models

To evaluate the models generated by the CART algorithm, there are two main tools: the gains chart and classification tables. Thanks to these tools, specificity and sensitivity analysis can be conducted, model stability can be investigated by comparing classification tables for the testing and learning sets, and how well models performs in terms of the ranking of the cases can be examined (74).

3.4.4.1 Classification Tables for the Models

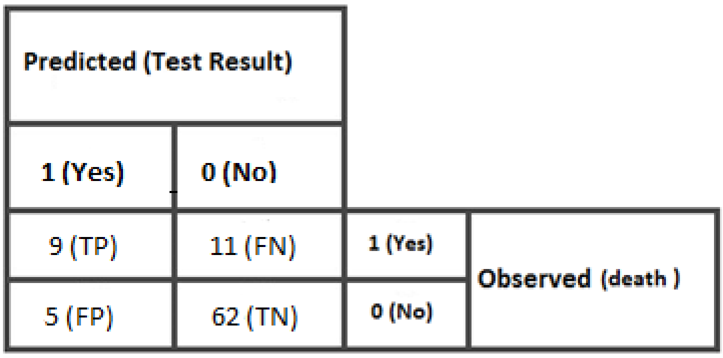
The general performance measure for a binary classifier system, such as decision trees, is the accuracy. However, for medical applications, two other measures are more frequently used than the accuracy: sensitivity and specificity. Indeed, high sensitivity and\or specificity of a classifier's answers is more important than high classification accuracy in many medical problems (12).

In general, accuracy shows the rate of correct values and refers to degree of fit between the model and the data. In addition, sensitivity measures the accuracy of predicting the target events, while specificity refers to the accuracy of predicting the target nonevent, for instance, the probability that the symptom is not present given that the person does not have the disease (78). These measures are calculated as follows:

- $$accuracy\ (proportion\ of\ correct\ predictions) = \frac{\mathrm{TP + TN}}{\mathrm{TP + TN + F\,P + FN}}$$
- $$sensitivity = \frac{instances\ with\ positive\ class\ correctly\ classified}{total\ number\ of\ positive\ instances} = \frac{TP}{TP + FN}$$
- $$specificity = \frac{Instances\ with\ negative\ class\ incorrectly\ classified}{Total\ number\ of\ negative\ instances} = \frac{\mathrm{TN}}{\mathrm{TN + FP}}$$

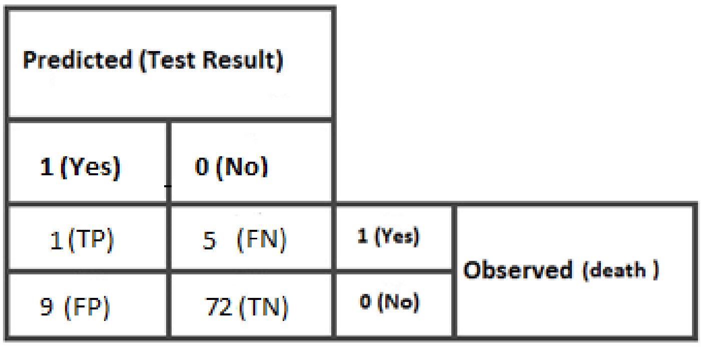
where TP, TN, FP and FN stands for true positives, true negatives, false positives and false negatives, respectively.

In our case, the values of TP, TN, FP and FN for each model generated are denoted on the tables below. See Table 3.8, Table 3.9 and Table 3.10.



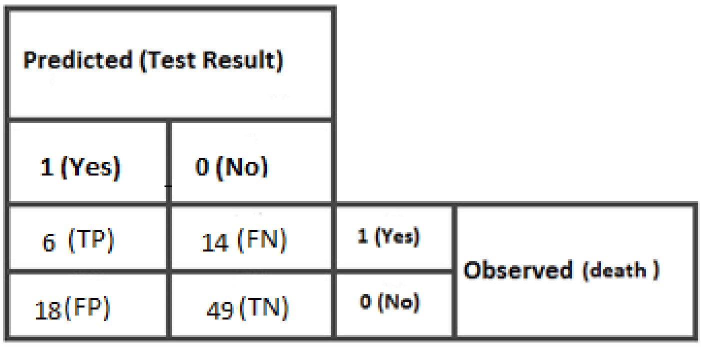
Predicted (Test Result)			
1 (Yes)	0 (No)		
9 (TP)	11 (FN)	1 (Yes)	Observed (death)
5 (FP)	62 (TN)	0 (No)	

Table 3.8 TP, TN, FP and FN values for Model I



Predicted (Test Result)			
1 (Yes)	0 (No)		
1 (TP)	5 (FN)	1 (Yes)	Observed (death)
9 (FP)	72 (TN)	0 (No)	

Table 3.9 TP, TN, FP and FN values for Model II



Predicted (Test Result)			
1 (Yes)	0 (No)		
6 (TP)	14 (FN)	1 (Yes)	Observed (death)
18(FP)	49 (TN)	0 (No)	

Table 3.10 TP, TN, FP and FN values for Model III

We found the accuracy measures for Model I, Model II and Model III as follows:

Model Name	**Accuracy**	**Sensitivity**	**Specificity**
Model I	0.816	0.45	0.925
Model II	0.862	0.166	0.888
Model III	0.63	0.3	0.731

Table 3.11 the accuracy measures for the models constructed

From the table above, it is obvious that Model III whose target variable addresses that if the patient will die in the study has the lowest accuracy: 0.63. That is, the rules generated from the Model III are 63% correct in recognising samples in the associated testing set. Additionally, sensitivity measure of Model II whose response variable indicates if the patient will die after a long stay in the hospital is 0.166, so the model is very likely to fail to recognise a patient who will die after a long stay in the hospital. Finally, as seen in the table, Model I has the highest specificity score: 0.925. Hence it is good at excluding negative test examples.

3.4.4.2 The Gains Chart

Gains chart (also known lift curves) is an effective visual tool to appreciate how well the model predicts and how well it ranks cases (74). The horizontal axis of a gains chart represents the percentage of the population examined and the vertical axis represents the corresponding cumulative gains. In addition, the diagonal line represents the null hypothesis that the tree gives no useful information about the target field. In our case, the null hypothesis assumes that all patients are the same. The vertical difference between the curved line and the diagonal line gives the gain .Thus the farther above the diagonal line a curve lies, the greater the gain, so a desirable chart will be a deeper bowl rather than a shallower bowl (65) (50).

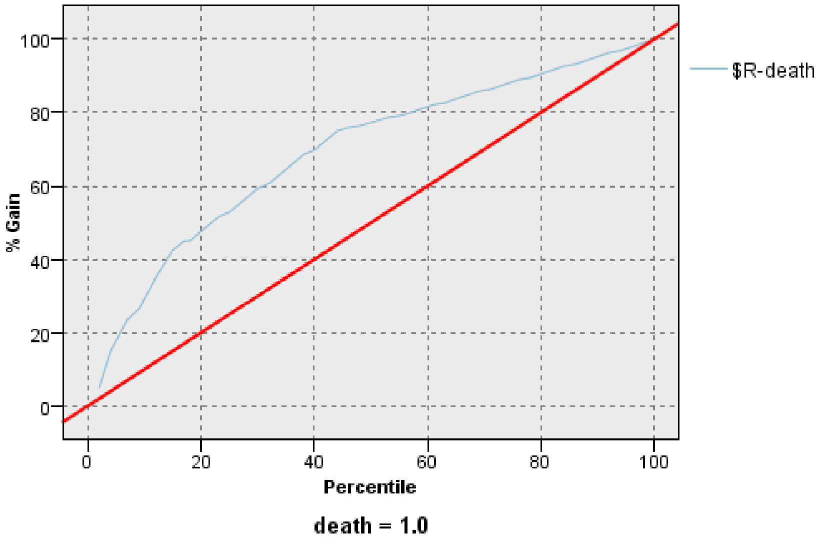


Figure 3.8 Evaluation of gains for Model I for the testing set

For Model I, the gains chart is above. Here, the x axis denotes the percentage of patients examined and the y axis denotes percentages of patients died in the study. According to this chart, for example, when canvassing the 60% of the patients in our study, we could expect to reach more than 80% of the total number of patients passing away in the study. When we consider Figure 3.9 which takes the training set into account, it can be concluded that Model I is relatively stable and does not suffer from overfitting, since the curved lines show similarity.

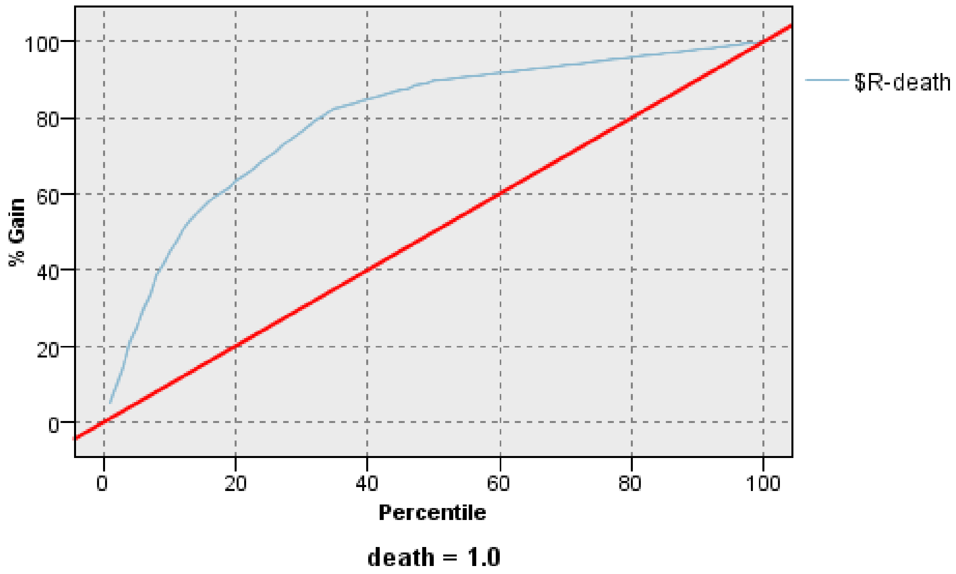


Figure 3.9 Evaluation of gains for Model I for the training set

3.5 Evaluation of the discovered knowledge

3.5.1 Model I Evaluation

The decision tree generated from Model I applied to the testing samples has six levels and ten terminal nodes (See Decision Tree I). With no analysis there is an about 17% rate of patients who has died in the hospital in the learning sample. However, simply by restricting the sample to those with urea higher than 13.95 mg/L, the hit rate increases to about 38%. To understand what have been explored more broadly, the terminal nodes can be examined individually. See Table 3.11

TN no	**Node no**	**death (%)**	**Number death**	**N**	**Rules**
1	3	22.7	10	44	(urea ≤ 13.95) and (sato2 ≤ 94.5)
2	5	80	8	40	(urea > 13.95) and (sato2 ≤ 93.5)
3	9	42.86	3	7	(urea ≤ 13.95) and (sato2 > 94.5) and (sodium ≤ 128.5)
4	18	3.54	4	113	(urea ≤ 13.95) and (sato2 > 94.5) (sodium > 128.5) and (crp

					>0.65)
5	19	10	3	30	(urea > 13.95) and (sato2 > 93.5) (wcc ≤ 1695) and (diasbp ≤ 89.5)
6	20	60	3	5	(urea > 13.95) and (sato2 > 93.5) (wcc ≤ 1695) and (diasbp > 89.5)
7	21	33.3	1	3	(urea > 13.95) and (sato2 > 93.5) (wcc > 1695) and (temp ≤ 35.75)
8	22	100	5	5	(urea > 13.95) and (sato2 > 93.5) (wcc > 1695) and (temp > 35.75)
9	23	66.66	2	3	(urea ≤ 13.95) and (sato2 > 94.5) (sodium > 128.5) and (crp ≤ 0.65) and (diasbp ≤ 81)
10	24	0	0	6	(urea ≤ 13.95) and (sato2 > 94.5) (sodium > 128.5) and (crp ≤ 0. 65) and (diasbp > 81)
Total	**--**	**--**	**39**	**187**	

Table 3.12 Terminal nodes of Model I with percentages of the events

The rules generated from Model I can be summarised as follows:

- As the root node split is considered to indicate the most important single variable (50), the urea level plays the most important role, when determining if a patient will die in the hospital.
- The risk of death of those with the urea less than 13.95 mg/L is quite low, about 10%.
- If a patient’s urea level is higher than 13.95 mg/L and the percentage of oxygen in his or her blood is less than 93.5, then he or she is very likely to die (80%).
- The most important variables to classification of death are urea level and oxygen percentage in the blood at first, whereas the third most important variable is either the amount of sodium in the blood or white blood cell count.
- Those with the urea less than 13.95 mg/L , higher than 94.5% oxygen saturation and normal serum sodium levels (between 135 - 145 mEq/L) have a 95% chance of living.

- If a patient’s urea level is higher than 13.95 mg/L, his or her oxygen saturation level is normal (between 96% - 100%) and his or her white blood cell count is higher than 17,000, then he or she has a 75% chance of dying.

See Figure 3.10 for Decision Tree I

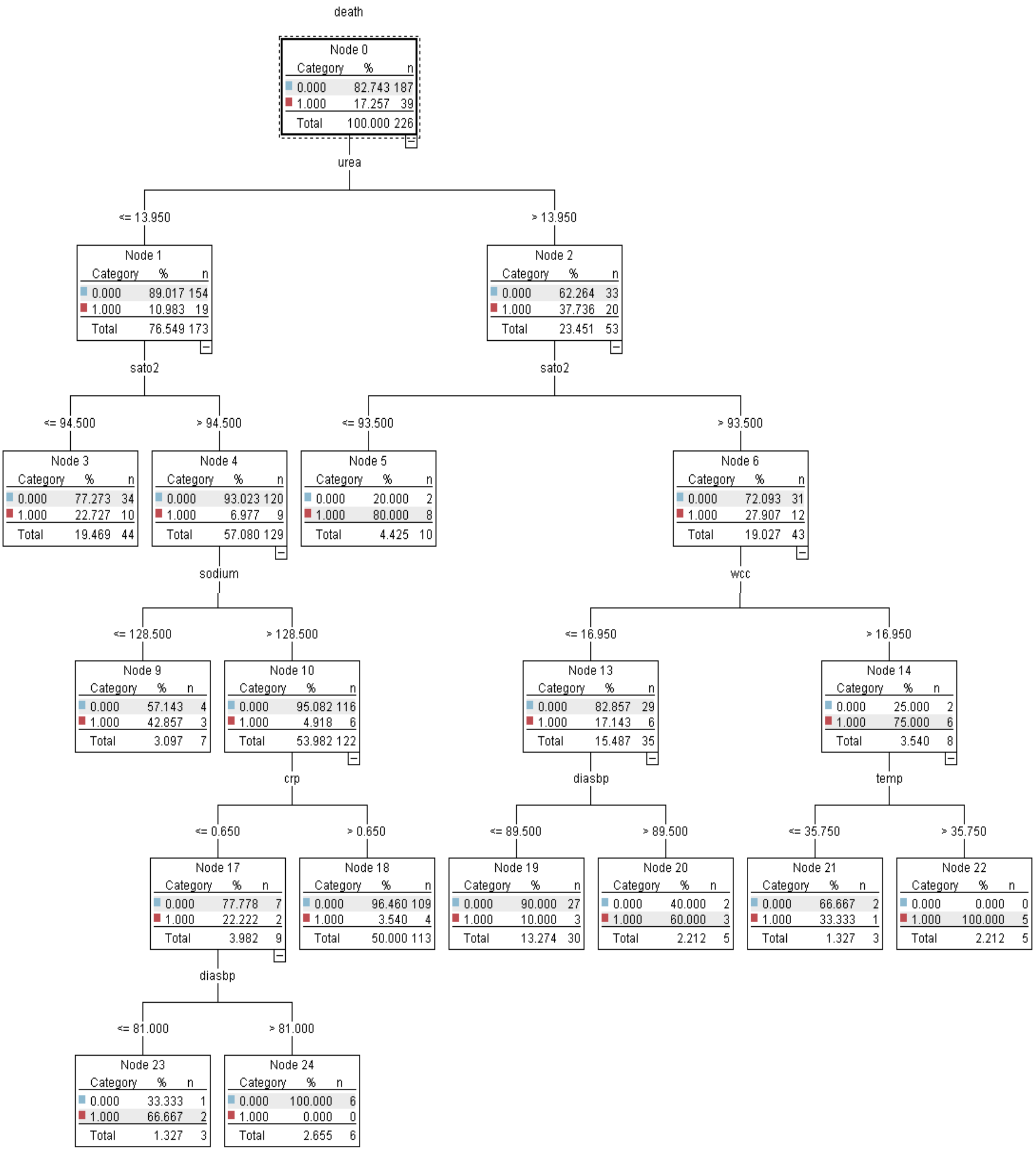


Figure 3.10 Decision Tree I

3.5.2 Model II Evaluation

Unlike the previous case, the decision tree generated from Model II whose output variable indicates that if the patient has died after a long stay in the hospital has eleven terminal nodes but six levels (See Decision Tree II). Interestingly, for both trees, the root split is on urea with the same cut-off value (13.95 mg/L). Apart from that, the best node for the tree generated from Model II of 63 observations with 96.825% event. However, all the terminal nodes of this tree have been represented on the table below.

TN no	Node no	deathlg (%)	Number death	N	Rules
1	6	0	0	10	(urea > 13.95) and (creatinine > 223.5)
2	7	0	0	5	(urea ≤ 13.95) and (sysbp ≤ 99.5) and (crp ≤ 10.3)
3	8	96	24	25	(urea ≤ 13.95) and (sysbp ≤ 99.5) and (crp > 10.3)
4	10	0	0	50	(urea ≤ 13.95) and (sysbp > 99.5) and (crp > 4.3)
5	11	0	0	7	(urea ≤ 13.95) and (creatinine ≤ 223.5) and (creatinine ≤ 113.5)
6	15	30.8	37	137	(urea ≤ 13.95) and (sysbp > 99.5) and (crp ≤ 4.3) and (crp ≤ 3.5)
7	17	96.825	61	63	(urea > 13.95) and (creatinine ≤ 223.5) and (creatinine > 113.5) and (sysbp ≤ 110.5)
8	21	96.296	26	27	(urea ≤ 13.95) and (sysbp > 99.5) and (crp ≤ 4.3) and (crp > 3.5) and (urea ≤ 7.65)
9	22	0	0	9	(urea ≤ 13.95) and (sysbp > 99.5) and (crp ≤ 4.3) and (crp > 3.5) and (urea > 7.65)
10	25	42.857	12	28	(urea > 13.95) and (creatinine ≤ 223.5) and creatinine > 113.5) and (sysbp > 110.5) and (sysbp ≤ 140)
11	26	85.714	48	56	(urea > 13.95) and (creatinine ≤ 223.5) and creatinine > 113.5) and (sysbp > 110.5) and (sysbp > 140)
Total	**--**	**--**	**208**	**417**	

Table 3.13 Terminal nodes of Model II with percentages of the events

Thanks to the Model II, we have reached the following results:

- In order of priorities, the most important variables to classification of death after a long stay in the hospital are urea, creatinine and sysbp (systolic blood pressure). Obviously, any increase in these variables may lead to a dramatic increase in the risk of death. Indeed, for example, it can be seen from the first split that the risk of death of those with urea level higher than 13.95 mg/L is more than double the risk of those with urea level less than 13.95 mg/L. Additionally, when comparing Node 11 and Node 12, it is clear that patients with normal creatinine levels (60 – 110 mmol/L) are not considered in the risk group, whereas under the same conditions, those with higher creatinine levels (> 113.5 mmol/L) have a very high risk of death (around 83%).

- Terminal Node 2 suggests that in the cases, where a patient has low urea level (< 13.95 mg/L), low systolic blood pressure (< 99.5 mmHg) and low C-reactive protein test score (< 10.3 mg/L), it is not expected that the patient will die after a week or more from his or her admission to the hospital.

- From Terminal Node 11, it is clear that If a patient has high systolic blood pressure (> 140 mmHg), high creatinine level (> 113.5 mmol/L) and at least relatively high urea level (> 13.95 mg/L), he or she is very likely to die (with 0.85 probability). Interestingly, this probability is 0.42 for the patients falling into Terminal Node 10 where the systolic blood pressure is less than 140 mmHg.

  See Figure 3.11 for Decision Tree II

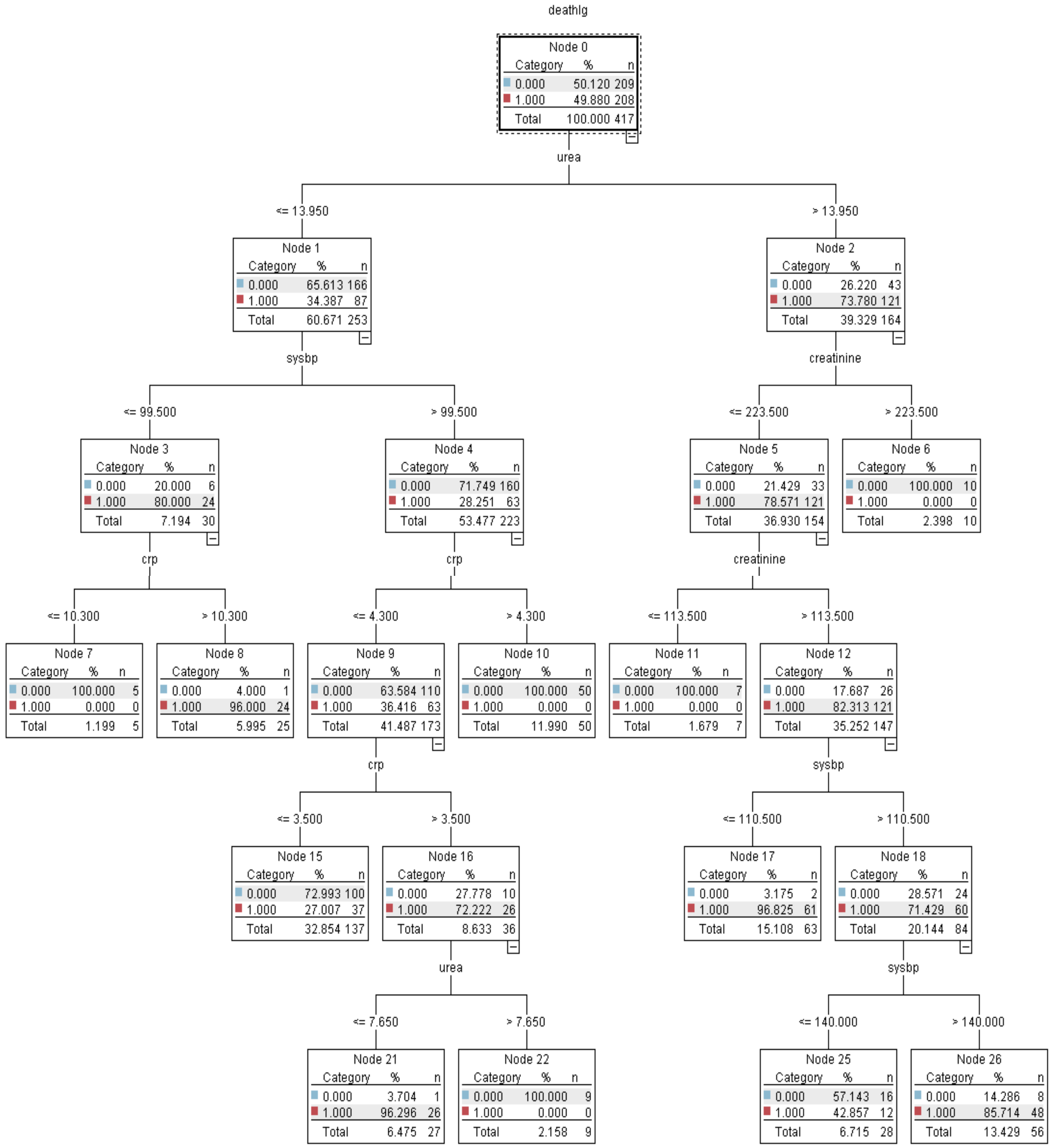


Figure 3.11 Decision Tree II

3.5.3 Model III Evaluation

Like the previous cases, this decision tree generated from Model III whose target variable shows if a patient has died during any time in his or her treatment has eleven terminal nodes and six levels (See Decision Tree III). Again, Model III uses the converted data set that is explained in the data understanding section.

The root split of this tree is on meansato2, which indicates the absolute difference between percentage of oxygen saturation found in the patient’s blood and the mean of the column involving these percentages (95.2%). Thus the oxygen saturation can be considered as the most important variable when the absolute differences between each predictor and its corresponding column mean are taken into account. On the other hand, the second split is made according to either C-reactive protein test score or urea level. Finally, terminal nodes of this decision tree and their characteristics are listed in the below table:

TN no	**Node no**	**deathlg (%)**	**Number death**	**N**	**Rules**
1	3	0	0	25	(meansato2 ≤ 2) and (meancrp ≤ 2.55)
2	6	90.476	38	42	(meansato2 > 2) and (meanurea > 9.8)
3	7	0	0	19	(meansato2 ≤ 2) and (meancrp > 2.55) and (meancreatinine ≤ 15.5)
4	9	0	0	12	(meansato2 > 2) and (meanurea ≤ 9.8) and (meanhb ≤ 0.4)
5	13	0	0	14	(meansato2 ≤ 2) and (meancrp > 2.55) and (meancreatinine ≤ 15.5) and (meansysbp ≤ 9)
6	19	0	0	11	(meansato2 ≤ 2) and (meancrp > 2.55) and (meancreatinine ≤ 15.5) and (meansysbp > 9) and (meansato2 ≤ 0.5)
7	20	60	54	90	(meansato2 ≤ 2) and (meancrp > 2.55) and (meancreatinine ≤ 15.5) and (meansysbp > 9) and (meansato2 > 0.5)
8	21	85.54	52	63	(meansato2 > 2) and (meanurea ≤ 9.8) and (meanhb > 0.4) and (meandiasbp ≤ 7.5) and (meanrr ≤ 6)
9	22	0	0	6	(meansato2 > 2) and (meanurea ≤ 9.8) and (meanhb > 0.4) and (meandiasbp ≤ 7.5) and (meanrr > 6)
10	23	32.877	24	73	(meansato2 > 2) and (meanurea

					≤ 9.8) and (meanhb > 0.4) and (meandiasbp > 7.5) and (meangcs ≤ 3.28)
11	24	100	19	19	(meansato2 > 2) and (meanurea ≤ 9.8) and (meanhb > 0.4) and (meandiasbp > 7.5) and (meangcs > 3.28)
Total	**--**	**--**	**187**	**374**	

Table 3.14 Terminal nodes of Model III with percentages of the events

The rules generated from Model III can be outlined as follows:

- It is clear from Terminal Node 1, if a patient's oxygen saturation percentage is between 93.2% and 97.2% and his or her C-reactive test score is between 3.85 mg/L and 6.95 mg/L include, and then he or she is expected to live.
- In the cases where a patient's oxygen saturation is less than 93.2% or higher than 97.2% and his or her urea level is less than 2.5 mg/L or higher than 22.1 mg/L, and then he or she is expected to die with around 0.9 probability.
- For creatinine, haemoglobin and systolic blood pressure predictors, the higher the absolute difference between the patient's value and the associated mean is, the higher death risk that appears. Indeed, when the absolute difference between a patient's creatinine level and the associated mean, which is 124 mmol/L in this case, is less than 15.5 mmol/L, the patient's chance of living is quite high. However, such difference becomes more than the threshold, 15.5 mmol/L, that chance starts to decrease dramatically (See Nodes 7th and 8th). Similarly, when a patient's haemoglobin value goes far from the patient's haemoglobin mean,12.28 g/dL, the death risk increases for that person (See Nodes 9th and 10th).
- Comparing with the other decision trees obtained in the previous sections, in this decision tree generated from the converted dataset, the distribution of the event to the nodes is much more homogenous. That is, patients with a high chance of dying have accumulated at certain nodes. Hence the generalisation ability of the model is quite high.

See Figure 3.12 for Decision Tree III

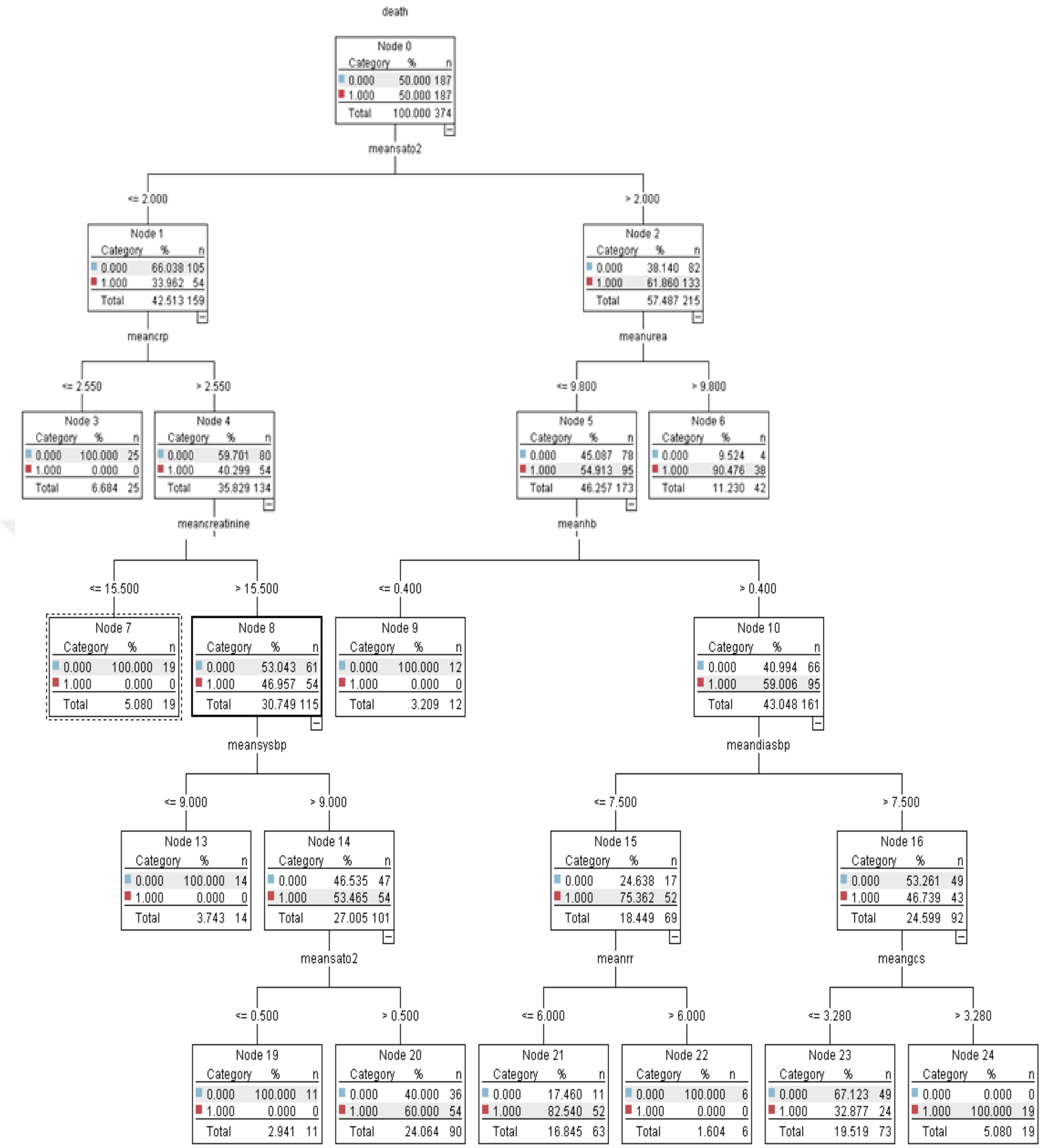


Figure 3.12 Decision Tree III

3.6 Conclusion

In this case study, we have constructed three different CART trees, namely Model I, Model II and Model III. Model I and Model II use the original dataset, while Model III uses the converted dataset. Among these models, Model I has the highest accuracy and the highest gain. Moreover, it is difficult to say that Model III's predictive performance is adequate.

Thanks to the models we have obtained some novel insights about the conditions that affect the mortality risk of patients over 90 years old. In this sense, we have discovered that the patients with renal failures constitute the most important risk group. Indeed, urea level plays the most important role, when determining a patient's mortality risk and high creatinine level is one of the most significant indicators for patients over 90 years old. Additionally, high systolic blood pressure (> 140 mmHg) is the other remarkable consideration which should be taken into account when estimating the mortality rate for a patient. In fact, a patient with relatively high urea level (> 13.95 mg/L), high creatinine level (> 113.5 mmol/L) and high systolic blood pressure (> 140 mmHg) is very likely to die (with 0.85 probability).

3.7 Future Work

There are several directions for future work. Firstly, instead of the default values of the CART algorithm explained in section 3.4.2, the parameters of the algorithm can be tuned to enhance the accuracy measures of the models. In this sense, for example, tuning the minimum change in impurity and/or the impurity measure for categorical target and/or the minimum record in a child branch may be useful.

Secondly, the original data set consist of 97 fields. However, in the feature subset selection of the preprocessing step, we have reduced that number to sixteen in the direction of the views of the medical expert from a practical point of view. At this point, the discarded fields can be included to further data mining applications in an attempt to find much more patterns in the data set.

Finally, as a practical application of this work, developing software based on the finding of this data mining application will be non-trivial extension.

Appendix A. Data Dictionary for the Oldest Old Dataset

The Oldest Old Dataset was derived from the gerontology department of three hospitals, namely Norwich Norfolk University Hospital (NNUH), Aberdeen Hospital and Woodend Hospital-Aberdeen. It consists of the clinical records of 393 patients who are over 90 years old and admitted to the hospitals in between 01/11/2008 to 06/06/2009. Each patient was recorded with 97 input variables including both categorical and quantitative values.

Data Accessibility: http://www.filefactory.com/file/b36d36e/n/OldestOld.xlsx

Number of Instances: 393

Number of Attributes: 16

Target Attribute: death, deathlg

1) **Data Field Name:** albumin

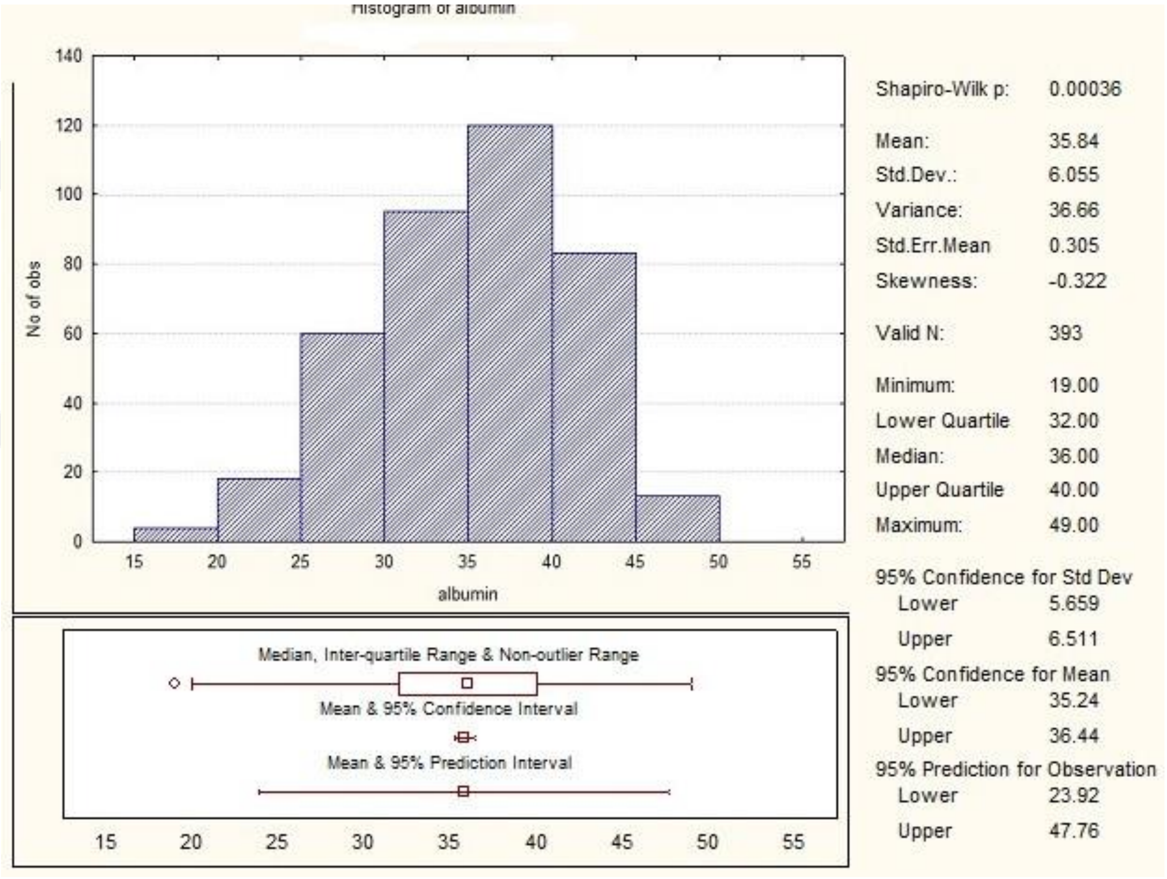
Short Description: Albumin test score. An albumin test measures the amount of albumin protein in the clear liquid portion of the blood, and it is used to check if the patient has a liver or kidney problems. The normal range is 35-50 g/L (1).

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 393

Missing N: 0



Shapiro-Wilk p:	0.00036
Mean:	35.84
Std.Dev.:	6.055
Variance:	36.66
Std.Err.Mean	0.305
Skewness:	-0.322
Valid N:	393
Minimum:	19.00
Lower Quartile	32.00
Median:	36.00
Upper Quartile	40.00
Maximum:	49.00
95% Confidence for Std Dev	
Lower	5.659
Upper	6.511
95% Confidence for Mean	
Lower	35.24
Upper	36.44
95% Prediction for Observation	
Lower	23.92
Upper	47.76

2) **Data Field Name:** creatinine

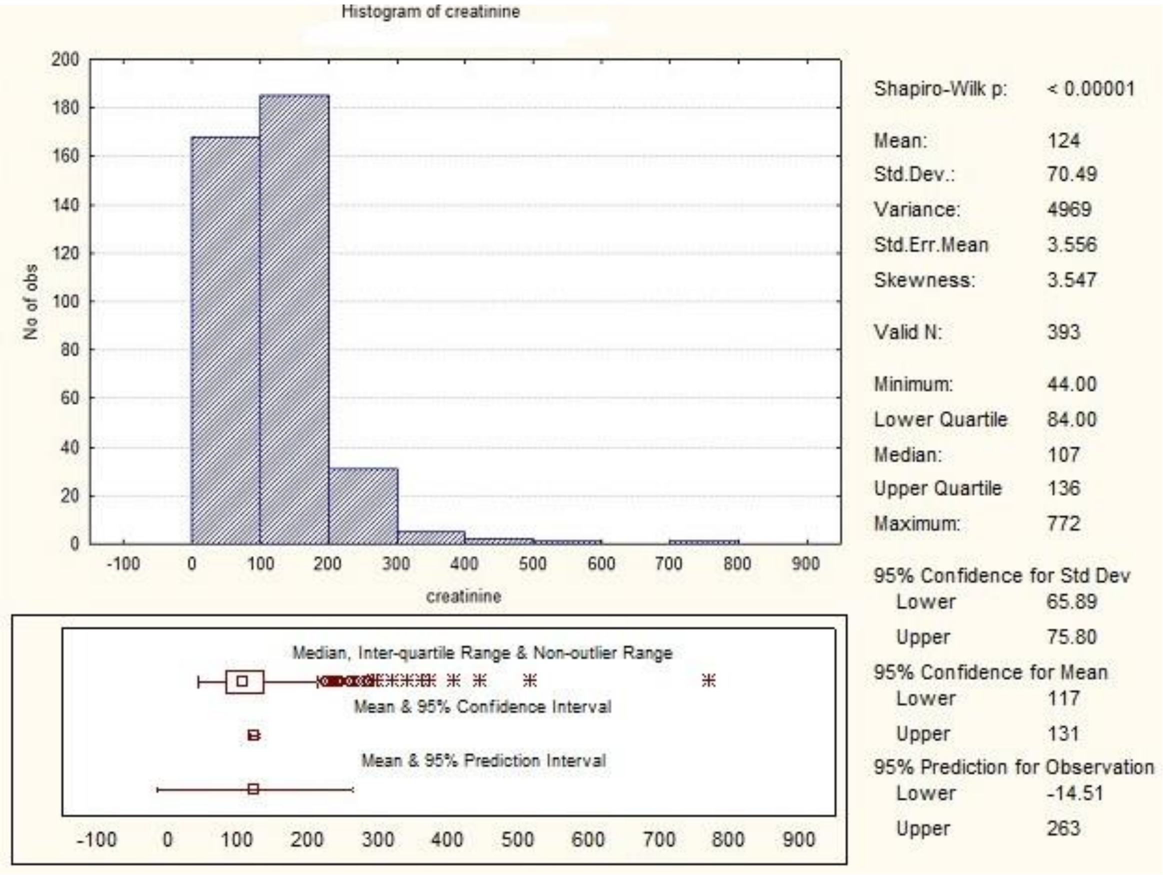
Short Description: Creatinine test score. The normal level of creatinine is usually 60 to 110 micromoles per litre.

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 393

Missing N: 0



Shapiro-Wilk p:	< 0.00001
Mean:	124
Std.Dev.:	70.49
Variance:	4969
Std.Err.Mean	3.556
Skewness:	3.547
Valid N:	393
Minimum:	44.00
Lower Quartile	84.00
Median:	107
Upper Quartile	136
Maximum:	772
95% Confidence for Std Dev	
Lower	65.89
Upper	75.80
95% Confidence for Mean	
Lower	117
Upper	131
95% Prediction for Observation	
Lower	-14.51
Upper	263

3) **Data Field Name:** crp

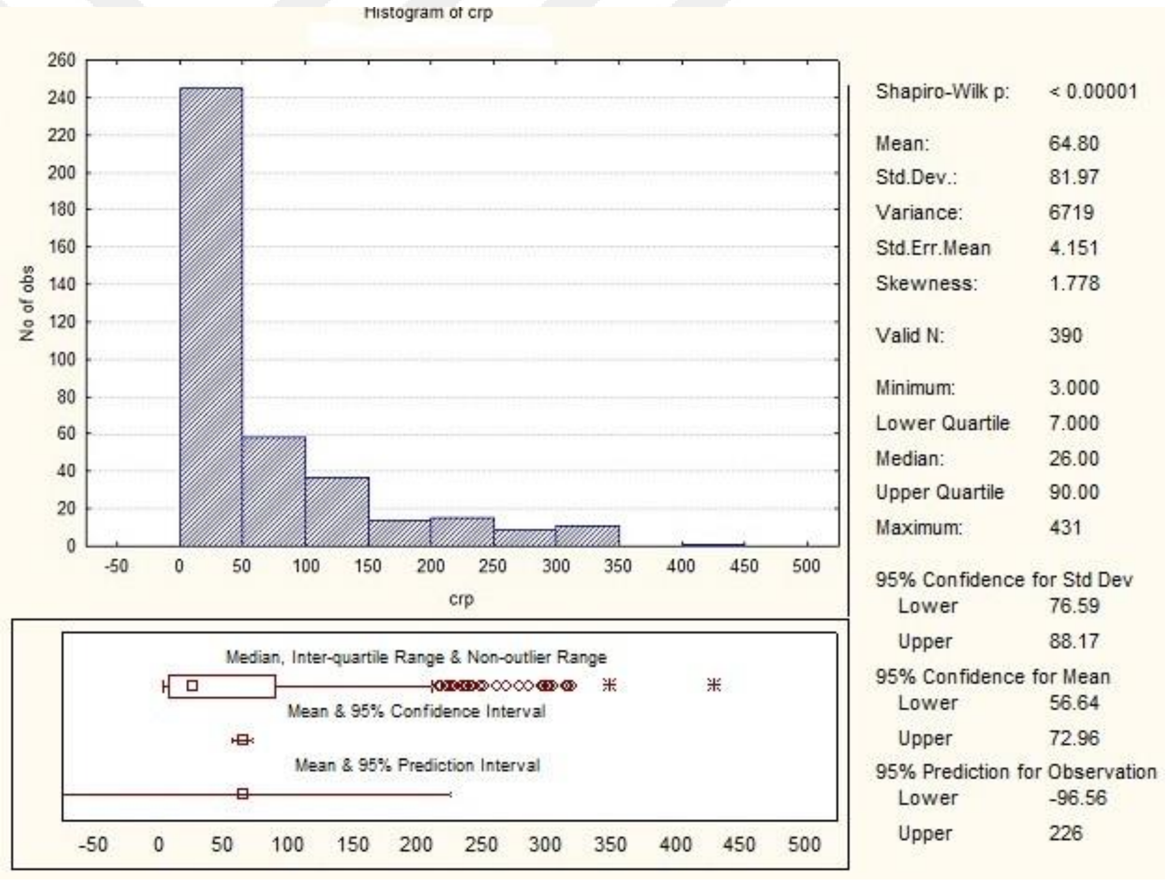
Short Description: C-reactive protein test score. C-reactive protein is a serum protein, produced by the liver and rises when an inflammation is occurring. CRP test can use to monitor heart attack risks, inflammatory bowel disease, arthritis and infection, normal range is between 10mg/L and 30 mg/L (2).

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 390

Missing N: 3 (0.77%)



```
Shapiro-Wilk p:          < 0.00001

Mean:                    64.80
Std.Dev.:                81.97
Variance:                6719
Std.Err.Mean             4.151
Skewness:                1.778

Valid N:                 390

Minimum:                 3.000
Lower Quartile           7.000
Median:                  26.00
Upper Quartile           90.00
Maximum:                 431

95% Confidence for Std Dev
   Lower                 76.59
   Upper                 88.17
95% Confidence for Mean
   Lower                 56.64
   Upper                 72.96
95% Prediction for Observation
   Lower                 -96.56
   Upper                 226
```

4) **Data Field Name:** diasbp

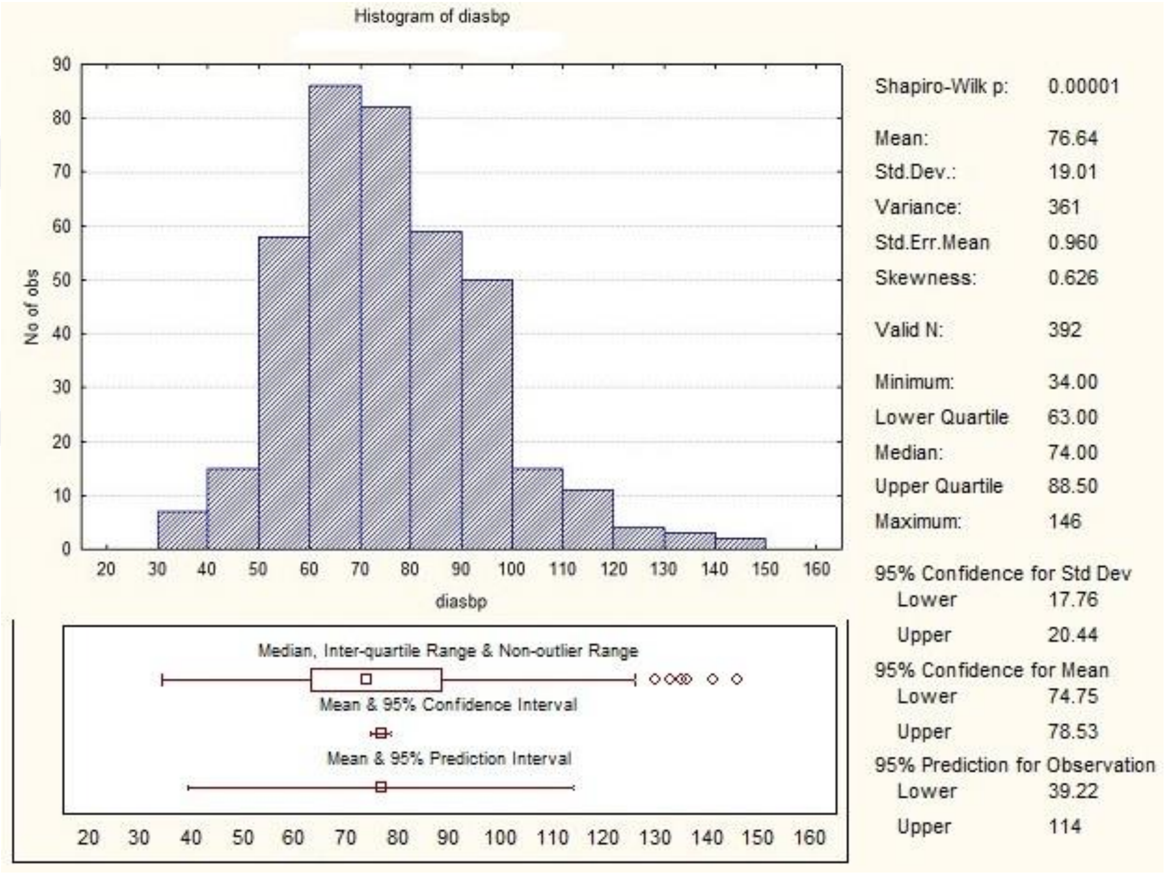
Short Description: Diastolic blood pressure of the patient, which is the pressure in the blood vessels, when the heart is resting between beats. The adult would be expected to have a diastolic pressure of around 80mmHg.

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 392

Missing N: 1 (0.3%)



```
Shapiro-Wilk p:     0.00001

Mean:               76.64
Std.Dev.:           19.01
Variance:           361
Std.Err.Mean        0.960
Skewness:           0.626

Valid N:            392

Minimum:            34.00
Lower Quartile      63.00
Median:             74.00
Upper Quartile      88.50
Maximum:            146

95% Confidence for Std Dev
   Lower            17.76
   Upper            20.44
95% Confidence for Mean
   Lower            74.75
   Upper            78.53
95% Prediction for Observation
   Lower            39.22
   Upper            114
```

5) **Data Field Name:** gcs

Short Description: Glasgow coma score. This measurement is used to assess the depth of coma or unconsciousness suffered by an individual (3).

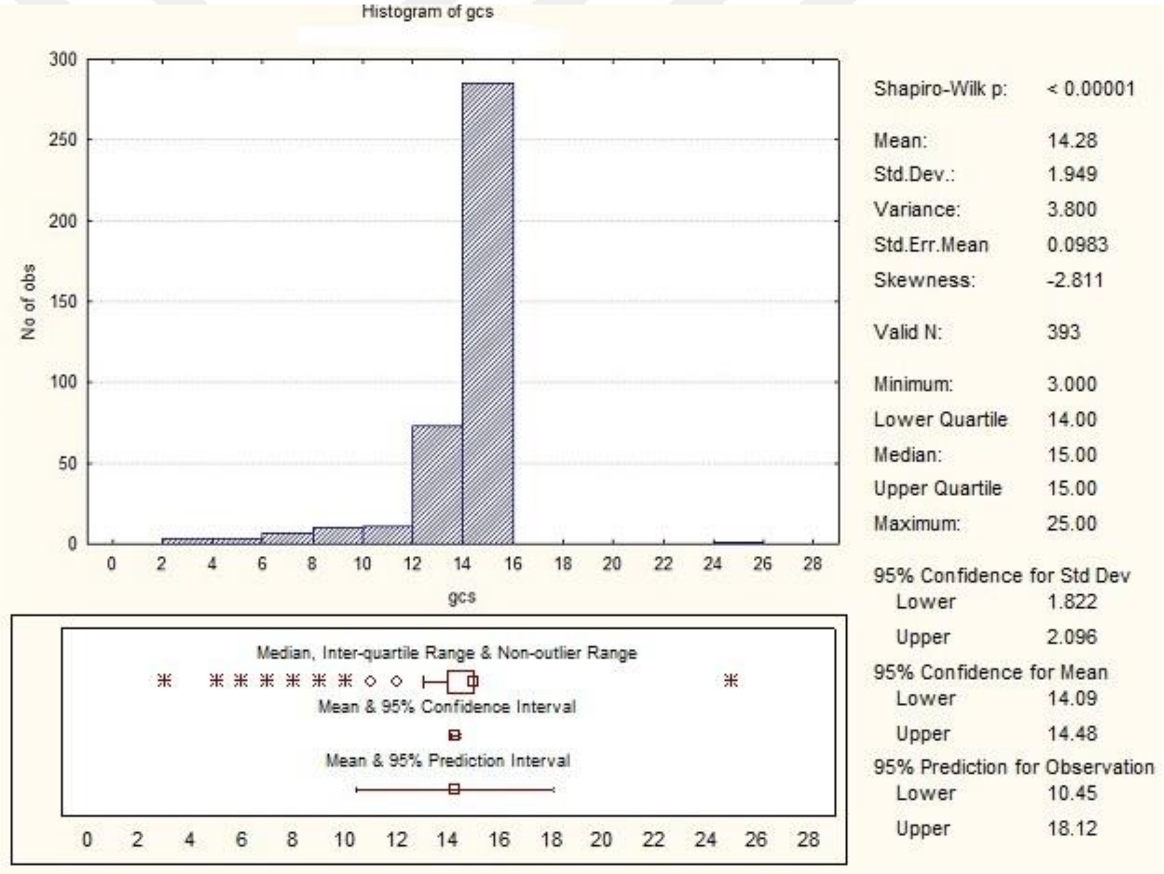
Glasgow coma score = (score for eye opening) + (score for best verbal response) + (score for best motor response). The minimum score is 3 which has the worst prognosis, whereas 15 is the maximum which has the best prognosis.

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 393

Missing N: 0



Shapiro-Wilk p:	< 0.00001
Mean:	14.28
Std.Dev.:	1.949
Variance:	3.800
Std.Err.Mean	0.0983
Skewness:	-2.811
Valid N:	393
Minimum:	3.000
Lower Quartile	14.00
Median:	15.00
Upper Quartile	15.00
Maximum:	25.00
95% Confidence for Std Dev	
Lower	1.822
Upper	2.096
95% Confidence for Mean	
Lower	14.09
Upper	14.48
95% Prediction for Observation	
Lower	10.45
Upper	18.12

6) **Data Field Name:** hb

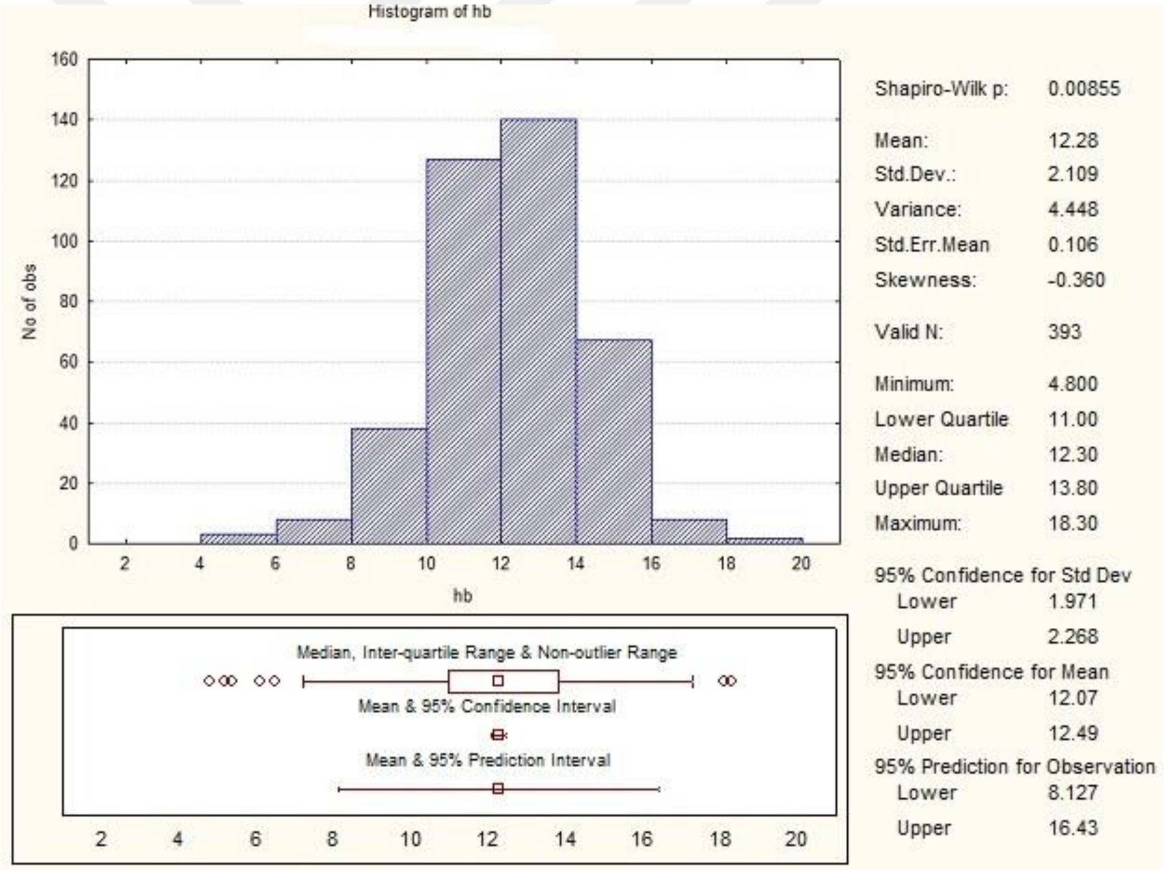
Short Description: Haemoglobin. Haemoglobin (Hb) is the oxygen caring pigment in the blood, and the result of its measurement is stated as the amount of haemoglobin in grams (gm) per decilitre (dl) of whole blood, a decilitre being 100 millilitres. The normal range is 12.4-14.9 g/dL for elderly females(over 65) and 11.7-13.8 g/dL for elderly males (over 65).

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 393

Missing N: 0



Shapiro-Wilk p:	0.00855
Mean:	12.28
Std.Dev.:	2.109
Variance:	4.448
Std.Err.Mean	0.106
Skewness:	-0.360
Valid N:	393
Minimum:	4.800
Lower Quartile	11.00
Median:	12.30
Upper Quartile	13.80
Maximum:	18.30
95% Confidence for Std Dev	
Lower	1.971
Upper	2.268
95% Confidence for Mean	
Lower	12.07
Upper	12.49
95% Prediction for Observation	
Lower	8.127
Upper	16.43

7) **Data Field Name:** pulse

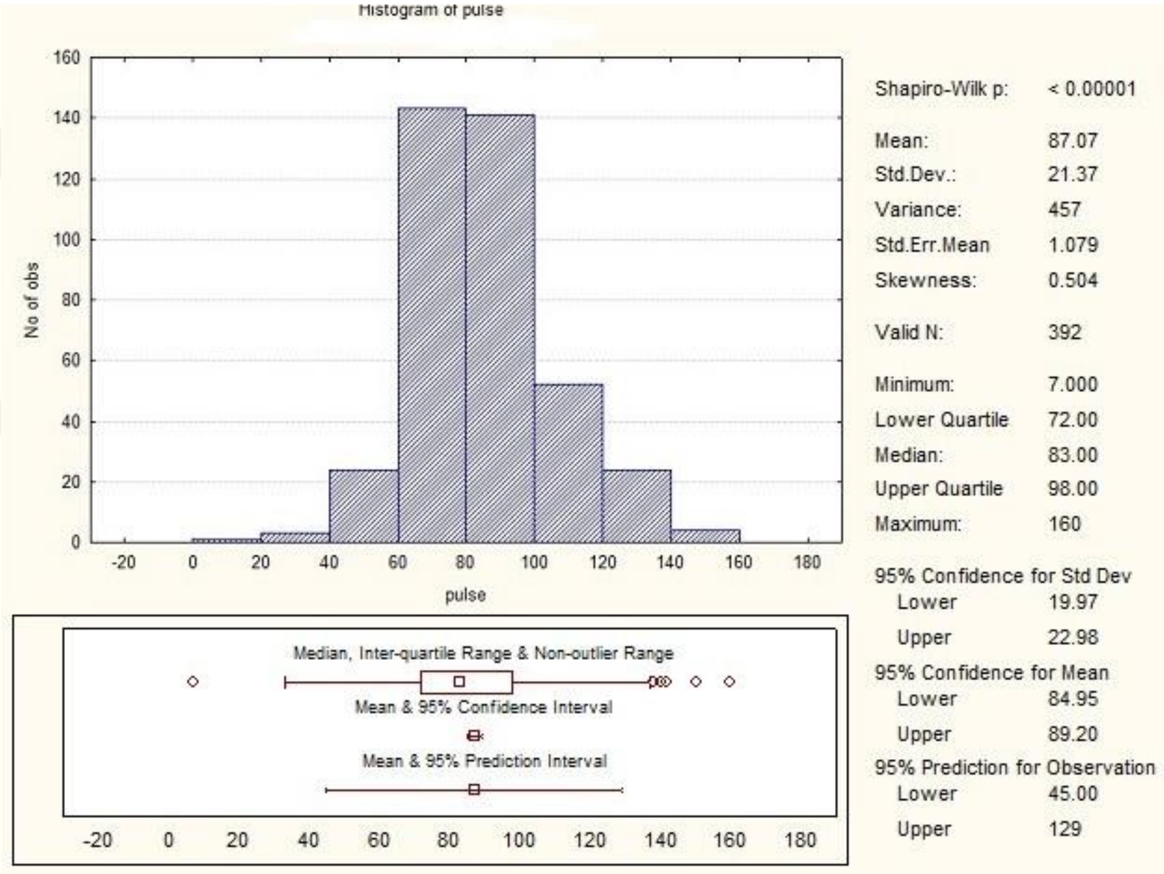
Short Description: The patient's pulse rate per minute. A normal pulse rate for a human, while resting, ranges from 60 to 100 beats per minute.

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 392

Missing N: 1 (0.3%)



```
Shapiro-Wilk p:     < 0.00001

Mean:               87.07
Std.Dev.:           21.37
Variance:           457
Std.Err.Mean        1.079
Skewness:           0.504

Valid N:            392

Minimum:            7.000
Lower Quartile      72.00
Median:             83.00
Upper Quartile      98.00
Maximum:            160

95% Confidence for Std Dev
   Lower            19.97
   Upper            22.98
95% Confidence for Mean
   Lower            84.95
   Upper            89.20
95% Prediction for Observation
   Lower            45.00
   Upper            129
```

8) Data Field Name: rr

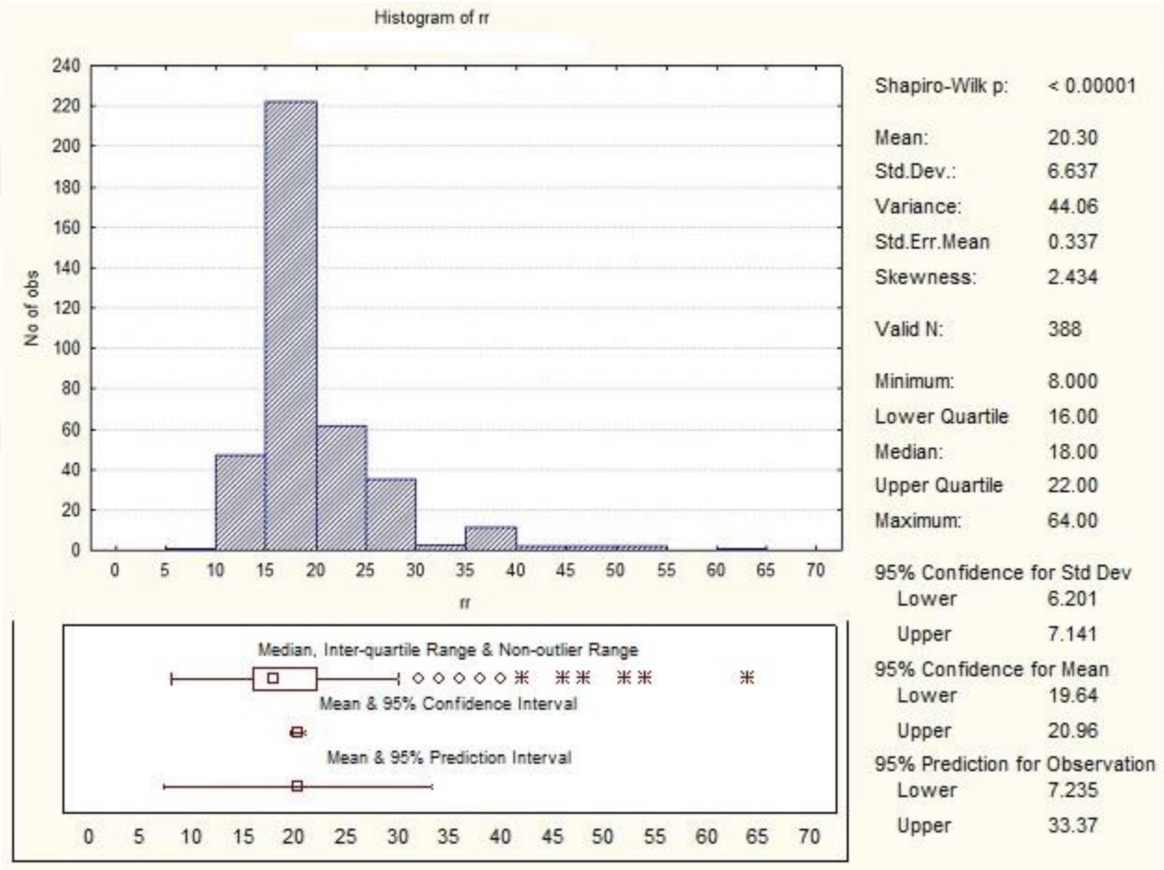
Short Description: The patient’s respiratory rate. The human respiration rate is the number of breath that the person takes in 1 minute, while resting. A normal respiratory rate for a people over 67 is between 16 to 25 breaths a minute.

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 392

Missing N: 1 (0.3%)



Shapiro-Wilk p:	< 0.00001
Mean:	20.30
Std.Dev.:	6.637
Variance:	44.06
Std.Err.Mean	0.337
Skewness:	2.434
Valid N:	388
Minimum:	8.000
Lower Quartile	16.00
Median:	18.00
Upper Quartile	22.00
Maximum:	64.00
95% Confidence for Std Dev	
Lower	6.201
Upper	7.141
95% Confidence for Mean	
Lower	19.64
Upper	20.96
95% Prediction for Observation	
Lower	7.235
Upper	33.37

9) **Data Field Name:** sato2

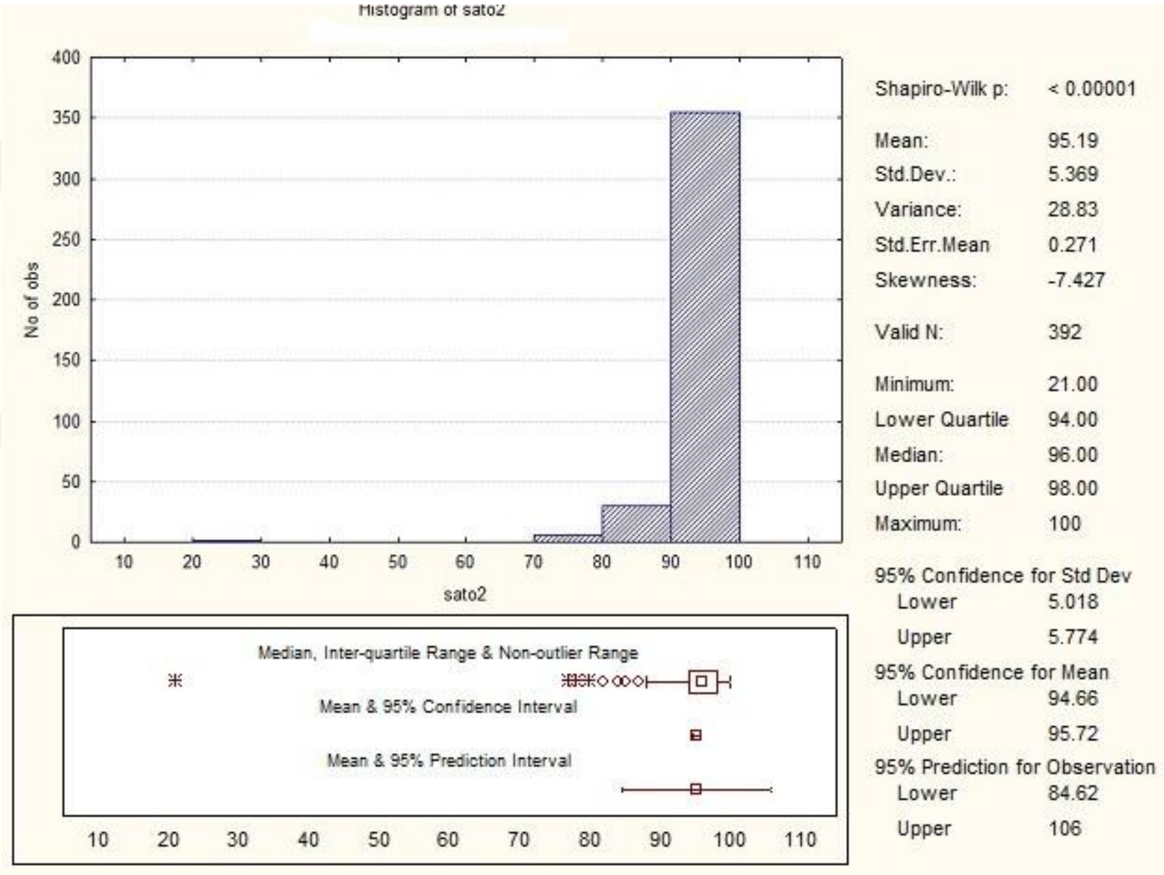
Short Description: Oxygen saturation. Oxygen saturation refers the percentage of oxygen found in the patient's blood which ranges between 96 and 100 percent for a healthy person.

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 392

Missing N: 1 (0.3%)



Shapiro-Wilk p: < 0.00001

Mean: 95.19
Std.Dev.: 5.369
Variance: 28.83
Std.Err.Mean 0.271
Skewness: -7.427

Valid N: 392

Minimum: 21.00
Lower Quartile 94.00
Median: 96.00
Upper Quartile 98.00
Maximum: 100

95% Confidence for Std Dev
Lower 5.018
Upper 5.774
95% Confidence for Mean
Lower 94.66
Upper 95.72
95% Prediction for Observation
Lower 84.62
Upper 106

10) **Data Field Name:** sodium

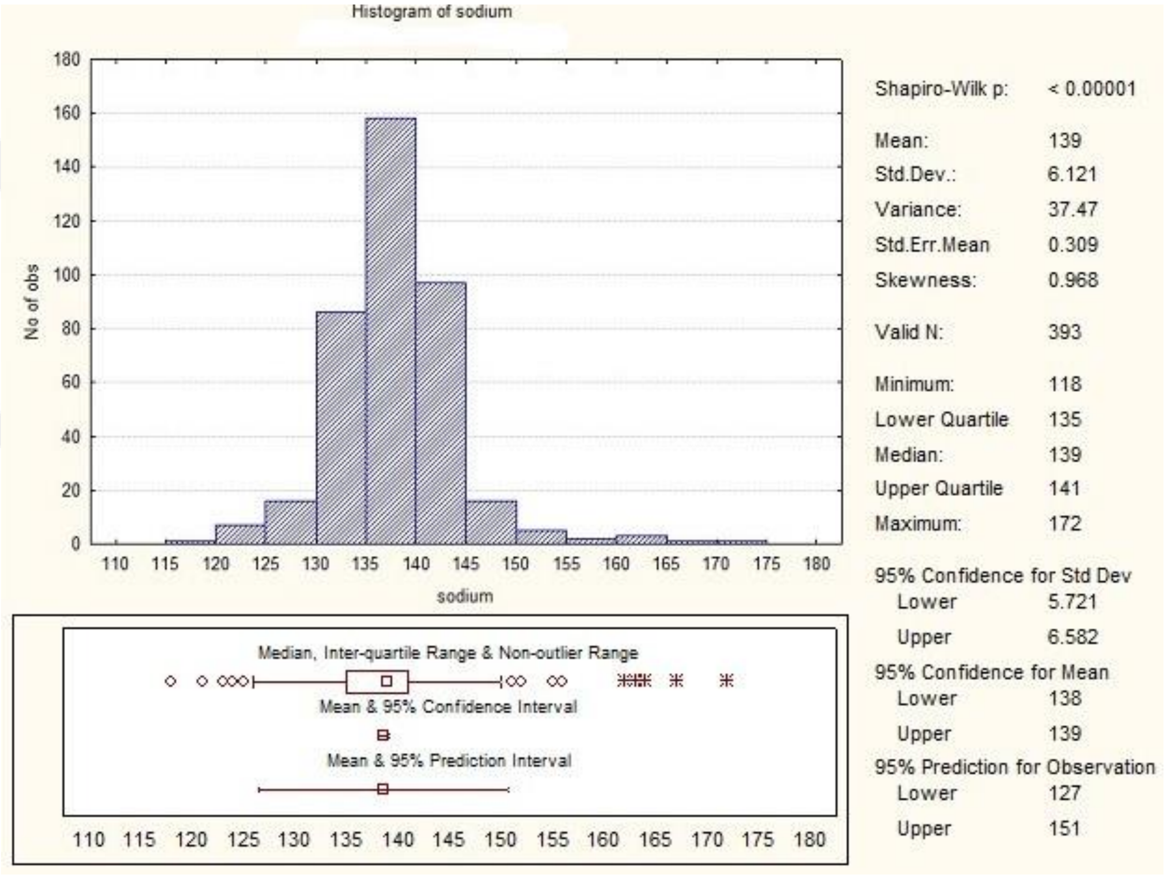
Short Description: The amount of sodium in the blood. Normal serum sodium levels are between 135- 145 milliequivalents per litre (mEq/L).

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 393

Missing N: 0



Shapiro-Wilk p:	< 0.00001
Mean:	139
Std.Dev.:	6.121
Variance:	37.47
Std.Err.Mean	0.309
Skewness:	0.968
Valid N:	393
Minimum:	118
Lower Quartile	135
Median:	139
Upper Quartile	141
Maximum:	172
95% Confidence for Std Dev	
Lower	5.721
Upper	6.582
95% Confidence for Mean	
Lower	138
Upper	139
95% Prediction for Observation	
Lower	127
Upper	151

11) **Data Field Name:** sysbp

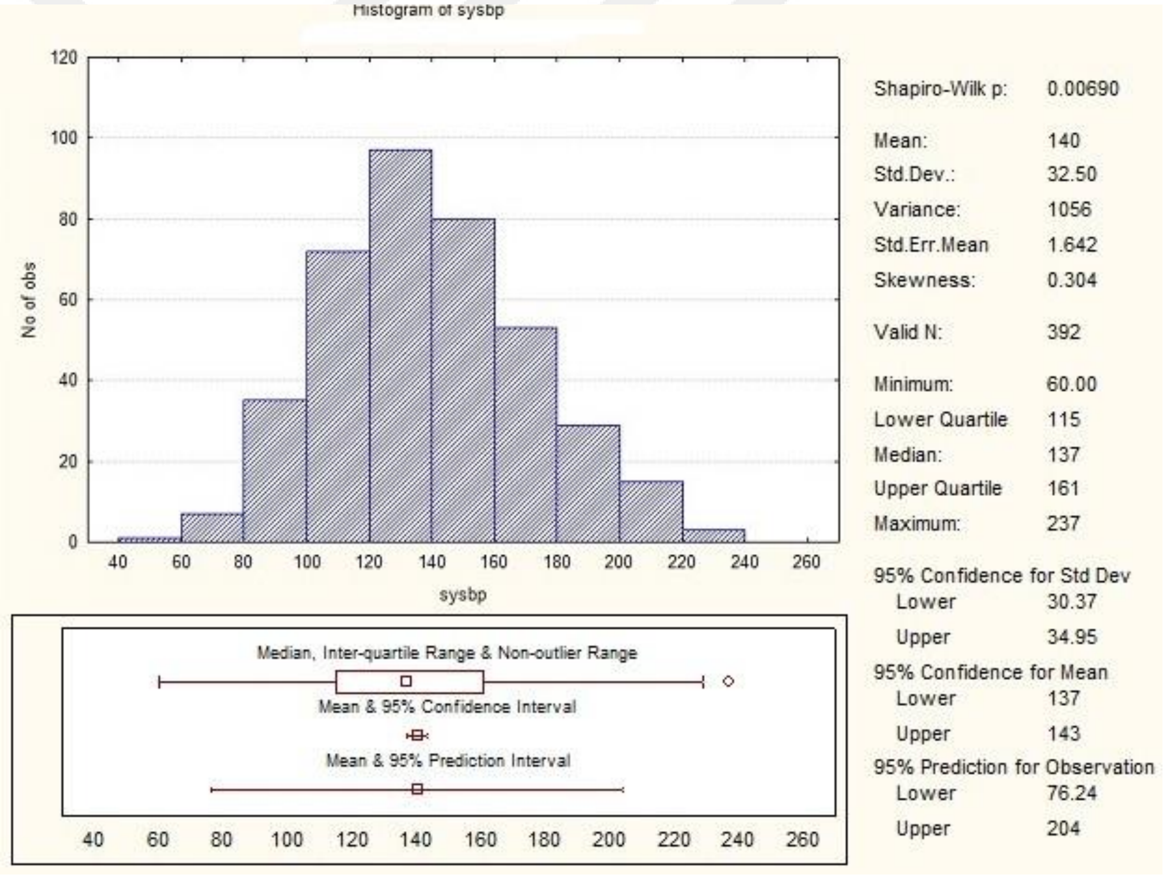
Short Description: Systolic blood pressure of the patient, which is the pressure in the arteries, when the heart contacts to pump the blood to the body. The adult would be expected to have a systolic pressure of about 120mmHg, but the pressure generally rises with age as the arteries get thicker and harder (4) (3).

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 392

Missing N: 1 (0.3%)



Shapiro-Wilk p:	0.00690
Mean:	140
Std.Dev.:	32.50
Variance:	1056
Std.Err.Mean	1.642
Skewness:	0.304
Valid N:	392
Minimum:	60.00
Lower Quartile	115
Median:	137
Upper Quartile	161
Maximum:	237
95% Confidence for Std Dev	
Lower	30.37
Upper	34.95
95% Confidence for Mean	
Lower	137
Upper	143
95% Prediction for Observation	
Lower	76.24
Upper	204

12) **Data Field Name:** temp

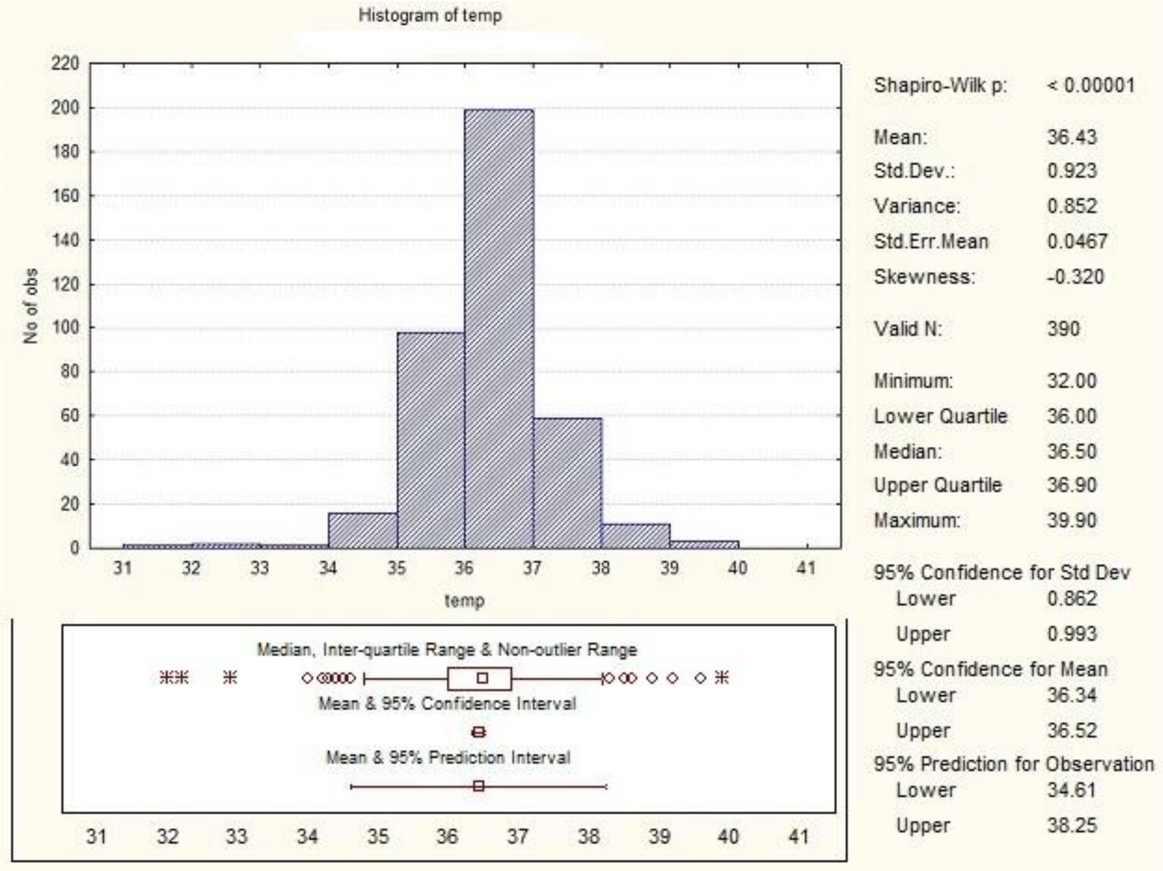
Short Description: Temperature of the patient (ºC). The normal range of human body varies between 36.3 and 37.1 degrees Celsius.

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 390

Missing N: 3 (0.8%)



Shapiro-Wilk p:	< 0.00001
Mean:	36.43
Std.Dev.:	0.923
Variance:	0.852
Std.Err.Mean	0.0467
Skewness:	-0.320
Valid N:	390
Minimum:	32.00
Lower Quartile	36.00
Median:	36.50
Upper Quartile	36.90
Maximum:	39.90
95% Confidence for Std Dev	
Lower	0.862
Upper	0.993
95% Confidence for Mean	
Lower	36.34
Upper	36.52
95% Prediction for Observation	
Lower	34.61
Upper	38.25

13) **Data Field Name:** urea

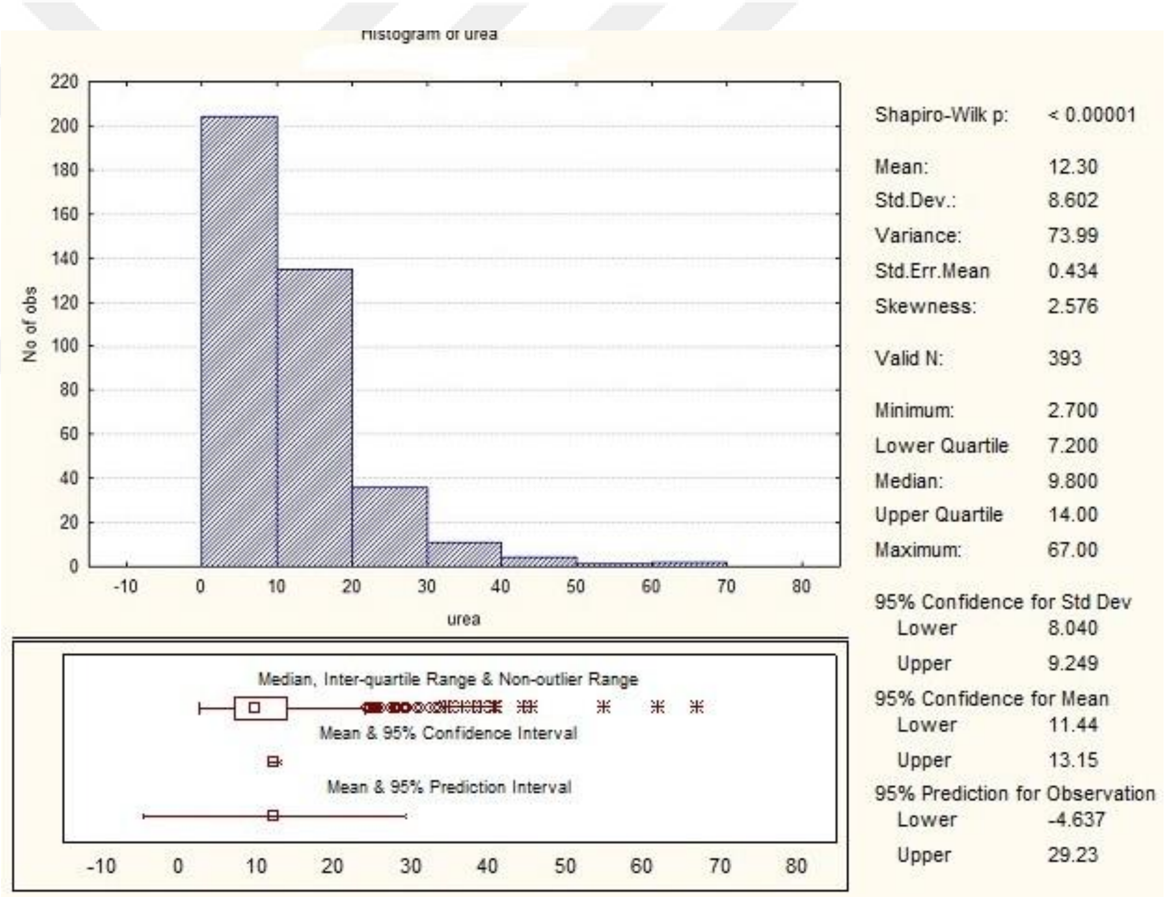
Short Description: Blood urea nitrogen (BUN) score. The normal range for blood urea nitrogen is 7 to 20 mg/dL and the score higher than the upper limit reveals there may be a problem with the patient's kidney, whereas the lower score may be a sign of liver damage (2).

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 393

Missing N: 0



Shapiro-Wilk p:	< 0.00001
Mean:	12.30
Std.Dev.:	8.602
Variance:	73.99
Std.Err.Mean	0.434
Skewness:	2.576
Valid N:	393
Minimum:	2.700
Lower Quartile	7.200
Median:	9.800
Upper Quartile	14.00
Maximum:	67.00
95% Confidence for Std Dev	
Lower	8.040
Upper	9.249
95% Confidence for Mean	
Lower	11.44
Upper	13.15
95% Prediction for Observation	
Lower	-4.637
Upper	29.23

14) **Data Field Name:** wcc

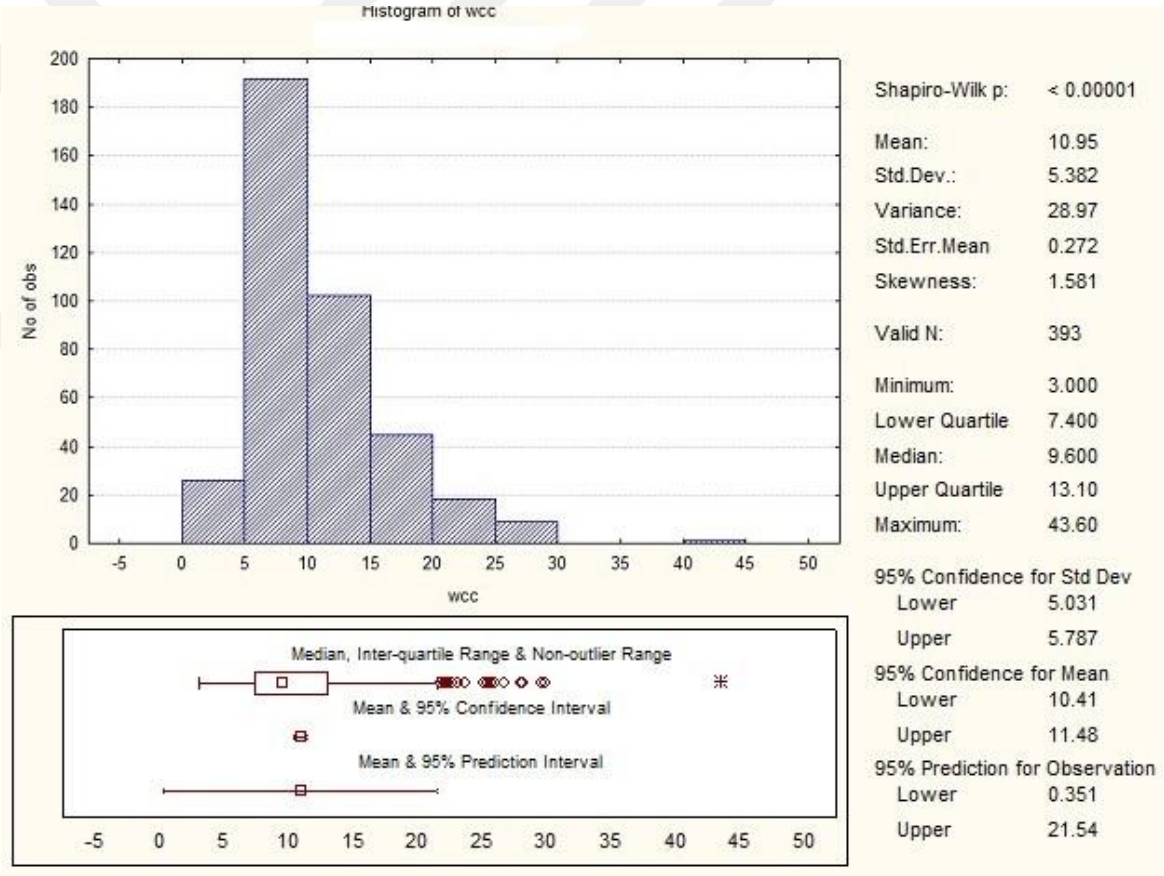
Short Description: White blood cell count. White blood cell (also known leukocyte) is any blood cell that contains a nucleus, and its normal range is between 4,000 and 11,000 per cubic millimetre of blood. The amount of white cell tends to increase in case of surgery or inflammation (4).

Field Type: Quantitative variable

Usage type: Input variable

Valid N: 393

Missing N: 0



Shapiro-Wilk p:	< 0.00001
Mean:	10.95
Std.Dev.:	5.382
Variance:	28.97
Std.Err.Mean	0.272
Skewness:	1.581
Valid N:	393
Minimum:	3.000
Lower Quartile	7.400
Median:	9.600
Upper Quartile	13.10
Maximum:	43.60
95% Confidence for Std Dev	
Lower	5.031
Upper	5.787
95% Confidence for Mean	
Lower	10.41
Upper	11.48
95% Prediction for Observation	
Lower	0.351
Upper	21.54

15) **Field Name:** death

Short Description: This variable indicates if the patient has died during any time in his or her the treatment or not. 1=yes, 0=No

Field Type: Dummy variable

Usage type: Output variable

Valid N: 393

Missing N: 0

death

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	322	81.9	81.9	81.9
	1	71	18.1	18.1	100.0
	Total	393	100.0	100.0	

16) **Field Name:** deathlg

Short Description: This variable indicates if the patient has died after a long stay, i.e. any stay equal to or greater than a week in the hospital. 1=yes, 0=no

Field Type: Dummy variable

Usage type: Output variable

Valid N: 393

Missing N: 0

deathlg

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	366	93.1	93.1	93.1
	1	27	6.9	6.9	100.0
	Total	393	100.0	100.0	

References

1. **A service of the U.S. National Library of Medicine.** MedlinePlus. [Online] July 4, 2010. http://www.nlm.nih.gov/medlineplus/.

2. **MayoClinic.** Diseases and Conditions. [Online] [Cited: July 4, 2010.] http://www.mayoclinic.com/health/DiseasesIndex/DiseasesIndex.

3. **Marcovitch, H.** *Black's Student Medical Dictionary.* s.l. : A & C Black Publishers Ltd, 2007. ISBN 13: 9780713687620.

4. **Martin, E A.** *Concise Colour Medical Dictionary.* s.l. : Oxford University Press, 2010. ISBN 13: 9780199557158.

Appendix B. Dictionary for the Conditions and treatments located in the Oldest Old Dataset

AAA: (Abbreviation for Abdominal Aortic Aneurysm) is a localised ballooning of the abdominal aorta due to progressive weakening of the aortic wall, may cause abdominal, back or side pain. **Abdominal discomfort**: Discomfort or tenderness in anywhere between the chest and the groin (1).

Abdominal pain: (also known stomach ache) is a sensation of discomfort or distress in anywhere between your chest and groin (1) (2).

Abdominal swelling: Swelling or bloating of the abdomen may be due to the accumulation of trapped intestinal contents within the bowel, the presence of asides within the abdomen.

Abnormal LFT's: is a condition in which there is an abnormal liver function.

Accidental fall: a fall due to slipping or tripping leading to injury (3).

Acidosis: is an abnormal condition in which acidity in the blood is increased, results in reduced alkalinity of the blood and tissues, so vomiting, general lassitude, thirst, can be observed (4) (5).

ACS: (Abbreviation for acute coronary syndrome) a form of chest pain owing to reduced oxygen supply to the heart muscle (6).

Acute abdomen: a sudden, severe abdominal pain due to inflammation, perforation, obstruction, infarction, or rupture of abdominal organs, and usually requiring emergency surgical intervention (7).

Acute Glaucoma: (also known angle-closure glaucoma) is a severe glaucoma which may result in permanent blindness, refers to a narrowing of the angle between your iris and sclera (8).

Acute on chronic renal failure: Any sudden onset kidney failure in patients with known chronic kidney failure (9).

Acute renal failure (ARF): is a sudden loss of kidney functions.

Acute pancreatitis: acute inflammation of the pancreas, leads to sudden and severe abdominal pain and can be life-threatening.

AF (Abbreviation for atrial fibrillation): is a common heart arrhythmia (abnormal heart rhythm) with rapid beating in upper chambers (atria) (10).

Agitation: An unpleasant state of intense arousal, increased tension and irritability (5).

Alcohol abuse: is a psychiatric diagnosis describing excessive use of alcoholic beverages (10).

Alzheimer's Disease: the most common form of the dementia. In Alzheimer's disease, healthy brain tissue degenerates so, the patient can suffer from memory loss, problems with abstract thinking, difficulty finding the right word, disorientation and loss of judgement (11).

Anaemia: is a condition in which the number of red blood cells or their oxygen-carrying capacity is inadequate to meet physiologic needs (5).

Angina: A severe, often constricting chest pain that can occur when the heart muscles is not receiving inadequate oxygen (12).

Anorexia: Loss of appetite especially when prolonged (5).

Aortic stenosis (AS): Pathologic narrowing of the aortic valve, which controls the direction of the blood from the left ventricle to the aorta, hence decreasing cardiac output (12) (13).

Aphasia: Inability to speak result from disease or injury to the cerebral cortex in the left half of the brain (for right-handed person) (4).

AR (Abbreviation for aortic regurgitation) (also known aortic insufficiency) is the leaking of the aortic valve of the heart that leads to the flow of blood from aorta back into the left ventricle during diastole (14).

Arrhythmia: an irregularity of the heartbeat.

Arthritiss: Inflammation of a joint or a state caused by Inflammation of joints (12).

Aspiration pneumonia: refers to inappropriate passage of oral or gastric contents, for instance food, water, salvia, or another foreign material into the lungs (3).

Atrial flutter: is a heart arrhythmia occurring in the atria in the heart where atria beat more often than the ventricles.

Back pain: A chronic or acute pain located in the back, usually below the cervical level (3).

Barrett's oesophagus: refers to the inflammation of the oesophagus usually due to repeated exposure to stomach acid, and usually seen in the reflux disease **(**11**).**

Bladder cancer: is a type of cancer occurring in the bladder, generally affects older adults and may lead to blood urine, frequent and painful urination and abdominal pain (11).

Blocked catheter: a blockage in the urinary catheter designed to be passed through the urethra into the bladder to drain it of urine (12).

Bone mets (bone metastases) : Cancer that has spread from a primary tumour to the bone, causes severe pain, bone fractures, spinal cord compression and hypercalemia (15).

Bowel cancer: is an umbrella term referring to the cancer that begins in the bowel, usually in the large bowel. Depending on the starting point, bowel cancer can sometimes be referred to as colon cancer, or rectal cancer (8).

BPH: (Abbreviation for benign prostatic hyperplasia) refers to the increase in size of the prostate gland to compress the urethra and lead to some overt urinary obstruction **(**10**).**

BPPV: (Abbreviation for benign paroxysmal positional vertigo) brief attacks of vertigo and nystagmus occurring when the head is placed in certain positions due to problems in the inner ear (16).

Bradycardia: Slowness of the heartbeat that is less than 60 beats per minute.

Bronchiectasis: is a chronic inflammatory of one or more bronchi or bronchioles resulting from dilatation and loss of elasticity of the bronchial walls (5).

Bronchospasm: refers to the contraction of the muscles in the bronchial tubes, causing breathing difficulty, usually associated with asthma and bronchitis.

Bullous pemphigoid (BP)is a chronic, autoimmune, blistering skin disease, sometimes involves mucous membranes and generally observed in people over 70.

Cataract: opacity of the lens due to the cloudy area on the cornea sufficient to cause visual impairment (4).

C Diff diarrhoea (Clostridium difficile diarrhoea) a type of diarrhoea caused by clostridium difficile bacterium and it is communicable.

Caecal tumour: a tumour located in the cecum which is a blind pouch-like commencement of the colon in the right lower quadrant of the abdomen at the end of the small bowel (3).

Cardiac event: usually refers to an heart attack (if troponin I is >=0.04) or angina (if troponin I is <0.04

Cardiac syncope: is a kind of syncope due to heart-related conditions.

CCF: (Abbreviation for congestive cardiac failure) (also known heart failure) a condition where there is inadequate pumping of the heart resulting in an accumulation of fluid in the lungs caused by any structural or functional cardiac disorder (3) (13).

Cellulitis: An acute and diffuse inflammation of loose connective tissue which is most commonly seen as a result of infection of a wound (1).

Chest infection: is a bacterial or viral infection of the airways leading down into the lungs, or of the lungs themselves. Main types are acute bronchitis and pneumonia (3).

Chest infection: is a general term of bacterial or viral infection of the airways leading down into the lungs, or of the lungs themselves, such as pneumonia and bronchitis (9).

Chest pain: Pressure, pain in the chest.

Chest tightness: The sensation of tightness located in the chest (10).

Chesty: Having or relating to a lot of mucus in the lungs (2).

Cholangitis: inflammation of the bile ducts.

Cholecystitis: inflammation of the gallbladder storing bile produced in the liver, causes pain in the upper abdomen as well as nausea and vomiting. Cholecystitis most often result from the gallstones becoming stuck in the cystic duct (9).

Claudication: (Limping): A lame walk with a yielding step usually caused by an inadequate supply of blood to affected muscles (12) (17).

CLL: (Abbreviation for chronic lymphocytic leukaemia) is a type of leukaemia, which is a disorder characterised by an increased number of white blood cells, and progresses more slowly than other types of leukaemia (11).

Coagulopathy: a disease affecting the coagulability of the blood, so there is a defect in the blood clotting mechanism (12) (3).

Coeliac disease: is a condition in which where a person is intolerant to the protein gluten, in case of taking food containing gluten, the immune system attacks the gluten, which can lead to the small intestine becoming damaged (8).

Coffee ground vomiting: Ejecting vomitus which looks like ground coffee due to the blood cells in the vomitus.

Colitis: is inflammation of the colon resulting in diarrhoea, discharge of mucus and blood, cramping, and abdominal pain (10).

Collapse: To break down in vital energy, stamina, or self-control through exhaustion or disease (5).

Colostomy: is surgery to divert a section of colon and to attach an opening in the abdominal wall in order to allow faecal matter to exit.

Community-acquired Pneumonia (CAP): is pneumonia caused by an infection currently present in the community (15).

Complete Heart Block: (also referred third-degree heart block) (third degree AV block) complete block of electric impulses originating in the atrium or the sinus node preventing them from reaching the AV node and vertilices (12).

Confusion: A mental state characterized by bewilderment, emotional disturbance, lack of clear thinking, and perceptual disorientation (1).

Constipation: inability to empty the bowels.

COPD: (Abbreviation for chronic obstructive pulmonary disease) is a chronic, ongoing, progressive disease of the lower respiratory tract in the lungs (10).

Cough: An illness that makes you cough a lot (2).

CRF: Abbreviation for chronic renal failure.

Cushing's Syndrome: the condition caused by excess amount of corticosteroid hormones in the body, may result in weight gain, reddening of the face, excess growth of the body and facial hair and hypertension (18).

CVA: (cerebral vascular attack) A sudden, nonconvulsive loss of neurologic function due to an ischemic or hemorrhagic intracranial vascular event, occurring the one side of the brain, whereas effects the other side of the body (10).

Cyanosed: Having a bluish skin caused by lack of oxygen in the bloodstream.

Decreased mobility: it is the state where the patient moves sluggishly or cannot move.

Dehydration: is an abnormal condition in which there is a fall in the water content of the body (4).

Delirium: An acute mental disorder characterised by reduced ability to maintain attention to external stimuli and disorganised thinking (3).

Dementia: chronic and progressive deterioration of behaviour and some intellectual functions due to organic brain disease (18).

Depression: a mental state caused by excessive sadness (18).

Deteriorating PD: worsening of Parkinson's disease.

Diabetes mellitus (DM): (also known diabetes) a metabolic disease in which a person has high blood sugar because of the lack insulin or cells do not respond the insulin produced, results in frequent urination, increased thirst and increased hunger (13).

Diarrhoea: Diarrhoea is the condition of having 3 or more loose or liquid stools per day which is not normal for the individual (19).

Digoxin toxicity: is a type of poisoning, result from over-accumulation of digitalis glycosides in the body, inadequate renal functions can contribute to the illness as well. Symptoms range from nausea, vomiting to blurred vision and cardiac arrhythmias (3).

Diverticular disease: is one of the most common digestive condition characterised by the formation of abnormal pouches in colon. These pouches can become inflamed and lead to stomach pain and feeling bloated (10).

Diverticulitis: is a disease of large intestine that is due to inflammation of diverticulum, which is a sac or pouch made at weak points in the walls of the alimentary tract , characterised by infection and causes lower abdominal pain[18].

Dizziness: impairment in spatial perception and stability (7).

Double incontinence: Inability to control of evacuation of both faeces and urine.

Drop attack: is sudden spontaneous fall occurring during standing or falling with an unknown reason (12).

Drowsiness: A state of impaired awareness associated with a desire tendency to sleep (3).

DVT: (Abbreviation for deep vein thrombosis) comprising a blood clot in the veins of the inner thigh or leg, may cause pain and swelling in the leg and may lead to complications such as pulmonary embolism (3) (8).

Dysarthria: Incomplete articulation of speech owing to disturbances of muscular control which caused by the central or peripheral nervous sytem[7].

Dysequilibrium: A derangement in equilibrium, which is the condition of being evenly balanced, mostly seen in the elderly (12) (20).

Dysphagia: Difficulty in swallowing.

Dysphasia: Loss of deficiency in the power to use understand language due to injury or disease of the brain (5).

Dyspnoea: Shortness of breath, breathing uncomfortably (13).

Dysuria: Painful or difficult urination.

Emphysema: is a type of chronic obstructive pulmonary disease where there is a pathological accumulation of air in the tissues, which leads to difficulty in breathing (10).

Epigastric pain: Pain in the upper middle part of the abdomen (14).

Epilepsy: is a condition in which a person suffers repeated fits and seizures and marked by abnormal electrical discharges in the brain (4) (5).

Epistaxis: Bleeding from the nose.

Exacerbation of Colitis: worsening of colitis.

Facial weakness: Dysfunction of the facial nerve resulting in paresis of facial movements, usually on one side.

Fall with soft tissue injury: A fall leading to injury to muscles, ligaments and tendons, it excludes fractures (9).

Fall: A coming down freely, usually under the influence of gravity (16).

Fe deficiency anaemia: is an anemia caused by lack of iron or low hemoglobin concentration or hematocrit value, for older people bleeding into the gut is a common cause (10).

Febrile: a condition in which the body temperature is above to the normal range of 36.5-37^{o}C.

Fluid overload: (also known Hypervolemia) is the medical condition where there is excessive volume of fluid in the blood.

Foot ulcers: A severe sore located in the skin of the foot, often characterised by diabetes.

Fracture clavicle: breakage of clavicle often caused by fall onto an outstretched upper extremity (13).

Fracture NOF: (fracture neck of femur) breakage of narrowed end of femur that is long bone between the hip and the knee (18).

Fracture tibia and fibula: breakage of tibia (the inner and larger bone of the lower leg) or fibula (the long thin outer bone of the lower leg) (18).

Gastric dilatation: refers the increase size of the abdomen usually with swallowed air.

Gastroenteritis: an acute inflammation of the lining membrane of the stomach, most frequently caused by a viral infection ,characterised by anorexia, nausea, diarrhoea and abdominal pain (3) (10).

General decline: An overall reduction in the quality of body functions.

General deterioration: The process or condition of becoming worse (12).

General frailty: a condition in which being poor in terms of both physically and mentally.

GERD: (or GORD) (Abbreviation for gastroesophageal reflux disease) is a chronic digestive disorder that occurs when stomach acid or bile flows back into the food pipe, may result in heartburn, chest pain, Dysphagia and so on **(**11**)**.

GI Bleeding: (Gastrointestinal bleeding): is a loss of blood in the gastrointestinal tract, from the mouth to the rectum and the degree of bleeding from almost undetectable to life-threatening (13).

Glaucoma is a disease of the eye marked by increased pressure within the eyeball (5).

Gout: A disorder of uric acid metabolism, occurring especially in men, generally caused by inflammatory arthritis, causes painful joints, often in the big toe[10].

GP (Abbreviation for general practitioner): a doctor working in the community who provides primary care to local area. GPs are regarded as a family doctor, and usually they are first port of the patients, then if they find it necessary, direct their patients to the related specialist (10).

Haematemesis: The vomiting of blood.

Haematuria: The passage of blood in the urine (18).

Head injury: A traumatic damage to the head caused by blunt or penetrating trauma of the skull (14).

Heart failure: a condition where the heart is unable to pump the blood adequately, with the result that congestion and oedema develop in the tissues (12).

Heel ulcer: A sore located in the heel often leads to lower limb amputation.

Hip replacements: refers to a surgical procedure that replace the hip joint with an artificial version to provide al long-term solution for the patients having worn or damaged hip joints (8).

Hospital-acquired pneumonia (HAP): (also known nosocomial pneumonia) refers to any pneumonia developing in people who have been hospitalised, usually after 48-72 hours of hospitalisation.

Hypercalcemia: refers to excessive level of calcium in the blood and patients with hypercalcemia are very likely to suffer from primary hyperparathyroidism and malignancy.

Hypertension: is an abnormal condition in which the blood pressure is above the normal range expected in a particular age group. However generally, the blood pressure higher than 130/80 mmHg is regarded as a hypertension.

Hypoglycaemia: insufficiency of glucose in the bloodstream, resulting in muscular weakness, mental confusion and sweating (18).

Hypokalemia: the presence in the blood of an abnormally low concentration of potassium, which may lead to cardiac arrhythmias, Muscle weakness, flaccid paralysis, rhabdomyolysis and abnormal renal function (18) (14).

Hypomania: a mild degree of mania, which leads to faulty judgement and behaviour lacks the usuakl social restraints (18).

Hyponatremia: abnormally low sodium level in the blood, associated with dehydration, may lead to kidney problems or congestive heart failure.

Hypothermic: A condition in which core body temperature is significantly lower than the normal range of 36.5-37.5 ^{0}C.

Hypothyroid: subnormal activity of the thyroid gland, leading to somnolence, subnormal temperature and muscle weakness (18).

IBS: (Abbreviation for irritable bowel syndrome) refers to a chronic gastrointestinal disorder characterised by recurrent crampy abdominal pain and diarrhoea, and affects the large intestine.

ILD: (Interstitial lung disease): refers to scarring of thickening of the lung tissues which may lead to breathlessness and having blood lack of enough oxygen.PF can be caused by pneumonia or tuberculosis (4).

Ileostomy: refers to a surgical procedure that attach the bottom of small intestine to an opening in the abdominal wall to allow faecal matter to exit (10).

Immobility: The absence of movement, or inability to move (12).

Increrased confusion: Growing in confusion level of the patient.

Infected groin wound: is a severe complication of vascular surgery.

Intervertebral disc prolapsed: refers to slipping of intervertebral discs lying between adjacent vertebrae in the spine due to so much pressure is put in the disc.

Intracranial bleeding: (also known intracranial hemorrhage) is a bleeding inside the scull, can be happened due to so many various reasons, however head injury, deteriorating of vein or artery and high blood pressure are chief reasons.

Ischaemic stroke: stroke characterised by thrombosis or embolism.

Ischaemic toe: is a lack of adequate arterial blood flow from the heart to the toe.

Jaundice: a yellow discoloration of the skin, indicating excess bilirubin in the blood (18).

Jerk: an involuntary muscular movement due to reflex action[3].

Laceration: refers to a skin wound from blunt and/or shearing injuries.

LACS (abbreviation for lacunar stroke) is a stroke that results from occlusion of one of the penetrating arteries that provides blood to the brain's deep structures (13).

Laryngeal Cancer: (also known cancer of the larynx) is a carcinoma of vocal cords or other portions of the larynx, which can cause hoarseness of the voice and swelling of the throat.

Left basal pneumonia: a pneumonia located in the left lower zone of the lungs.

Left femoral embolectomy: Surgical removal of a clot or foreign material which has been transported left femur from a distant vessel by the bloodstream (1).

Left hypochondrial pain: (also known left upper quadrant pain) is a pain located in the left upper abdominal region and generally associated with colonic diseases or splenic diseases.

Left sided CVA: a cerebral vascular attack occurring the left side of the brain, whereas effects the right side of the body.

Left Sided weakness: Losing main functions of the left side of the body that can be related directly to the dysfunction of its nervous (14).

Leg pain: Unpleasant sensation in the leg.

Lethargy: A state of deep and prolonged unconsciousness , abnormal drowseness (5).

Liver mets: (liver metastases) refers to cancer which originated somewhere else and spread to the liver.

Longstanding confusion: a confusion that have existed for a long time.

Lower back pain: (also known lumbago) A pain in between the bottom of the ribs and the top of the legs.

LRTI: (abbreviation for lower respiratory tract infection) is a type of respiratory tract infection associated with below the vocal cords, such as bronchiolitis, pneumonia and tracheitis, leading to shortness of breath, weakness, high fever, coughing and fatigue (14) (13).

Lumbar disc prolapsed: is a slipped lumbar disc which causes severe low back pain and generally resulting from a sudden physical activity such as heavy lifting.

Lung opacity : is a lesion in the lung resulting from an inflammation, infection or a neoplasm.

LV aneurysm (Abbreviation for left ventricular aneurysm) refers to arise in one or more bubbles in the left ventricle that may prevent the blood flow to the body, generally occurs after a heart attack.

LVF (Abbreviation for left ventricular failure)(also known heart failure) Congestive heart failure demonstrated by signs of pulmonary congestion and oedema.

LVF secondary to AF: a left ventricular failure due to atrial fibrillation.

Macular degeneration: is a medical condition in which macula area of the retina, which is the centre of the inner lining of the eye, suffers atrophy and thinning, which may lead to loss of central vision.

Malignancy: a malignant state; resistant to treatment which occurs in severe form and frequently fatal (12).

Mechanical fall: a fall caused by gravity rather than medical reason, for example syncope.

Medication side-effects: a disease resulting from undesirable effects of the drug or the treatment applied.

Medication toxicity: The state of being poisonous due to the treatment applied (12).

Melena (British Melaena) : Black, terry stool composed largely of blood that has been acted on by gastric juices.

Mesenteric ischaemia: refers inflammation and injury of the small intestine caused by inadequate blood supply, frequently presents with severe abdominal pain, and generally observed in elderly people (13).

Mastectomy: is surgery to remove the whole breast which has been affected by cancer.

MI: (abbreviation for myocardial infarction)(also known heart attack) a medical condition in which the heart is not receiving enough oxygen due to interruption of blood supply, causing heart cells to die.

MR: (Abbreviation for mitral regurgitation) is a condition in which the mitral valve does not close properly when the heart pumps the blood out that may lead to congestive heart failure (13).

Musculoskeletal chest pain: it is a sharp or dull pain due to radiation of pain from the thoracic spine or bone injury such as fractured rib and aggravated by chest wall movement such as chest expansion with a deep breath and lying on the side.

Musculoskeletal pain: A type of pain affecting the muscles, ligaments (a band or sheet of fibrous tissue that connects two or more bones, or other structures), and tendons, along with the bones (12).

Myelodysplasia: any malformations of the spinal canal and spinal cord, characterised by abnormal blood cells produced by the bone marrow.

Myeloma: a malignant disease of the bone marrow, due to cancerous plasma cells that build up in the bone marrow (18).

Myelosuppression: reduction in blood cell production by the bone marrow, can lead to anaemia, infection, and abnormal bleeding (18).

Myocardial contusion: is a bruise of the heart muscle due to rupture hemorrage of small vessels in the myocardium (21).

Myocardial ischaemia (also known ischaemic heart disease) a disorder of cardiac function caused by reduced blood supply to the heart muscle, generally due to coronary artery disease. MI is more common in men (3) (13).

Nausea: Feeling of impending vomiting (22).

Nephrotic syndrome: refers the kidney disorders characterised by low serum albumin, oedema, large amount of protein in the urine and usually increased blood cholesterol (3).

Nil: (also known minimal change disease) Nil disease is a glomerular disease leading to heavy proteinuria caused by lack of obvious histologic glomerular changes on light microscopy (10).

Non-Hodgkins lymphoma (NHL) is the most common cancer type in the lymphatic system characterised by enlarged lymph nodes, fever, night sweats, and weight loss (5).

Non sustained VT: (non sustained ventricular tachycardia) type of ventricular tachycardia, which is a heart rate exceeding 100 beats per minute originating from an ectopic ventricular focus, where
the nst rhythm self-terminates within 30 seconds (13) (12).

NSTEMI: (Abbreviation for Non-ST elevation myocardial infarction) is a type of heart attack where only small part of the artery is blocked by the clot.

OA: (Abbreviation for osteoarthritis) (also known degenerative joint disease) occurring chiefly in older persons, characterised by degenerative disease caused by wear of the articular cartilage (18).

Oedema: Abnormally large amounts of fluid in the tissues such as around the ankles, lower leg or sacral area (22).

Off legs: Unsteadiness and difficulty with walking, dizziness, lethargy (9).

Opioid toxicity: The toxic reaction of the body to the narcotic substance, generally via allergic reaction or overdose[17].

Orthopnoea: Discomfort in breathing which occurs when lying flat (12) (13).

Osteoarthritis: Noninflammatory degenerative disease of joints characterised by gradual loss of cartilage and resulting in the development of bony spurs and cysts at the margins of the joints (20).

Osteomyelitis: refers to an infection of the bone marrow, that is often of bacterial origin and may result in death of bone tissue (5) (9).

Osteoporosis: refers to a total loss of bone, resulting from depletion of bone calcium and protein.

Painful foot: The foot with unpleasant sensory.

Painful right shoulder: A state of pain in the right shoulder.

Painless obstructive jaundice: hepatic disorder caused by obstruction to the flow of bile into the duedonum (12).

Pallor: Extreme paleness of the skin, such as in anaemia or after blood loss[1].

Palpitations: Abnormally rapid and irregular beating of heart such that the person becomes conscious of its action (4).

Paranoid: Having delusions of persecution (12).

Parkinson's disease: is a persistent disorder of part of the brain and is marked by tremor of resting muscles, rigidity, slowness of movement, impaired balance, and a shuffling gait (5).

PE (Abbreviation for pulmonary embolism) is an obstruction of a blood vessel in the lungs, generally due to embolism of a blood clot (thrombus) from the veins in the legs (20).

Pedal oedema: An oedema located in the foot.

Pericarditis: Inflammation of the sac surrounding the heart (inflammation of the pericardium) (10).

Peripheral neuropathy: is a condition in which there is a damage to one or more of the peripheral nerves, so the connection between the central and the peripheral nervous system become slow. Diabetes is the most common cause of this disease (9).

Peripheral vascular disease (PVD): (also known peripheral artery disease (PAD)) PVD is a common circulatory problem including all disease caused by the obstruction of large arteries in the arms and legs and it is very likely to cause either acute or chronic ischemia (11) (13).

Peritoneal metastasis: is the result of seeding of the peritoneal cavity by the tumour prior to surgery, and it is the most frequent reason of tumour progression in advanced gastric cancer (23).

Peritonitis: An inflammation of the lining of the abdominal cavity, usually caused by infection (24).

Pleural effusion: Increased amounts of fluid within the pleural cavity surrounding the lungs, it may result in a difficulty in breathing by limiting the expansion of the lungs during inspiration (13) (12).

Pleuritic chest pain: Chest pain pertaining to pleurisy which is the inflammation of the pleura.

PMR (Abbreviation for polymyalgia rheumatica) : A form of inflammatory rheumatism characterised by gross early-morning stiffness, which tends to slow down during the day , and pain in the shoulders and sometimes around the hips, generally occurs in women (4).

PND: (Abbreviation for Paroxysmal Nocturnal Dyspnoea) A form of dyspnoea, occurring suddenly and usually after an hour or two hour after the individual has fallen asleep (17).

Pneumonia: inflammation of the lung caused by bacteria in which the air sacs filled with inflammatory cells (18).

Polycythemia rubra vera : (also known primary polycythemia) is a blood disorder in which the bone marrow produces excess red blood cells and may lead to headache , vertigo, difficulty in breathing (11).

Polypharmacy: (for the elderly) refers comorbid complication due to the amount of medication necessary to treat the conditions that become more prevalent with age, occurs when the medications interact with each other, with food or alcohol (10).

Polyuria: urinating too much liquid in a given period, a characteristic of diabetes (3).

Poor compliance of medication: a disease appearing when the patient does not fulfil the requirements of his or her medication.

Post thrombotic syndrome: refers to the long-term effects that can occur after venous thrombosis, which refers a clot within the venous vasculate and impairs blood flow back to the right part of the heart (13)

Postural hypotension: (also known orthostatic hypotension) is a condition in which the blood pressure falls rapidly after a change in body position. The decrease is typically greater than 20/10 mm Hg and may be related to hydration status, drug side effect or dysautonomia (24) (10) (13).

PPM: permanent pacemaker.

PR bleeding: Bleeding from the prosthion (PR) which is the point on the maxillary alveolar process that projects most anteriorly in the midline (3).

Presyncope: Unlike syncope, which is actually fainting, presyncope is near-fainting which may include lightheadedness, dizziness caused by cardiovascular problem (15).

Productive cough: A cough accompanied by expectoration (sputum) (12).

Prostate Cancer: is a type of cancer beginning in the male prostate.

Pseudogout: is an acute inflammation causes in the joints: red, tender, and swollen joints that may resemble gouty arthritis (13).

Pulmonary oedema: abnormal accumulation of fluid in the lungs usually resulting from mitral stenosis or left ventricular failure (LVF) and leads to impaired gas exchange and may result in respiratory failure (13) (5).

Pyrexia: An abnormal raise of body temperature, usually as a result of a pathologic process (1).

Rash: Skin eruption usually with little or no elevation above the surface (5).

Recent UTI: (Abbreviation for Urinary Tract Infection), an acute infection that affects any part of the urinary tract caused by a microbe, or a bacteria (12).

Rectal bleeding secondary to prostate Ca: a bleeding from the rectal area due to prostate cancer.

Recurrent collapses: a collapse occurring several times.

Recurrent falls: A fall happening time after time.

Reduced appetite: (also known poor appetite) Loss or reduction in appetite for food.

Reduced mobility: Reducing of capability of moving.

Reduced urine output: Total lack of urine.

Reflux: is a common digestive disorder that occurs when stomach acid or bile flows back into the food pipe, which may result in heartburn, chest pain, Dysphagia (11).

Renal cancer: (also known kidney cancer) refers to proliferative malignant disorder of kidney cells.

Rhabdomyolysis: The rapid destruction of skeletal muscle cells with the release of myoglobin and other toxic cell components (e.g. potassium), caused by traumatic injury, excessive exertion, or stroke (18) (5).

Rheumatoid arthritis (RA): The second most common type of arthritis, affects the connective tissues, causes painful and swollen joints and more observed in women (18).

Rib fracture: breakage of rib.

Right ankle soft tissue injury: is the damage of soft tissues such as muscles, ligaments or tendons in the right ankle, usually result in pain, swelling or brusing (13).

Right facial droop: Sagging on the right side of the face, which usually indicates paralysis of facial muscles due to trauma, infection or tumour removal near or at the facial nerve (15).

Right knee swelling: Protuberant located in the right knee.

Right PACS: (right partial anterior circulation stroke syndrome) refers to the symptoms of a patient who clinically appears to have suffered from a right partial anterior circulation infarct (PACI), which is a type of cerebral infarction affecting part anterior circulation supplying right side of the brain (13).

Right sided weakness: Losing main functions of the right side of the body.

Rigor: Sudden chill, accompanied by severe shivering (22).

Rotator cuff tendonitis : is an irritation and inflammation of the tendons of the shoulder usually result from shoulder injury or overuse of the shoulder (5).

Sarcoma of thigh: Malignant tumour from connective tissue located in the thigh (14).

SCC (Abbreviation for Squamous cell carcinoma) is a malignant tumour arising from the keratinising cells of the epidermis or its appendages, may occur in many different organs, such as skin, bladder, lungs.

Scoliosis: Abnormal lateral curvature of the spine to one side and the most common regions to be affected by scoliosis are the chest are and the lower part of the back (8).

SDH: (Abbreviation for subdural hematoma) is a kind of traumatic brain injury in which there is an extravasation of blood between the dura (the outer protective of the brain) and arachnoid (middle layer of the meninges) (12).

Seizure: A sudden attack; especially the physical manifestations (as convulsion, sensory disturbance) are caused by abnormal electrical discharges in the brain (5).

Sepsis: The presence of pathogenic microorganisms or their toxins in the blood or in tissues (1).

Septic arthritis: a type of arthritis caused by bacteria, rickettsiae, mycoplasmas, viruses, fungi, or parasites (1).

Shingles: is an infection of a nerve and the area around it,characterized by inflammation, pain, and skin eruptions and caused by the herpes varicella-zoster virus (8).

Simple faint: is a condition in which the patient falls to the ground and he or she is unconscious for less than two minutes, and recovers rapidly (10).

Sinus bradycardia: slowness of the heartbeat which is less than 60 beats per minute that originates from the sinus node, which is the impulse-generating tissue located in the right atrium of the heart (10).

Slurred speech: Unusual speech in which words are not pronounced clearly or completely but are run together or partially eliminated (14).

Small bowel obstruction: An obstruction of the small intestine which prevents the free transit of the products of digestion, may cause postoperative adhesions and surgical operation (20).

SOB: (Abbreviation for Short Of Breath) difficulty in breathing.

Social crisis: is the crisis keeping an individual out of the social life, any psychological trauma may lead to social crisis.

Sputum: Material containing mucus, cellular debris, microorganisms or blood coughed up from the lungs and expectorated via the mouth (1).

Stokes-Adams Syndrome: refers to a sudden collapse without warning, associated with loss of consciousness for a couple of seconds, occurs due to heart rhythm problems and after the attack the patient is very likely to feel hot and flushed (9) (10).

Stroke POCS: (Abbreviation for posterior circulation stroke) a type of stroke that refers to the symptoms of a patient who clinically appears to have suffered from a posterior circulation infract, type of cerebral infarction that affects the posterior circulation supplying one side of the brain, but who was not yet had any diagnostic imaging to confirm the diagnosis (13).

Stroke: A medical condition, in which blood flow to the brain suddenly is interrupted, often affects one side of the body and causes loss of the ability to speak or to move particular muscles (18)

Suprapubic pain: pain in the lower central part of the abdomen.

SVT: (Abbreviation for supraventricular tachycardia) an abnormally accelerated rhythm originating upper heart chambers. Rates may be in the rage of 150-250 beats per minute (3) (10) (13).

Swollen legs: Protuberant or abnormally distended legs (5).

Syncope (fainting): a temporary loss of consciousness and postural tone resulting from reduced blood flow to the brain (i.e., brain ischemia) (1).

Tachycardia: is an abnormal condition where the heart rate is more than 100 per minute, which is much more than normal range.

TACS: (abbreviation for total anterior circulation stroke) is a stroke which occurs in the patient who has suffered from a total anterior circulation infarct, but who has not yet had any diagnostic imaging to confirm the diagnosis (13).

Temporal arteritis: (also known giant cell **CA))** refers to an inflammation and damage to blood vessels, in particular the artery that brings blood to the optic nerve and can cause blindness (13).

Thyroidectomy: Surgical removal of all or part of the thyroid gland.

TIA: (Abbreviation for transient ischaemic attacks) (also known mini stroke) is a set of symptoms that usually lasts ten minutes or less, and occurs due to a temporary lack of blood to part of the brain (9).

Tiredness: Exhausted of strength or energy; fatigued (2).

TKR (abbreviation for total knee replacement) (also known arthroplasty) is a type of surgery where the damaged, worn or diseased knee is replaced with an artificial joint **(**8**)**.

Tuberculosis (TB) : is a contagious disease caused by a bacteria called Mycobacterium tuberculosis, which generally affects the lungs but can affect almost any tissue of the body such as brain, kidney and bones (10) (12).

TURP (abbreviation for transurethral resection of the prostate) is urological operation used to treat BPH.

UGI Bleeding: (Upper gastrointestinal bleeding) (Upper GI bleeding) is a hemmorrage (loss of blood) in the upper gastrointestinal tract commonly caused by peptic ulceration and it about 4 times as common as bleeding from the lower GIT (9) (13).

Unclear: No diagnosis given.

Uncontrolled BM: uncontrolled bowel movement.

Unresponsiveness: A total lack of response to neurologic stimuli.

Unstable angina: recurrent episodes of angina which is increasing in severity, duration or frequency (3).

Unwell: being in poor health, feeling sick.

Uraemia: accumulation of urea and other nitrogenous waste compounds, which are usually excreted in the urine, generally occurs in severe kidney disease (18) (5).

Urinary frequency: (also known frequent urination) needing to urinate more often than usual which is four to eight times a day.

Urinary incontinence: Urinary incontinence is any involuntary leakage of urine (13).

Urinary retention: The inability to urinate.

Urosepsis: a sepsis caused by unprocessed urinary matter backing up into the bloodstream.

URTI: (Abbreviation for upper respiratory tract infections) is a respiratory tract infection involving the upper respiratory tract, such as nose, sinuses, pharynx and larynx, may lead to nasal congestion, cough, fever or running nose (10).

UTI : (Abbreviation for urinary tract infection) an infection that affects the structures of the body participating secretion and elimination of the urine, i.e., the kidney, the urethra and the ureters4.

Vascular dementia: a type of dementia, due to cerebrovascular disorders, including cerebral infarction, and conditions associated with chronic brain ischemia, with a stepwise deteriorating course and a 'patchy' distribution of neurologic deficits (1) (10).

Vasovagal syncope: a type of syncope occurring as a part of normal physiologic response to stress, resulting from cerebral ischemia, secondary to decreased cardiac output, peripheral vasodilation, the patient may lose consciousness for several seconds (3) (14).

Vestibulitis (also known vulvar vestibulitis) The inflammation of the opening of the female genital organ.

Vomiting: An act of ejecting food from the stomach through the mouth.

Weakness: A condition of being feeble, fragile or lacking physical strength, energy or vigor (14).

Bibliography

1. Medical Dictionary Online. [Online] http://www.online-medical-dictionary.org/.

2. **Cambridge.** *Cambridge Dictionary Online.* [Online] http://dictionary.cambridge.org/.

3. **Mondofacto.** *Mondofacto Online Medical Dictioanry.* [Online] http://www.mondofacto.com/dictionary/.

4. **Marcovitch, Harvey.** *Black's Student Medical Dictionary.* s.l. : A & C Black Publishers Ltd, 2007. ISBN 13: 9780713687620 .

5. **National Institutes of Health.** A service of the U.S. National Library of Medicine. [Online] http://www.nlm.nih.gov/medlineplus/.

6. **American Heart Association.** Acute Coronary Syndrome. [Online] http://www.americanheart.org/presenter.jhtml?identifier=3010002.

7. *Dorland's Medical Dictionary for Health Consumer.* s.l. : Elsevier, 2007.

8. **NHS.** Health A-Z - Conditions and treatments. [Online] http://www.nhs.uk/conditions/Pages/hub.aspx.

9. **Patient UK.** Information for Patients. [Online] www.patient.co.uk.

10. **Wrong Diagnosis.** Diseas and Symptoms. [Online] www.wrongdiagnosis.com.

11. **Mayo Clinic.** Diseases and Conditions. [Online] http://www.mayoclinic.com/.

12. *Stedman's Medical Dictionary for the Health Professions and Nursing: Standard.* s.l. : Stedman's, 2007.

13. **Wikipedia.** [Online] http://www.wikipedia.org/.

14. *Mosby's Dictioanary of Medicine, Nursing, Health Professions.* s.l. : Elsevier Health Sciences, 2008. ISBN 13: 9780323049375 .

15. *McGraw-Hill Concise Dictionary of Modern Medicine.* s.l. : The McGraw-Hill Companies Inc., 2002.

16. *Mosby's Medical Dictionary.* s.l. : Elsevier Health Sciences, 2008. ISBN 13: 9780323052900.

17. *The American Heritage Medical Dictionary.* s.l. : Houghton Mifflin, 2008. ISBN 13: 9780618824359.

18. **Oxford.** *Concise Colour Medical Dictionary.* s.l. : Oxford University Press, 2010. ISBN 13: 9780199557158 .

19. **World Health Organisation.** Health topics. [Online] http://www.who.int/topics/en/.

20. *The Gale Encyclopedia of Medicine.* s.l. : Gale , 2002. ISBN 13: 9781414403687.

21. *Myocardial Injury: Contrasting Infarction and Contusion.* **Pooler, c.** s.l. : American Association of Critical-Care Nurses, 2002.

22. **Brooker, Chris.** *Churchill Livingstone Medical Dictionary.* s.l. : Elsevier Health Sciences, 2008.

23. **Chandrasoma, Para.** *Gastrointestinal pathology.* s.l. : Simon &Schuster, 1999.

24. [Online] www.about .com.

References

1. **U.S Census Bureau .** International Data Base (IDB) World Population by Age and Sex. *U.S Census Bureau.* [Online] 2010. http://www.census.gov/ipc/www/idb/worldpopinfo.html.

2. **Haub, C** *World Population Data Sheet.* s.l. : Population Reference Bureau, 2006.

3. Ageing Fasting increase in the 'oldest old'. *Office for National Statistics.* [Online] 2010. http://www.statistics.gov.uk/cci/nugget.asp?id=949.

4. Life expectancy at birth and at age 65 by local areas in the United Kingdom. *Office for National Statistics.* [Online] http://www.statistics.gov.uk/statbase/Product.asp?vlnk=8841.

5. **Department of Health.** *National Service Framework for Older People.* 2001.

6.**General Register Office for Scotland.** *Household Projections for Scotland 2008 - based.* 2010.

7. **Office for National Statistics.** *General Household Survey.* 2007.

8. Older People Living Arrengements. *Office for National Statistics.* [Online] http://www.statistics.gov.uk/cci/nugget.asp?id=1264.

9. **Sinclair A J.** *Diabetes in Old Age.* s.l. : Wiley, 2009. 978-0-470-06562-4.

10.**Bridgwood A, Lilly R,Thomas M.** *Living in Britain.* s.l. : Office for National Statistics Social Survey Division, 1999.

11.**Teich J M, Waeckerle J F.** *Emergency Medical Informatics.*s.l. : Emergency medical informatics, 1997, Vol. 30.

12. **Lavrac N.** *Selected techniques for data mining in medicine.* s.l. : Artificial Intelligence in Medicine, 1999. Vol. 16.

13. **Bellazzi R, Zupan B.** *Predictive data mining in clinical medicine: Current issues.* s.l. : International journal of medical informatics, 2008, Vol. 77.

14. **Peek N, Combi C, Tucker A.***Biomedical Data Mining.* 2009 : Methods Inf Med, Vol. 3.

15. Data Mining Applications in 2008 (Dec 2008) . *KDnuggets.* [Online] [Cited: July 20, 2010.] http://www.kdnuggets.com/polls/2008/data-mining-applications.htm.

16. **Konenenko I.** Comparasion of Inductive and Naive Bayesian Learning Approaches to Automatic Knowledge Acquisation. [book auth.] Bob Wielinga. *Current trends in knowledge acquisition.* s.l. : IOS, 1990.

17. **Pazzani M J.** Searching for Dependencies in Bayesian Classifiers. [book auth.] Hans-Joachim Lenz Douglas H. Fisher. *Learning from data: artificial intelligence and statistics V.* s.l. : Springer, 1996.

18.**Han J,Kamber M.** *Data mining: concepts and techniques.* s.l. : Morgan Kaufmann, 2006. ISBN 10: 1-55860-901-6.

19.**Kukar M, Kononenko I, Silvester T.** Prognosing the femoral neck fracture recovery with machine learning. University of Ljubljana : Citeseerx.

20. **Lu D F, Street W N, Delaney C.** Knowledge discovery: Detecting elderly patients with impaired mobility. [book auth.] Peter Murray, Connie Delaney Hyeoun-Ae Park. *Studies in Health Technology and Informatics.* s.l. : IOS Press, 2006.

21.**Riberio J, Neves J, Sanchez J, Delgado M.** *Wine Vinification prediction using Data Mining tools.*s.l. : COMPUTING and COMPUTATIONAL INTELLIGENCE, 2009.

22. **Delen D, Walker G, Kadam A.** *Predicting breast cancer survivability: a comparison of three data mining methods.* s.l. : Artificial Intelligence in Medicine, 2005, Vol. 34.

23. **Bartosch-H¨arlid A, Andersson B, Aho U, Nilsson J, Andersson R.** *Artificial neural networks in pancreatic disease.*s.l. : British Journal of Surgery, 2008, Vol. 95.

24. **Gil D, Johnsson M , Chamizo J M G, Paya A S.** *Application of artificial neural networks in the diagnosis of urological dysfunctions* s.l. : Expert Systems with Applications, 2009, Vol. 36.

25. **Unay D, Soldea O, Ekin Ahmet, Cetin M, Ercil A.** Automatic Annotation of X-Ray Images: A Study on Attribute Selection. *Lecture Notes in Computer Science.* s.l. : Springer, 2010.

26. **Brown M P S, Grundy W N, Lin D, Cristianini N.** Support Vector Machine Classification of Microarray Gene Expression Data. s.l. : UCSC-CRL-99-09, 1999.

27. **Noble W S***What is a support vector machine.* s.l. : NATURE BIOTECHNOLOGY , 2006, Vol. 24.

28. **Rahman M, Desai B C, Bhattacharya P.***Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion.* s.l. : Computerized Medical Imaging and Graphics, 2008, Vol. 32.

29. **Khan M, Ding Q, Perrizo W.** *k -nearest Neighbor Classification on Spatial Data Streams Using P-trees.*s.l. : Advances in Knowledge Discovery and Data Mining, Springer, 2002, Vol. 2336.

30. **Liu Y, Dellaert F, Rothfus W E, Moore A, Schneider J, Kanade T** *Classification-Driven Pathological Neuroimage Retrieval Using Statistical Assymetry Measures.***.** s.l. : Medical Image Computing and Computer-Assisted Interventio, Springer, 2010, Vol. 2208.

31. **Liu C L, Lee C H, Lin P M.** *A fall detection system using k-nearest neighbor classifier.* s.l. : Expert Systems with Applications, 2010, Vol. 37.

32. **Tsumato S.** *Problems with Mining Medical Data.* s.l. : 24th International Computer Software and Applications Conference, 2000.

33. **Tan K C, Yu Q, Heng C M, Lee T H.** *Evolutionary computing for knowledge discovery in medical diagnosis.*s.l. : Artificial Intelligence in Medicine, 2003, Vol. 27.

34. **Krzystof J C.** *Medical Data Mining and Knowledge Discovery.* s.l. : Physica - Verlag, 2001. ISBN: 3-7908-1340-0.

35. **Barrera J, Cesar R M, Ferreira J E , Gubitoso M D.** *An environment for knowledge discovery in biology.*2004 : Computers in Biology and Medicine, Vol. 34.

36. **Chen H, Fuller S S, Friedman C, Hersh W.** *MEDICAL INFORMATICS Knowledge Management and Data Mining in Biomedicine* . s.l. : Springer, 2005. ISBN-10: 0-387-2438 1-X (HB).

37. **Sujansky W.** *Heterogeneous Database Integration in Biomedicine.***.** s.l. : Journal of Biomedical Informatics, 2001, Vol. 34.

38. **Rosales R E , Rao R B.** *Guest Editorial: Special Issue on impacting patient care by mining medical data.*s.l. : Data Mining and Knowledge Discovery - Springer, 2010, Vol. 20.

39. **Richards G, Rayward-Smith VJ, Sönksen PH, Carey S, Weng C.** *Data mining for indicators of early mortality in a database of clinical records.* s.l. : Artificial Intelligence in Medicine, 2001, Vol. 22.

40. **Bath P A.** *The name assigned to the document by the author. This field may also coData Mining in Health and Medical Information.* s.l. : Annual Review of Information Science and Technology, 2004.

41. **Cios K J, Moore G W.** *Uniqueness of medical data mining.***.** s.l. : Artificial Intelligence in Medicine, 2002, Vol. 26.

42. **Rector A L.** *Clinical Terminology: Why is it so hard?* s.l. : Methods of information in medicine, 1999, Vol. 38.

43. **Agrawal R, Srikant R.** *Privacy-preserving data mining.* s.l. : ACM SIGMOD Record , 2000, Vol. 29.

44. **Terry H.** *Security Issues for Implementation of E-Medical Records.***.** s.l. : Communications of the ACM , 2001, Vol. 44.

45. **Berman J J.***Confidentiality issues for medical data miners.* s.l. : Artificial Intelligence in Medicine, 2002, Vol. 26.

46. **Kantarcioglu M, Jin J,Clifton C.** *When do data mining results violate privacy?***.** s.l. : Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.

47. **Brankovic L, Estivill-Castro V.** *Privacy Issues in Knowledge Discovery and Data Mining.*Melbourne, Victoria, Australia : In Proc. of Australian Institute of Computer Ethics Conference (AICEC99), 1999.

48. **Bertino E,Ooi B C,Yang Y, Deng R H.** *Privacy and Ownership Preserving of Outsourced Medical Data.* 2005 : Proceedings of the 21st International Conference on Data Engineering (ICDE 2005), IEEE.

49. **Chae J J, Mukherjee D, Manjunath B S.** *A Robust Data Hiding Technique Using Multidimensional Lattices.* s.l. : Proceedings of the Advances in Digital Libraries Conference, IEEE, 1998.

50. **Larose D T.** *Discovering Knowledge in Data.* s.l. : Wiley, 2005. ISBN: 0-471-66657-2.

51.**Refaat M.** *Data Preparation for Data Mining Using SAS.* s.l. : Morgan Kaufmann , 2007. ISBN: 0-12-373577-7 .

52. **Pearson R K.** *Mining Imperfect Data.* s.l. : SIAM, 2005. ISBN 0-89871-582-2.

53. **Wang J.** *Data mining: opportunities and challenges.* s.l. : IRM Press, 2003. ISBN: 1-931777-83-7.

54. **Zhu X, Davidson I.** *Knowledge Discovery and Data Mining: Challenges and Realities* . s.l. : Information Science Reference, 2007. ISBN 978-1-59904-252-7.

55. **Feelders A.** Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation? s.l. : Lecture Notes in Computer Science, Springer, 1999. Vol. 1704.

56. Standard Random Partitioning. *Resambling Stats.* [Online] [Cited: Augst 10, 2010.] http://www.resample.com/xlminer/help/Partition/Partition.htm.

57. **Shmueli G,Patel N R, Bruce P C.** *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner.* s.l. : Wiley, 2010. ISBN: 978-0-470-52682-8.

58. **Kantardzic M.** *Data Mining: Concepts, Models, Methods, and Algorithms* . s.l. : John Wiley & Sons, 2003. ISBN:0471228524.

59.**Nisbet R, Elder J,Miner G.** *Handbook of Statistical Analysis and Data Mining Applications.* s.l. : Elsevier Science & Technology UNITED KINGDOM, 2009. ISBN:9780123747655.

60. **Estabrooks A, JO T, Japcowics N.** *A Multiple Resambling Method for Learning from Imbalanced Datasets.*s.l. : Computational Intelligence, 2004, Vol. 20.

61. **Kubat M, Matwin S.** *Addressing the curse of imbalanced training sets: one-sided selection.*s.l. : In Proc. 14th International Conference on Machine Learning, 1997.

62. **Liu Y, Chawla N V , Harper M P, Shriberg E , Stolcke A.** *A Study in Machine Learning from Imbalanced Data for Sentence Boundary Detection in Speech.* s.l. : Computer Speech and Language, 2006, Vol. 20.

63. **Japkowicz N.** *The Class Imbalance Problem: Significance and Strategies.* s.l. : in Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000), 2000.

64. **Questier F, Put R, Coomans D, Walczak B, Heyden Y V.***The use of CART and multivariate regression trees for supervised and unsupervised feature selection.* 45-54, s.l. : Chemometrics and Intelligent Laboratory Systems, 2005, Vol. 76.

65. **IBM PASW Modeler 13.** Help Topics. 1994-2009.

66. **Loh W Y.** Classification and Regression Trees (CART). s.l. : Free article on Wiley Online Library, 2010.

67. **Bittencourt H R, Clarke R T.** *Feature Selection by Using Classification and Regression Trees (CART).* 66-70 , s.l. : International Archieves of Photogrammetry Remote Sensing and Spatial Information Sciences, 2004, Vol. 35.

68.**Breiman L, Freidman J, Stone C J, Olshen R H.** *Classification and regression trees.* s.l. : Chapman & Hall, 1984.

69. **Goel P K, Prasher S O, Patel R M, Landry J A.** *Classification of hyperspectral data by decision trees and artificial neural networks to identify weed stress and nitrogen status of corn.* 67-93, s.l. : Computers and Electronics in Agriculture , 2003, Vol. 39.

70. **Steinberg D.** CART: Classification and Regression Trees. [book auth.] V Kumar X Vu. *The top ten algorithms in data mining.* 2009 : CRC Press.

71. **Hand D J, Mannila H, Smyth P.** *Principles of data mining.* s.l. : Massachusetts Institute of Technology, 2001. ISBN: 0-262-0890.

72. **StatSoft .** Classification and Regression Trees (C&RT). *StatSoft Electronic Statistics Textbook.* [Online] [Cited: Augst 15, 2010.] http://www.statsoft.com/textbook/classification-and-regression-trees/v/.

73. **Elhadi E M, Zomrawi N.** *Change Detection Analysis By Using Ikonos And Quick Bird Imageries* 10, s.l. : Marsland Press /Nature and Science, 2009, Vol. 7. ISSN: 1545-0740.

74. **Kolyshkina I, Brookes R.***Data mining approaches to modelling insurance risk.* s.l. : PricewaterhouseCoopers, 2002.

75. **Hinkle D E, Wiersma W, Jurs S G.** *Applied statistics for the behavioral sciences.* s.l. : Houghton Mifflin, 2003.

76. **Wikipedia.** Statistical significance. [Online] [Cited: Augst 17, 2010.] http://en.wikipedia.org/wiki/Statistical_significance.

77. **Rose C, Smith M.** *Mathematical Statistics With Mathematica.* s.l. : Springer-Verlag, 2002. ISBN: 0-387-95234-9.

78. **Matignon R.** *Data Mining Using SAS Enterprise Miner.* s.l. : Wiley, 2007. ISBX: 978-0-470-14901-0 .