DATA MINING FOR HEART FAILURE: An Investigation into the Challenges in Real Life Clinical Datasets

Lisa Kirke

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science

Department of Computer Science

The University of Hull

June 2015

ABSTRACT

Clinical data presents a number of challenges including missing data, class imbalance, high dimensionality and non-normal distribution. A motivation for this research is to investigate and analyse the manner in which the challenges affect the performance of algorithms. The challenges were explored with the help of a real life heart failure clinical dataset known as *Hull LifeLab,* obtained from a live cardiology clinic at the Hull Royal Infirmary Hospital. A Clinical Data Mining Workflow (CDMW) was designed with three intuitive stages, namely, descriptive, predictive and prescriptive. The naming of these stages reflects the nature of the analysis that is possible within each stage; therefore a number of different algorithms are employed. Most algorithms require the data to be distributed in a normal manner. However, the distribution is not explicitly used within the algorithms. Approaches based on Bayes use the properties of the distributions very explicitly, and thus provides valuable insight into the nature of the data.

The first stage of the analysis is to investigate if the assumptions made for Bayes hold, e.g. the strong independence assumption and the assumption of a Gaussian distribution. The next stage is to investigate the role of missing values. Results found that imputation does not affect the performance as much as those records which are initially complete. These records are often not outliers, but contain problem variables. A method was developed to identify these. The effect of skews in the data was also investigated within the CDMW. However, it was found that methods based on Bayes were able to handle these, albeit with a small variability in performance. The thesis provides an insight into the reasons why clinical data often causes problems. Even the issue of imbalanced classes is not an issue, for Bayes is independent of this.

ACKNOWLEDGEMENT

First and foremost I would like to take this opportunity to express my deepest gratitude to those who supported me during the course of the PhD. I am especially thankful to Dr. Chandra Kambampati for his tireless support as my supervisor, his guidance and constructive criticism during the course of this research. Chandra has supported me academically through the rough road to finishing this thesis and without his persistent help this thesis would not have been possible.

I would like to acknowledge Nongnuch Poolsawad, to whom I first spoke when I began the PhD in 2011. Thank you for reassuring me when I had doubts about using MATLAB. Thank you for enlightening the first glance of research during the writing of my first ever research paper, especially during those late nights we were working together before deadlines and for the fun we have had in the last three years.

My sincere thanks go to my wonderful partner, Mark and amazing son, Samuel for their patience and support throughout this journey, both emotionally and mentally, especially when I needed reclusion. Samuel, your warm smile always reassured me that I will see the end, even when I didn’t see or believe it.

Lastly, I would like to acknowledge my dad, who has encouraged, and inspired me and sacrificed himself to help my pursuit of a higher education. Thank you for your unconditional love, trust and encouragement through the years.

DECLARATION

Parts of the work reported in this thesis were published in the following research papers:

1. Poolsawad, N., **Moore, L**., Kambhampati, C. & Cleland, J. G. F. 2012. Handling Missing Values in Data Mining - A Case Study of Heart Failure Dataset. The 2012 8th International Conference on Natural Computation (ICNC'12) and the 2012 9th *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'12)* Chongqing, China.
2. **Moore. L**. and Kambhampati. C., The effect of features using Feature Selection for Bayesian Classifier. *Proceedings of IEEE Systems, Man & Cybernetics 2013,* 6th-9th Oct, Manchester, UK
3. Poolsawad, N., **Moore, L**., Kambhampati, C. & Cleland, J. G. F. 2012. Performance Metrics for Classification in Clinical Dataset. *The 19th International Conference on Neural Information Processing* (ICONIP2012) Doha, Qatar.
4. Poolsawad, N., **Moore, L**., Kambhampati, C. & Cleland, J. G. F. 2014. Issues in the Mining of Heart Failure Datasets. *International Journal of Automation and Computing,* vol. 11. no.2. pp.162-179
5. **Moore. L**., Kambhampati. C., Cleland, J. G. F. 2014. Classification of real live Heart Failure clinical dataset- Is TAN Bayes better than other Bayes? *Proceedings of IEEE Systems, Man & Cybernetics 2014.* 5th-8th Oct, San Diego, CA, USA

LIST OF ABBREVIATIONS

ACC	Accuracy
ANN	Artificial Neural Network
AODE	Averaged One-Dependence Estimation
CDS	Clinical Decision Support
CDM	Clinical Decision Making
CDMW	Clinical Data Mining Workflow
CFS	Correlation Feature Selection
CMCI	Concept Most Common value Imputation
CRISP-DM	Cross Industry Standard Process for Data Mining
EHRs	Electronic Health Records
EM	Expectation Maximisation imputation
EMRs	Electronic Medical Records
EPRs	Electronic Patient Records
FKM	Fuzzy *k*-Means clustering imputation
FN	False Negative
FP	False Positive
ID3	Iterative Dichotomiser 3
KADS	Knowledge Acquisition and Documentation Structuring
KDD	Knowledge Discovery in Data
KDE	Kernel Density Estimation

KMI	*K*-Means clustering Imputation
KNNI	*K*-Nearest Neighbour Imputation
MAR	Missing at Random
MCI	Most Common value Imputation
MCAR	Missing Completely at Random
MD	Missing Data
MSE	Mean Squared Error
MNAR	Missing Not at Random
MLP	Multilayer perceptron
MWST	Maximum Weighted Spanning Tree
NPV	Negative Predictive Value
PEFR	Peak Expiratory Flow Rate
PMML	Predictive Data Mining Markup Language
PPV	Positive Predictive Value
SEMMA	Sample, Explore, Modify, Model, Assess
SHFM	Seattle Heart Failure Model
SOM	Self-Organising Map
SEN	Sensitivity
SQL	Structured Query Language
SMOTE	Synthetic Minority Over-Sampling Technique
SPEC	Specificity
SVMI	Support Vector Machine Imputation

TAN	Tree Augmented Naïve Bayesian Network
TN	True Negative
TP	True Positive
XML	Extensible Markup Language

NOTATIONS

X	=	Dataset
	=	$\{x_{i,}\}, i = 1, 2, 3, \dots, n ; j = 1, 2, 3, \dots, m$
	=	$(X_1, X_2, \dots, X_N)$
	=	$X_i \in X \subseteq \mathbb{R}^n; i = 1, \dots n$
$x_{i,j}$	=	each data object, each data element
μ_{ij}	=	mean of each variable
σ_{ij}	=	standard deviation of each variable
$\sigma^2{}_{ij}$	=	variance of each variable
n	=	Number of dataset attribute
N	=	Number of records or samples
$P(.)$	=	Probability
$f(x)$	=	Probability density function
$\underline{x}$	=	a representation of a reduced subset of feature
Y	=	outcome of fully observed data
∞	=	Infinite
D_t	=	Distance

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1-INTRODUCTION

1.1 Motivation and research problem

The wealth of electronic data available has made it almost impossible to collect, sift through, analyse and gain knowledge from Electronic Health Records (EHRs) (Blumenthal and Tavenner, 2010, Noteboom *et al.*, 2014). This makes it challenging for clinicians to capture a patient's entire clinical history, especially if it is spread out over a number of different healthcare systems. A study (Arch-int and Arch-int, 2011) proposed an architecture of semantic information integration for Electronic Patient Records (EPRs), using ontology and web service models to safely allow interoperability between EPR systems. This enables one to rapidly discover patient information dispersed over different healthcare systems, thereby enhancing care coordination between clinicians (Burton *et al.*, 2004) and thus improving patient care. The term EPRs, EHR and Electronic Medical Records (EMRs) are interchangeable synonyms used in healthcare informatics, each with a slightly different definition (Boonstra and Broekhuis, 2010). However EHR will be used in this thesis. EHRs have extensively presented their 'meaningful use' (Kapoor and Kleinbart, 2012) from decision making to achieving specified improvement in care delivery. However, there is also a lack of acceptance and adoption of EHRs in the healthcare domain. A recent study (Gajanayake *et al.*, 2013) presents the contributing factors, categorised into eight types namely; financial, technical, time, psychological, social, legal, organisational and change process (Boonstra and Broekhuis, 2010). As clinicians have a great impact on the overall use and acceptance level of EHRs, it is required that they actively support and apply EHRs to benefit from them while considering these factors.

The presence of this quantity of data requires tools in order to discover relationships and new knowledge. Data mining methods allow for this to happen. This is a process of extracting and discovering meaningful knowledge from large quantities of data. Data mining consists of the main components: Classification, Clustering and Association rule learning (Kesavaraj and Sukumaran, 2013, Batra *et al.,* 2013, Han and Kamber, 2006, Larose, 2014). Data mining has assisted clinicians in both medical decisions and in creating a framework for evidence-based medicine, for example in providing valuable insights into how to detect a particular disease early and thus make the clinical process more efficient. This has a further advantage in that care can be tailored to the specific needs of patients with the following set of aims:

- To improve quality of life and clinical outcomes by involving patients in their care, for example, patients using instruments at home that do not necessarily require specialised or expert skills; in this way unplanned hospital admissions will be reduced.
- To develop predictive models that will help in the design of personalised care and the planning of care.
- To discover new knowledge, useful and lifesaving information to improve treatments.

However, its application to clinical data is very challenging (Batra *et al.,* 2013).

The effective use of data mining methods for rapid clinical decision making requires the availability of high quality clinical data (Shahriar and Anam, 2008, Lu and Su, 2010). In this thesis, the data is obtained from a live cardiology clinic at the Hull Royal Infirmary Hospital. This data poses a number of challenges, which create problems for developing appropriate classification and prediction algorithms. Thus the motivation for this research is not only to develop a methodology for mining this rich source of data, but also to explore and investigate the challenges posed by such real life clinical data while improving the performance of classification algorithms (El Ayadi and Plataniotis, 2010, Guo, 2010, Naidu *et al.*, 2014, Kanj *et al.*, 2012). This would involve an exploration of the underlying properties of the data such as errors, and variability in the data. These factors which accumulate over a period of time and across large number of patients, often contribute towards the challenges of clinical data. The challenges are; missing data (Farooq *et al.*, 2013), high dimensionality (Xue-min *et al.*, 2011), class imbalance (Li *et al.*, 2010) and non-normal distribution (Korkusuz *et al.*, 2011). Numerous methods have been both highlighted and implemented in literature to handle these challenges (Moore *et al.*, 2014, Moore and Kambhampati, 2013, Poolsawad *et al.*, 2012, Poolsawad *et al.*, 2014b, Zhang *et al.*, 2012, Poolsawad *et al.*, 2011, Bohacik *et al.*, 2013b, Poolsawad *et al.*, 2014a, Bohacik *et al.*, 2013a)

In general, classification methods such as Bayes classifier (Moore and Kambhampati, 2013, Moore *et al.*, 2014, Hani *et al.*, 2010) are often employed to classify the data and develop prediction algorithms. However, an investigation of the data found that the primary assumption made for Bayes, namely, that the continuous values associated with each class are in a Gaussian distribution, is not satisfied by the real life clinical data available. Hence, there is a great interdependency between the challenges, the methods applied and the final result. Bayes also delivers results based on a strong independence (Frank *et al.*, 2002, Zhang, 2004) assumption, where variables (in the remainder of the thesis, variables, features and attributes will be interchangeably used) are conditionally independent of each other given the class (McCallum and Nigam, 1998). However, a significant amount of research has been conducted on relaxing the Bayes independence assumption in order to improve its performance. Friedman and co-authors (Friedman *et al.*, 1997) present methods such as Tree Augmented Naïve Bayesian (TAN) classifier and Averaged One-Dependence Estimation (AODE) (Webb *et al.*, 2002) that achieves this by estimating dependences or increasing the number of parameters that are estimated. An advantage is that the independence assumption allows parameters for each variable to be learned separately, mainly when the number of variables is large. In addition, relaxing the assumption allows relevant correlations to be captured and thus improves classification accuracy.

In this thesis, the underlying processes of the Bayes algorithm allow the data to be explored in greater detail. Furthermore, literature suggests that Bayes classifier is robust and less sensitive to missing data and high dimensionality (Peng *et al.*, 2005, Lei *et al.*, 2005, Blomberg and Ruiz, 2013, Shi and Liu, 2011, Tillander, 2012, Tillander, 2013). This is evident from results in literature and those shown later in this thesis (in the case of missing data), where evaluating the effect of different imputation approaches on classification finds that imputation methods could improve the accuracy of the classifier. An exploration and investigation of the classification performance presented on a confusion matrix (Costa *et al.*, 2007, Lalkhen and McCluskey, 2008, Eriksen *et al.*, 2003) are carried out to understand the properties and data space of the real life clinical data.

1.2 Clinical dataset

There are two drivers behind the development of computer based clinical diagnostics aids. One is the change in demographics towards a more aging population, and the other is the prevalence of chronic ailments such as heart failure. Numerous models exist to estimate the risk of patients with heart failure (Cleland *et al.*, 1999), for example, the Seattle Heart Failure Model (SHFM) (Levy *et al.*, 2006), its application in clinical practice to improve the prediction of heart failure (Jong *et al.*, 2012), and Shelton and colleagues (Shelton *et al.*, 2010) proposed risk score method to predict the occurrence of persistent atrial fibrillation in patients with heart failure. The key feature of SHFM is that it collects data regularly without the need for experts. However, Cleland and co-authors (Cleland *et al.*, 1999) state that during the Framingham heart study (Ho *et al.*, 1993) the prognosis of heart failure had not improved between 1975 and 1988. Cleland and colleagues present the Hull LifeLab dataset (Pearson and Cowie, 2005), which serves the purpose of improving and understanding diagnosis, treatment, delivery of care to patients, natural history, and the mechanism and markers of heart failure (Bohacik *et al.,* 2014, Jacobs *et al.,* 2014).

Hull LifeLab is a large, epidemiologically representative, information-rich clinical dataset consisting of patients with possible heart failure referred to a cardiology out-patient clinic. The dataset is from a clinic which serves a mixed urban/rural community of about 550,000 people in Kingston-Upon-Hull and East Riding of Yorkshire between 2000 and 2012. All referred patients were invited to participate in the study and 98% gave informed consent for their information to be retained and used for research purposes. Consenting patients received a comprehensive clinical assessment and those found to have heart failure were followed up with further outpatient assessments at regular intervals (typically every 4-6 months). The dataset is composed of 463 continuous and categorical variables and 2,032 patient records including quality of life. This thesis considers 61 important variables associated with blood chemistry. The reason for considering blood chemistry is that the other variables are either well understood or are essentially categorical, whose values are subjective and dependent on the interpretation of the patients' record by either the nurse or a clinician. These additional variables are also prone to having missing data greater than 20%.

1944 patient records are considered from the live dataset, as the remainder had variables where more than 20% (Acuña and Rodriguez, 2004) was missing. These 1944 patient records were collected at four different time points for example at 3, 6 12 and 18 months. After 18 months, 1459 patients had no record of death and had attended an outpatient clinic therefore classed as alive. The remainder were classed as dead (485) as there was a record of death present for each one of those patient. The classes will be referred to as alive and dead class in the following chapters.

1.2.1 Missing data

Missing data are ubiquitous in datasets, particularly in clinical datasets (Cismondi *et al.,* 2013) due to different ways of collecting the data, for example the transfer of data from diverse health care systems and the transcript of data from EHRs. As a result clinical datasets contain noise, missing data and outliers during data recording or entry due to incorrect measures and mixed variable types (Weitschek *et al.,* 2013). In addition, in cases where clinical data are collected as part of a clinical trial, the medical report pro forma allows certain variables to be left blank. This is usually due to the ailment being treated or perhaps the patient may not wish to disclose certain information, such as whether he or she is a smoker (Zhang *et al.,* 2012).

In order to understand the nature of the data which is missing, it helps to categorise these into three types of missing data, namely (a) Missing Completely at Random (MCAR), (b) Missing at Random (MAR), and (c) Missing Not at Random (MNAR) (Dziura *et al.,* 2013, Blomberg and Ruiz, 2013, Pérez *et al.,* 2002, Little and Rubin, 1987) as shown in table 1.2.

Missing mechanism	**Examples**
MCAR	The missing data type does not depend on observed and unobserved data from the dataset. Data that are MCAR do not exist or are not recorded. This is usually due to a random failure of an experimental instrument or a dropped test tube in the laboratory which may lead to missing data.
MAR	MAR depends only on some other observed data from the dataset caused by the variable of the study design. For example an individual's gender is recorded as male and the variable 'pregnancy' is left blank.
MNAR	A variable with MNAR depends on the non-observation of the target variable. Examples are cases where the patient may be too ill for clinicians to collect the remaining sample for a specific test or a patient is not responding to treatment or may have dropped out because they believe the treatment is not effective.

Table 1.2: Examples of the three missing data mechanisms

Some data mining methods such as Bayes classifier and decision tree are tolerant to missing data (John and Langley, 1995, Kohavi, 1995). However, a number of methods require a complete dataset. Missing data imputation methods can be implemented to achieve the latter. Imputation methods are a key pre-processing step in modelling missing data and are an important data preparation task for data mining applications. Following are some well-known missing data imputation methods: Concept Most Common value Imputation (CMCI), Most Common value Imputation (MCI) (Grzymala-Busse and Hu, 2001), Expectation Maximisation Imputation (EMI) (Gupta and Chen, 2011), Fuzzy *k*-Means clustering Imputation (FKMI) (Li *et al.,* 2004), *K*-Means clustering Imputation (KMI) (Patil *et al.*, 2010), *K*-Nearest Neighbour Imputation (KNNI) (Zhang, 2012, Silva and Hruschka, 2013), and Support Vector Machine Imputation (SVMI) (Yang *et al.,* 2012).

1.2.2 High dimensionality

Clinical data are accumulated with hundreds to thousands of variables (Guyon and Elisseeff, 2003, Balakrishnan *et al.,* 2008, Xiaoyan *et al.,* 2008). High dimensionality in clinical datasets presents the issue of diverse features, such as data containing too many or irrelevant variables. These prevent common data organization strategies from being efficient and thus affect the application of data mining methods. The ‘curse of dimensionality’ refers to numerous phenomena that occur when performance depends on model and computational complexity, data dimensional space and interrelationship among sample size (Clarke *et al.*, 2008). The ‘curse of dimensionality’ in high dimensional data analysis is addressed through the application of data pre-processing methods, such as feature selection and feature extraction (Lee *et al.*, 2013, Balakrishnan *et al.,* 2008, Clarke *et al.,* 2008), for example, the backward search approach (Balakrishnan *et al.*, 2008) and Principal Component Analysis (PCA) (Bishnu and Bhattacherjee, 2012) respectively. Selecting a small number of variables has been shown to be beneficial for classification tasks such as in building prediction models and improving predictive accuracy (Balakrishnan *et al.*, 2008). Dimensionality reduction of data has the advantage, that while reducing dimensionality, the data complexity is also reduced and thus predictive accuracy and the integrity of the data are maintained (Houle *et al.*, 2010).

1.2.3 Class imbalance

Class imbalance is among the leading challenges that reduce the performance of classification and prediction algorithms. Class imbalance occurs when the number of instances of one class is heavily under-represented (minority class) relative to another class (majority class), resulting in an unequal number of observation of classes. Almost all classification algorithms (learning or otherwise) require an even balance of classes in order to be able to identify models and thus increase the performance of the algorithms (Menardi and Torelli, 2014). Class imbalance can also be reflected in the skewed distributions (Shuo and Xin, 2012, Longadge and Dongre, 2013, García *et al.*, 2007) of variables. Hence a key challenge would be to develop a classifier that can provide good accuracy for the minority class prediction as instances are more likely to be misclassified than the majority class instances.

Strategies for dealing with class imbalance have been proposed in literature, such as over sampling and under sampling (Xiaoyuan *et al.*, 2011). Both techniques are known to re-sample the training dataset so that during classification the classifier algorithm receives an equal share of instances per class (Orriols *et al.*, 2005). A recent review (Longadge *et al.*, 2013) suggests that misclassification of a case can result in a major problem particularly in medical and clinical application. For example, in a case of heart failure based on two classes; high risk and low risk, misclassifying the high risk group to low risk group may lead to some additional clinical testing. However, misclassifying low risk as high risk leads to stress and anxiety of the patient as well as a re-evaluation of the application of the classification method.

1.2.4 Non-normal distribution

Clinical data often suffers from non-normal distribution and therefore it is extremely important to consider the degree of non-normality. When faced with data containing non-normal distribution (and skewed), two actions should be considered: (a) identify the cause of non-normality and (b) address non-normality by attempting to transform the data into a normal distribution or applying non-parametric inferential statistical analysis. Non-normality is caused by various factors, for example:

- Outliers (Fleishman, 1978); are usually represented as too many extreme values in the dataset resulting in skewed distributions or a small proportion of the data having a variance greater than the remainder of the population.
- Overlap of two or more processes; for example in the case of medical/clinical settings healthcare practitioners may overlap more than one data during data entry or two or more frequent values.
- Data values close to zero or a natural limit; this causes the data distribution to skew to the left or right.

Often practitioners reach a point in research where the need to adequately perform statistical analyses requires normally distributed data. Common methods of transformation can be applied, such as; logarithmic, square root, reciprocal transformations and Box Cox (Counsell *et al.*, 2011). These methods have the benefit of reducing skewness and introducing equal spreads in the distribution. Also a skewness value of zero indicates a symmetric (normal) distribution and thus the tails on both sides of the mean balance out (Li, 1999).

1.3 Research aim and objectives

The challenges present in clinical datasets have been briefly described above. These challenges affect the outcomes of the algorithms, and also if used within a decision support system could lead to a change of care. Thus the primary aim of this thesis is to investigate the challenges of real life clinical data, including missing data, high dimensionality, class imbalance, and non-normal distribution. A secondary aim is to determine the manner in which the challenges affect classification algorithms in order to improve performance.

1.3.1 Objectives

1. To identify challenges associated with a real life clinical dataset as applied to clinical practice and the most appropriate set of algorithms for the dataset. This would include the following:
    - i Investigate and identify methods for handling missing data
    - ii Investigate the relationship between methods for missing data with a view to develop prediction models and improve classification performance
    - iii Develop an integrated solution using Bayes methods for missing data.
2. Investigate ways of improving classifiers to enhance performance for better clinical prediction models and decision support systems.

The aims and objectives will be assessed through the application of performance evaluation measures (see chapter 4 for the detailed metrics for comparison and performance evaluation) to determine their success and failure.

1.4Thesis overview

This thesis will present data mining methods to investigate the challenges of a real life dataset with a main focus on missing data and distribution of the data. These challenges will be dealt with within the next seven chapters of this thesis. Chapter 2 introduces and discusses a Clinical Data Mining Workflow (CDMW) tailored to real life clinical data. This workflow consist of six steps namely, 1) raw clinical dataset, 2) data exploration, 3) data preparation, 4) Modelling, 5) Evaluation and 6) CDM. The workflow also consists of three stages, namely, descriptive, predictive and prescriptive. Chapter 3 discusses the first stage, the descriptive stage. This stage involves exploring the descriptive statistics of the data such as the distribution and statistical measures of the original data and imputed data. The data is imputed with seven different missing data imputation methods, namely, Concept Most Common value Imputation (CMCI), Expectation Maximisation Imputation (EMI), Fuzzy k-Means clustering Imputation (FKMI), k-Means clustering Imputation (KMI), k-Nearest Neighbour Imputation (KNNI), Most Common value Imputation (MCI) and Support Vector Machine Imputation (SVMI). Chapter 4 considers the information learnt about the data to predict future outcomes of the heart failure data. This chapter presents the different types of performance metrics used by naïve Bayes and TAN classifiers to present the classification performance of the original dataset and imputed data. Based on these results, chapter 5 assesses the different classes using the different imputation methods and then combines them into one dataset known as a hybrid imputed dataset e.g. SVM and EM. The class data record and posterior probabilities are explored to understand why misclassification occurred. Euclidean distance is also applied to determine the similarity in the data records and what variable is contributing the most. In chapter 6, other classification algorithms such as Bayes classifier based on Kernel Density Estimation (KDE), beta distribution based Bayes classifier, decision tree (C4.5) and Multilayer Perceptron (MLP) are discussed and implemented. The algorithms will be applied on the original data and SVM and EM hybrid imputed data for comparative analysis. The classification outcome in both types of data will be discussed and an explanation offered as to why these algorithms were not initially considered. The thesis is concluded in chapter 7 with a summary of the main contribution of the thesis and suggestions for future work. All experiments carried out in this thesis were performed using software provided within MATLAB (MathWorks, 2005) and WEKA (Hall *et al.*, 2009, Witten and Frank, 2005).

CHAPTER 2-FRAMEWORKS FOR MINING DATA

2.1 Introduction

The cornerstone for successfully mining data to gather more and new information is the ability to collect data. The more data that is available the better. However, the exponential growth of data often comes with a number of challenges (as discussed in chapter 1). Given these challenges, mining the data is often performed within a framework, which consists of a cyclical sequence of steps, namely (a) data pre-processing, (b) modelling and (c) prediction (d) evaluation and decision making. Often these four steps are a collection of smaller sub-steps. Depending on the nature of the problem and the various sub-steps a number of different frameworks have been developed such as Cross Industry Standard Process for Data Mining (CRISP-DM) (Wirth and Hipp, 2000) and Sample, Explore, Modify, Model and Assess (SEMMA) (Obenshain, 2004, Cerrito, 2006).

This chapter will outline these steps and also discuss some of the more important frameworks for data mining. However, these frameworks deal more with business oriented problems and data, which is not applicable to the clinical problem. Thus these are modified for the application at hand. This modification results in a Clinical Data Mining Workflow (CDMW) more suited for clinical data gathered from live clinics. This workflow consists of six steps, namely, 1) raw clinical dataset, 2) data exploration, 3) data preparation, 4) modelling, 5) evaluation and 6) Clinical Decision Making (CDM). From the discussion it can be seen that most frameworks have three broad stages: descriptive, predictive and prescriptive (Hand *et al.,* 2001, Lejeune, 2001, Delen and Demirkan, 2013, Kaisler *et al.,* 2013). These are intuitive stages, and most frameworks often use and follow these stages with varying degrees of detail or methodology within them.

2.2 Data mining frameworks

Data mining frameworks provide a methodology not only to deal with challenges posed by the data, but also to develop a deeper understanding of the domain of application through machine learning methods (Han and Kamber, 2006). In this section two popular frameworks are discussed, namely, CRISP-DM and SEMMA. These standards have been selected as they define the process of data mining for knowledge discovery in various topics (Potamias and Moustakis, 2001, Huifang and Ding, 2010, Bosnjak *et al.*, 2009)

2.2.1 CRISP-DM

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a comprehensive process model for carrying out data mining projects. The process model aims to make large data mining projects cost effective, faster, reliable, repeatable and manageable. Studies report that the CRISP-DM process model is not only beneficial but also provides an overview of the life cycle of a data mining project applicable in many industry sectors (Wirth and Hipp, 2000). Recent studies suggest that there has been limited application within the healthcare domain (McGregor *et al.,* 2012, Huang *et al.,* 2014). However, successful mining applications have been implemented in the healthcare field, three of which are: hospital infection control, ranking hospitals and identifying high-risk patients (Obenshain, 2004).

CRISP-DM is presented as a cyclical process that comprises of six steps (fig 2.1) (Chapman *et al.*, 2000), namely,1) business understanding, 2) data understanding, 3) data preparation, 4) modelling, 5) evaluation and 6) deployment. These steps define the inputs, outputs and general strategies to be applied in each step. The following steps are explained in detail.

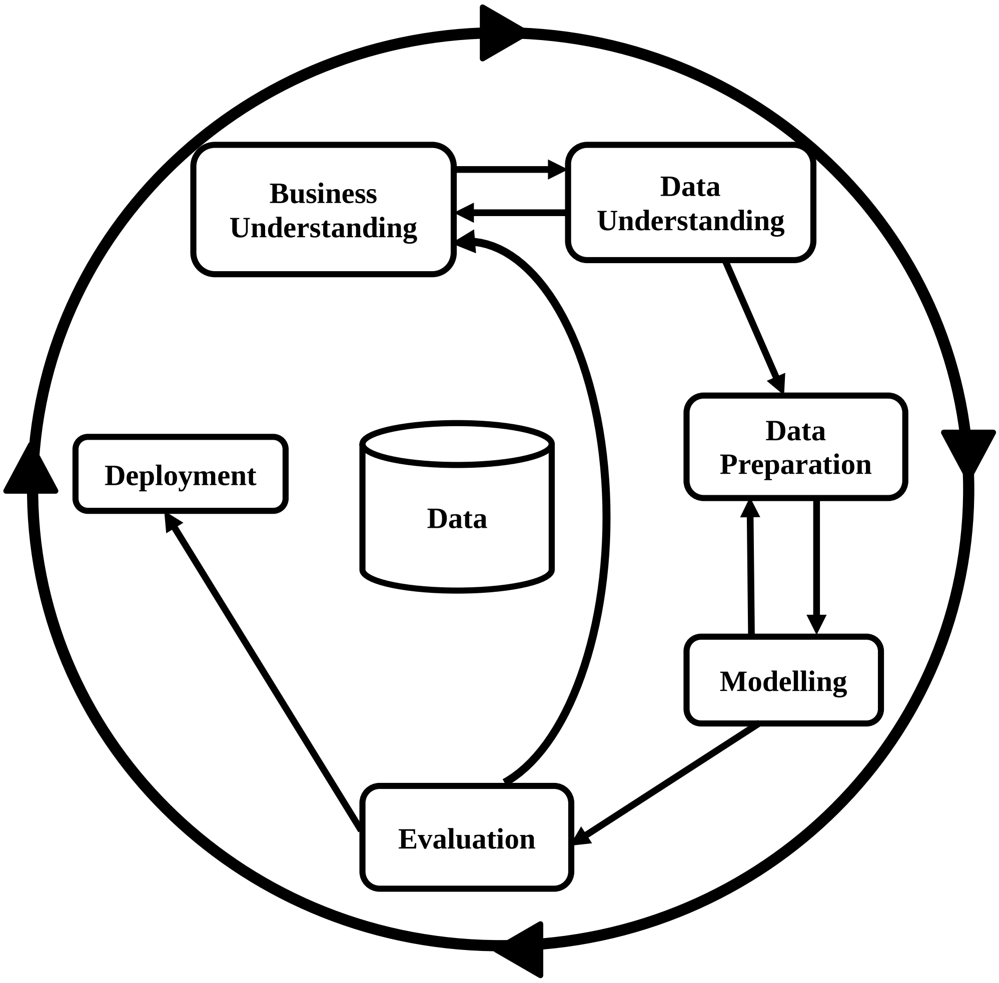


Figure 2.1 Steps of the CRISP-DM process model

1) **Business understanding:** focuses on firstly understanding the objectives and requirements of the project from a business perspective and secondly, using this knowledge to determine data mining goals and lastly propose a project plan to achieve the set objectives.

2) **Data understanding:** This step involves and considers data requirements such as the initial data collection, description, exploration and quality of the data, for example using descriptive modelling such as clustering to identify the data quality and discover insights into the data.

3) **Data preparation:** Data cleaning, variable selection, data integration and transformation are performed in this step; this is to successfully feed the data into modelling tools.

4) **Modelling:** Various predictive modelling (Weiss, 1998) techniques could be selected and applied in this stage. Some models have specific requirements on the form of data; therefore the analyst may often re-visit the data preparation stage if necessary. The models are then applied to analyse and predict the probability of a desired outcome.

5) **Evaluation:** This step is the review process and evaluation of results, for example data mining results and models are assessed to be certain that the model properly achieved the business objectives. The stage also includes the application of predictive modelling where the decision and actions of the evaluated results/models are also expressed by taking advantage of the predictions made.

6) **Deployment:** The knowledge gained from the data needs to be organised and presented in a way that is beneficial to the consumer, i.e. healthcare practitioners. In order to achieve this, firstly a deployment plan is created, which includes necessary steps and how to perform them; secondly the final report is produced.

The tasks follow each other as a sequence of steps, but within this main stream, many iterative cycles can be observed. This can be explained by the fact that the output of each phase influences the next methodological step. For example, after the data understanding step, the user often has to return to the business understanding and reconsider the aims, objectives and reasons for Knowledge Discovery in Data (KDD) (Bosnjak *et al.*, 2009). Similarly, after the data modelling step, a new data pre-processing may be required in the data preparation step in order to improve the data models and develop additional ones.

2.2.2 SEMMA

SEMMA is a methodology oriented process that clarifies the data mining process through an analysis cycle. SEMMA was developed by the SAS Institute Inc. (Cerrito, 2006, Obenshain, 2004, Azevedo and Santos, 2008, SAS). The acronym SEMMA represent Sample, Explore, Modify, Model, Assess which are the five data mining processing steps. Figure 2.2 shows the SEMMA steps in the SEMMA analysis cycle (Obenshain, 2004) and these steps can be performed iteratively as needed.

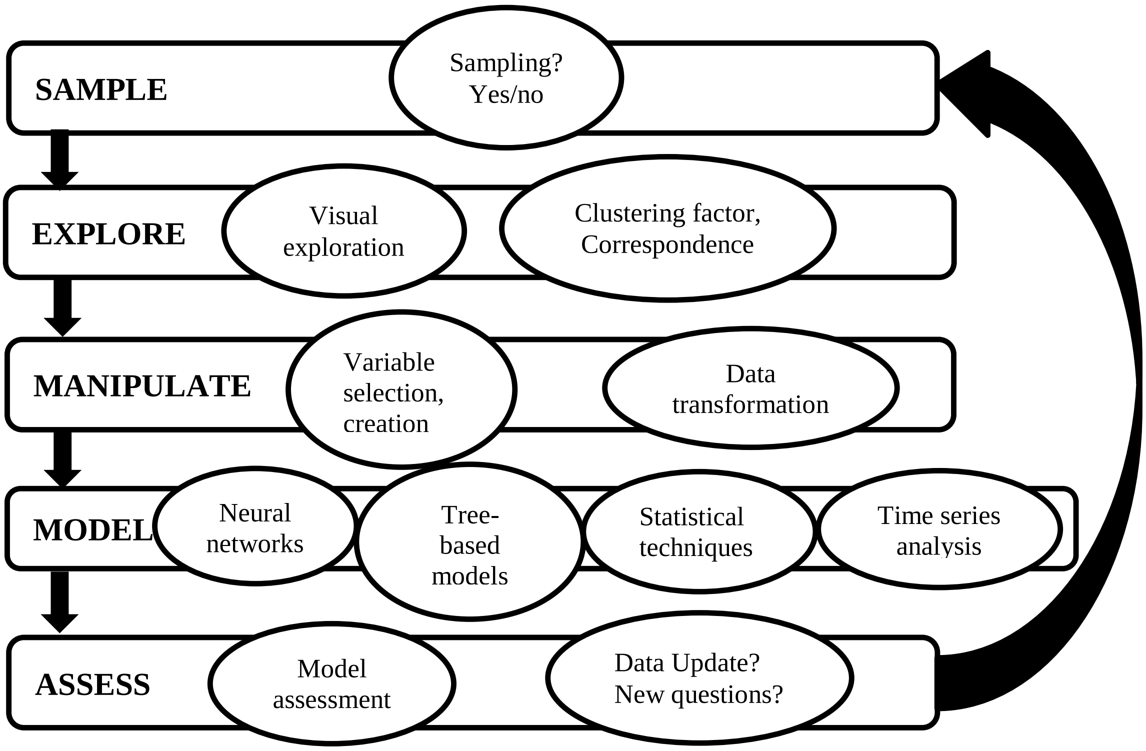


Figure 2.2 The SEMMA analysis cycle

1) **Sample:** This step involves the sampling of the data; extraction of a large portion of the data that contains significant information but yet small enough to compute.
2) **Explore:** Exploration of the data involves searching for trends, patterns and anomalies in order to gain deeper understanding and idea, for example descriptive modelling such as clustering methods to group observations for better knowledge.

3) **Manipulate:** Data quality is essential for data mining; data records with underlying challenges such as missing data for one or more variables could obstruct some of the patterns. Therefore data modification is paramount; this involves creating, selecting and transforming the variables to focus the model selection process.

4) **Model:** This step involves using predictive modelling (Hand *et al.*, 2001)such as analytical tools, for example neural network and decision tree to search for patterns that predict a desired outcome.

5) **Assess:** This is a prescriptive modelling step that involves the evaluation of the usefulness and reliability of findings from the data mining process. This step allows the user to assess the performance of the model. This is commonly executed by applying the model to a different dataset and repeating steps 2, 3 4 and 5, if the model is valid, it will work for the data as well as for the sample data used to construct the model. However, if the model does not work, the user can repeat the entire process again, starting with sampling.

The SEMMA process provides an easy to understand process, allowing adequate development and maintenance of data mining projects. In contrast to the CRISP-DM process, SEMMA also allows the user to return to previous steps in the process and focus mostly on the application to exploratory statistical and visualization-based data mining techniques (Bellazzi and Zupan, 2008).

2.2.3 Clinical Data Mining Workflow

CRISP-DM and SEMMA present solutions to business problems and achieve business goals. However, in order to be applicable within a clinical setting, they both have to be tailored and thus a clinical workflow has been developed. Although the inspiration behind the workflow is due to the objectives and challenges outlined in chapter 1, this workflow is generic and is well suited to most clinical applications, e.g. home telemonitoring and in situations where data is streamed fast and in reasonable quantities. Figure 2.3 shows an iterative data mining process for implementing machine learning methods on the Hull LifeLab dataset in order to support CDM for personalised care. As mentioned earlier the frameworks are often developed to present a cyclical set of steps. This is a natural consequence of the manner in which the performance of the various components of the framework is tested. This result in sets of concentric steps presented below. The six steps in the workflow can be categorised into the following stages: (a) Descriptive, where exhaustive exploration of the data is carried out, (b) Predictive, where modelling of the data is carried out through the use of predictive models and (c) Prescriptive, where the full methodology is evaluated and then modifications are made to either the data or the modelling strategy. These are further discussed in the next section when the three stages are compared on the frameworks and workflow.

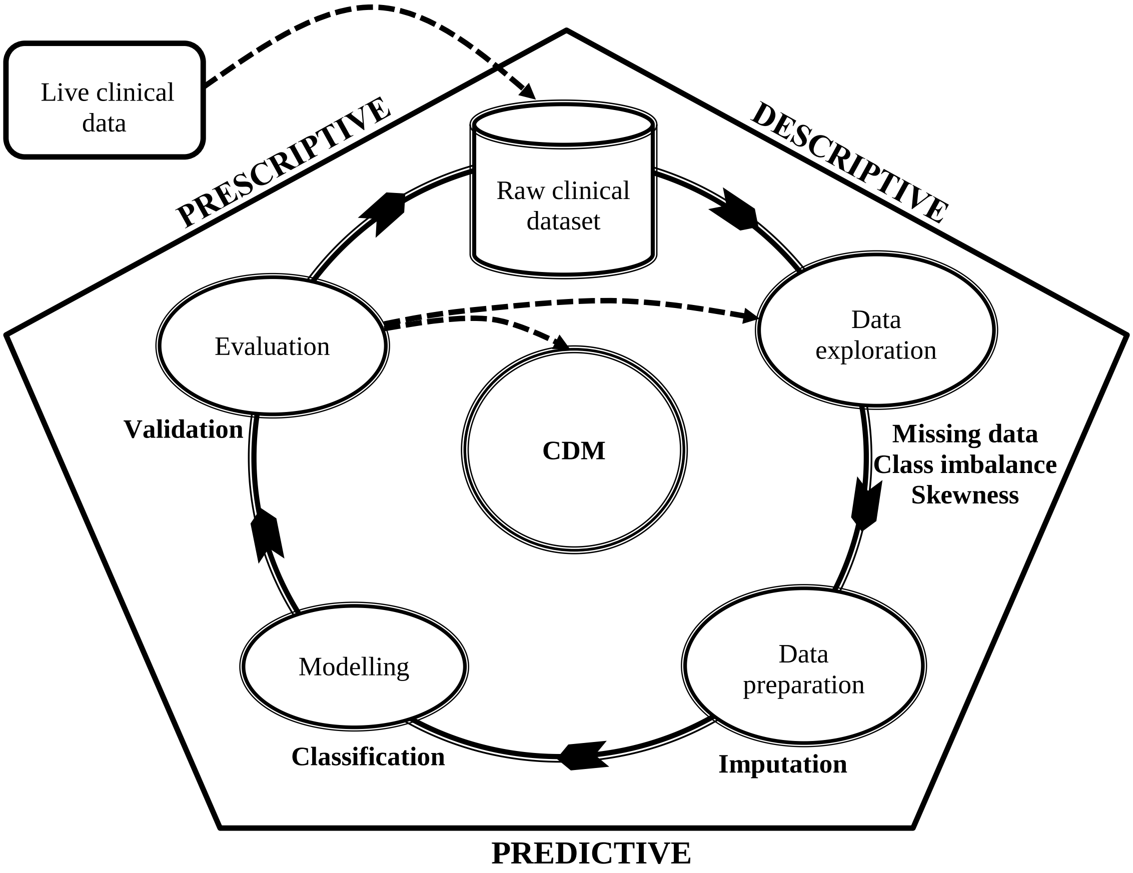


Figure 2.3 Clinical Data Mining Workflow (CDMW)

1) **Raw clinical dataset:** The first step of the workflow is to sample the related data from the available sources. The raw clinical dataset is sampled from a live clinical data provided by clinicians and healthcare providers based in Castle Hill Hospital, East Riding of Yorkshire, United Kingdom. The term ‘raw’ and ‘original’ data are interchangeably used in this thesis. The clinical data is real time data, which includes blood chemistry and categorical variables and the various acronyms used in the data. A large portion of the data containing the significant information was extracted from live clinical data and irrelevant variables were removed, such as patient personal information, for example, patient id, gender and a blood chemistry variable, namely, NT-proBNP. This was because the majority of the data for this particular variable was missing; as a result the variable was eliminated. Gender was a simple binary 0, 1 and as a result we were told by the

clinicians not to consider this. However, the main goal is to generate a population risk model and not a risk prediction based on gender. This step is similar to the *‘sample’* step of SEMMA.

2) **Data exploration:** This step is similar to the ‘explore’ step of SEMMA and *‘data understanding’* step of CRISP-DM. The data is explored to observe characteristics of the dataset such as those discussed in chapter 1: missing data, high dimensionality, class imbalance and non-normal distribution. The descriptive stage is applied to gain better insight of the properties of the dataset, for example, exploring the data distribution such as the different types of distribution represented in the data and the number of missing data present.

3) **Data preparation:** The task of this step shares a similarity with the *‘manipulate’* step of SEMMA and *‘data preparation’* step of CRISP-DM. Pre-processing is the main process for addressing the challenges of the dataset, such as those identified during the data exploration step, for example, the application of missing data imputation and feature selection methods to impute the missing data variables and to select relevant variables in order to improve classification accuracy.

4) **Modelling:** Once the data is prepared, a predictive model is constructed to explain patterns and extract useful knowledge from the data in order to predict future outcomes. This step is a common and crucial step in data mining frameworks as it is also present in SEMMA and CRISP-DM to predict outcomes in business projects. The development of the model is dependent on the data exploration and data preparation steps. For example other challenges that were not explored during the data exploration step may contribute to the performance of the model and as a result another pre-processing technique will be required to tackle the challenge. In this step of the workflow, a classification approach such as Bayes classifier can then be applied to the pre-processed data. Classification is a fundamental issue in

machine learning and data mining that classifies a set of given observations into existing classes (Kesavaraj and Sukumaran, 2013).

5) **Evaluation:** Involves assessing the classification model to present an optimal performance of the classification task through the use of evaluation metrics, for instance a confusion matrix which is a table layout that assesses accuracy, precision, characterises errors and aids to refine statistical measures for classification test (Foody, 2002). The task of this step is equivalent to the task of the *'assess'* step of SEMMA and the *'evaluation*' step of CRISP-DM.

6) **Clinical Decision Making (CDM):** This step involves the presentation of the entire data mining process to clinicians and healthcare practitioners in order to successfully support them in the CDM process. This involves a strategic plan of the outcome of each step and how each step has influenced the final result and its benefit for personalised care. This step shares a similar task to that of the *'deployment'* step of CRISP-DM, which requires a plan in order to organise and present the findings and the purpose they serve for business.

2.3 Comparison of the data mining frameworks

It can be seen above that all three frameworks, CRISP-DM, SEMMA frameworks and CDMW employ the three stages: descriptive, predictive and prescriptive. Although some of the steps do not share an identical title, however their task is very similar. For example, the 'data understanding' step of CRISP-DM, the 'explore' of SEMMA and the 'data exploration' step of CDMW all share the same task and represent the descriptive stage. Figure 2.4 presents the differences between the three business analytic stages (Delen and Demirkan, 2013) and table 2.1 shows where the three stages are represented in the frameworks.

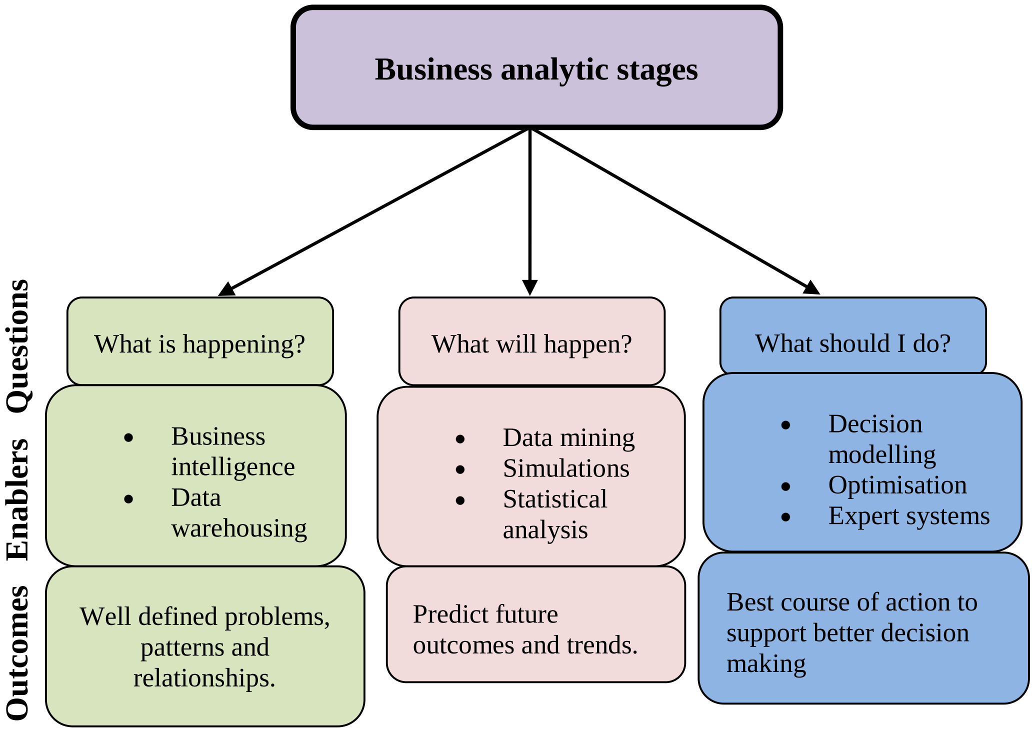


Figure 2.4 The three stages of business analytics

CRISP-DM		SEMMA		CDMW	
Step	**Stage**	**Step**	**Stage**	**Step**	**Stage**
Business understanding	Descriptive	Explore	Descriptive	Data exploration	Descriptive
Data understanding	Descriptive	Model	Predictive	Data preparation	Descriptive
Modelling	Predictive	Assess	Prescriptive	Modelling	Predictive
Evaluation	Prescriptive	-	-	Evaluation	Prescriptive

Table 2.1: Comparison of the three stages in CRISP-DM, SEMMA and CDMW

The descriptive stage allows the data analyst to investigate the important aspects that represent and describe the data by answering the question of '*what is happening and/or current in the data?*' It includes data warehousing and business intelligence to outline and identify problems of the data and thus assist in understanding the data. Types of strategies applied in the descriptive stages include models describing the relationship between variables, such as cluster analysis and segmentation (Hand *et al.,* 2001).

The predictive stage uses statistical analysis and data mining techniques such as classification and regression to discover trends and patterns representing the relationships between data inputs and outputs to predict future outcomes and *'what will happen'.*

The prescriptive stage evaluates the full methodology to determine the beast course of action and then further modifications are made to support better decision making and outcomes. This involves contribution from decision modelling, expert knowledge and systems to answer the question, *'what should happen or what should I do?'*

2.4 Summary

CRISP-DM, SEMMA and CDMW frameworks make data mining more effective and efficient for CDM. The frameworks follow the intuitive stages shown in fig 2.4. However, what is different is the relationship between the three stages, which are often tailored for the application and nature of data. In contrast there are two key challenges involved in data mining steps: 1) agreeing on a data preparation method such as a data cleaning or pre-processing technique so that data mining methods are successfully implemented and 2) agreeing on a predictive model to predict future outcomes. In spite of this, the main goal of the frameworks consists of a particular course of action, to understand, evaluate and compare data which are mainly intended to achieve a result. Many other frameworks exist to achieve the same goals, such as Knowledge Acquisition and Documentation Structuring (KADS) (Wielinga *et al.*, 1992), Knowledge Discovery in Databases (KDD) (Azevedo and Santos, 2008) and Predictive Data Mining Markup Language (PMML) (Bellazzi and Zupan, 2008).

CHAPTER 3-THE DESCRIPTIVE STAGE

3.1 Introduction

As discussed in chapter 2, frameworks and workflow for mining of data comprise three stages namely; (a) descriptive, (b) predictive and (c) prescriptive (see figure 2.4). The following chapters of this thesis will use these stages to discuss the challenges in mining clinical datasets. Thus this chapter will focus on the descriptive stage. The descriptive stage consists of two steps: 1) data exploration and 2) data preparation. The data exploration step explores the data to find key information on the space occupied by the data, e.g. mean, standard deviations, median and skews of the variable. On the other hand, the data preparation step is more active, in that dependent on the methods used for preparation, some of the properties mentioned earlier can change. Essentially this step comprises pre-processing methods such as missing data imputation to impute missing data. The key to this step is to ensure that the properties of the data distribution such as the mean and standard deviation are not altered significantly.

3.2 Data exploration

The Hull Lifelab dataset applied in this thesis consist of 463 variables comprising categorical, continuous and the clinicians' summarisation of patients' past records. However, not all variables are applicable to the problem at hand; as a result 60 variables (continuous) are considered. The 60 variables are shown in table 3.1. These variables are recommended by clinicians and are a summation of the variables in the Seattle heart failure dataset (Levy *et al.*, 2006), the Framingham heart study (Tsuji *et al.*, 1994, Ho *et al.*, 1993) and the Mid-Atlantic Group of Interventional Cardiology (MAGIC) congenital heart disease (Everett *et al.*, 2006).

The statistical measures of the data are presented to capture and understand the properties of the data distribution. In addition, the minimum (min) and maximum (max) values, percentage of missing data (% of MD), mean (μ), standard deviation (σ), median and skew values of the variables are presented.

No.	**Variable**	**No.**	**Variable**
1	Age (years)	31	MR-proADM
2	Sodium (mmol/L)	32	CT-proET1
3	Potassium (mmol/L)	33	CT-proAVP
4	Chloride (mmol/L)	34	PCT
5	Bicarbonate (mmol/L)	35	ECG (bpm)
6	Urea (mmol/L)	36	QRS width msec)
7	Creatinine (umol/L)	37	QT (msec)
8	Calcium (mmol/L)	38	LVEDD (cm)
9	Adj Calcium (mmol/L)	39	LVEDD (Hgt indexed)
10	Phosphate (mmol/L)	40	BSA (m^2)
11	Bilirubin (umol/L)	41	Aortic Root (cm)
12	Alkaline Phosphatase (iu/L)	42	Left Atrium (cm)
13	ALT (iu/L)	43	Left Atrium (BSA Indexed)
14	Total protein (g/L)	44	Left Atrium (Hgt indexed)
15	Albumin (g/L)	45	Aortic Velocity (m/s)
16	Uric acid (mmol/L)	46	E
17	Glucose (mmol/L)	47	Height (m)
18	Cholesterol (mmol/L)	48	Weight (kg)
19	Triglycerides (mmol/L)	49	BMI (kg/m^2)
20	Haemoglobin (g/dL)	50	Pulse (bpm)
21	White Cell Count (10^9/L)	51	Systolic BP (mmHg)
22	Platelets (10^9/L)	52	Diastolic BP (mmHg)
23	MCV (fL)	53	Pulse BP (mmHg)
24	Hct (fraction)	54	FEV1 (L)
25	Iron (umol/L)	55	FEV1 Predicted (L)
26	Vitamin B12 (ng/L)	56	FEV1 Predicted (%)
27	Ferritin (ug/L)	57	FVC (L)
28	CRP (mg/L)	58	FVC Predicted (L)
29	TSH (mU/L)	59	FVC Predicted (%)
30	MR-proANP	60	PEFR (L)

Table 3.1: Variables of the Hull LifeLab dataset.

3.2.1 Distributions of clinical dataset

Data distribution describes the characteristics of data. Understanding the characteristics of the data provides a deeper understanding of how missing data imputation methods affect the distribution and thus performance of the classifier.

The distributions of data for all the variables were looked at, and in what follows are three key variables (sodium, creatinine and uric acid). The distribution of the full set can be seen in Appendix I. The three variables are used by clinicians in diagnosing heart failure (Schrier, 2008, Ochiai *et al.*, 2005, Chamorro *et al.*, 2002, Shelton *et al.*, 2010, Zamora *et al.*, 2007, Cowie *et al.*, 2000).

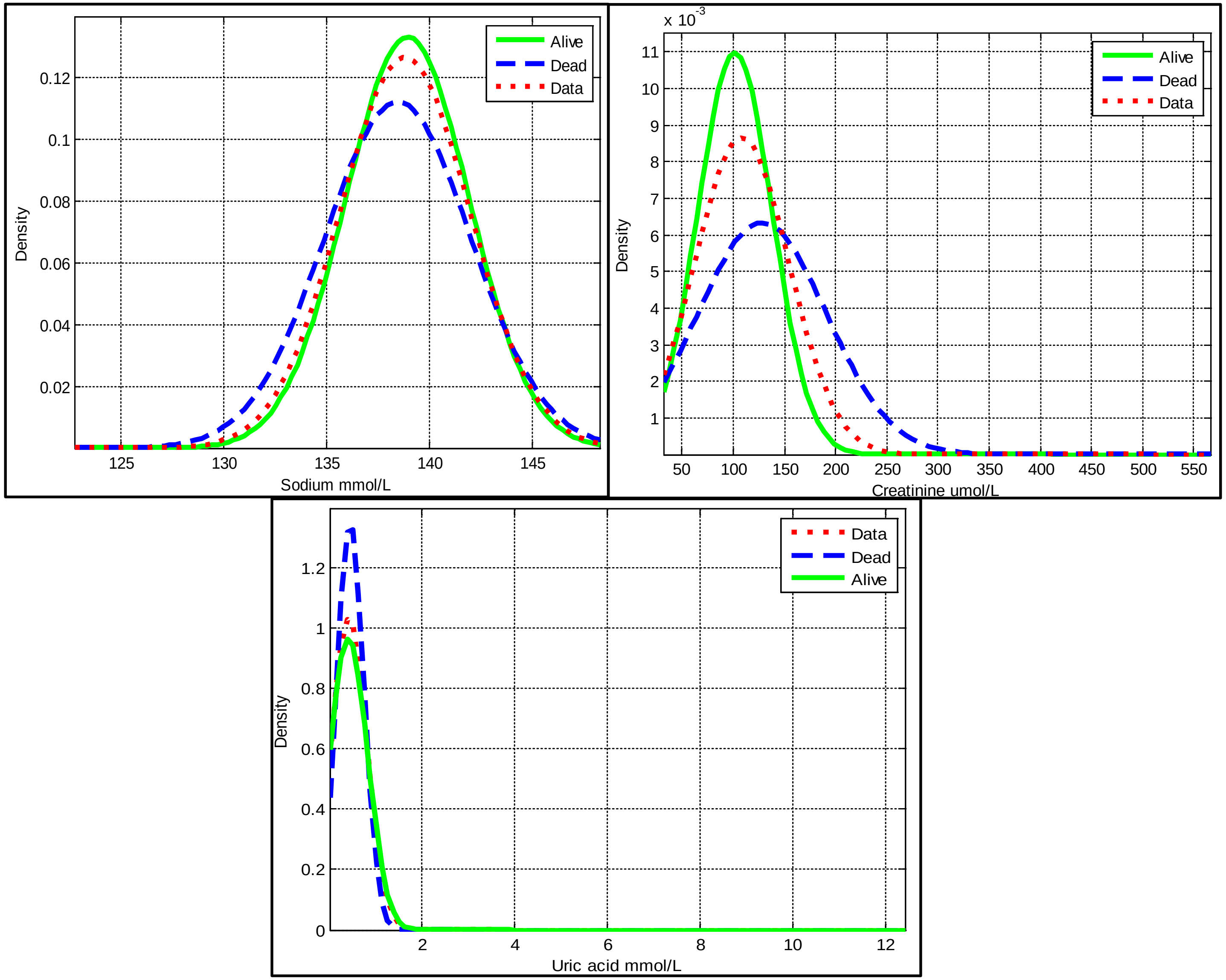


Figure 3.1 Distributions of the data and classes for sodium, creatinine and uric acid variables.

The graphs in figure 3.1 show the distribution of the overall data, dead class and alive class for the sodium, creatinine and uric acid variables. It can be seen that the sodium and creatinine variables show the most spread in their distribution for all three data groups (overall data, dead and alive class) whereas uric acid presents a more compact distribution. This could be due to the values of the variable, where the values of uric acid are almost identical or similar and often within a small narrow range. The alive class shows a tighter distribution in sodium and creatinine when compared to the overall data and dead class, whereas the dead class shows a large spread and variation. The reason for this could be due to the class imbalance present, where the alive class is represented by more records (1459) than the dead class (485); hence, the distribution for the alive class is much tighter than that of the dead class. The skew values are also larger in the alive class (table 3.2) due to the large representative of samples. Thus the distributions for the alive class data will lean towards the overall data distributions. Creatinine and uric acid also show extreme values, i.e. outliers in their distributions see appendix I, indicated by arrows. These outliers are particularly identifiable in uric acid due to the compactness of the distribution.

Overall data							
Variables	**Min**	**Max**	**% of MD**	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	**Median**	**Skew value**
Sodium	123	148	3	138.81	3.16	139	-0.88
Creatinine	37	561	3	108.45	46.03	98	3.27
Uric acid	0.13	12.3	18	0.42	0.39	0.39	22.68
Dead class							
Variables	**Min**	**Max**	**% of MD**	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	**Median**	**Skew value**
Sodium	125	148	2	138.45	3.56	139	-0.71
Creatinine	37	561	2	128.22	62.99	112	2.65
Uric acid	0.17	5.7	14	0.45	0.29	0.43	13.67
Alive class							
Variables	**Min**	**Max**	**% of MD**	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	**Median**	**Skew value**
Sodium	123	148	4	138.93	3.00	139	-0.93
Creatinine	38	512	4	101.75	36.30	94	3.09
Uric acid	0.13	12.3	19	0.42	0.41	0.38	23.34

Table 3.2: Statistical measures of sodium creatinine and uric acid for the overall data, dead class and alive class.

Table 3.2 presents the statistical measures of the overall dataset, dead and alive class. It can be seen that the skew value is dominant in the alive class for all three variables except creatinine of the overall dataset. The alive skew values are also similar to those of the overall data this could be due to the class imbalance problem as mentioned previously. The alive statistical measures also lean towards the overall data. Uric acid shows the highest skew value in all three groups of data, particularly in the alive. This is reflected and visually seen from the graphs shown in figure 3.1, where the distribution is shown to be very compact and thus highly skewed. This is also reflected in the low μ and σ values shown, where the dead class has a σ value of 0.29. This indicates a lack of variation and dispersion in the variable, whereas creatinine shows a relatively high σ value in all three groups of data.

Previous studies (von Hippel, 2005, Sematech, 2006) have applied the relationship of the mean and median to show how negative and positive skew distributions can be identified. Where the mean is greater than or to the right of the median, it is known as positive skew while negative skew is indicated when the mean is less than or to the left of the median. Sematech (Sematech, 2006) further states that the distribution is symmetric if the mean is equal to the median. These characteristics have been applied here to show what type of distribution the variables represent. Table 3.3 shows a comparison of the skew distribution types in the three groups of data.

Overall data			
	Sodium	Creatinine	Uric acid
Skew	−0.88	+3.27∗	+22.68
Dead class			
	Sodium	Creatinine	Uric acid
Skew	−0.71	+2.65	+13.67
Alive class			
	Sodium	Creatinine	Uric acid
Skew	−0.93	+3.09	+23.34

Table 3.3: Comparison of skew distribution types

It can be seen that creatinine and uric acid are both positively skewed in all three groups of data with the mean values greater than the median values while sodium is negatively skewed with the mean values less than the median value.

3.3Data preparation

There are a number of methods that can be applied to improve the underlying challenges of the clinical dataset for data mining algorithms, for example, missing data imputation for handling missing data, feature selection for high dimensionality, under- and over sampling for class imbalance and transformation methods for non-normal distribution (discussed in chapter 1). Also it has been acknowledged in chapter 1 that Bayes classifier is less sensitive to missing data and high dimensionality. However, this section and the remainder of this thesis will consider seven different missing data imputation methods. Seven imputation methods are implemented to maintain the richness of the dataset and as a result given the number of variables present we can get closer to the expected mean of the whole population. We could eliminate variables with 10%-20% missing data but this would mean discarding important information about the data and therefore a poor judgement will be made about the ailment in question. While the focus is on missing data, there is a great interdependency between missing data and the other challenges such as class imbalance and non-normal distribution, as well as methods applied and the final result. This focus will aid in understanding the mechanics of incorrect classification of records, the relationships between imputation methods and how imputation methods affect the statistical measures (mean, standard deviation, median and skew values). Thus the distribution of the data and class problem will be considered.

There are various missing data imputation methods available; however seven have been found to be most useful for clinical data (Zhang *et al.*, 2012). These are: Most Common value Imputation (MCI)(Zhang *et al.*, 2012), Concept Most Common value imputation (CMCI) (Grzymala-Busse and Hu, 2001), Expectation Maximization Imputation (EMI) (Musil *et al.*, 2002, Dempster *et al.*, 1977), *k*-Nearest Neighbour Imputation (KNNI) (Batista and Monard, 2003), *k*-Means clustering Imputation (KMI) (Li *et al.*, 2004, Žalik, 2008), Fuzzy *k*-Means clustering Imputation (FKMI) (Sarkar and Leong, 2001, Liao *et al.*, 2009) and Support Vector Machine Imputation (SVMI) (Pelckmans *et al.*, 2005, Gunn, 1998, Honghai *et al.*, 2005). A brief discussion of each method will be presented to understand their task. Distributions of the three variables (sodium, creatinine and uric acid) after implementation of the missing data imputation methods will be shown as well as their statistical measures.

3.3.1 Most common value imputation

Most Common value Imputation (MCI) is one of the simplest methods to implement amongst existing methods (Zhang *et al.*, 2012). Depending on the attribute data type, there are differences in the manner in which MCI replaces missing data. For example, for a nominal attribute MCI imputes missing data with the mode; the most common value mode of the attribute, for numerical attributes the missing data is replaced with the mean value of the attribute. Whereas for symbolic attributes, every missing attribute value is replaced by the most common attribute value (Kantardzic, 2011). A disadvantage of the approach is that it can severely distort the distribution for this variable, leading to complications with summary measures including, notably, underestimates of the standard deviation. Moreover, mean imputation distorts relationships between variables by pulling estimates of the correlation toward zero (He, 2010, Little, 1992, Pigott, 2001).

3.3.2 Concept most common value Imputation

Concept Most Common value imputation (CMCI) is similar to MCI. However CMCI imputes missing data by taking into account the most common value of the attribute but uses attributes belonging to the given class instead of applying global most common value (Grzymala-Busse and Hu, 2001). In cases where the attribute data type is nominal the missing data is replaced by the mode, numerical attributes are replaced by a mean value and symbolic attributes are replaced by the most common attribute value that occurs for the class.

3.3.3 Expectation maximization imputation

Expectation Maximization Imputation (EMI) imputes missing data through two iterative steps, namely, the Expectation (E) step and the Maximization (M) step (Gold and Bentler, 2000, Dempster *et al.*, 1977). The former step computes the conditional expected log likelihood value of the X data using the observed data and the current parameter estimates. The expected value of x_1, given the measurement y_1 and based upon the current parameter estimates, is computed as per Moon (Moon, 1996).

$$x_1{}^{[k+1]} = E[x_1|y_1, p^{[k]}] \tag{3.1}$$

where:

$p^{[k]}$, indicate the estimate of p after the kth iteration, $k = 1,2,\ldots$

In the latter step, the expected log likelihood obtained in the E step is maximised by maximum likelihood to obtain and update the model parameter estimates (Musil *et al.*, 2002). This maximised data is used to impute the missing data. Moon (Moon, 1996) presents the steps of the EM algorithm for imputing missing data which is outlined in figure 3.2. The algorithm is iterated until convergence is achieved in the final step, such as when the parameter estimates converge to some criterion (Dempster *et al.*, 1977).

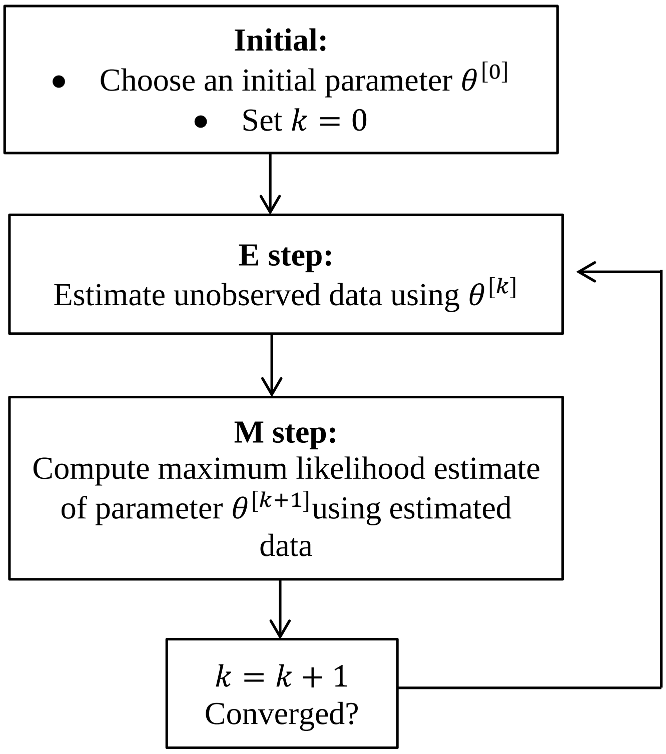

Initial:
● Choose an initial parameter $\theta^{[0]}$
● Set $k = 0$
E step:
Estimate unobserved data using $\theta^{[k]}$
M step:
Compute maximum likelihood estimate of parameter $\theta^{[k+1]}$using estimated data
$k = k + 1$
Converged?

Figure 3.2 The EM algorithm

3.3.4 K-nearest neighbour imputation

k-nearest neighbour imputation (KNNI) imputes missing data using values calculated from the k nearest neighbours (Jonsson and Wohlin, 2004) . The most similar neighbours are found by minimising a distance function such as Euclidean distance eq.3.2; however other distances are also used depending on the nature of the attributes.

$$E(a,b) = \sqrt{\sum_{i \in n}(x_{ai} - x_{bi})^2} \qquad (3.2)$$

where

- $E(a,b)$ is the distance between the two cases a and b
- x_{ai} and x_{bi} are the values of attribute i in cases a and b respectively
- n is the set of attributes without missing data in both cases.

There are two benefits associated with the use of this approach. One is that KNNI can predict the most frequent value among the k-nearest neighbour (qualitative attributes) and the mean among the k-nearest neighbour (quantitative attributes). The other is there is no need to create a predictive model for each attribute with missing data (Batista and Monard, 2003). The KNN algorithm is robust in that it can be easily adapted to work with any attribute as class by simply modifying the attributes to be considered in the distance metric. However, the approach has a few drawbacks that limit its use for Knowledge Discovery from Databases (KDD). For example the algorithm searches through the entire dataset for the most similar instances, making it computationally expensive. Several efforts are presented both by Batista and Monard for solving this limitation (Batista and Monard, 2003).

3.3.5 K-means clustering imputation

The general purpose of clustering is to divide the dataset into groups based on similarity of objects and to minimize intra-cluster dissimilarity (Li *et al.*, 2005, Li *et al.*, 2004). *K*-Means clustering is a simple and fast method applied in many areas such as data analyses, image processing and pattern recognition (Žalik, 2008). Unlike CMCI and KNNI, which use some form of a measure of similarity, *K*-Means clustering Imputation (KMI) focuses on dissimilarity. KMI computes the intra-cluster dissimilarity based on the summation of distances between the objects x_t (also called input data points) and the centroid of the cluster they are assigned to. A cluster centroid (also called cluster centre) represents the mean value of the objects in the cluster (Žalik, 2008, Li *et al.*, 2004). Data objects that belong to the same cluster are taken to be nearest neighbours of each other, and KMI applies a nearest neighbour algorithm to replace missing data in a similar to KNNI (Li *et al.*, 2004).

Numerous forms of the k-means algorithm are presented in literature. However, Zalik (Žalik, 2008) introduces an efficient version of the algorithm shown in table 3.4 Nobjects $x_1, x_2, \dots, x_N$ into k disjoint subsets c_i $i = 1, \dots, k$, each containing n_i objects, $0 < n_i < N$, minimizes the following Mean Square Error (MSE) cost function:

$$J_{MSE} = \sum_{i=1}^{k} \sum_{x_t \in C_i} \|x_t - c_i\|^2 \qquad (3.3)$$

where:

- x_t is a vector representing the $t - th$ data point or object in the cluster C_i
- c_i is the geometric centroid of the cluster C_i.

K-means algorithm allocates an object x_t into the ith cluster if the cluster membership function $I(x_t, i)$ is 1.

$$I(x_t, i) = \begin{Bmatrix} 1 \quad if\ i \quad = arg\,min\left(\left\|x_t - c_j\right\|^2\right)\ j = 1, \dots, k \\ 0\ otherwise \end{Bmatrix} \tag{3.4}$$

K-means algorithm is divided into three steps. These are shown in table 3.4.

Step	K-means clustering algorithm
1	Use random sampling to select k cluster centres $c_1, c_2, \dots, c_k$ in the input dataset For each input data point x_t and all k clusters, steps 2 and 3 are repeated until all centres converge.
2	Calculate the cluster membership function eq. (3.6) and decide the membership for each input data point in one of the k clusters whose cluster centre is closest to that point.
3	For all k cluster centres, set c_i to be the centre of mass of all points in cluster C_i

Table 3.4: The k-means algorithm

Although k-means is widely applied in various areas, it has two main limitations: 1) The number of k clusters must be fixed and known in advance and 2) the results of k-means algorithm depend on the selected cluster centres.

3.3.6 Fuzzy k-means clustering imputation

Fuzzy clustering presents a better tool than the overall objective of clustering and k means clustering, especially when the clusters overlap and may be trapped in local minimum if the initial points are not selected properly (Liao *et al.*, 2009). In fuzzy k-means clustering, each data object x_i uses a membership function to describe the degree to which the data object belongs to certain cluster v_k, thus making the resulting algorithm less susceptible to get stuck in local minimum (García *et al.*, 2015). The membership function is defined as eq. (3.5) rather than the k-means clustering MSE cost function in eq. (3.3).

$$U(v_k, x_i) = \frac{d(v_k,x_i)^{-2(m-1)}}{\sum_{j=1}^{k} d(v_j,x_i)^{-2/(m-1)}} \tag{3.5}$$

where:

$m > 1$ is the fuzzifier

$\sum_{j=1}^{k} U\left(v_j, x_i\right) = 1$ for any data object $x_i (1 \leq i \ \leq N)$

The membership degree of each data object is considered to compute the cluster centroid, the formula for computation is:

$$v_k = \frac{\sum_{i=1}^{N} U(v_k,x_i) * x_i}{\sum_{i=1}^{N} U\ (v_k, x_i)} \tag{3.6}$$

The algorithm comprises three steps, which is outlined below.

Step	**Fuzzy k-means clustering algorithm**
1	Select evenly distributed Kcentroids to avoid local minimum situation
2	Update the membership functions and centroids until the overall distance meets the user-specified distance threshold ε. Note that each data object is assigned to all K clusters with different membership degrees.
3	Non-reference attributes are imputed for each incomplete data object, x_i based on the membership degrees and the values of the cluster centroid.

3.3.7 Support vector machine imputation

Support vector machine (SVM) imputes missing data through the use of the following steps (Honghai *et al.*, 2005).

1) Select the examples that have no missing data, i.e. the complete examples as the training dataset.

2) Set the condition attributes (input attribute); whose values are missing as the decision attributes (output attribute) and uses the decision attributes as the condition attributes.
3) SVM regression is then used to predict the condition attribute values

SVM uses the value of the attribute being imputed as the target value rather than the original classification value and ignores attributes with missing data when generating the new training data. SVMI uses either regression or classification to impute continuous attributes (Gunn, 1998). For example in the case of classification, the continuous attribute is classified with each of the SVM models and the value corresponding to the SVM that classifies the example as positive is selected. If more than one SVM generates a positive classification, one value is selected randomly.

3.4 Distributions after missing data imputation

This section presents sodium, creatinine and uric acid distributions after the application of the seven missing data imputation methods (CMCI, EMI, FKMI, KMI, KNNI, MCI, and SVMI) discussed in section 3.3. The mean(μ), standard deviation(σ), median and skew values of the original data and after imputation are also presented for comparison.

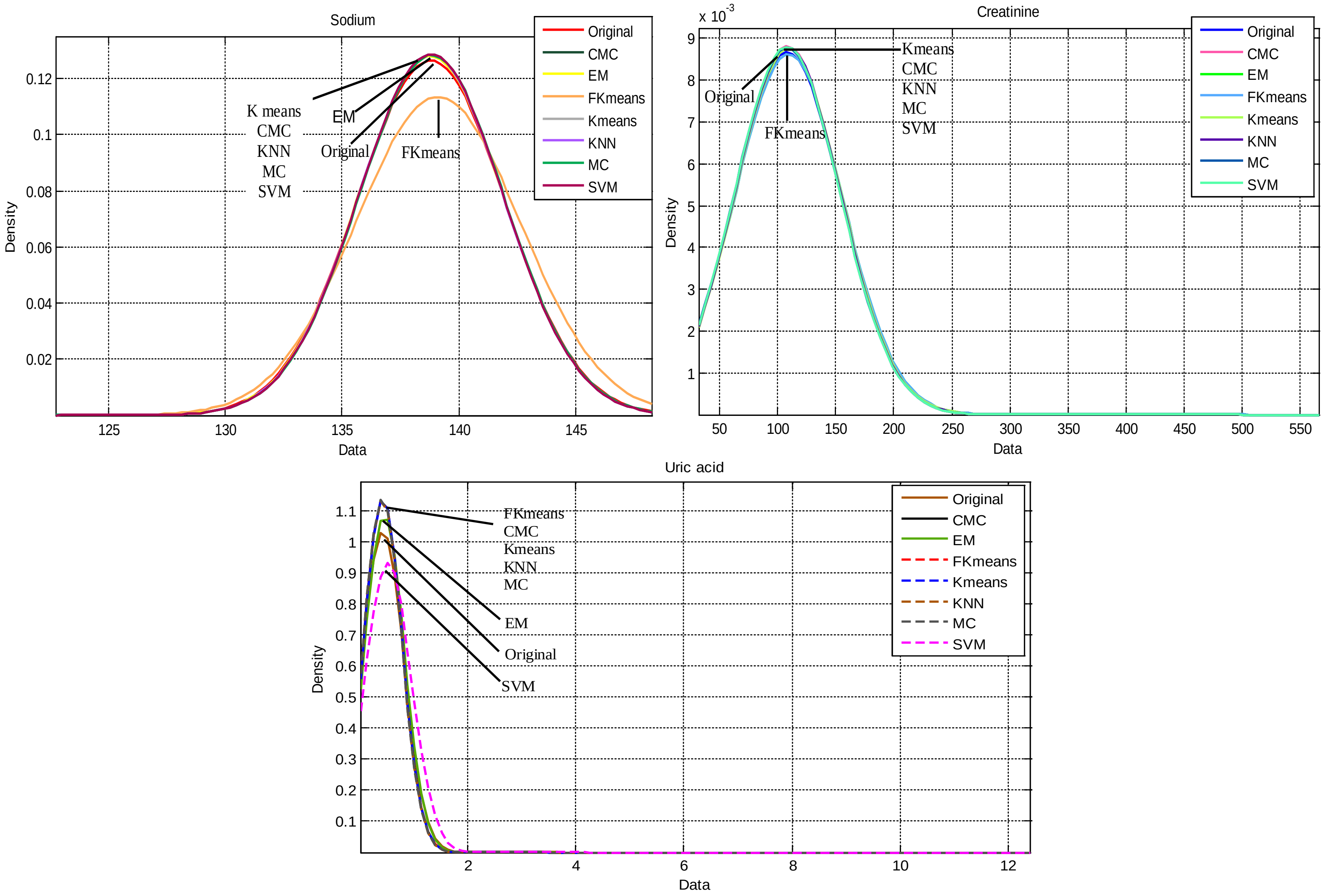


Figure 3.3 Original and imputation distributions of the overall dataset

Overall dataset												
	Sodium				Creatinine				Uric acid			
Statistical measures	μ	σ	Median	Skew value	μ	σ	Median	Skew value	μ	σ	Median	Skew value
Original	138.81	3.16	139	-0.88	108.45	46.03	98	3.27	0.42	0.39	0.39	22.68
CMC	138.81	3.10	139	-0.90	108.35	45.28	98.32	3.33	0.42	0.35	0.40	24.90
EM	138.80	3.12	139	-0.86	109.32	46.25	98.50	3.15	0.45	0.37	0.41	21.48
FKM	139.12	3.51	139	-0.29	108.69	45.29	99	3.30	0.42	0.35	0.39	24.92
KM	138.81	3.10	139	-0.90	109.01	45.54	99	3.24	0.43	0.35	0.41	24.94
KNN	138.80	3.11	139	-0.89	108.50	45.34	98	3.30	0.42	0.35	0.39	24.89
MC	138.84	3.11	139	-0.92	107.88	45.38	97	3.33	0.42	0.35	0.39	24.91
SVM	138.81	3.10	139	-0.90	107.98	45.32	97	3.34	0.52	0.43	0.42	13.63

Table 3.5: Original and imputation statistical measures of the overall dataset

Figure 3.3 presents distribution of the original data and the seven missing data imputation methods of the overall dataset and table 3.5 presents the original and after imputation statistical measures of the overall dataset. It can be seen that a the μ, σ, median and skew values of the imputation methods show a similarity or are identical to the original statistical measures, with the exception of FKM which shows a subtle increase with a μ value of 139.12 and σ value of 3.51 in the sodium. This can be visually seen from the graph in figure 3.3 where there is a shift in the distribution due to the increase of the μ value and the distribution has a large spread compared to the other imputation methods. FKM shows a skew value of -0.29 in the sodium variable. Comparing this to the original skew value of -0.88 indicates that this has decreased. EM and SVM skew values both in creatinine and uric acid are also reduced compared to the original skew values, especially SVM which shows a decrease in value of 13.63 for uric acid. SVM also show an increase in σ value of 0.43 in uric acid. This is also reflected in figure 3.3 where the distribution shows a wider spread, whereas the remaining σ values are reduced.

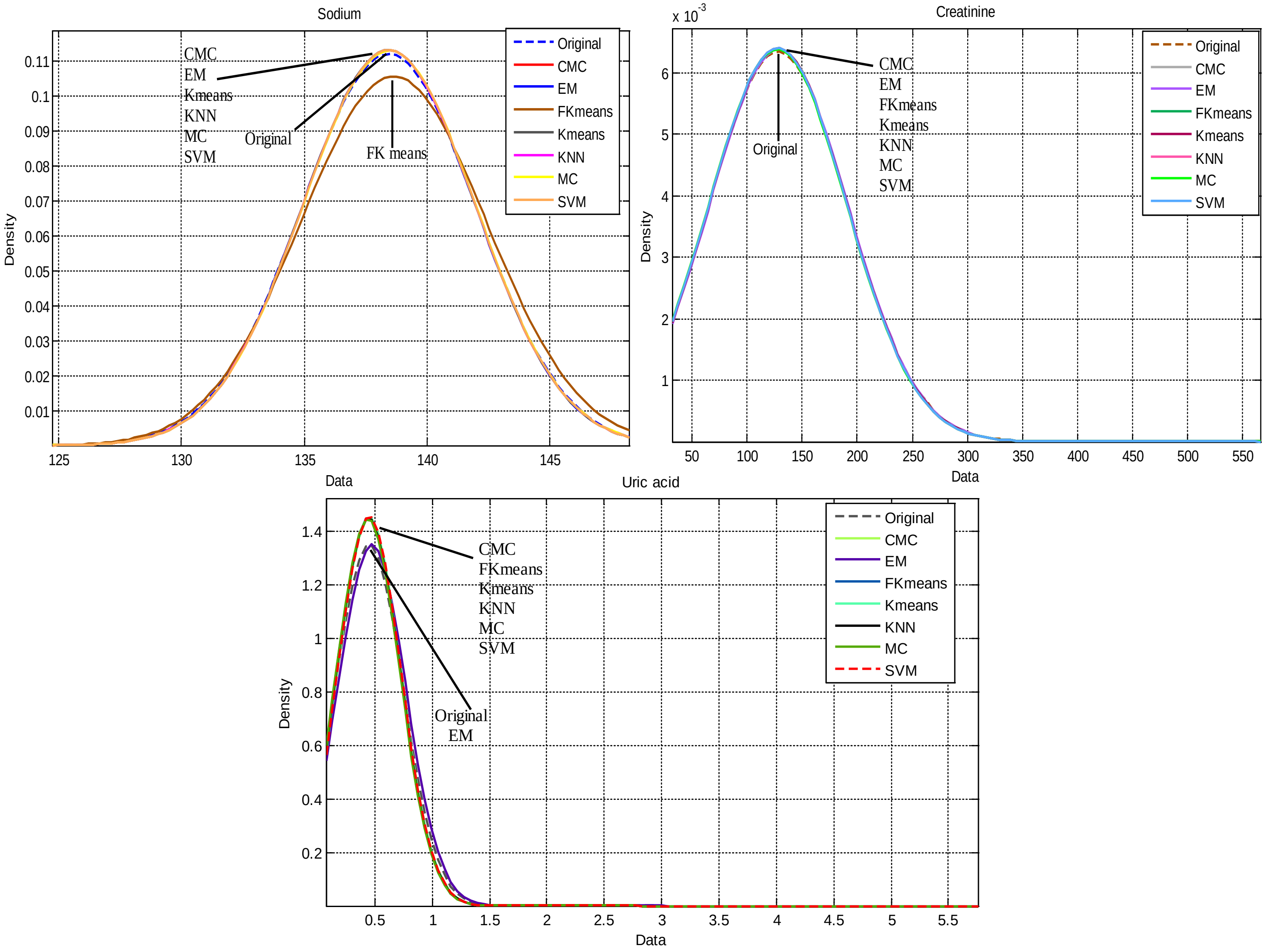


Figure 3.4 Original and imputation distributions of the dead class

Dead class												
	Sodium				Creatinine				Uric acid			
Statistical measures	μ	σ	Median	Skew value	μ	σ	Median	Skew value	μ	σ	Median	Skew value
Original	138.45	3.56	139	-0.71	128.22	62.99	112	2.65	0.45	0.29	0.43	13.67
CMC	138.45	3.53	139	-0.72	128.23	62.34	113	2.68	0.45	0.27	0.43	14.66
EM	138.45	3.53	139	-0.71	128.70	62.57	113	2.64	0.47	0.30	0.44	11.73
FKM	138.64	3.78	139	-0.40	127.92	62.39	112	2.69	0.44	0.27	0.41	14.64
KM	138.45	3.53	139	-0.72	128.20	62.44	112.27	2.67	0.45	0.27	0.43	14.72
KNN	138.46	3.53	139	-0.72	127.86	62.40	112	2.69	0.44	0.27	0.42	14.63
MC	138.47	3.53	139	-0.73	127.50	62.55	111	2.68	0.44	0.28	0.41	14.60
SVM	138.46	3.53	139	-0.72	127.86	62.39	112	2.69	0.45	0.27	0.44	14.71

Table 3.6: Original and imputation statistical measures of the dead class

The graph in figure 3.4 presents the original and the seven missing data imputation methods distributions for the dead class. It can be seen that creatinine missing data imputation distributions all show the same distribution. This is also reflected in the statistical measures presented in table 3.6. It can be seen that there are only subtle changes in the statistical measure values. FKM (-0.40) and EM (11.733) both show reduced skew values in sodium and uric acid respectively when compared to both their original values. This reduced skew is reflected in both their distributions as shown in the graph, where their distributions look less skewed. However, in general skewness is dominant in uric acid, which is visually shown in the graph and the skew values in table 3.6. FKM in the sodium shows a subtle increased μ value of 138.64 and σ value of 3.78, therefore causing a shift in the distribution and a larger spread in the graph shown in figure 3.4.

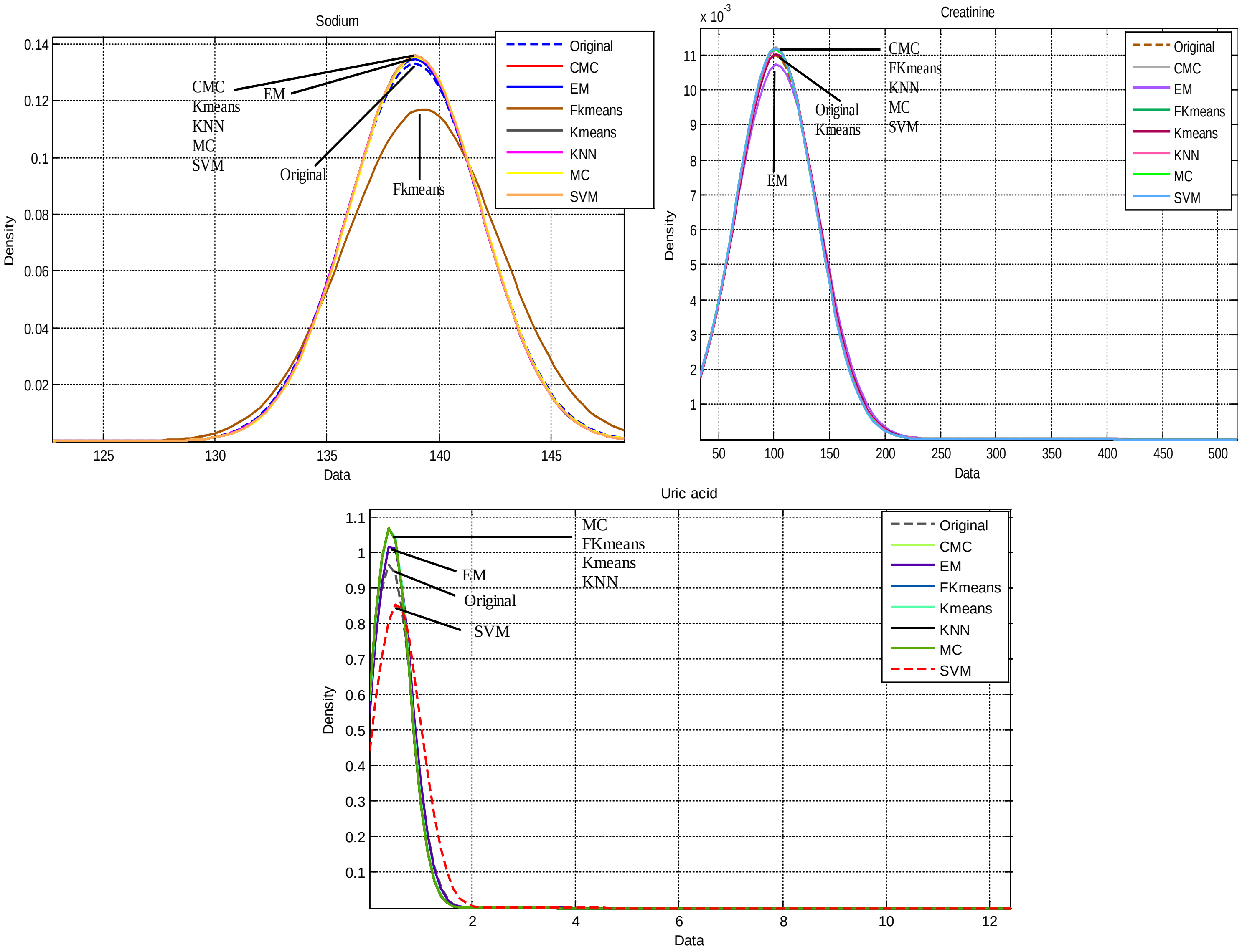


Figure 3.5 Original and imputation distributions of the alive class

Alive class												
	Sodium				Creatinine				Uric acid			
Statistical measures	μ	σ	Median	Skew value	μ	σ	Median	Skew value	μ	σ	Median	Skew value
Original	138.93	3.00	139	-0.93	101.75	36.30	94	3.09	0.42	0.41	0.38	23.34
CMC	138.93	2.94	139	-0.95	101.74	35.59	95	3.15	0.42	0.37	0.40	25.84
EM	138.92	2.96	139	-0.89	102.88	37.22	95	2.93	0.44	0.39	0.41	22.58
FKM	139.28	3.41	139	-0.20	102.30	35.74	95	3.07	0.41	0.37	0.38	25.84
KM	138.93	2.94	139	-0.94	102.63	36.15	95	3.00	0.42	0.37	0.41	25.84
KNN	138.92	2.94	139	-0.93	102.06	35.79	95	3.08	0.41	0.37	0.38	25.83
MC	138.96	2.94	139	-0.97	101.36	35.71	93	3.15	0.41	0.37	0.38	25.85
SVM	138.93	2.94	139	-0.95	101.37	35.63	93	3.17	0.54	0.47	0.42	13.00

Table 3.7: Original and after imputation statistical measures of the alive class

The graph in figure 3.5 presents the original and seven missing data imputation methods distributions of the three variables for the alive class while table 3.7 presents the original and after imputation statistical measures of the alive class. In table 3.7 FKM shows a reduced skew value of -0.20 in sodium while SVM shows a reduced skew value of 13.00 in uric acid. FKM, EM and SVM show an increased σ value of 3.41, 37.22 and 0.47 in sodium, creatinine and uric acid respectively. This is also reflected in their distributions shown in figure 3.5, where each distribution deviates from the distributions of the other imputation methods. It also indicates a large spread and variation in their distributions.

It can also be seen in figure 3.5 that the original distribution of sodium and creatinine variables does not hugely deviate from the imputation distributions with the exception of FKM and EM respectively. This indicates that CMC, KM, KNN, MC, SVM imputations have not had a great effect on the alive class. However, in uric acid the change is marked, where all the imputation distributions deviate from the original. However, MC, FKM, KM and KNN all show identical distributions. This is reflected in table 3.7 where their μ and σ are identical.

3.5 Summary

This chapter has discussed the descriptive stage of the proposed workflow, which involved two steps: 1) the data exploration step and 2) the data preparation step. Both steps entailed the distribution of the overall data, the dead class and the alive class as well as their statistical measures such as the mean, standard deviation, median and skew values. The statistical measures are important descriptive statistics and provide more information in understanding the statistical characteristics and properties of the dataset. For example, the relationship between the mean and standard deviation is used to measure how far the statistical variation and dispersion are from the mean. A high or low skew value indicates how far the variable distribution deviates from a normal distribution. The relationship between the mean and median allows an identification of the distribution type the variables and the imputation methods represent. It can be seen that in all three data groups (overall data, dead class and alive class), creatinine and uric acid distributions all represent a positive skew, as their mean values are greater than their median values, while sodium distributions are negatively skewed as their mean values are less than their median values.

The data preparation step involved seven missing data imputation methods which are shown to alleviate the missing data issue in clinical data. However, the main purpose is to understand their mechanisms and determine their effect on the dataset primarily the statistical measures. In all three data groups, FKMI, EMI and SVMI show the most change in the distributions and the statistical measures. Ideally the nature of the distribution should not change by a large margin as this will change classification performance. However, Luengo and co-authors (Luengo *et al.*, 2012) states that MCI, KNNI, EMI are commonly used, while FKMI and SVMI are suggested as the best methods for imputing missing data and in improving the behaviour of classification. For example C4.5 behaves better when missing data is imputed with SVMI while FKMI works best, no matter what classification method is chosen (Luengo *et al.*, 2012).

CHAPTER 4-THE PREDICTIVE STAGE

4.1 Introduction

The previous chapter discussed the descriptive stage of the workflow. This involved the exploration of the Hull LifeLab dataset. The distribution of the data for each variable was analysed which allowed an investigation of the properties of the data, including missing data and the problem of skews in the dataset. Of the challenges posed by such a dataset, missing data has a special place. The reason is that an incorrect imputation can change the distribution characteristics of the dataset. Most classification techniques that are based on probabilistic distribution can cause a dramatic change in distribution and have a detrimental effect on the performance of the classifier.

This chapter discusses the predictive stage of the workflow. The stage involves the use of information learned about the data in the descriptive stage of the workflow (chapter 3), for example, the exploration of the variable distributions and statistical measures of the data and the application of seven missing data imputation methods. This chapter uses the original and imputed datasets to construct a predictive model in order to extract and explain useful knowledge about the data in order to predict future outcomes. The primary focus is to predict '*what will happen*' through the application of Bayes classifier. This stage also explores the effect of the classifier on the different datasets, the relationship between the imputation methods and how the information can assist clinicians in clinical decision making.

Bayes classifier learns the underlying probabilistic model of the data, i.e. by modelling the interactions between the variables and the missing data imputation methods. The approach only requires an estimate of a few parameters, therefore having a lower variance for the parameter estimates (Hand, 1992). Due to its simplicity, the approach has been applied in many studies and produced improved predictive performance (John and Langley, 1995, Friedman *et al.,* 1997, Hand and Yu, 2001, Mani *et al.,* 1997), compared to various methods such as Artificial Neural Networks (ANNs) (Wang, 2003) that perform equally well and are capable of building predictive models from clinical data. Bayes classifier accommodates explanation and model transparency while ANN is considered as a 'black box' which does not provide thorough understanding of the mechanisms that govern the outcome (Bellazzi and Zupan, 2008).

The following sections will discuss the different types of performance metrics applied, Bayes classifier and its classification performance results. This will include the naïve Bayes performance of the original clinical data and the seven missing data imputation methods discussed in chapter 3. Based on the performance outcome of these results, such as if the performance is not good enough, an augmented naïve Bayes will be applied, known as Tree Augmented Naïve (TAN) Bayesian classifier (Shi and Huang, 2002, Friedman and Goldszmidt, 1996, Friedman *et al.,* 1997), to improve classification accuracy (Chow and Liu, 1968, Cheng and Greiner, 1999). The time, space and structural complexities of both classifiers are also briefly discussed in order to understand how their complexities impact classification accuracy.

4.2 Performance evaluation

The purpose of performance evaluation is to determine the effectiveness and usefulness of any classifier. Most performance evaluation measures such as a confusion matrix are used to determine a classifier's ability to identify classes correctly (Gu *et al.,* 2009).

This section will discuss the application of a confusion matrix and the many common performance metrics derived from it such as: *True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN)*. These metrics will be used to illustrate how to calculate five evaluation measures and how to use the results for prediction. The evaluation measures estimated are: *Positive Predictive Value (PPV), Negative Predictive Value (NPV), sensitivity (SEN) also known as recall, specificity (SPEC) and overall accuracy (ACC).*

4.2.1 Confusion matrix and evaluation measures

An objective method for evaluating and assessing the performance of classification algorithms is the use of a confusion matrix. This confusion matrix is a table that represents the performance of classification and summarises the learning system's performance (Fawcett, 2006, Gu *et al.,* 2009, Powers, 2011, Marschollek *et al.,* 2008, Subasi *et al.,* 2006, Hess *et al.,* 2006, Karbing *et al.,* 2007, Stehman, 1997) . In this thesis a two by two layout is applied as shown in table 4.1. This table shows a confusion matrix of a two class problem: positive and negative. The first column presents the true class outcome and the first row presents the predicted class outcome. The table consist of four possible predictor outcomes; *TP, TN, FP* and *FN* (Costa *et al.*, 2007). *TP* is when instances are correctly classified as positive, *TN* is when instances are correctly classified as negative. *FP* is the number of misclassified negative instances classified as positive while *FN* is the number of misclassified positive instances classified as negative. In the Hull LifeLab dataset, *TP* and *TN* are considered as alive and dead instances correctly classified as alive and dead respectively, while *FP* and *FN* are dead and alive instances incorrectly classified as alive and dead respectively.

Performance measure		**Predicted**	
		Positive	Negative
True	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 4.1: Confusion matrix table

The four outcomes are commonly used by five evaluation measures for determining and evaluating the effectiveness of the classifier. The evaluation measures are; *Positive*

Predictive Value (PPV) (equivalent to precision), Negative Predictive Value (NPV), sensitivity (SEN) also known as recall, specificity (SPEC) and overall accuracy (ACC). Their equations are shown below in equations 4.1 – 4.5 respectively.

$$PPV\ (precision) = \frac{TP}{(TP+FP)} \tag{4.1}$$

$$NPV = \frac{TN}{(TN+FN)} \tag{4.2}$$

$$Sensitivity\ (recall) = \frac{TP}{(TP+FN)} \tag{4.3}$$

$$Specificity = \frac{TN}{(TN+FP)} \tag{4.4}$$

$$Accuracy = \frac{TP+TN}{(TP+FP+FN+TN)} \tag{4.5}$$

PPV, NPV, SEN, SPEC and ACC are all measures of accuracy of the Hull LifeLab dataset and their percentages will be given. PPV estimates are the percentage of positive examples correctly predicted as positive, while NPV are the percentage of negative examples correctly predicted as negative. Based on our dataset, the positives are the alive class and the negatives are the dead class as the alive class is a representation of a positive outcome and the dead class is a representative of a negative outcome in the heart failure dataset. Ford and colleagues (Ford *et al.*, 2007) state that in using population-based data for risk factor analyses it is important that identified cases are true cases. Therefore, high PPV should be achieved. This is also key in this research where PPV and NPV evaluation measures are both identified as true positive and negative cases respectively and high percentages of both predictions are important. Sensitivity (SEN) is the proportion of actual positive examples that are correctly identified as positive. In a clinical context and in machine learning, sensitivity is regarded as primary as it aims to identify all real positive cases and focuses on how confident the classifier is (Powers, 2007). Specificity (SPEC) is the actual negative examples that are correctly identified as negative. Accuracy

(ACC) measures the percentage of the overall accuracy of correct predictions and effectiveness of the classifier (Costa *et al.*, 2007, Gu *et al.*, 2009, Powers, 2007). A perfect evaluation measure would be described as 100% therefore it is important to achieve a high performance so that suitable clinical decisions are made.

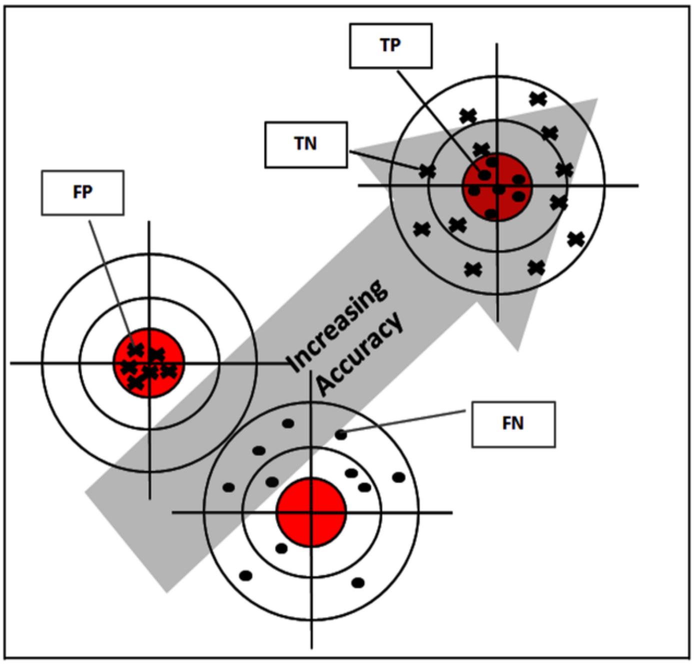


Figure 4.1 Relationship of performance indicators

Figure 4.1 shows the relationship of the performance indicators. It can be seen that classification accuracy should be increased in *TP* and *TN* rates; this does not mean that the accuracy of *FP* and *FN* are not important. However the number of misclassified dead examples classified as alive *(FP)* and misclassified alive examples classified as dead *(FN)* should be relatively low. This is crucial for clinical reasons as both outcomes can result in false decision making for personalised care. However, chapter 5 of this thesis will only consider the number of *FP* results to be extremely low (Pepe *et al.*, 2004), as a high *FP* prediction may have more serious consequences than a *FN* prediction. For example, a high *FP* prediction can provoke anxiety, increase costs and cause morbidity. It is also important to clinicians to obtain the true healthy individuals from the population during screening. This indicates a successful screening and thus allows better planning for personalised care for the sick population. In addition, in other clinical studies such as breast cancer, techniques are needed in order to decrease *FP* results while maintaining high sensitivity (Elmore *et al.*, 1998).

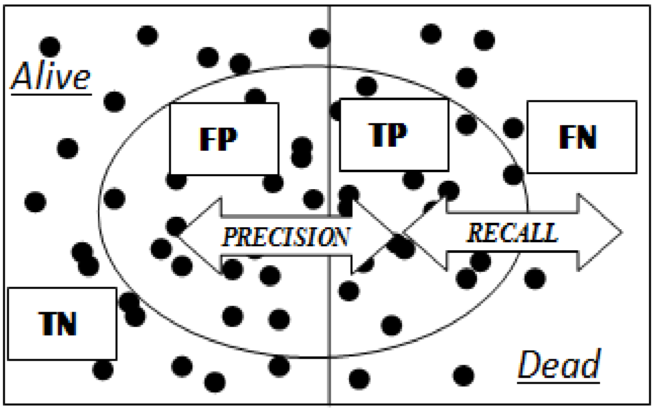


Figure 4.2 Relationship between precision and recall value of classification

Figure 4.2 presents the relationship between precision and recall in the target class; alive and dead class respectively. It can be seen that precision is required in both classes, but especially in the alive class. Precision is a combination of the correct classified examples *(TP)* and the misclassified examples *(FP)*, thus the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. Recall is a combination of *TP* and *FN*, meaning the number of relevant records retrieved to the number of relevant records in the dataset. It is important to distinguish and understand the differences between each prediction outcome and evaluation measure so that classification performance is interpreted correctly for clinical decision making.

4.3 Naïve Bayes classifier

The naïve Bayes approach allows us to capture uncertainty about the assumed underlying probabilistic model by determining probabilities of the outcomes. Prior knowledge about the data and observed data are both combined in order to provide a useful perspective for understanding and learning the predictive task. For example it is applied in clinical settings to solve diagnostic and predictive problems such as a case of heart failure where a group of patients are identified either as high risk or low risk; where the high risk group represents 485 patients and low risk 1459 patients. If an observed patient who is ‘known’ to be low risk has a posterior probability greater than that of the high risk posterior probability, the patient will be classified as low risk and if not greater, the said patient will be classified as high risk. Thus a naïve Bayes classifier is a simple probabilistic approach that presents clear semantics for learning probabilistic knowledge, with the independent assumption of variables within each class (Karlık and Öztoprak, 2012).Two assumptions are made, when using a naïve Bayes classifier.

Assumption 1: It is assumed that the predictive variable X_i is conditionally independent given the class variable C as shown in figure 4.3 (John and Langley, 1995, Muhammed, 2012).

This is also known as the strong naïve independence assumption between the variables. The conditional probability can be determined using Bayes’ theorem thus the conditional probability of each class given the observed values is:

$$p(C = c|X = x) = \frac{p(X=x|C=c)p(C=c)}{p(X=x)} \qquad (4.6)$$

where:

- C is the random variable denoting the class of an instance
- X is a vector of random variables denoting the observed variable values
- c represents a particular class label while x represents a particular observed variable value vector.

Assumption 2: Within each class the numeric variables are in a Gaussian (normal) distribution.

Thus knowing the mean (μ) and standard deviation (σ) allows us to determine the distribution for each variable. Further, given that we are interested in the probability given a class, the data is separated into classes and for each class the mean and standard deviation are estimated. Thus the conditional probability of an observed value is:

$$P(X_i|C = c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} exp^{\left(-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}\right)} \tag{4.7}$$

where:

- μ_{ij} is the mean of the variable values in x_i associated with class $C = c_j$
- σ_{ij} is the standard deviation of the variable values in x_i, associated with class $C = c_j$

Table 4.2 gives an outline of the naïve Bayes algorithm.

No.	Steps
0	get c_j class and x_i feature
1	Model the Gaussian distribution (4.7) to compute the mean μ_i and standard deviation σ_i for x_i
2	Use the above steps to compute the posterior probabilities (4.7)
3	Calculate the conditional probabilities $p(c|x_1, \dots, x_n)$ for x_i (4.6)

Table 4.2: Naïve Bayes classifier algorithm

In chapter 3, Hull LifeLab clinical dataset was explored and it was found that not all variables within the dataset have a classic Gaussian distribution; in other words skew distributions are presented. Indeed, this is a characteristic of real life clinical data. As discussed earlier (section 4.1), naïve Bayes is applied because it provides an insight into the structure of the data space and identifies potential problems associated. In addition, as will be seen in this chapter and in later chapters, Bayes processes will enable an analysis of the data space in order to answer: (a) Can missing data imputation methods improve prediction accuracy of outcome of heart failure and the classifier learned? (b)

What are the underlying factors yielding poor performance? (c) How can better classification accuracy be achieved? Modelling the true distributions of the data prevents the above from being achieved and will change the true nature of the data. However, in later chapters (see chapter 6), methods which use the existing data distribution are discussed.

Naïve Bayes has many advantages in that unlike many other classifiers it only requires a small amount of training data to perform analysis and at the same time it is computationally more efficient. The time complexity of the algorithm is $O(Nn)$, where N is the number of training instances and n is the number of attributes (Yan *et al.*, 2011). During training time, naïve Bayes requires only two tables: one to store the class probability estimates and the other to store the conditional attribute-value probability estimates. Therefore the resulting space complexity is $O(Nn\mu_{ij})$, where μ_{ij} is the mean number of values per attribute (Webb *et al.,* 2005).

Figure 4.3 presents the graphical structure of naïve Bayes (Liangxiao *et al.,* 2009) which indeed is a direct representation of '*Assumption 1*'. Liangxiao and colleagues state that naïve Bayes is the simplest form of Bayesian network to construct, where all arcs are directed from the class C node as the parent to the predictive variable $x_1 \ldots x_n$ nodes.

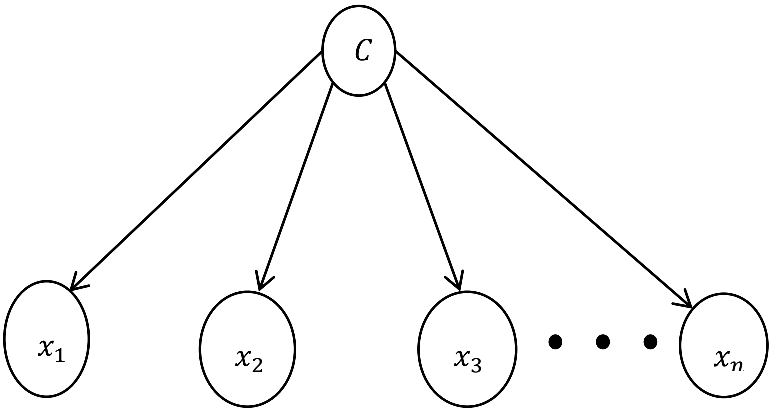


Figure 4.3 A naïve Bayes structure in which variables are conditionally independent given the class variable.

4.4Naïve Bayes performance of clinical data

The naïve Bayes performance of the original clinical dataset and the dataset imputed with seven missing data imputation methods will be presented. A confusion matrix will be used to show their classification performances and evaluation measures (PPV, NPV, SEN, SPEC and ACC) which will be presented as percentages.

4.4.1 Naive Bayes performance of original clinical data

In order to understand the performance of the algorithm and the data structure, the naïve Bayes method was first tested on the original dataset. This further enables bench marking of the algorithm.

	Predict		**Evaluation measure**	**%**
True	Alive	Dead	PPV	85
Alive	1178TP	281FN	NPV	50
Dead	211FP	274TN	SEN	81
			SPEC	56
			ACC	75

Table 4.3: Naïve Bayes performance of the original clinical dataset

Table 4.3 shows the naïve Bayes performance of the original clinical data and it can be seen that the *PPV* and *SEN* are 85% and 81% respectively. *PPV* and *SEN* measure different, yet complementary aspects of the performance. *PPV* measures the number of incorrectly classified positives, while *SEN* measures *FN*. Thus the naïve Bayes algorithm with a *PPV* of 85% can be interpreted as predicting 15% as being alive when they should be classed as dead. Similarly, a *SEN* of 81% can be interpreted as the algorithm predicting 19% dead when they should have been predicted as alive. While *NPV* and *SPEC* are 50% and 56% respectively, these represent the *TN* result. However both measures are different and yet complement each other. *NPV* measures the number of incorrectly classified negatives, while *SPEC* measures the number of incorrectly classified positives. Therefore it can be interpreted as the algorithm predicting 50% as dead when they should be alive and 44% as alive when in fact they should have been predicted as dead.

4.4.2 Naive Bayes performance of imputation methods

Table 4.4 presents the naïve Bayes performance of the seven imputation methods.

CMC		**Predict**		**Evaluation measure**	**%**
	True	Alive	Dead	PPV	86
	Alive	1198TP	261FN	NPV	52
	Dead	202FP	283TN	SEN	82
				SPEC	58
				ACC	76

EM		**Predict**		**Evaluation measure**	**%**
	True	Alive	Dead	PPV	85
	Alive	728TP	731FN	NPV	33
	Dead	132FP	353TN	SEN	50
				SPEC	73
				ACC	56

FKM		**Predict**		**Evaluation measure**	**%**
	True	Alive	Dead	PPV	85
	Alive	1207TP	252FN	NPV	52
	Dead	213FP	272TN	SEN	83
				SPEC	56
				ACC	76

KM		**Predict**		**Evaluation measure**	**%**
	True	Alive	Dead	PPV	85
	Alive	1196TP	263FN	NPV	51
	Dead	212FP	273TN	SEN	82
				SPEC	56
				ACC	76

KNN		**Predict**		**Evaluation measure**	**%**
	True	Alive	Dead	PPV	85
	Alive	1201TP	258FN	NPV	52
	Dead	211FP	274TN	SEN	82
				SPEC	56
				ACC	76

MC		**Predict**		**Evaluation measure**	**%**
	True	Alive	Dead	PPV	85
	Alive	1201TP	258FN	NPV	51
	Dead	216FP	269TN	SEN	82
				SPEC	55
				ACC	76

SVM		**Predict**		**Evaluation measure**	**%**
	True	Alive	Dead	PPV	86
	Alive	1204TP	255FN	NPV	53
	Dead	196FP	289TN	SEN	83
				SPEC	60
				ACC	77

Table 4.4: Naïve Bayes performance of the imputation methods

It can be seen that there is a similarity in classification performance in CMC, FKM, KM, KNN, MC and SVM imputation methods. The naïve Bayes algorithm presents CMC and SVM with *PPV* of 86%, similarly FKM, KM, KNN and MC shows 85%, which can be interpreted as predicting 14% and 15% as being alive when they should be classed as dead. The evaluation performance of CMC, KM, KNN and MC also show *SEN* to be similar to *PPV,* with each imputation method showing *SEN* of 82%, while FKM and SVM show 83%. This means that 18% and 17% respectively are predicted as dead when they should have been predicted as alive. The table also shows that CMC, FKM, KM, KNN and MC all share the same *ACC* of 76% while SVM shows a similar *ACC* of 77%.

On the contrary, EM presents *SEN* of 50%. This can be interpreted as predicting 50% as being dead when they should be predicted as alive. Similarly this is also shown in *NPV* of CMC, FKM and KNN where *NPV* is 52%, while KM and MC show 51% and SVM 53%, which indicates that 48%, 49% and 47% respectively, are incorrectly classified negatives and therefore predicted as being dead when they should be classed as alive. The algorithm also shows *SPEC* of 56% in FKM, KM and KNN, 55% in MC while CMC shows 58%. It can be interpreted as the algorithm predicting 44%, 45% and 42% respectively as alive when they should have been predicted as dead. However *NPV* and *SPEC* of the EM are 33% and 73% respectively. Since *NPV* measures the number of incorrectly classified negatives *(FN)*, it can be interpreted as the algorithm predicting 67% as being dead when they should be alive. Similarly, *SPEC* can be interpreted as predicting 27% alive when they should have been predicted as dead. The *NPV* result shows a poor performance as more than half of it shows a large number of the population to be incorrectly classified. If, for example, the number of *TNs (NPV)* increased, thus the 33%, this would automatically reduce the portion of the falsely classified results.

These similarities in classification performance may be due to the similar task shared amongst some of the imputation methods. For example, KM, FKM and KNN all impute missing data based on a k number and with CMC all four imputation methods use the in class mean for imputation. Although KNN and KM impute missing data based on some measure of similarity and dissimilarity respectively, however their task for imputing missing data is very similar. For example KM applies a nearest neighbour algorithm to replace missing data in a similar way to KNN. CMC is an extension of MC and their task is relatively similar in that they both use mean estimates for imputation (Luengo *et al.*, 2012). Luengo and colleagues state that FKM and SVM perform best when using Bayes classifiers, this is shown in table 4.4 where SVM presents the best classifier, but not by a large margin.

It can also be seen in the table that *SEN* and *SPEC* are inversely proportional, indicating that as the *SEN* increases *SPEC* decreases and vice versa. *PPV* and *NPV* are related to the prevalence of heart failure in the population; therefore as the percentage of *PPV* increase, *NPV* decreases.

The classification performances in table 4.3 and 4.4 are very similar. This suggests that despite the challenges present in the data such as missing data, the naïve Bayes performance of the original clinical data (table 4.3) is able to produce good results. This is an indication that naïve Bayes is not too sensitive to missing data. The similarities in performance in table 4.4 also indicate that some of the imputation methods have not modified the data space by a large margin. However, the results do indicate that improvement can be made, in that the number of *FPs* and *FNs* can be reduced. This also leads to the associated question, if they cannot be improved upon, what is it in the data that is resulting in this lack of improvement? In other words, ideally the aim is to be able to obtain optimal classification accuracy. As a result, TAN will be implemented in the next section to achieve this, as well as learn the classification model efficiently.

4.5 Tree augmented naïve Bayesian network

TAN is an extension of naïve Bayes that constructs a tree-like structure Bayesian network (Friedman *et al.,* 1997). TAN weakens and manipulates the strong conditional independence assumption presented by naïve Bayes (Jiang *et al.*, 2012), in order to find correlation among attributes. Figure 4.4 presents a TAN Bayesian network structure. The dashed arcs are required by the naïve Bayes classifier, with one arc from the class C node (as the parent) connecting to all attributes $x_1 \dots x_n$, while the solid arcs represent the correlation and dependences between attributes, forming an undirected tree making it possible to learn the classification model effectively (Friedman *et al.*, 1997). The structure has the advantage of preventing over fitting problems (Friedman *et al.*, 1997, Bouhamed *et al.*, 2012). However, its disadvantage is that it restricts the number of parents to only a single parent (the class node) for each attribute required and at most one other attribute.

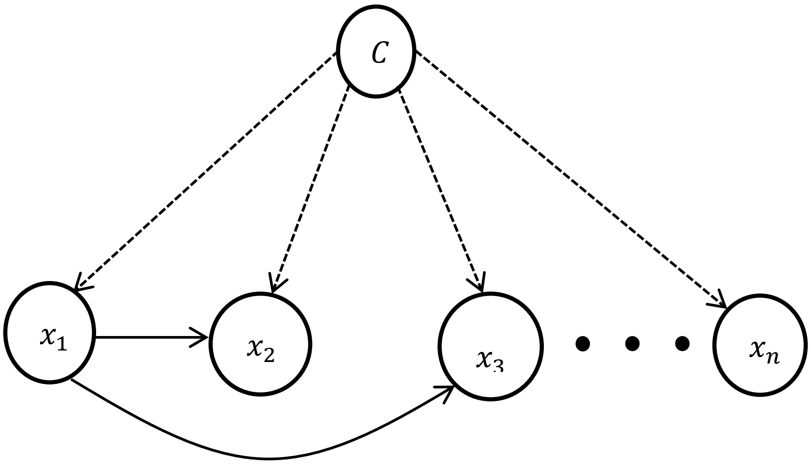


Figure 4.4 A Tree Augmented Naïve Bayesian network

Table 4.5 shows a four step learning algorithm for constructing a tree Bayesian network (Friedman *et al.,* 1997, Jiang *et al.,* 2005, Chow and Liu, 1968). The tree is constructed based on a procedure described by Chow and Liu (Chow and Liu, 1968) which involves implementing mutual information relationships between two features X, Y which measures how much information Y provides about X. The approach finds a tree that maximises the likelihood given the data.

$$I_{\hat{P}_D}(X; Y) = \sum_{x,y} \hat{P}_D(x, y)\, log\, \frac{\hat{P}_D(x,y)}{\hat{P}_D(x)\hat{P}_D(y)} \qquad (4.8)$$

Step	**Learning algorithm for constructing a tree**
	Input: a training dataset $X_1, \dots, X_m$
1	Compute the mutual information $I_{\hat{P}_D}\left(X_i;\ X_j \vert C\right)$ between each pair of features, $i \neq j$ defined by eq.4.8
2	Build a complete undirected tree where nodes are features $x_i, \dots, x_m$. Annotate the weight of an arc connecting x_i to x_j by $I_{\hat{P}_D}\left(X_i;\ X_j\right)$
3	Construct an undirected Maximum Weighted Spanning Tree (MWST). An example is shown in fig. 4.5. MWST time complexity is $O(n^2 \log n)$, n is the number of node in the graph (Friedman *et al.,* 1997) .
4	Transform the undirected tree to a directed tree by choosing a root feature and setting the direction of all arcs to be outward from it. For example x_1 is chosen as the root node (fig. 4.6)

Table 4.5: Learning algorithm for constructing a tree

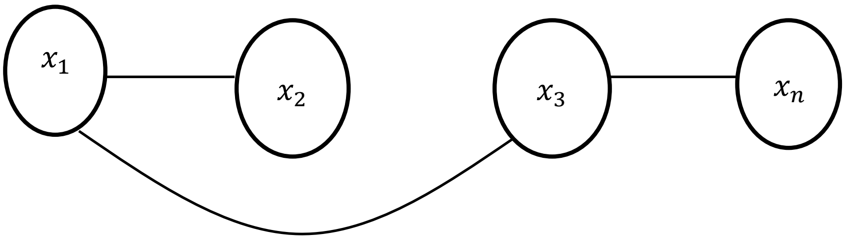


Figure 4.5 An undirected tree

The Maximum Weighted Spanning Tree (MWST) shown in step 3 is constructed by selecting a subset of arcs from a graph which constitute the tree and the sum of weights attached to the selected arcs are maximised.

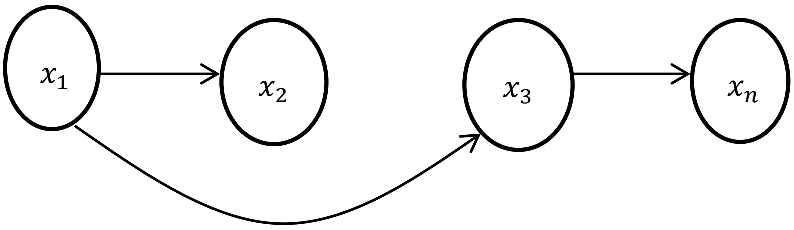


Figure 4.6 A directed tree

Step	**Learning algorithm for constructing TAN**
	Input: a training dataset $X_1, \dots, X_m$
1	Compute the conditional mutual information $I_{\hat{P}_D}\ (X_i;\ X_j \mid C)$ between each pair of features, $i \neq j$ defined by eq. 4.9.
2	Build a complete undirected tree where nodes are features $x_i, \dots, x_m$. Annotate the weight of an arc connecting x_i to x_j by $I_{\hat{P}_D}(X_i;\ X_j)$
3	Construct an undirected Maximum Weighted Spanning Tree (MWST). An example is shown in fig. 4.5. MWST time complexity is $O(n^2 \log n)$, n is the number of node in the graph (Friedman *et al.,* 1997) .
4	Transform the undirected tree to a directed tree by choosing a root feature and setting the direction of all arcs to be outward from it. For example x_1 is chosen as the root node (fig. 4.6)
5	Build a TAN model by adding a node labelled C and adding the arc from C to each X_i(dashed lines); fig 4.4
	Output: Returns a naïve Bayes network augmented with a tree (TAN)

Table 4.6: Learning algorithm for constructing TAN

Table 4.6 shows a five step learning algorithm for constructing TAN. The algorithm is similar to that of the algorithm for constructing a tree shown in table 4.5 except that instead of using the mutual information between two features, it considers conditional mutual information between features X, Y given the class variable C.

$$I_{\hat{P}_D}\ (X;Y|C) = \sum_{x,y,c} \hat{P}_D(x,y,c) log \frac{\hat{P}_D(x,y|c)}{\hat{P}_D(x|C)\hat{P}_D(y|c)} \qquad (4.9)$$

The time and space complexity is $O(n^2N)$ and $O(n^2)$ respectively, where n is the number of variables, N is the number of samples in the training set (Friedman *et al.,* 1997). Construction of the algorithm indicates that its memory requirement is quadratic in the number of attributes. Therefore the space requirement is higher than that of naïve Bayes. This also makes it unfeasible to apply high dimensional data (Shi and Huang, 2002). TAN stores the probability estimates for each attribute-value conditioned by the parent selected for that attribute, and the class (Webb *et al.*, 2005). A previous study by Meila (Meila, 1999) investigates a way of reducing the space requirements by fitting tree distributions to high dimensional sparse data. This led to an acceleration of the tree learning algorithm, which in turn increases computational complexity.

4.5.1 Tree augmented naïve Bayes performance of original clinical data

	Predicted		**Evaluation measures**	**%**
True	Alive	Dead	PPV	87
Alive	1278TP	181FN	NPV	62
Dead	185FP	300TN	SEN	88
			SPEC	62
			ACC	81

Table 4.7: Tree augmented naïve Bayes performance of the original clinical data

Table 4.7 presents the TAN Bayes performance of the original clinical dataset. It can be seen that although PPV and SEN measures are different, their performance is similar. The TAN algorithm with a PPV of 87% can be interpreted as predicting 13% as being alive *(FP)* when they should be classed as dead. Similarly, a SEN of 88% can be interpreted as the algorithm predicting 12% dead when they should have been predicted as alive. NPV and SPEC show a performance of 62% which can be interpreted as the algorithm predicting 38% as FN (dead) and FP (alive) respectively. The algorithm shows an overall ACC of 81%. This result shows that there is an improvement when compared to the naïve Bayes performance of the original clinical data shown in table 4.3, as TAN shows a greater performance than that of naïve Bayes, which shows PPV, NPV, SEN, SPEC and ACC to be 86%, 52%, 82%, 58% and 76% respectively. The numbers of incorrectly classified positives and negatives are also reduced in TAN. For example the percentage of *FPs* (PPV) in naïve Bayes and TAN are 15% and 13% respectively and *FNs* (SEN) are 19% in naïve Bayes and 12% in TAN Bayes.

4.5.2 Tree augmented naïve Bayes performance of imputation methods

Table 4.8 shows the TAN performance of the seven missing data imputation methods. It can be seen from their evaluation measures that the classification performances have improved when compared to that of the naïve Bayes in table 4.4. For example SVM shows the most improvement, with a PPV of 96% in TAN and 86% in naïve Bayes. The number of *FPs* is also reduced by a difference of 10%. For example as PPV measures the number of incorrectly classified positives (FPs), the TAN Bayes algorithm with a PPV of 96% can be interpreted as predicting 4% as being alive when they should be classed as dead, while the number of incorrectly classified positives in naïve Bayes is 14%.

However, the EM imputation shows a low SPEC of 51%, while naïve Bayes shows 73%. This can also be interpreted as TAN Bayes algorithm predicting 49% as being alive therefore increasing the number of *FPs* when compared to FPs of 27% in naïve Bayes. The reason for this could be due to the maximisation relationship shared by the TAN and EM algorithms. The maximisation step of the EM algorithm maximises the expected maximum likelihood found in the expectation step to impute missing data, while TAN maximises the log likelihood in step 3 of the learning algorithm.

Just like in the naïve Bayes performance FKM, KM, KNN and MC also present similar TAN performances and when compared to the TAN performance of the original clinical data. For example the imputation methods and original data have a PPV (except MC) and ACC of 87% and 81% respectively, similar NPV, SEN and SPEC measures are also shown. This indicates that the imputation methods did not change the data space;

however there is an improvement when compared to naïve Bayes performance due to the difference in their learning algorithm. However CMC performance strays from their performance in that CMC shows a better performance where PPV, SEN, NPV, SPEC and ACC are 93%, 92%, 77%, 79% and 89% respectively, therefore, predicting 7%, 8%, 23%, 21% and 11% as FP and FN respectively. The improvement in TAN Bayes is due to the augmentation during learning of the algorithm. For example the algorithm relaxes the independent assumption in naïve Bayes by estimating the conditional mutual information in order to capture correlations between the variables given the class variable. However, an advantage is that the independent assumption allows parameters for each variable to be learned separately, especially when the number of variables is large.

CMC		**Predict**		**Evaluation measure**	**%**	**EM**		**Predict**		**Evaluation measure**	**%**
	True	Alive	Dead	PPV	93		**True**	Alive	Dead	PPV	84
	Alive	1347TP	112FN	NPV	77		Alive	1283TP	176FN	NPV	58
	Dead	101FP	384TN	SEN	92		Dead	238FP	247TN	SEN	88
				SPEC	79					SPEC	51
				ACC	89					ACC	79
FKM		**Predict**		**Evaluation measure**	%	**KM**		**Predict**		**Evaluation measure**	%
	True	Alive	Dead	PPV	87		**True**	Alive	Dead	PPV	87
	Alive	1282TP	177FN	NPV	62		Alive	1275TP	184FN	NPV	62
	Dead	198FP	287TN	SEN	88		Dead	187FP	298TN	SEN	87
				SPEC	59					SPEC	61
				ACC	81					ACC	81
KNN		**Predict**		**Evaluation measure**	%	**MC**		**Predict**		**Evaluation measure**	%
	True	Alive	Dead	PPV	87		**True**	Alive	Dead	PPV	88
	Alive	1267TP	192FN	NPV	61		Alive	1277TP	182FN	NPV	62
	Dead	186FP	299TN	SEN	87		Dead	182FP	303TN	SEN	88
				SPEC	62					SPEC	62
				ACC	81					ACC	81

SVM		**Predict**		**Evaluation measure**	%
	True	Alive	Dead	PPV	96
	Alive	1389TP	70FN	NPV	85
	Dead	60FP	425TN	SEN	95
				SPEC	88
				ACC	93

Table 4.8: Tree augmented naïve Bayes performance of the seven imputation methods

4.6 Summary

Overall the experiments in this chapter show that TAN outperforms naïve Bayes, maintains robustness and improves classification accuracy. The results show that TAN Bayes is least sensitive to missing data as the algorithm shows the best performance on the seven imputation methods. Out of the seven imputation methods SVM performed the best. Although, Bayes is robust in the presence of missing data, and this is shown in this chapter, however this is true only to some extent. This chapter has also shown that imputation can improve the accuracy of the Bayes classifier and thus the prediction accuracy. In addition, the results have also shown and allowed us to understand the influence the imputation methods have on the classification model.

The requirement for time and space complexity is greater in TAN than in naive Bayes. Figures 4.7 and 4.8 shows a section of a graph screen shot of naïve Bayes and TAN structures of the original clinical dataset obtained from WEKA. Their structural complexities differ in that the structure of naïve Bayes is simply straightforward (fig. 4.7), whereas TAN (fig. 8) at first glance may appear intimidating and complicated; however this helps learn the TAN model effectively. For example the graph shows correlation amongst attributes by measuring how much information one attribute provides about another attribute. This is illustrated where C-Reactive Protein (CRP) is dependent on albumin and ECG and albumin is dependent on calcium and total protein.

During the learning stage, TAN embodies a good trade-off between the quality of the approximation of correlations among attributes and the computational complexity. This is an indication that the high complexity required to learn a TAN model is necessary as it has improved classification accuracy, which is one of the main purpose of this research.

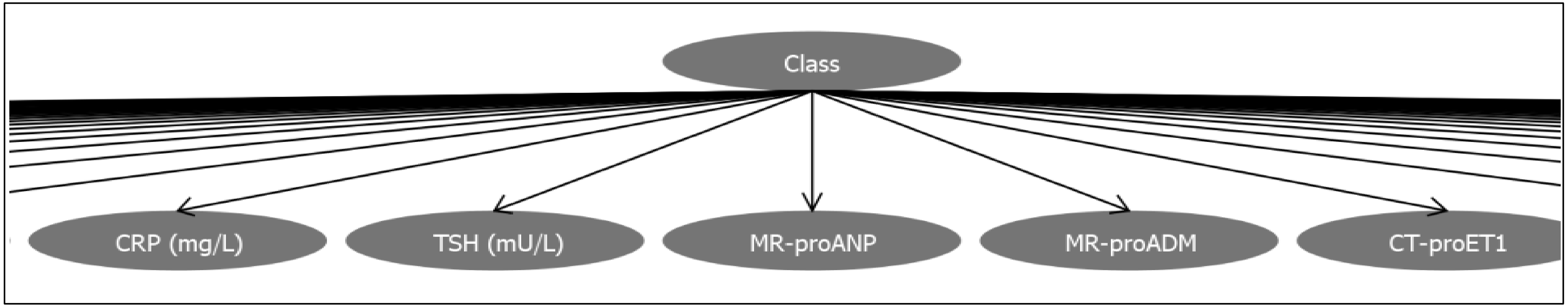


Figure 4.7 Naïve Bayes structure of the original clinical data

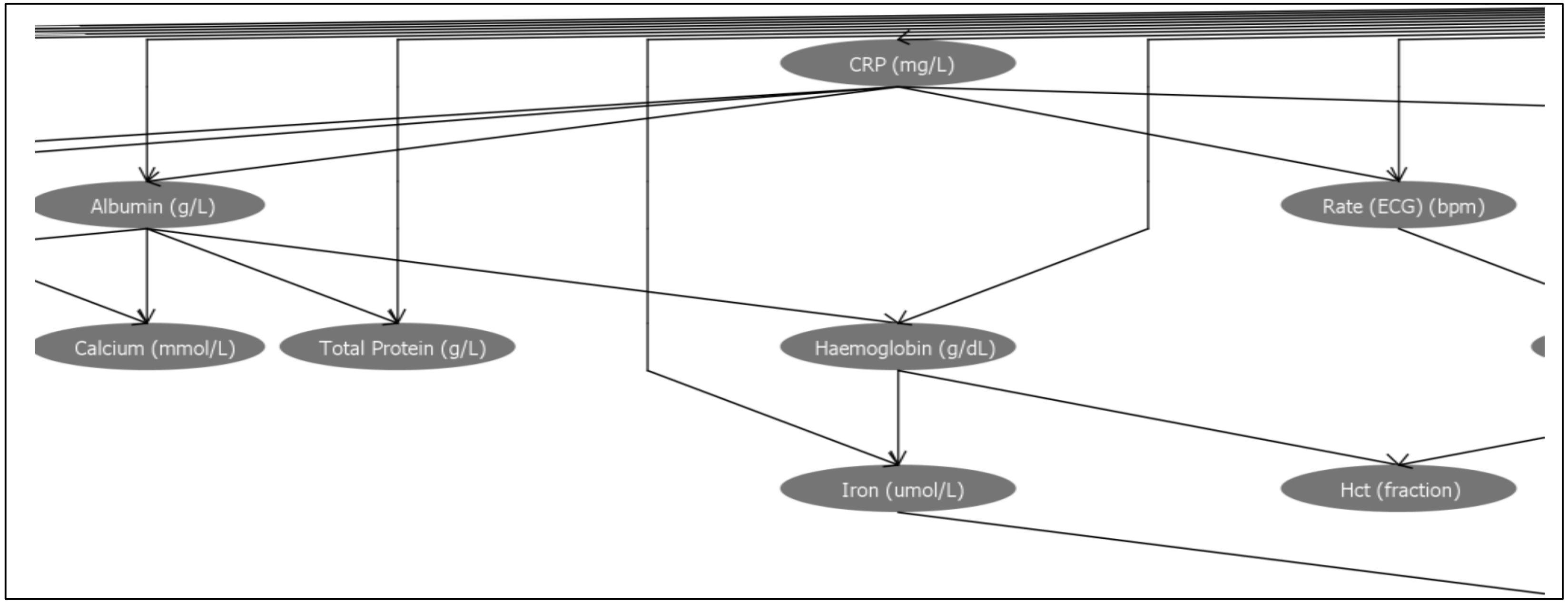


Figure 4.8 TAN structure of the original clinical data

Naïve Bayes performances of the imputation methods show different performance outcomes. For example, PPV of SVM and SPEC of EM imputation methods show the maximum percentage outcome with 86% and 73% respectively (table 4.4). Based on this difference those performances with maximum percentage outcomes will be combined. Therefore the SVM alive class and EM dead class will be combined in order to obtain a better classification performance. Chapter 5 will implement naïve Bayes and TAN classification in order to determine the performance of the new combined dataset.

CHAPTER 5-THE PRESCRIPTIVE STAGE

5.1 Introduction

In earlier chapters, not only was the clinical workflow discussed, but in chapter 4, the performance metrics and evaluation measures used in the predictive stage were also discussed. At the same time the performance of two Bayes-based classification algorithms on different imputation schemes was presented. Although it is said that Bayes is immune to missing values, given the imbalance in the classes, imputation was carried out. Imputation helps maintain the richness of the data so that the expected mean of the whole population is obtained; this is also used as a model for the data space. Experiments using naïve Bayes and TAN Bayes have shown that TAN achieves better classification accuracy than naïve Bayes. The drawback, however, is that TAN has greater space and time complexity compared to naïve Bayes (see chapter 4). In this chapter, using the results from chapter 4, a detailed analysis of the data space is carried out in order to better understand the nature of the data and perhaps to identify a process by which problematic variables and records could be identified.

An important component of Bayes methods is the determination of the means and standard deviation of the data for each class (see chapter 3). This implies that it is possible to use different methods for imputing data for each class. In chapter 4, results for various imputation schemes were presented. Based on these results, this chapter looks into the possibility of obtaining data for different classes using different imputation schemes and then combining them into one dataset. The new dataset (s) will be referred as a hybrid-imputed dataset in this thesis. For example as shown in chapter, the performance of naive Bayes on data imputed using SVM shows the highest percentage PPV of 86% while the EM presents the highest SPEC of 73%. As a result, since PPV is a positive outcome and represents the alive class, while SPEC is a negative and represents the dead class, the alive class of the SVM imputed dataset will be combined with the dead class of the EM dataset. As will be seen in the coming sections, this approach provides better results; however, there is still a significant number of *FPs* and *FNs*. In order to understand why and how these come about, an investigation of the class posterior probabilities of both the *FP* and *TN* records is conducted to determine the probabilities associated with prediction. The number of missing data present in each *FP* record will also be investigated to determine if they will affect classification performance.

Euclidean distance is applied to determine the distance and similarity in the *FP* and *TN* records such as the maximum and minimum Euclidean distances of the dead class and their corresponding alive class distances. This will also help to identify the variable contributing the most through investigating the ∞ -norm distance of the *FP* and *TN* records.

5.2 Performance of hybrid imputed datasets

Based on the performance of classification using the different imputation schemes, a hybrid dataset was created as discussed above. Thus the different imputation methods were combined (see 5.2.1) and the performances evaluated. For example, the alive class where the missing data was imputed using FKM imputation and the dead class where EM imputation was used, were combined to form a new dataset. Other classes imputed with different imputation methods were also combined in that order, alive and dead respectively, and can be seen in this section. At the same time TAN Bayes is also applied to the new hybrid datasets to determine whether better classification accuracy can be achieved when compared to naïve Bayes performance.

5.2.1 Naïve Bayes performance of hybrid dataset

Based on the results in chapter 4, four hybrid imputed datasets were created. Table 5.1 shows the performance of naïve Bayes classification on these datasets.

SVM and EM		**Predicted**		**Evaluation measures**	%
	True	Alive	Dead	PPV	95
	Alive	1447TP	12FN	NPV	97
	Dead	79 FP	406TN	SEN	99
				SPEC	84
				ACC	95
FKM and EM		**Predicted**		**Evaluation measures**	%
	True	Alive	Dead	PPV	95
	Alive	1448TP	11FN	NPV	97
	Dead	79FP	406TN	SEN	99
				SPEC	84
				ACC	95
FKM and SVM		**Predicted**		**Evaluation measures**	%
	True	Alive	Dead	PPV	86
	Alive	1209TP	250FN	NPV	54
	Dead	192FP	293TN	SEN	83
				SPEC	60
				ACC	77
SVM and FKM		**Predicted**		**Evaluation measures**	%
	True	Alive	Dead	PPV	86
	Alive	1220 TP	239 FN	NPV	55
	Dead	194 FP	291 TN	SEN	84
				SPEC	60
				ACC	78

Table 5.1: Naïve Bayes performance of the four hybrid imputed datasets

It can be seen that SVM and EM and FKM and EM hybrid datasets show the most improvement and also the measures for evaluation are identical. For example the algorithm shows a PPV of 95% in both hybrid datasets; therefore it can be interpreted as predicting 5% as alive *(FP)* when they should be classed as dead. The classification has not only improved accuracy but also reduced the number of incorrectly classified positives. This is also reflected in NPV, SEN, SPEC and ACC measures, which are 97%, 99%, 84% and 95% respectively, which indicates that classification accuracy has improved. This can be interpreted as predicting 3% and 1% as dead when they should be alive, and 16% as alive when they should be dead. In table 4.4 of chapter 4, the incorrectly classified positives and negatives were higher in the naïve Bayes performance of the imputed datasets. For example PPV predicted 14% to be predicted as live when they should be dead while SEN predicted 18% as dead when they should be alive.

On the contrary, the naïve Bayes performance of FKM and SVM and SVM and FKM hybrid datasets show similar measures of evaluation and the classification performance does not show much improvement when compared to SVM and EM, FKM and EM hybrid datasets. PPV and SPEC show identical measures of 86% and 60% respectively, while NPV shows 54% and 55%, and SEN 83% and 84% in the respective hybrid datasets. This could be due to the similarities in classification performance of the datasets individually imputed with FKM and SVM in table 4.4 of chapter 4. For example FKM and SVM share similar PPV of 85% and 86% and similar SPEC of 56% and 60%. However, the classification performance of SVM and EM in table 4.4 is quite different. For example EM shows a SPEC of 73% while SVM shows a PPV of 86%; this means a difference of 13%. SVM PPV is also similar to FKM PPV of 85%. As a result, when combined with the EM dead class this causes the measure of evaluation for both hybrid datasets to be identical as shown in table 5.1.

5.2.2 Tree augmented naïve Bayes performance of hybrid dataset

TAN Bayes classifier was also applied on the hybrid datasets shown in table 5.1. Table 5.2 presents the TAN performance of the four hybrid datasets.

SVM and EM		**Predicted**		**Evaluation measures**	%
	True	Alive	Dead	PPV	97
	Alive	1432TP	27FN	NPV	94
	Dead	49FP	436TN	SEN	98
				SPEC	90
				ACC	96
FKM and EM		**Predicted**		**Evaluation measures**	%
	True	Alive	Dead	PPV	96
	Alive	1426TP	33FN	NPV	93
	Dead	52FP	433TN	SEN	98
				SPEC	89
				ACC	96
FKM and SVM		**Predicted**		**Evaluation measures**	%
	True	Alive	Dead	PPV	96
	Alive	1401TP	58FN	NPV	88
	Dead	59FP	426TN	SEN	96
				SPEC	88
				ACC	94
SVM and FKM		**Predicted**		**Evaluation measures**	%
	True	Alive	Dead	PPV	95
	Alive	1379TP	80FN	NPV	84
	Dead	75FP	410TN	SEN	95
				SPEC	85
				ACC	92

Table 5.2: TAN Bayes performance of the four hybrid imputed datasets

It can be seen that all four hybrid datasets show an improvement when compared to the naïve Bayes performance of the hybrid datasets in table 5.1. The number of *FPs* are reduced, particularly in the SVM and EM, and FKM and EM hybrid datasets, where the *FPs* are 49% and 52% respectively. This is due to the conditional mutual information between the variables, which is not present in naïve Bayes. Where naïve Bayes predicts outcomes based on independence between variables, this is shown in tables 4.3 and 4.4. In contrast, TAN predicts outcomes based on the information shared amongst variables. This means that TAN correctly predicted 30 records from the naïve Bayes *FP* records as dead in SVM and EM, 27 records in FKM and EM, 133 records in FKM and SVM, and 120 records in SVM and FKM. This further indicates that the records misclassified by naïve Bayes shared mutual relations between variables and therefore are correctly predicted as dead (TN).

The dataset which shows the best performance will be explored. For example the *FP* records of SVM and EM will be explored to identify the underlying task of the prediction, probabilities and identify records shared amongst naïve Bayes and TAN performance.

5.3 Exploration of data records

Although the metrics for evaluating the classification methods indicate an improvement in tables 5.1 and 5.2, it can be seen in both naïve Bayes and TAN Bayes performance that there are still a number of false positive and false negative results. In a model which is likely to be making life and death decisions, it is imperative to reduce the number of false positives and false negatives. The key to Bayes classification is the use of posterior probabilities, i.e.

$$If\ posterior(alive) > Posterior(dead), class = alive$$

$$Or\ else\ class = dead$$

The first step in looking at the reason for false positives is to look at the degree of difference between the two sets of posteriories. In this case, the 79 and 49 *FP* records of *SVM and EM* hybrid dataset shown in table 5.1 and 5.2 will be considered. Secondly, the number of missing data present in each record will be determined. The number of missing data has been considered to determine whether this plays a role in the prediction outcome. Thirdly the number of *FP* records shared by the 79 records of naïve Bayes and 49 records of the TAN Bayes performance will be identified. Their *TN* records will also be explored in the same way as the FP records and used as a reference to compare to the *FP* records. Finally, an investigation is carried out as to why different methods of imputation result in varied numbers of *FPs* and *FNs* (see table 5.1).

5.3.1 Effect of missing data and hybrid imputation on class posterior probabilities (naïve Bayes)

Table 5.3 and 5.4 presents the alive and dead posterior probabilities, the number of Missing Data (MD), the true and predicted outcomes of five *FP* and *TN* records. The five records were randomly selected from the 79 *FP* records shown in appendix II and 406 *TN* (not shown) records of the naïve Bayes performance of the SVM and EM hybrid dataset shown in table 5.1.

No.	Record no.	Alive post	Dead post	MD	True	Predicted
3	1486	1	0	0	Dead	Alive
16	1686	0.864	0.136	0	Dead	Alive
25	1725	1	0	0	Dead	Alive
36	1800	1	0	7	Dead	Alive
45	1865	0.993	0.0075.33E-106	1	Dead	Alive

Table 5.3: Alive and dead class posterior probabilities, number of MD, the true and predicted outcomes of the 79 FP records

No.	Record no.	Alive post	Dead post	MD	True	Predicted
108	1568	0	1	28	Dead	Dead
123	1583	0	1	6	Dead	Dead
142	1602	0	1	15	Dead	Dead
219	1679	0	1	23	Dead	Dead
361	1821	0	1	1	Dead	Dead

Table 5.4: Alive and dead class posterior probabilities, number of MD, the true and predicted outcomes of the 406 TN records

It can be seen in table 5.3 that the alive posterior values are greater than that of the dead posterior. This simply indicates that patients have been predicted as alive when in fact they should be dead *(FP).* This is also the same outcome in table 5.4, where the dead class posterior values of the *TN* records are greater than those of the alive class, indicating that these records are correctly classified as dead. It can also be seen that the dead and alive posterior probabilities of the *FP* records nearly overlap when compared to the *TN* records' probabilities and there are no clear distinction, therefore causing misclassification. In contrast, in the TN records, the dead and alive posterior probabilities are far apart and very distinctive.

The number of missing data present in both *FP* and *TN* records are different in that the *TN* records consists of more missing data in their records, while in the *FP* records, the majority of the records have no missing data, five records have 1 missing data and three records have 7 missing data; the full 79 records are shown in appendix II. This comparison indicates that the presence of missing data in the records does not produce misclassification of the class; rather, those records with less or no missing data are the problem. As will be seen later on, it is often the presence of incorrectly obtained measurements that are a key cause for misclassification.

5.3.2 Effect of missing data and hybrid imputation on class posterior probabilities (TAN Bayes)

Tables 5.5 and 5.6 present the alive and dead posterior probabilities (alive post and post), the number of Missing Data (MD), the true and predicted outcomes of five *FP* and *TN* records respectively. Five records were also selected at random from the 49 *FP* records shown in appendix III and 436 *TN* records (not shown) of the SVM and EM imputation TAN performance shown in table 5.2.

No.	Record no.	Alive post	Dead post	MD	True	Predicted
14	1729	0.996	0.004	0	Dead	Alive
26	1861	1	0	7	Dead	Alive
30	1869	0.934	0.066	0	Dead	Alive
43	1899	0.521	0.479	7	Dead	Alive
47	1914	0.976	0.024	0	Dead	Alive

Table 5.5: Alive and dead class posterior probabilities, number of MD, the true and predicted outcomes of the 49 *FP* records.

No.	Record no.	Alive post	Dead post	MD	True	Predicted
170	1629	0	1	9	Dead	Dead
240	1699	0	1	4	Dead	Dead
261	1720	0	1	8	Dead	Dead
310	1769	0	1	23	Dead	Dead
453	1912	0	1	27	Dead	Dead

Table 5.6: Alive and dead class posterior probabilities, number of MD, the true and predicted outcomes of the 436 *TN* records

It can be seen in table 5.5 that there are fewer missing data in the five *FP* records when compared to the five *TN* records in table 5.6, which shows more missing data in the records. For example, appendix III shows 49 *FP* records, where a majority of the records have no missing data, two records have 1 missing data and three records have 7 missing data, while in table 5.6, the *TN* five records show 9, 4, 8, 23 and 27 missing data in the five records respectively. *These 49 FP records are also present in the 79 FP records of the naïve Bayes performance.* ***The shared records are highlighted in bold in appendix II of the FP records.***

The comparison between the *FP* and *TN* records further indicates that the higher number of missing data present in the *FP* records does not affect misclassification; rather it is those records with less missing data. This essentially indicates that those records with zero missing data are observed incorrectly, as a majority of them have been incorrectly predicted as alive. As a result, those records with zero and one missing data in the 79 and 49 *FP* records will be removed in order to determine whether classification performance will be improved. Naïve Bayes and TAN Bayes classifiers will be performed on the modified dataset. This will be shown in section 5.4 of this chapter.

5.3.3 Performance of modified SVM and EM hybrid dataset

Naïve Bayes and TAN classification performance of the SVM & EM hybrid data and original clinical dataset after discarding *FP* records with 0 and 1 missing data is examined in order to determine whether classification accuracy will improve. As shown in table 5.3 and appendix II there are more records with 0 and 1 missing data. Hence, these are records with more observed data than imputed data.

76 *FP* records which had 0 and 1 missing data were removed from the dataset in order to achieve better accuracy. In doing so, the number of dead class records was reduced from 485 to 409, thus reducing the number of records in the dataset to 1868.

	Predicted		**Evaluation measure**	**%**
True	*Alive*	*Dead*	PPV	100
Alive	1459TP	0FN	NPV	100
Dead	5FP	404TN	SEN	100
			SPEC	99
			ACC	100

Table 5.7: Naïve Bayes performance of SVM and EM hybrid dataset, 1868 records

	Predicted		**Evaluation measure**	**%**
True	*Alive*	*Dead*	PPV	100
Alive	1459TP	0FN	NPV	100
Dead	4FP	405TN	SEN	100
			SPEC	99
			ACC	100

Table 5.8: TAN performance of SVM and EM hybrid dataset, 1868 records

	Predicted		**Evaluation measure**	**%**
True	*Alive*	*Dead*	PPV	87
Alive	1212TP	247FN	NPV	48
Dead	178FP	231TN	SEN	83
			SPEC	56
			ACC	77

Table 5.9: Naïve Bayes performance of the original dataset, 1868 records

	Predicted		**Evaluation measure**	**%**
True	*Alive*	*Dead*	PPV	88
Alive	1301TP	158FN	NPV	60
Dead	172FP	237TN	SEN	89
			SPEC	58
			ACC	82

Table 5.10 TAN Bayes performance of original dataset, 1868 records

Tables 5.7 and 5.8 show the naïve Bayes and TAN performance of the SVM and EM hybrid dataset. It can be seen in both tables that classification performance has significantly improved, where PPV, NPV, SEN and ACC are 100%. The difference in *FP* values is by 1, where the naïve Bayes algorithm shows FP of 4, while TAN Bayes algorithm shows *FP* of 5. This indicates that one record has been correctly classified as dead *(TN)* by TAN. The conditional mutual information implemented in TAN allows for dependencies between the variables to be located and therefore attempts to classify the true prediction of the data based on this.

On the contrary, tables 5.9 and 5.10 present naïve Bayes and TAN Bayes performance of the original clinical data after removing the 76 *FP* records with 0 and 1 missing data. It can be seen in both tables that their performance is different when compared to the hybrid imputed dataset shown in tables 5.7 and 5.8. These results reflect a good performance of the original clinical dataset. However, the elimination of the 76 *FP* records has not affected or changed classification accuracy by a large margin when compared to the naïve and TAN Bayes performance of the full original clinical data shown in tables 4.3 and 4.7, where similar performances are shown.

These results indicate that the hybrid imputed datasets provide a better classification of the Hull LifeLab dataset than records with complete observed data and incomplete data records. Thus it is better to have records with missing data so that suitable missing data imputation methods are applied. This also further confirms that the observed data are incorrect and correctly observed variables are needed. In addition, this process can be used to identify variables and records of concern at the stage of data collection especially when the clinicians and model have different opinions. The next section will explore and investigate the properties and problems that influence the classification of the *FP* outcome, through the application of *K*-means clustering and Euclidean distance.

5.4 Euclidean distance

In this section Euclidean distance is applied to better understand the properties of the dataset. Firstly, this involves computing the means for the alive class and the dead class. Euclidean distance uses a metric to determine the distance and similarity between the class mean and a data point which is the 79 *FP* and 406 *TN* records of the SVM and EM hybrid dataset.

Consider two points $a = (a_1, a_{2,} \dots, a_d)$ is the mean vector for the alive class, and $x = (x_1, x_{2,} \dots, x_d)$ is the data point. Thus if D_t is the distance between the two points, where t is a positive number, is given by $D_t(a, x) = (\sum_{i=1}^{d} |a_{i-}x_i|^t)^{1/t}$ for $1 \leq t < \infty$. The Euclidean distance is obtained when $t = 2$, and the distance in the infinite norm is obtained when $t = \infty$. The Euclidean distance is useful to determine an overall perspective of the distance of any point from another given point. However, it does not provide much information as to the contributions of the various dimensions. This is important, for it allows for determining which variable is contributing the most to the mismatch of classes. This is possible when $t = \infty$: thus $D_\infty = \max_{1 \leq i \leq d} |a_i - x_i|$

5.4.1 Euclidean distance of FP and TN records

The dead and alive Euclidean distance of the *FP* records are shown; the Euclidean distance of the *TN* records are also determined and used as a reference. Appendix IV shows the Euclidean distance of the 79 *FP* records. From this, records with the maximum and minimum Euclidean distances of a point from the alive mean and dead mean are selected (highlighted in bold). Two records with the maximum and minimum Euclidean distance are also selected from the 406 TN records (not shown).

No.	**Record no.**	**Euclidean distance range**	**SVM & EM Euclidean distance**	
FP records				
			Dead	*Alive*
8	1543	Max	1162.57	343.29
58	1886	*Min*	762.76	436.93
TN records				
			Dead	*Alive*
341	1801	Max	1170.95	1950.70
252	1709	Min	92.64	969.99

Table 5.11: SVM and EM class Euclidean distance

Table 5.11 presents the maximum and minimum Euclidean distances in the dead class and their corresponding alive class distances in the *FP* and *TN* records. It can be seen in the maximum and minimum *FP* records that the alive class shows the smallest distance when compared to the dead class. The table shows the maximum Euclidean distance in the dead and alive class to be 1162.57 and 343.29 respectively while the minimum Euclidean distance in the dead and alive classes are 762.76 and 436.93 respectively. This indicates that the data points are closer to the alive mean and further away from the dead mean; therefore the record is classed as alive when they should be classified as dead.

Similarly, it can be seen in the *TN* records that the maximum and minimum Euclidean distance is smaller in the dead class when compared to the alive class. The maximum Euclidean distance in the dead and alive class is 1170.95 and 1950.70 respectively, while the minimum Euclidean distance in the dead and alive class are 92.64 and 969.99 respectively. This means that these records are correctly classified as dead and further indicates that the data points are closer to the dead mean than the alive mean. The Euclidean distance results and misclassification of the dead class may have been influenced by several factors. For example, the presence of variability in the observed data which are presented as outliers may have been introduced during data collection.

5.4.2 *Investigating the* $\infty -$ *norm distance of FP and TN records*

Table 5.12 presents the maximum difference (max diff) of the variable contributing the most in the dead class and their corresponding alive class maximum difference in the *FP* and *TN* records.

No. of record	**Record no.**	**Variables and maximum difference**			
		FP records			
		Dead		*Alive*	
		Max diff	Variable	Max diff	Variable
8	1543	786	PEFR	179.29	PEFR
58	1886	486	PEFR	120.71	PEFR
		TN records			
		Dead		*Alive*	
		Max diff	Variable	Max diff	Variable
341	1801	0	PEFR	606.71	PEFR
252	1709	0	PEFR	606.71	PEFR

Table 5.12: Maximum difference of variable contributing the most.

It can be seen in the *FP* records, that records 8 and 58 present the maximum difference of 786 and 486 in the dead class and the corresponding variable is Peak Expiratory Flow Rate (PEFR) (Vaughan *et al.,* 1989) which is the variable contributing the most, while the corresponding alive class shows a low maximum difference of 179.29 and 120.71. The high maximum difference presented by the dead class indicates how far the records’ data point is from the alive mean; hence the reason why the dead records are classified as alive. This indicates that PEFR is a problematic variable and therefore causing the misclassification of the dead class. In comparison to the *TN* records, the maximum difference of records 341 and 252 is 0, which is less than that of the alive class, showing a maximum difference of 606.71. This indicates that the records data points are closer to the dead class and therefore they are correctly predicted as dead. Appendix V shows the full dead and alive maximum difference of the PEFR variable, where records 8 and 58 are highlighted in bold.

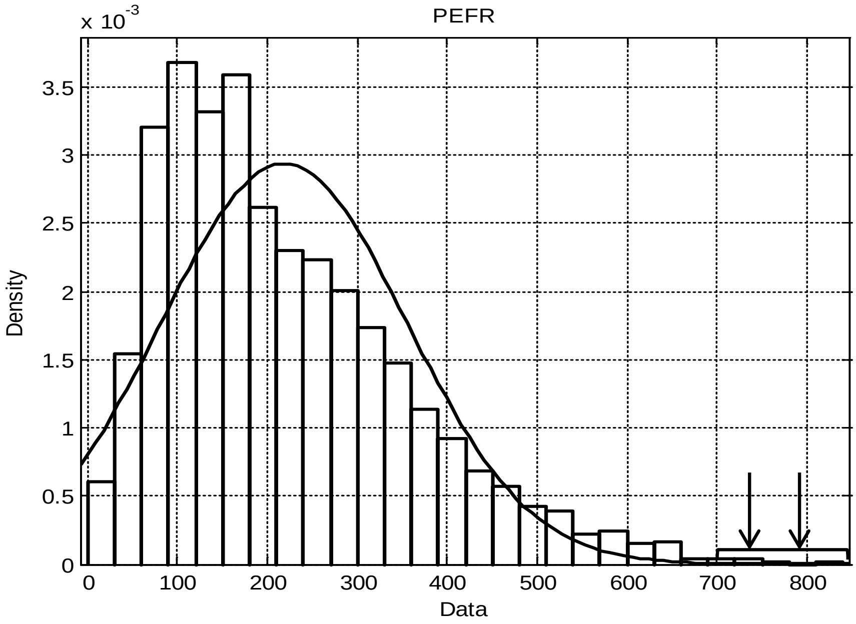


Figure 5.1 Data distribution of the PEFR variable

Figure 5.1 shows the distribution of the PEFR variable. It can be seen that the data of the PEFR variable is varied, with some are presented as extreme values (indicated by the arrows). As a result, the distribution of the variable is positively skewed, where most of the distribution is concentrated on the left of the figure. This could be due to the nature of data collection.

PEFR is a lung functioning test that is used to determine pulmonary impairment. For the purpose of clinical assessment new PEFR measurements are repeatedly taken for comparison with the old measurement to determine any significant changes and whether they agree sufficiently to be replaced with an old measurement. Bland and Altman (Bland and Altman, 1986) suggest that this investigation is usually carried out by using correlation coefficients which are known to be misleading, which may have contributed to the variation of the variable.

5.5 Summary

For most data driven methodologies, the presence of abundant data is a prerequisite. A reason for this is that the law of large numbers ensures that the larger the size of the sample means and standard deviations are as close as possible to the expected means and standard deviations. The dataset under consideration has a large number of data points, albeit imbalanced. However, methods based on Bayes are robust to these imbalances as long as there is sufficient number of representative samples. This is the case with the dataset under consideration. However, the results were not as expected. Thus this chapter presented results, using the inherent transparency afforded by Bayes methodologies. Therefore it is possible to analyse why some records are being misclassified and allows the challenges of the clinical data to be explored in a greater detail.

In this chapter, the performance of hybrid imputed datasets was assessed. It showed that irrespective of the imputation or hybridisation of the dataset, there was a significant number of *FPs* that still remained. As a first step towards understanding the causes for this, an investigation of the class posterior probabilities was carried out. The next stage was to look for records with imputation and the number of missing data present within the *FP* records in order to determine if the imputation and missing data were causing the problem. Once this exploration was done, a further investigation was carried out to explain the reason for misclassification of the *FP* records, such as determining the problematic variable(s) in the data, through the application of Euclidean distance. This involved estimating the distance between the class mean and the 79 *FP* data points.

The results in section 5.3 show that *FP* records with no imputed data are the ones which are incorrectly classified. Section 5.4 shows an investigation of the various records using the ∞ norm (see appendix V for full dead and alive list of the 79 *FP* records). It was possible to determine the variable which was at the root of the problem of misclassification. The PEFR variable presented varied measurements and contributed the most variation in the data. These results have shown that however robust a method is, the nature of the data present is crucial in order to make effective clinical decisions. In clinical settings, issues surrounding the quality of clinical data usually include poor data handling processes and errors during the migration process, e.g.

- When transferring data from one system to another;
- Failure to stick to data entry and maintenance procedures;
- Failure to update instruments used for a particular test,
- Instruments of multiple versions may vary in subtle ways in cases where multiple instruments are used for a particular test.

These issues can occur at any stage of data collection and initial processing and thus as a result could be the problem with the real life heart failure clinical dataset. Therefore, it is crucial that these measures are considered during data collection.

The PEFR variable and other variables shown in appendix I such as creatinine and uric acid flag up extreme values also known as outliers. Although there are several techniques available in literature for detecting and eliminating such values in clinical data; the current approach explores the data structure to determine the presence of bad data records. Often, in real life clinical datasets, it is always possible for two records to have similar values but be classified in different classes, mainly due to the experience of the clinicians and the nature of data collection. Exploring the *FP* records, the number of missing data present in records, their posterior probabilities and data distribution has undoubtedly aided in understanding the data space and properties. In addition, it has allowed an investigation of the data from a broad overview to a fine structure in order to impact effective quality of data for better healthcare.

CHAPTER 6 OTHER CLASSIFICATION ALGORITHMS

6.1 Introduction

There are two aspects to data mining. One is to understand the nature of the data and the other is the performance. The key is to understand the relationship between the two. Thus chapters 4 and 5 have shown that classification accuracy can be improved through the application of an augmented naïve Bayes, imputation methods, the proposed hybrid imputed dataset and discarding problematic records. At the same time the methods provide tools which can look into the reasons for poor performance. There are other sophisticated methods available (e.g. J48, ANNs etc.); however these often remain as black boxes, in the sense that they do not yield the internal relationships between the data and the method.

This chapter will discuss four other classification algorithms, namely, (a) Bayes classifier based on Kernel Density Estimation (KDE) (b) Beta based Bayes classifier, (c) decision tree (C4.5) and (d) Multi-layer Perceptron (MLP). The algorithms will be applied on the original data and SVM and EM hybrid data for comparative analysis. Results will be presented in two ways: 1) as a training set and 2) with the 10-fold cross validation and compared to the naïve Bayes performance of the original data and SVM and EM hybrid imputed data shown in table 4.3 and 5.1 respectively. The classification outcome in both results will be discussed and an explanation offered as to why these algorithms were not initially considered.

6.2 Bayes classifier based on kernel density estimation

Kernel Density Estimation (KDE) is a technique applied in data mining for solving the smoothing problem posed by real life clinical data (Guidoum, 2014). Most classification methods are parametric, including naïve Bayes. There is, however, a class of Bayes classifiers which are not parametric. These are classed as kernel density methods. Kernel applies a kernel density estimator (eq. 6.1) rather than a Gaussian distribution. The advantage here is that unlike the naïve Bayes classifier, which assumes that continuous variables are in a Gaussian distribution, the data exist in a non-Gaussian distribution. Thus for naïve Bayes a Gaussian distribution is fitted irrespective of the actual distribution. For example this can be seen in figure 6.1 where kernel, Gaussian and the actual data distribution of the potassium variable is shown. The normal blood potassium level is between 3.5-5.1mmol/L, the two distinctive peaks shows that the majority of the population are within this range (Parikh and Webb, 2012).

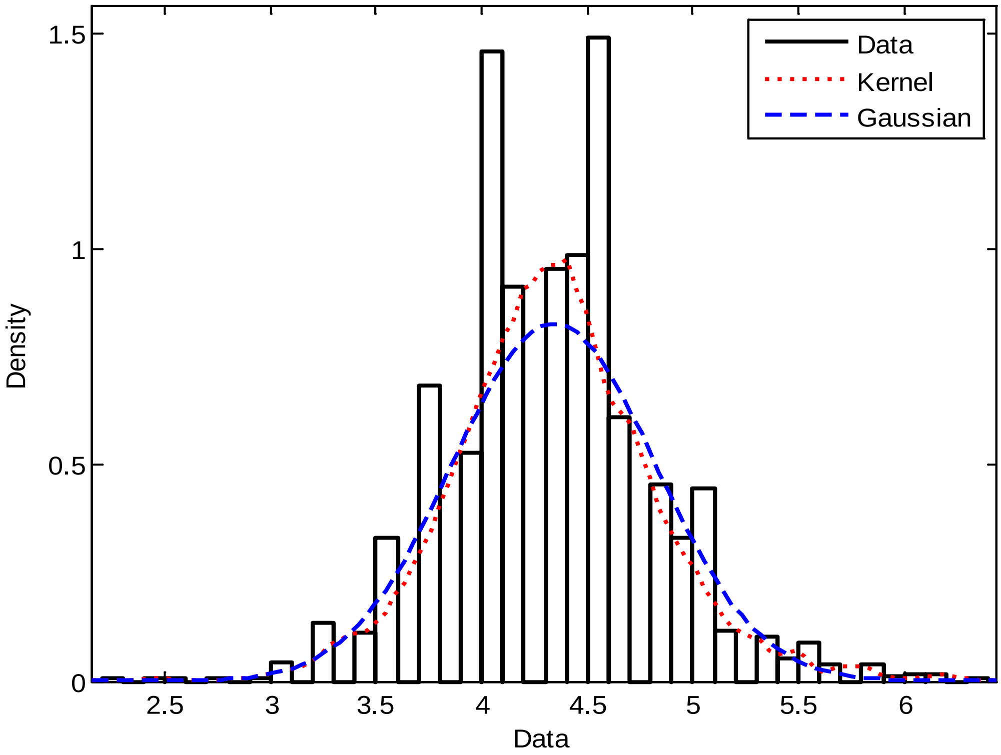


Figure 6.1 Naive Bayes (Gaussian), Kernel and data distribution of the potassium variable

KDE is a non-parametric method that estimates the probability density function $f(x)$ of the continuous variables X using kernels (Pérez *et al.*, 2009). It should be noted that KDE is a flexible estimator in that, for modelling the conditional density no assumptions are made on the shape of the probability density or the number of kernels. Let

$(X_1, X_2, \ldots, X_n)$ be n samples of an independent continuous random variable X, with an unknown density function. Thus the density function $f(x)$ is as follows (eq. 6.1.)

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \tag{6.1}$$

where:

$K_h(.)$ is the kernel function,

n is the number of samples

h is a bandwidth matrix; a smoothing parameter which controls the degree of smoothing applied to data.

Thus KDE is characterised by the kernel density K selected and the bandwith h. The kernel based density estimate $\hat{f}_h(.)$ is determined by averaging n kernel densities $K_h(.)$ placed at each observation x_i.

KDE is also known as a flexible naïve Bayes algorithm (John and Langley, 1995). John and Langley state that kernel estimation with Gaussian kernels is similar to naïve Bayes Gaussian, except that the estimated density is averaged over a large set of kernels $p(X = x|C = c) = \frac{1}{n}\sum_i g\,(x, \mu_i, \sigma_c)$, where i ranges over the training point of variables X in class c and $\mu_i = x_i$, as applied in this thesis.

Its flexibility presents a few disadvantages, in that a significant amount of storage space is required for storing continuous attribute values during training, and could increase exponentially as the number of variables increases. At the same time KDE computes the probability measure of n variables in order to obtain $P\left(X_i = x_i \middle| C = c_j\right)$ one per observed value of X in class C, and hence if the number of samples N is large, the computational and space complexity will increase (Li *et al.*, 2006, Sinha and Gupta, 2008).

6.2.1Kernel density estimation performance via Bayesian network classifier

A confusion matrix was used for a more detailed analysis of the class attribute distribution. Table 6.1 presents the kernel based Bayesian network classifier performance of the original and hybrid datasets. It can be seen that the performance with the original data is similar to the naïve Bayes performance with the original data shown in table 4.3, while the performance with the hybrid data is similar to that of the naïve Bayes performance with the hybrid data shown in table 5.1. However, if the results from the cross validation are compared, there is greater similarity between the two sets of algorithms (KDE and naïve Bayes). This could be due to cross validation assessing how the results will generalise to independent data, so that over-fitting problems are limited. This generalisation is similar to the independence assumption posed by naïve Bayes, which is also retained by KDE during computation. The results indicate that the added complexity of the KDE method does not yield any appreciable improvement in the performance.

Original dataset									
Training set					***Cross validation***				
	Predict		***Evaluation measure***	***%***		***Predict***		***Evaluation measure***	***%***
True	Alive	Dead	PPV	86	***True***	Alive	Dead	PPV	84
Alive	1249TP	210FN	NPV	57	Alive	1202TP	257FN	NPV	50
Dead	201FP	284TN	SEN	86	Dead	225FP	260TN	SEN	82
			SPEC	59				SPEC	54
			ACC	79				ACC	75
SVM and EM hybrid dataset									
Training set					***Cross validation***				
	Predict		***Evaluation measure***	***%***		***Predict***		***Evaluation measure***	***%***
True	Alive	Dead	PPV	95	***True***	Alive	Dead	PPV	95
Alive	1459TP	0FN	NPV	100	Alive	1448TP	11FN	NPV	97
Dead	80FP	405TN	SEN	100	Dead	81FP	404TN	SEN	99
			SPEC	84				SPEC	83
			ACC	96				ACC	95

Table 6.1: Kernel based Bayesian network classification performance of the original data and hybrid data.

6.3 Beta based Bayes classifier

Beta distribution is a method applied to continuous random variables to control the shape and behaviour of continuous probability distributions. Beta distribution for X, where two shape parameters, alpha, α and beta, β are unknown (Gupta and Nadarajah, 2004), is computed by Maximum Likelihood Estimates (MLEs) (Gnanadesikan *et al.*, 1967), (eq. 6.2). MLEs are the values of the parameters that maximize the likelihood function for fixed values of X. If $x_1, \dots, x_n$ are independent variables each having a beta distribution, the joint log likelihood function for n *independent and identically distributed (iid)* observations is:

$$\ln \mathcal{L}(\alpha, \beta | X) = \sum_{i=1}^{n} \ln(\mathcal{L}_i(\alpha, \beta | x_i)) \tag{6.2}$$

The shape parameters α and β are determined by maximising the likelihood function, which involves estimating the values of the parameters that give the highest likelihood given the data X. The likelihood function is determined in a similar way to the beta Probability Density Function (PDF) shown in equation 6.3. However for the pdf, the parameters are considered as the normalising constants and the variable as x. The likelihood function reverses the roles of the variables, where the observed sample values are fixed while the variables are unknown parameters.

$$y = f(x|\alpha, \beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x) \tag{6.3}$$

where:

$B(.)$ is the beta function acting as the normalising constant

(α, β) are two positive shape parameters that control the shape of the distribution.

$I_{(0,1)}(x)$ is the indicator function that ensures only values of x in the range (0,1) have nonzero probability.

In this thesis, the MATLAB function '*betafit*' (MathWorks, 2005) is applied to estimate the beta parameters. The function produced an error, suggesting that the data must be within the closed interval of [0,1]. As a result, the Hull LifeLab data was normalised to remain within this interval.

The mean (μ) and variance (σ^2) beta distribution with parameters α and β are computed (eq. 6.4 and 6.5) for each class and then applied to the naïve Bayes algorithm outlined in table 4.2. Thus the mean and variance are given by:

$$\mu = \frac{\alpha}{\alpha+\beta} \tag{6.4}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{6.5}$$

The estimates of the two shape parameters are dependent on the properties of the data and depending on these estimates different beta distribution can be obtained. For example if one or both parameters are smaller than 1, the probability is concentrated at x-values of the distribution between 0 and 1. Whatever way the data is structured will reflect on the resulting shapes and distribution (Zhanyu and Leijon, 2011).

6.3.1 Beta distribution classification results

The Hull LifeLab data have been normalised so that the data is within the same scale, between 0 and1 and manageable for the application of beta naïve Bayes in order to achieve both data integrity and performance.

Table 6.2 presents the beta based naïve Bayes classifier of the original data and hybrid data. For most of the methods discussed in this thesis, missing data was not a major computational issue during cross validation. However, with beta functions cross validation failed and as a result the missing data were replaced with 0.001. Therefore, the original data result shown in the table below is a product of the missing data replaced with 0.001 in both the training set and cross validation.

It can be seen that the training set result of the original data is poor when compared to the naïve Bayes performance with the original data in table 4.3. For example the training set shows a SPEC of 36% due to the high number of *FPs*. However, the cross validation results show a better performance, similar to the naïve Bayes performance in table 4.3. These results suggest that beta has not changed or improved the classification performance due to the complexity posed by the Hull LifeLab data, since the estimated parameters are dependent on the data space and properties. On the contrary, the same performance in the training set and cross validation results are presented by the hybrid data and the performance is similar to the naïve Bayes performance with the hybrid data shown in table 5.1. However, beta distribution shows a subtle improvement, but not by a large margin.

Original dataset									
Training set					***Cross validation***				
	Predict		***Evaluation measure***	***%***		***Predict***		***Evaluation measure***	***%***
True	Alive	Dead	PPV	81	***True***	Alive	Dead	PPV	88
Alive	1291TP	168FN	NPV	51		1124TP	335FN	NPV	50
Dead	309FP	176TN	SEN	88		151FP	334TN	SEN	77
			SPEC	36				SPEC	69
			ACC	75				ACC	75
SVM and EM hybrid dataset									
Training set					***Cross validation***				
	Predict		***Evaluation measure***	***%***		***Predict***		***Evaluation measure***	***%***
True	Alive	Dead	PPV	95	***True***	Alive	Dead	PPV	95
	1459TP	0FN	NPV	100		1459TP	0FN	NPV	100
	81FP	404TN	SEN	100		81FP	404TN	SEN	100
			SPEC	83				SPEC	83
			ACC	96				ACC	96

Table 6.2: Beta distribution naïve Bayes classification performance of the original and hybrid data

6.4 Decision tree classifier (C4.5)

The methods considered so far made a number of assumptions on either the distribution of the data or the dependency of the variables. However there is a class of methods which do not make these assumptions. In what follows are two applied methods; a decision tree and feed forward network algorithms.

Decision tree is a non-parametric classification method applied in machine learning and data mining for classifying examples and prediction of values (Balakrishnan and David, 2010). The method generates a tree structure consisting of either a leaf, indicating a class, or a decision node of a test with one branch and a subtree for each possible outcome of the test conditions (Quinlan, 1986, Quinlan, 2014). The tree structure is used to classify an example by simply starting at the root of the tree and proceeding through it until a leaf is encountered.

In this thesis, the decision tree is generated using the C4.5 algorithm developed by Ross Quinlan (Quinlan, 1986, Quinlan, 2014). C4.5 is an extension of the Iterative Dichotomiser 3 algorithm (ID3) which converts the ID3 algorithm trained tree into sets of if-then rules. The rules are then presented in the form of a J48 pruned tree. The tree will be generated for the original data and hybrid data to show the rules of the pruned tree. Detailed steps of the C4.5 algorithm are presented in table 6.3.

Step	Description
Summary of tree	**INPUT:** Training data
	OUTPUT: Decision tree
1	X contains all the samples in the dataset, belonging to a single class C_j The decision tree for X is a leaf identifying class C_j
2	Determine the class to be associated with the leaf by using the concept of information entropy (eq. 6.6). If X is any set of samples in X, let $freq(C_j, X)$ stand for the number of samples in X that belong to class C_j, out of the k possible classes, $C_1, C_2, \dots C_k$ and $\|X\|$ is the number of samples in the set X. Then the entropy of the set S: $Entropy(X) = -\sum_{j=1}^{n} \frac{freq(C_j,X)}{\|X\|} log_2 \left(\frac{freq(C_j,X)}{\|X\|}\right)$ (6.6) where: n= number of attributes Entropy measures the average amount of information needed to identify the class of a case in X.
3	X is partitioned into subsets of samples $X_{1,} X_2, \dots, X_N$ where X_i contains all the samples in X that have outcome of the chosen test node. Therefore the decision tree for X consists of a decision node identifying the test and one branch for each possible outcome. The criterion is to select an attribute with the highest gain value to make the decision.
4	Repeat the process for each branch until all examples have the same class

Table 6.3: C4.5 (decision tree) algorithm

The computation of the decision tree algorithm can be over-complex and cause overfitting on the training data, particularly on data with a large number of attributes. The cost at each node involves searching through the attributes $O(n_{attributes})$ to locate the attribute that offers the largest reduction in entropy. This includes the cost of number of attributes, samples and logarithmic number of samples $O\left(n_{attributes}N_{samples}\log\left(N_{samples}\right)\right)$ at each node. As a result, this leads to a total cost over the entire trees by summing the cost at each node $O(n_{attributes}N^{2}{}_{samples}\log\left(N_{samples}\right))$ (Dumont *et al.,* 2009, Breiman *et al.,* 1984). However, to reduce the complexity of the tree during construction of the decision tree, pruning is introduced.

Pruning reduces the number of nodes by eliminating rules that provide little contribution to classifying instances or improving accuracy. It is an important element as it reduces the computational complexity of the final classifier, avoids over-fitting of the data and improves the tree structure by controlling the size of the tree (Drazin and Montag, 2012, Maimon and Rokach, 2008). On the other hand, the initial computational complexity can be reduced and over-fitting avoided by setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree.

6.4.1 J48 pruned tree results

The C4.5 algorithm is often used to implement a j48 pruned tree (e.g. WEKA machine learning software (Hall *et al.,* 2009, Witten and Frank, 2005)) on the original and hybrid datasets. The outcome is a set of rules. The number of rules and leaves generated by the original dataset is 219 and 110 respectively, while the hybrid imputed dataset generated 61 rules and 31 leaves (see appendix VI and VII for full rules). It can be seen that the number of leaves is reduced in the hybrid dataset. This indicates that the tree structure is dependent on the data. In this case it is due to the richness of the data space after imputation, although decision tree is able to process erroneous and incomplete datasets (Bhargava *et al.,* 2013). However, implementation of such data is computed differently from a complete dataset. For example in step 3 of the C4.5 algorithm, J48 divides instances with missing data for the split attribute up into fractional parts proportional to the frequencies of the observed non-missing data (Witten and Frank, 2005). Therefore missing attribute data are ignored during the measure of information entropy in step 2.

Despite the reduced number of rules and leaves in the hybrid data, there is still the problem of generating a biased tree due to the class imbalance present in the Hull LifeLab dataset where the alive class dominates the dead class. This imbalance creates a biased predictor and imbalanced decision tree and thus causes the problem of excessive testing time (Ramanan *et al.,* 2007). A simple solution to this will be to balance the classes to reduce the bias, and tree size, and thus improve accuracy. However, this does not allow the properties of the data to be examined in great detail and thus the aim of this thesis would not be achieved.

Figure 6.2 and table 6.4 present the decision tree and rules of the original data.

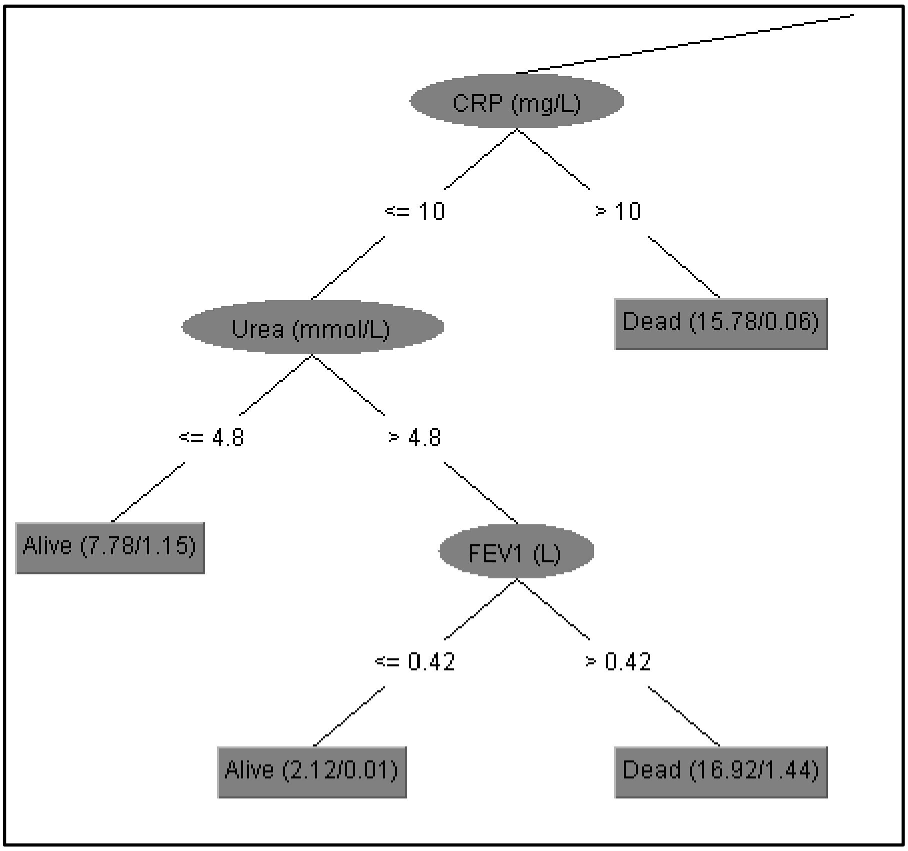


Figure 6.2 Decision tree (j48) of the original data

Original data

```
Urea (mmol/L) <= 9.5
|  FEV1 (L) <= 0.92
|  |  BMI <= 22.838625
|  |  |  CRP (mg/L) <= 10
|  |  |  |  Urea (mmol/L) <= 4.8: Alive (7.78/1.15)
|  |  |  |  Urea (mmol/L) > 4.8
|  |  |  |  |  FEV1 (L) <= 0.42: Alive (2.12/0.01)
|  |  |  |  |  FEV1 (L) > 0.42: Dead (16.92/1.44)
```

Table 6.4: Rules of the original data

It can be seen that there is a hierarchical division of the attributes. Table 6.4 shows urea to be the highest gain. Due to the size of the tree and the limited space in the thesis, only a small section of the tree is shown in figure 6.2, which shows the *CRP* variable as the first node. It can be seen that if the *CRP* variable is greater than 10 it is classified as dead. The numbers in parenthesis represent the number of instances correctly classified as dead/the number of instances incorrectly classified as dead which are 15.78/0.06 respectively. The first number is usually larger than the second number as the algorithm is designed to obtain the best possible number of correct classifications. However, if *CRP* is less than 10, the algorithm proceeds to urea and if the urea is less than 4.8 the algorithm classifies the variable as alive, where 7.78 is correctly classified as alive and 1.15 is incorrectly classified as alive. However, if urea is greater than 4.8, the algorithm continues to the next node *(FEV1).* If the *FEV1* is less than 0.42 then the algorithm classifies the variable as alive, where 2.12 is correctly classified as alive and 0.01 is incorrectly classified as alive. Similarly if *FEV1* is greater than 0.42, the algorithm classifies the variable as dead, where 16.92 is correctly classified as dead and 1.44 is incorrectly classified as dead.

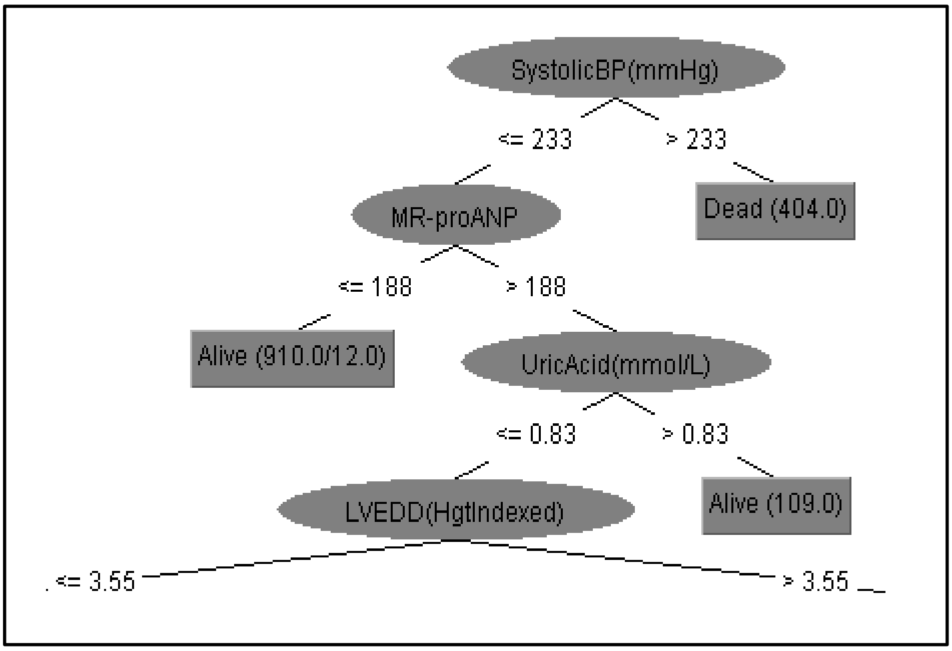


Figure 6.3 Decision tree (j48) of the hybrid dataset

Hybrid imputed dataset

```
SystolicBP(mmHg) <= 233
|  MR-proANP <= 188: Alive (910.0/12.0)
|  MR-proANP > 188
|  |  UricAcid(mmol/L) <= 0.83
|  |  |  LVEDD(HgtIndexed) <= 3.55
```

Table 6.5: Rules of the hybrid dataset

Figure 6.3 and table 6.5 also present a small section of the decision tree and rules of the hybrid dataset. It can be seen in the table that the *SystolicBP* variable has the highest gain and contains the most information. For this reason it has been selected as the first split criterion. Therefore, if the *SystoliBP* variable is less than 233, the algorithm proceeds to the next split which is *MR-proANP* and 910.0 is correctly classified as alive while 12.0 is incorrectly classified as alive. The number of correctly classified outcomes shown in table 6.5 is larger than those in the original data. This indicates that decision tree produces a better result when implemented on complete data.

6.4.2 Decision tree classification performance

The decision tree classification performance of the original and SVM and EM hybrid dataset are presented in table 6.6. It can be seen that accuracy is greater in the hybrid imputed dataset when compared to incomplete dataset. For example the training set of the hybrid data shows a significant performance, where the NPV and SEN measures of evaluation are 100% while PPV, SPEC and ACC are 98%, 95% and 99% respectively. The performance of the original dataset (training set) is better when compared to naïve Bayes and TAN Bayes classifier results of the same dataset shown in figure 4.3 and 4.7 respectively. Similarly, the hybrid results show a better performance when compared to the hybrid result of the naïve Bayes shown in table 5.1. The improvement in classification performance generated by the imputed dataset, is due to the measure of entropy, which estimates the average value (expected) of the information contained in the data. As a result, with richer data and more information presented by the data, the expected value is achieved.

The decision tree algorithm results provide an output number of classified and misclassified outcomes. However, the algorithm does not give an insight as to how and why misclassification occurred. Therefore decision tree does not have the properties to analyse the cause of misclassification and to further reduce the number of *FPs* to the minimum.

Original dataset									
Training set					***Cross validation***				
	Predict		***Evaluation measure***	***%***		***Predict***		***Evaluation measure***	***%***
True	Alive	Dead	PPV	92	***True***	Alive	Dead	PPV	80
Alive	1434TP	25FN	NPV	93		1213TP	246FN	NPV	42
Dead	126FP	359TN	SEN	98		305FP	180TN	SEN	83
			SPEC	74				SPEC	37
			ACC	92				ACC	72
SVM and EM hybrid dataset									
Training set					***Cross validation***				
	Predict		***Evaluation measure***	***%***		***Predict***		***Evaluation measure***	***%***
True	Alive	Dead	PPV	98	***True***	Alive	Dead	PPV	96
	1458TP	1FN	NPV	100		1408TP	51FN	NPV	89
	26FP	459TN	SEN	100		66FP	419TN	SEN	97
			SPEC	95				SPEC	86
			ACC	99				ACC	94

Table 6.6: Decision tree classification performance of the original data and hybrid data.

6.5 Multilayer perceptron

Multilayer perceptron (MLP) is a feedforward artificial neural network applied to learn classification problems (Silva *et al.*, 2008). The training process involves a combination of three layers, namely, the input layer, one or more hidden layers, and the output layer. Each layer is connected and information flows from one layer to the next. The architecture of a multilayer perceptron is shown in figure 6.4. (Gardner and Dorling, 1998, Autio *et al.*, 2007, Vaughn, 1996).

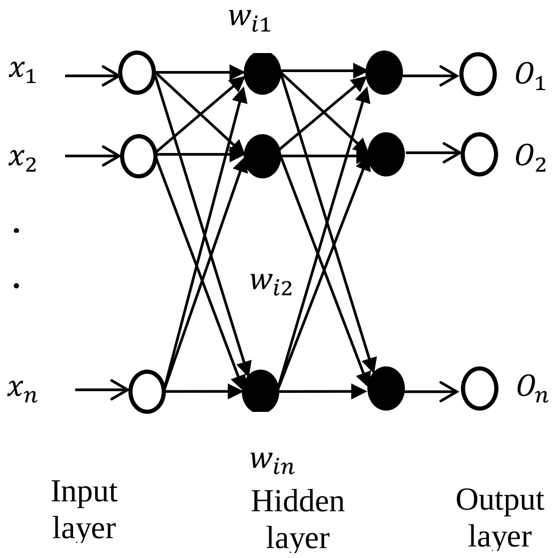


Figure 6.4 The architecture of a multilayer perceptron

The architecture shown in figure 6.4 is a nonlinear mapping between the inputs, the variable of the dataset, and the outputs. During training of the network, the inputs $x_1, x_2 \ldots x_n$, are fed to the input layer. The nodes of the input layer does not process the information, but distribute them to the next layer. The nodes at the hidden layer, i, first generate a weighted sum of inputs $\sum x_i w_i$which is then passed through an activation function that determines the activation level of the processing neuron (Vaughn, 1996). This information is then subsequently passed on to the nodes of the next hidden layer, where the information is processed again in the same manner as before. Thus each hidden layer (if more than one) passes on its outputs to the next layer until the output $O_1, O_2 \dots O_n$ is obtained.

The network is designed by a popular training algorithm, namely, the back-propagation algorithm (Chen and Jain, 1994, Leung and Haykin, 1991, Gardner and Dorling, 1998, Bishop, 1995). The general outline of the algorithm is shown in table 6.7.

Step	Description
1	Initialise network weights
2	Present inputs from training data to the network
3	Propagate the inputs through the network to calculate actual output
4	Calculate an error signal by comparing the actual output $O_i$to the target output T_i
5	Propagate error signal back through the network
6	Adjust weights to minimise overall error
7	Repeat steps 2-7 with next input vector, until overall error is satisfactorily small

Table 6.7: Back-propagation algorithm

The algorithm minimises the error between outputs of the network and the target output. The cost function used is a key criterion, namely, the Mean Squared Error (MSE) function (Nitta, 1997):

$$J(x, o) = \sum_{i=1}^{N} \frac{1}{2} (O_i - T_i)^2 \tag{6.7}$$

where:

N is the number of neurons in the output layer

O_i are the outputs of the network

T_i are the target outputs to be reached

The weights are then modified iteratively according to the gradient of the cost function.

The complexity associated with the training of an MLP is dependent on the number of nodes and the corresponding number of weights. The computational complexity of back propagation is $O(W)$ and the numerical gradient computation is of the order $O(W^2)$ (where W is the number of weights in the network). Thus, the larger the network, the greater the complexity of the network. For example, in this case with 60 variables, two hidden layers with 10 nodes in each layer and two output nodes, the complexity is of the order $O((60 * 10 + 10 * 10 + 10 * 2)^2) = O(720^2)$. Thus the training algorithm's complexity increases with the topology of the network, and the degree of approximation required.

6.5.1 Multilayer perceptron classification performance

Table 6.8 presents the multilayer perceptron classification performance of the original and hybrid datasets. It can be seen that the performance of the hybrid data is better than the performance of the original dataset. The training set result shows NPV and SEN to be 100% while PPV, SPEC and ACC are 99%, 95% and 99% respectively. The hybrid performance also outperforms the naïve Bayes result shown in table 5.1. This could be due to the fact that MLP is able to adjust the weights of algorithm using the error signal so that a minimum error is obtained. Although MLP produces outstanding classification performance, a key and perhaps the most important disadvantage, from a clinical perspective, for the Hull LifeLab clinical dataset is that the algorithm is simply a black box that does not allow information about the data to be learnt in great detail.

Original dataset									
Training set					***Cross validation***				
	Predict		***Evaluation measure***	***%***		***Predict***		***Evaluation measure***	***%***
True	Alive	Dead	PPV	97	***True***	Alive	Dead	PPV	81
Alive	1455TP	4FN	NPV	99		1235TP	224FN	NPV	47
Dead	30FP	455TN	SEN	100		284FP	201TN	SEN	85
			SPEC	94				SPEC	41
			ACC	98				ACC	74
SVM and EM hybrid dataset									
Training set					***Cross validation***				
	Predict		***Evaluation measure***	***%***		***Predict***		***Evaluation measure***	***%***
True	Alive	Dead	PPV	99	***True***	Alive	Dead	PPV	96
	1459TP	0FN	NPV	100		1421TP	38FN	NPV	92
	22FP	463TN	SEN	100		62FP	423TN	SEN	97
			SPEC	95				SPEC	87
			ACC	99				ACC	95

Table 6.8: Multilayer perceptron classification performance of the original and hybrid dataset

6.6 Analysis of results

The results in this chapter have shown that the classification algorithms obtained different predictive accuracy with the original data and the hybrid data. However, for Bayes classifier based on KDE shown in table 6.1 and beta based Bayes classifier shown in table 6.2, the results using the hybrid data are very similar. This could be due to the fact that some of the assumptions required for Bayes classifier are shared by the other classifiers. For example, KDE retains the independence assumption but eschews the Gaussian distribution assumption. Moreover, the MLEs of the parameters that index the beta distribution can be severely biased. As a result, there is no appreciable difference and not much of an improvement in performance is shown when compared to naïve Bayes performance on the original data shown in table 4.3 and SVM and EM hybrid data in table 5.1.

The advantage posed by Beta and KDE Bayes classifiers is that they take into account skews and the shape of the distribution, which can be controlled to smooth the data distribution. In cases where one is not certain about skews, these classifiers would be recommended as a first step, rather than naïve Bayes. However, the performance presented by both methods does not show much of an improvement and in cases where the data is noisy and contains a lot of issues, this will be reflected in the distribution. On the contrary, it is possible to obtain some idea about the data by altering the bandwidth, h value. For example, if $h \to \infty$, a Gaussian distribution will be obtained and if h is small, too many kernels will be obtained for each data point. Parameters of the data also play a part in changing the shape of the distribution, such as the mean and standard deviations. For example, Table 6.9 presents glucose, ferritin, MR-proANP and PEFR variable mean values for the original data, SVM and EM hybrid data and the SVM and EM hybrid data using the α and β parameters to compute the mean beta distribution. Figure 6.5 shows the corresponding graphical representation of the means.

Overall data				
	Glucose	***Ferritin***	***MR-proANP***	***PEFR***
Original data	6.698	115.386	225.647	218.533
SVM & EM	6.974	136.709	301.583	353.375
Beta	7.255	146.177	364.564	618.593
Dead class				
	Glucose	***Ferritin***	***MR-proANP***	***PEFR***
Original data	7.014	119.488	316.841	180.050
EM	8.333	221.770	621.213	727.316
Beta	8.491	185.644	722.823	806.223
Alive class				
	Glucose	***Ferritin***	***MR-proANP***	***PEFR***
Original data	6.588	114.042	195.332	230.773
SVM	6.523	108.433	195.332	229.070
Beta	6.647	127.174	194.77	228.360

Table 6.9: Mean values of the four variables for the original data, SVM & EM data and beta distirbution

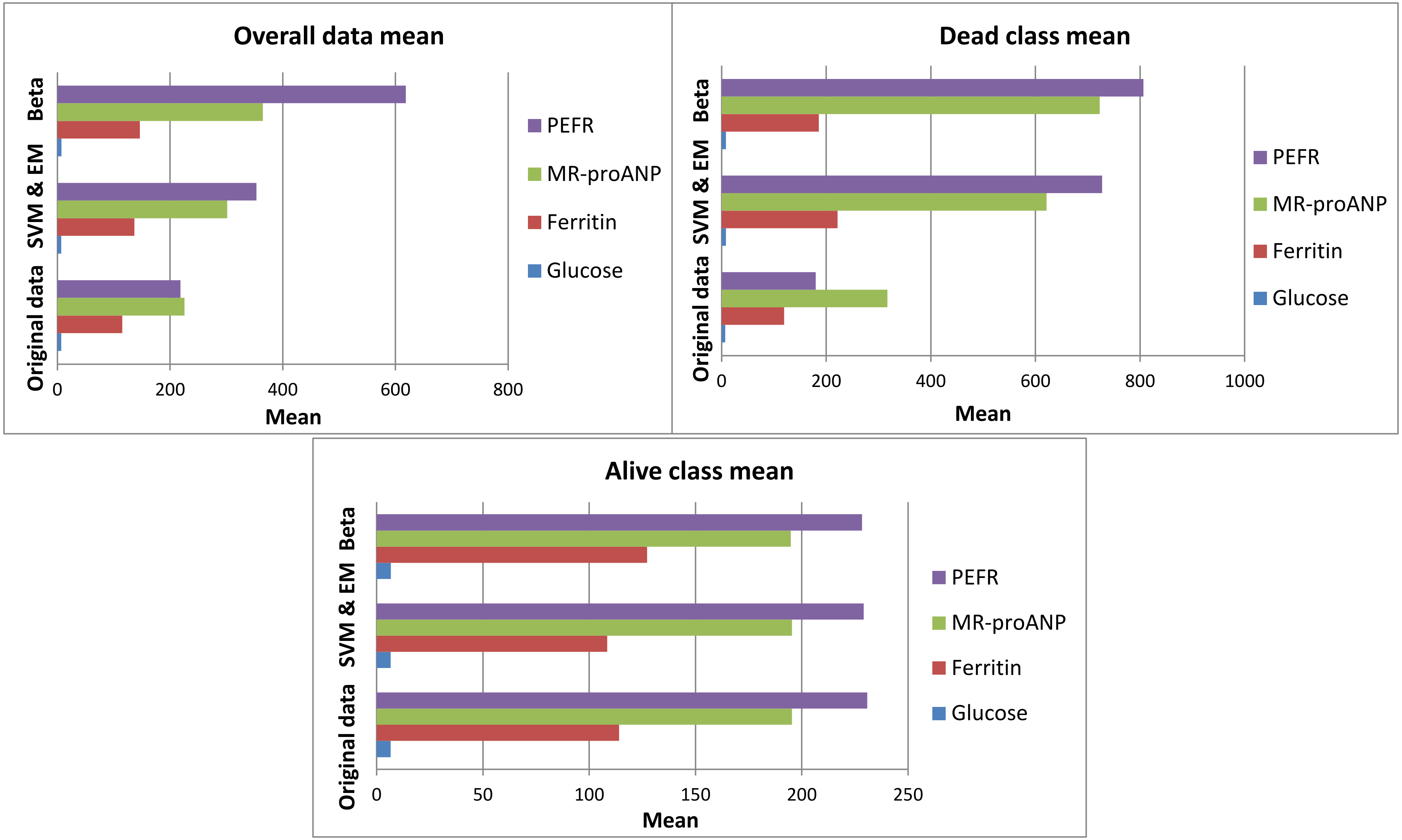


Figure 6.5 Glucose, ferritin, MR-proANP and PEFR mean values of the original data, SVM & EM data and beta SVM & EM dataset

The four variables were chosen due to their being the variables that presented poor distribution as shown in appendix I. It can be seen in table 6.9 that the beta distribution shows the most increased mean in all four variables when compared to the original data and SVM & EM imputed data. This is also reflected in the graph shown in figure 6.5, where there is a steep growth from the original data means to the beta distribution means especially for the PEFR variable in the overall data mean and dead class mean, whereas the glucose mean values are consistent in all three data groups.

The dead class shows a different pattern of mean values. For example ferritin presents an increase in the SVM and EM data and then a slight drop for the beta mean, where their mean values are 221.770 and 185.644 respectively. The PEFR variable shows a mean difference of approximately 500 between the original data and the SVM and EM imputed data, which illustrates the sudden increase in the dead class graph. This is also reflected in the MR-proANP variable which presents a difference of approximately 300.

On the contrary, the alive class presents similar mean values for each variable. For example the PEFR means for the original data, SVM & EM data and beta distribution are 230.773, 229.070 and 228.360 respectively. The similarity is also present in glucose, ferritin, MR-proANP and PEFR variables and reflected in figure 6.5, where a consistency is shown.

The increase in the mean values presented by the SVM and EM algorithm and beta distribution indicates a shift in the means and therefore the distribution fits a more appropriate curve for the dataset. In contrast, the low mean values presented by the original data indicates that the means of the variables are located on the left side of the distribution as shown in appendix I. The increased mean by SVM and EM hybrid data, especially for the overall data and dead class is due to the maximisation step of the EM algorithm and the further increased mean by beta distribution is because of the MLEs of the parameter estimates for the beta distribution. The results in table 6.9 further indicate that the original dead class is causing the most problem in the data due to the large mean difference between the original and SVM & EM hybrid dataset. The reason for this is because the dead class has less number of missing data to be imputed. Table 6.10 present the number of missing data present in the different variables for the alive and dead class.

No. of missing data			
Alive class			
Glucose	Ferritin	MR-proANP	PEFR
195	291	0	85
Dead class			
Glucose	Ferritin	MR-proANP	PEFR
46	102	0	48

Table 6.10: Number of missing data present in the variables of the alive and dead class

It can be seen that the number of missing data present in the alive class is greater than those presented by the dead class. This suggests that imputing the missing data generated consistent mean values. In contrast, the dead class shows less number of missing data therefore less imputed data. This indicates that the observed data variables are causing a huge variation in the dead mean values, especially in the PEFR variable. This also provides further evidence and answers the question posed in chapter 5 as to what variable is causing misclassification of the dead class and contributing the most?; the varied measurements of the PEFR variable as mentioned in chapter 5 thus affects the mean values presented in this chapter.

It can be seen that the decision tree and MLP performance shown in table 6.6 and table 6.8 respectively show similar performance with the hybrid dataset. The performance also improved when compared to the KDE and beta performance and fewer misclassifications (*FP* and *FN*) are presented when compared to the SVM and EM hybrid results in table 5.1. Neither MLP nor decision tree based classifiers use the distribution of the data explicitly, although for the MLP the data is scaled, whereas for the decision tree no such requirement is present. The complexity in both can be controlled using pruning strategies. However in each case this strategy is different. For example, decision tree uses pruning strategies to adjust the size of the tree based on the desired accuracy and hence entropy to specifically measure the amount of information needed to identify the class of a sample. However, a cautionary note is required in that variation present in the data can generate a different decision tree (especially when the variables are close to each other in value). This is the case in some of the variables in the Hull LifeLab dataset such as *'Age'* and *'Pulse bpm'*. In contrast, MLP uses the back propagation algorithm to minimise the cost function between the actual outputs and target output.

Decision tree prefers balanced classes and over fits the training data. However balancing methods have generalisation problems and for clinical application it is an issue. Moreover, pruning and tuning the pruning procedures are required to avoid overfitting issues during each implementation. In contrast, decision tree is found useful for understanding the structure of the decision making process which makes the usefulness of MLP much less clear for data interpretation. However, both algorithms lack the ability to explore the data in great detail. Although good results are produced, an explanation as to why misclassification has occurred and information about the nature of the data is not offered.

The *FP* records of the SVM and EM hybrid imputed data generated by the four classifiers are shown in appendix VIII. The records were identified to determine whether the same records are present in the 79 *FP* records of the naïve Bayes performance shown in table 5.1 of chapter 5. KDE (table 6.1), beta (table 6.2), decision tree (table 6.6) and MLP (table 6.8) show 80, 81, 26 and 22 *FP* records respectively, whereas 79 *FP* records of the KDE and beta and all 26 and 22 records were present in the 79 *FP* records of the naïve Bayes classification performance. This indicates that the data of the records is poorly collected.

6.7 Summary

KDE and beta Bayes based classifiers are both data dependent and thus good predictive performance is based on the data properties, characteristics and tuning of the data distribution. The complexity of the classifiers is an important factor and this chapter has provided an insight into how they can influence performance. KDE requires a significant amount of memory to store the probability estimates but does not improve results when compared to the naïve Bayes results in chapters 4 and 5. All the classifier algorithms are fairly complex, with naïve Bayes having the lowest complexity (both in terms of time and space complexity)

Unlike the classifier algorithms discussed above, naïve Bayes classifier is simple, easy to train and allows one to explore the outcome of the prediction using the mean and standard deviation parameters obtained during the design of experiments. The assumption made about the type of the distribution density function introduces prior information into the classifier’s design process. In cases where this additional information is correct, it can reduce the classification error, otherwise the classification error will be large. Although this assumption is not met by most variables of the Hull LifeLab data, nevertheless the data performed surprisingly well; perhaps not to the standards of a computer scientist but good enough to provide an insight regarding predictive accuracy. The independence assumption has allowed us to benefit from understanding the nature of the clinical data in that it permits parameters for each variable to be learned separately. The algorithm was also able to deal with the challenges posed by the dataset. For example, unlike decision tree, which requires balanced classes and ignores missing data, Bayes classifier is not sensitive to these challenges, including irrelevant attributes. In general the advantages outweigh the disadvantages of Bayes for clinical data. However, implementing other statistical techniques to guide towards the problematic attributes and records has been beneficial for the considered data and in understanding the underlying properties.

CHAPTER 7 CONCLUSION AND FUTURE RESEARCH

7.1 Introduction

The purpose of this research was to contribute towards tackling the challenges posed by the real life clinical datasets when used for mining, while simultaneously improving classification performance of the data. There are four challenges posed by the real life heart failure clinical dataset, such as missing data, high dimensionality, class imbalance and non-normal distribution. Having investigated data mining frameworks, e.g. CRISP-DM and SEMMA, a workflow, Clinical Data Mining Workflow, was developed (see chapter 2). This workflow provided an outline within which classification, and data could be analysed. Thus this thesis follows the three distinctive stages presented in the workflow, namely, the descriptive, predictive and prescriptive stages. The stages are interlinked and each plays a part in understanding the properties of the data by providing clear answers as to why misclassification occurred. The findings in this thesis do not present an outlier problem but a problem in the manner clinical data was gathered by clinicians and tools to identify the problems in the data. This chapter concludes the thesis with a summary of the main contributions of the thesis and gives suggestions for future research.

7.2 Contributions of the research

This thesis explores the underlying challenges through the application of Bayes methodologies and analyses the manner in which the challenges affect the performance of the algorithm. Although methods based on Bayes are robust to the challenges, i.e. missing data, the performances are poor. The workflow allowed for an analysis of the reasons for this poor performance, at the same time Bayes utilises a set of tools which further enhances the analysis. This allows for a deeper understanding of the nature of the data. A number of tests were carried out to understand the importance of missing values in the dataset. This resulted in the investigation of the use of a hybrid imputation dataset, which improved classification performance. The hybrid imputation dataset also allowed for a more detailed exploration of the posterior probabilities associated with the classes. This investigation resulted in an analysis of missing data and records. It was found that the records with missing data were classified correctly and that it was records with no imputation that were causing the problem. A simple test of discarding the full records while leaving records with imputed data resulted in substantially improving classification performance. However, this did not locate where the real problem occurred, as a result the infinite-norm approach was used to determine the problem variable (see chapter 5). This section outlines the objectives of the research and the corresponding chapter (s) that present the solution of the objectives.

Objective 1: To identify challenges associated with a real life clinical dataset as applied to clinical practice and the most appropriate set of algorithms for the dataset.

In chapter 1 the four main challenges posed by real life clinical datasets, such as missing data, class imbalance, high dimensionality and non-normal distribution are identified. This chapter discusses how these challenges occur, such as during the collection of EHRs in clinical practice and how data mining processes can extract meaningful information for the purpose of clinical decision making and thus improve quality of life. Bayes has been suggested as the appropriate data mining algorithm to classify the Hull LifeLab data and develop prediction algorithms.

Objective 2: Investigate and evaluate methods for handling missing data

In chapter 2 the proposed Clinical Data Mining Workflow (CDMW) specifically tailored to real life clinical datasets that builds the flow of the thesis was presented. The three stages of the workflow are descriptive, predictive and prescriptive stages, which are discussed in chapters 3-5. Chapter 3 introduces the descriptive stage which involves two steps: 1) data exploration and 2) data preparation. The latter step involves the application of seven missing data imputation methods, namely, Most Common value Imputation (MCI), Concept Most Common value imputation (CMCI), Expectation Maximization Imputation (EMI), *k*-Nearest Neighbour Imputation (KNNI), *k*-Means clustering Imputation (KMI), Fuzzy *k*-Means clustering Imputation (FKMI) and Support Vector Machine Imputation (SVMI). These imputation methods were implemented to understand their mechanisms and determine their effect on the statistical measures and distribution of the data.

Objective 3: Investigate the relationship between methods for missing data with a view to develop prediction models and improve classification performance.

In chapter 4, the predictive stage was presented, which includes the naïve Bayes classification performance of the seven imputed datasets. An augmented naïve Bayes, known as TAN Bayes was also implemented to improve the naïve Bayes classification accuracy. The performance of the imputed datasets was compared, their tasks were taken into consideration and the time, space and structural complexities of the Bayes methods were also considered to determine their impact on classification accuracy. The results showed TAN to outperform naïve Bayes; this is because TAN weakens the strong independence assumption of naïve Bayes. Despite the high computational complexity posed by the TAN algorithms, the algorithm improved classification accuracy. This indicates that the high complexity is required to learn the TAN algorithm. On the other hand, naïve Bayes has allowed the variables of the data to be learnt independently and considered the accuracy of the classes separately.

Objective 4: Develop an integrated solution using Bayes methods for missing data.

In chapter 5 the prescriptive stage of the workflow was discussed. This chapter assesses the performance of the proposed SVM and EM hybrid imputed dataset by exploring the posterior probabilities of the 79 *FP* records. The dataset was developed based on the performance of the different classes using different imputation schemes and then combining them to form one dataset. This approach improved classification accuracy. However, to understand why misclassification was generated, an investigation of the class posterior probabilities of both the *FP* and *TN* records was conducted to determine the probabilities associated with prediction. The records of dead patients who were incorrectly classified as alive (that is FP) were investigated to see if imputing missing data was the cause for the incorrect classification. It was found that these records had no missing data or data was not imputed.

The proposed SVM and EM hybrid imputed dataset is *an extension of the CMC* imputation method discussed and applied in chapter 3. CMC imputes missing data using the in class mean of the dead and alive class respectively, while the hybrid imputed data imputes the missing data in the two classes using two different imputation schemes, i.e. SVM for the alive class and EM for the dead class. This proposed method has allowed the properties of the classes to be understood separately, and thus improved classification accuracy. For example in chapter 6, table 6.9 and figure 6.5 shows a sudden mean increase with the SVM and EM data and beta distribution for the dead class. This is due to a maximisation approach applied by the EM algorithm and during the α and β parameter estimates in the beta distribution. In the EM algorithm, the maximisation step maximises the expectation of the complete data log likelihood while beta distribution computes the means based on the MLE parameters α and β. As a result, using the MLE algorithm on the already maximised expected value of the data further improved the beta distribution mean values.

Objective 5: Investigate ways of improving classifier to enhance performance for better clinical prediction models and decision support systems.

In chapter 6 other classification methods are discussed. The considered methods are, Bayes classifier based on KDE, beta based Bayes classifier, decision tree and MLP. MLP showed the most improved classification performance with reduced number of records in the *FPs* and *FNs* of the training set (table 6.8). However, the algorithm does not allow misclassification of the data to be understood in great detail. The beta distribution bases Bayes classifier also improved classification due to the nested MLE process which allowed proposed SVM and EM hybrid imputed data where the EM was used on the dead class. The application of the EM algorithm on the dead class maximises the mean beta distribution

Overall the choice of the classifier for generating predictive model is a complex task, however it is required. The selection of a correct data mining algorithm depends on not only the goal of an application, but also on the dataset. Although in some cases the data is not compatible with the assumptions of the considered method, it allows the data to be explicitly explored to achieve the set aims and objectives.

In the light of the findings presented in this thesis and the conclusion drawn, contributions to the area of investigating challenges of a real life clinical data through the application of data mining methods are as follows:

- The application of the proposed Clinical Data Mining Workflow (CDMW) to assist in the data analysis of real life clinical data
- Implementation of data mining methods such as missing data imputation methods and Bayes classifier to determine the effect of the challenges on classification performance
- Implementation of the proposed methods for classification, i.e. hybrid imputed data

- Determination of the cause of misclassification, exploring the posterior probabilities of misclassified records.

7.3 Future research

Clinical data also present the challenge of high dimensionality. Although Bayes is not sensitive to the issue, it should be considered for future research, particularly in clinical datasets where irrelevant data exist. As a result, recommendations for future research are as follows:

1) Investigate feature selection and feature extraction methods with a view to develop prediction models and decision support systems.
    - The application of feature selection methods such as Correlation Feature Selection (CFS) to select relevant features
    - Feature extraction such as PCA to reduce dimensions of a clinical data.
2) Investigate the relationships of the seven imputation methods on the different classes using other non-Bayes classifier
3) Recursive Bayes classifier can be applied to estimate the PDF of the parameters recursively, each time a new data is introduced, new set of parameters are estimated to replace the old ones. This is particularly useful for tele-monitoring.

BIBLIOGRAPHY

Acuña, E. & Rodriguez, C. 2004. The Treatment of Missing Values and its Effect on Classifier Accuracy. *In:* Banks, D., Mcmorris, F., Arabie, P. & Gaul, W. (eds.) *Classification, Clustering, and Data Mining Applications.* Springer Berlin Heidelberg.

Arch-Int, N. & Arch-Int, S. 2011. Semantic Information Integration for Electronic Patient Records Using Ontology and Web Services Model. Information Science and Applications (ICISA), 2011 International Conference on, 26-29 April 2011, pp. 1-7.

Autio, L., Juhola, M. & Laurikkala, J. 2007. On the neural network classification of medical data and an endeavour to balance non-uniform data sets with artificial data extension. *Computers in biology and medicine,* 37**,** pp. 388-397.

Azevedo, A. & Santos, F., Manuel 2008. KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining***,** pp. 182-185.

Balakrishnan, K. & David, M. J. 2010. Significance of Classification Techniques in Prediction of Learning Disabilities. *arXiv preprint arXiv:1011.0628*.

Balakrishnan, S., Narayanaswamy, R., Savarimuthu, N. & Samikannu, R. 2008. SVM ranking with backward search for feature selection in type II diabetes databases. Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on, 12-15 Oct. 2008. 2628-2633.

Batista, G. E. A. P. A. & Monard, M. C. 2003. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence,* 17**,** pp. 519-533.

Batra, S., Parashar, H. J., Sachdeva, S. & Mehndiratta, P. 2013. Applying data mining techniques to standardized electronic health records for decision support. Contemporary Computing (IC3), 2013 Sixth International Conference on, 8-10 Aug. pp. 510-515.

Bellazzi, R. & Zupan, B. 2008. Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics,* 77**,** pp. 81-97.

Bhargava, N., Sharma, G., Bhargava, R. & Mathuria, M. 2013. Decision Tree Analysis on J48 Algorithm for Data Mining. *International journal of advance research in computer science and software engineering,* 3.

Bishnu, P. S. & Bhattacherjee, V. 2012. A dimension reduction technique for K-Means clustering algorithm. Recent Advances in Information Technology (RAIT), 2012 1st International Conference on, 15-17 March pp. 531-535.

Bishop, C. M. 1995. *Neural networks for pattern recognition*, Oxford university press.

Bland, J. M. & Altman, D. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet,* 327**,** pp. 307-310.

Blomberg, L. C. & Ruiz, D. D. A. 2013. Evaluating the influence of missing data on classification algorithms in data mining applications. *SBSI 2013: Simpósio Brasileiro de Sistemas de Informação*.

Blumenthal, D. & Tavenner, M. 2010. The “Meaningful Use” Regulation for Electronic Health Records. *New England Journal of Medicine,* 363**,** pp. 501-504.

Bohacik, J., Kambhampati, C., Davis, D. N. & Cleland, J., G. F. 2014. Prediction of mortality rates in heart failure patients with data mining methods. *Annales UMCS, Informatica,* 13**,** pp. 7-16.

Bohacik, J., Kambhampati, C., Davis, D. N. & Cleland, J. F. G. 2013a. Analysis of Fuzzy Decision Trees on Expert Fuzzified Heart Failure Data. Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, 13-16 Oct. pp. 350-355.

Bohacik, J., Kambhampati, C., Davis, D. N. & Cleland, J. G. 2013b. Prediction of mortality rates in heart failure patients with data mining methods. *In:*Annales UMCS, Informatica, pp. 7-16.

Boonstra, A. & Broekhuis, M. 2010. Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. *BMC health services research,* 10**,** 231.

Bosnjak, Z., Grljevic, O. & Bosnjak, S. 2009.CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data. Applied Computational Intelligence and Informatics, 2009. SACI '09. 5th International Symposium on, 28-29 May, pp. 509-514.

Bouhamed, H., Masmoudi, A., Lecroq, T. & Rebaï, A. 2012. A new approach for Bayesian classifier learning structure via K2 Algorithm. *Emerging Intelligent Computing Technology and Applications.* Springer.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. 1984. Classification and regression trees. Wadsworth. *Belmont, CA.*

Burton, L. C., Anderson, G. F. & Kues, I. W. 2004. Using Electronic Health Records to Help Coordinate Care. *Milbank Quarterly,* 82**,** pp. 457-481.

Cerrito, P. B. 2006. *Introduction to data mining using SAS Enterprise Miner,* SAS Institute.

Chamorro, Á., Obach, V., Cervera, Á., Revilla, M., Deulofeu, R. & Aponte, J. H. 2002. Prognostic significance of uric acid serum concentration in patients with acute ischemic stroke. *Stroke,* 33**,** pp. 1048-1052.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. 2000. Crisp-Dm 1.0: Step-by-step data mining guide. SPSS. Inc.

Chen, D. S. & Jain, R. C. 1994. A robust backpropagation learning algorithm for function approximation. *Neural Networks, IEEE Transactions on,* 5**,** pp. 467-479.

Cheng, J. & Greiner, R. 1999. Comparing Bayesian network classifiers. *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence.* Stockholm, Sweden: Morgan Kaufmann Publishers Inc.

Chow, C. & Liu, C. 1968. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on,* 14**,** pp. 462-467.

Cismondi, F., Fialho, A. S., Vieira, S. M., Reti, S. R., Sousa, J. M. C. & Finkelstein, S. N. 2013. Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine,* 58**,** pp. 63-72.

Clarke, R., Ressom, H. W., Wang, A., Xuan, J., Liu, M. C., Gehan, E. A. & Wang, Y. 2008. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer,* 8**,** pp. 37-49.

Cleland, J. G. F., Gemmell, I., Khand, A. & Boddy, A. 1999. Is the prognosis of heart failure improving? *European Journal of Heart Failure,* 1**,** pp. 229-241.

Costa, E., Lorena, A., Carvalho, A. & Freitas, A. 2007. A review of performance evaluation measures for hierarchical classifiers. Evaluation Methods for Machine Learning II: papers from the AAAI-2007 Workshop, pp. 1-6.

Counsell, N., Cortina-borja, M., Lehtonen, A. & Stein, A. 2011. Modelling psychiatric measures using Skew-Normal distributions. *European Psychiatry,* 26**,** pp. 112-114.

Cowie, M., Wood, D., Coats, A., Thompson, S., Suresh, V., Poole-Wilson, P. & Sutton, G. 2000. Survival of patients with a new diagnosis of heart failure: a population based study. *Heart,* 83**,** 505-510.

Delen, D. & Demirkan, H. 2013. Data, information and analytics as services. *Decision Support Systems,* 55**,** pp. 359-363.

Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)***,** pp. 1-38.

Drazin, S. & Montag, M. 2012. Decision tree analysis using WEKA. *Machine Learning-Project II, University of Miami***,** pp. 1-3.

Dumont, M., Marée, R., Wehenkel, L. & Geurts, P. 2009. Fast multi-class image annotation with random subwindows and multiple output randomized trees. International Conference on Computer Vision Theory and Applications (VISAPP).

Dziura, J. D., Post, L. A., Zhao, Q., Fu, Z. & Peduzzi, P. 2013. Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale Journal of Biology and Medicine,* 86**,** 343-58.

EL ayadi, M. & Plataniotis, K. N. 2010. Improving classification performance of linear feature extraction algorithms. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 14-19 March, pp. 2166-2169.

Elmore, J. G., Barton, M. B., Moceri, V. M., Polk, S., Arena, P. J. & Fletcher, S. W. 1998. Ten-Year Risk of False Positive Screening Mammograms and Clinical Breast Examinations. *New England Journal of Medicine,* 338**,** pp. 1089-1096.

Eriksen, B. O., Hoff, K. R. & Solberg, S. 2003. Prediction of acute renal failure after cardiac surgery: retrospective cross-validation of a clinical algorithm. *Nephrology Dialysis Transplantation,* 18**,** pp. 77-81.

Everett, A. D., Ringel, R., Rhodes, J. F., Doyle, T. P., Owada, C. Y., Holzer, R. J., Cheatham, J. P., Ringewald, J., Bandisode, V., Chen, Y.-L. & Lim, D. S. 2006. Development of the MAGIC Congenital Heart Disease Catheterization Database for Interventional Outcome Studies. *Journal of Interventional Cardiology,* 19**,** pp. 173-177.

Farooq, K., Peipei, Y., Hussain, A., Kaizhu, H., Macrae, C., Eckl, C. & Slack, W. Efficient clinical decision making by learning from missing clinical data.

Computational Intelligence in Healthcare and e-health (CICARE), 2013 IEEE Symposium on, 16-19 April, pp. 27-33.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters,* 27**,** pp. 861-874.

Fleishman, A. 1978. A method for simulating non-normal distributions. *Psychometrika,* 43**,** pp. 521-532.

Foody, G. M. 2002. Status of land cover classification accuracy assessment. *Remote sensing of environment,* 80**,** pp. 185-201.

Ford, J. B., Roberts, C. L., Algert, C. S., Bowen, J. R., Bajuk, B. & Henderson-Smart, D. J. 2007. Using hospital discharge data for determining neonatal morbidity and mortality: a validation study. *BMC health services research,* 7**,** 188.

Frank, E., Hall, M. & Pfahringer, B. 2002. Locally weighted naive bayes. Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence, 2002. Morgan Kaufmann Publishers Inc., pp. 249-256.

Friedman, N., Geiger, D. & Goldszmidt, M. 1997. Bayesian network classifiers. *Machine learning,* 29**,** pp. 131-163.

Friedman, N. & Goldszmidt, M. 1996. Building classifiers using Bayesian networks. Proceedings of the national conference on artificial intelligence, 1996. pp. 1277-1284.

Gajanayake, R., Sahama, T. & Iannella, R. 2013. The role of perceived usefulness and attitude on electronic health record acceptance. e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on, 9-12 Oct, pp. 388-393.

García, S., Luengo, J. & Herrera, F. 2015. *Data preprocessing in data mining*, Springer.

García, V., Sánchez, J. S., Mollineda, R. A., Alejo, R. & Sotoca, J. M. 2007. The class imbalance problem in pattern classification and learning.

Gardner, M. W. & Dorling, S. R. 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment,* 32**,** pp. 2627-2636.

Gnanadesikan, R., Pinkham, R. & Hughes, L. P. 1967. Maximum likelihood estimation of the parameters of the beta distribution from smallest order statistics. *Technometrics,* 9**,** pp. 607-620.

Gold, M. S. & Bentler, P. M. 2000. Treatments of Missing Data: A Monte Carlo Comparison of RBHDI, Iterative Stochastic Regression Imputation, and Expectation-Maximization. *Structural Equation Modeling: A Multidisciplinary Journal,* 7**,** pp. 319-355.

Grzymala-Busse, J. & Hu, M. 2001. A Comparison of Several Approaches to Missing Attribute Values in Data Mining. *In:* ZIARKO, W. & YAO, Y. (eds.) *Rough Sets and Current Trends in Computing.* Springer Berlin Heidelberg.

Gu, Q., Zhu, L. & Cai, Z. 2009. Evaluation Measures of the Classification Performance of Imbalanced Data Sets. *In:* Cai, Z., Li, Z., Kang, Z. & Liu, Y. (eds.) *Computational Intelligence and Intelligent Systems.* Springer Berlin Heidelberg.

Guidoum, A. C. 2014. Kernel Estimator and Bandwidth Selection for Density and its Derivatives.

Gunn, S. R. 1998. Support vector machines for classification and regression. *ISIS technical report,* 14.

Guo, Q. 2010. An Effective Algorithm for Improving the Performance of Naive Bayes for Text Classification. Computer Research and Development, 2010 Second International Conference on, 7-10 May, pp. 699-701.

Gupta, A. K. & Nadarajah, S. 2004. *Handbook of beta distribution and its applications,* CRC Press.

Gupta, M. R. & Chen, Y. 2011. Theory and Use of the EM Algorithm. *Found. Trends Signal Process.,* 4**,** pp. 223-296.

Guyon, I. & Elisseeff, A. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research,* 3**,** pp.1157-1182.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter,* 11**,** pp. 10-18.

Han, J. & Kamber, M. 2006. *Data Mining: Concepts and Techniques,* Massachusetts, Morgan kaufmann.

Hand, D. 1992. Statistical methods in diagnosis. *Statistical Methods in Medical Research,* 1**,** pp. 49-67.

Hand, D. J., Mannila, H. & Smyth, P. 2001. *Principles of data mining,* London, MIT press.

Hand, D. J. & Yu, K. 2001. Idiot's Bayes—Not So Stupid After All? *International Statistical Review,* 69**,** pp. 385-398.

Hani, A. F. M., Nugroho, H. A. & Nugroho, H. 2010. Gaussian Bayes classifier for medical diagnosis and grading: Application to diabetic retinopathy. Biomedical Engineering and Sciences (IECBES), 2010 IEEE EMBS Conference on, Nov. 30-Dec. 2, pp. 52-56.

He, Y. 2010. Missing Data Analysis Using Multiple Imputation: Getting to the Heart of the Matter. *Circulation: Cardiovascular Quality and Outcomes,* 3**,** pp. 98-105.

Hess, K. R., Varadhachary, G. R., Taylor, S. H., Wei, W., Raber, M. N., Lenzi, R. & Abbruzzese, J. L. 2006. Metastatic patterns in adenocarcinoma. *Cancer,* 106**,** pp.1624-1633.

Ho, K. K., Anderson, K. M., Kannel, W. B., Grossman, W. & Levy, D. 1993. Survival after the onset of congestive heart failure in Framingham Heart Study subjects. *Circulation,* 88**,** pp. 107-15.

Honghai, F., Guoshun, C., Cheng, Y., Bingru, Y. & Yumei, C. 2005. A SVM regression based approach to filling in missing values. *In:* Knowledge-Based Intelligent Information and Engineering Systems, 2005. Springer, pp. 581-587.

Houle, M., Kriegel, H.-P., Kröger, P., Schubert, E. & Zimek, A. 2010. Can Shared-Neighbor Distances Defeat the Curse of Dimensionality? *In:* Gertz, M. & Ludäscher, B. (eds.) *Scientific and Statistical Database Management.* Springer Berlin Heidelberg.

Huang, W., Mcgregor, C. & James, A. 2014. A comprehensive framework design for continuous quality improvement within the neonatal intensive care unit: Integration of the SPOE, CRISP-DM and PaJMa models. *In:* IEEE-EMBS

International Conference on Biomedical and Health Informatics (BHI), 2014, 1-4 June, pp. 289-292.

Huifang, Z. & Ding, P. 2010. A knowledge discovery and data mining process model in E-marketing. Intelligent Control and Automation (WCICA), 2010 8th World Congress on, 7-9 July, pp. 3960-3964.

Jacobs, L., Thijs, L., Jin, Y., Zannad, F., Mebazaa, A., Rouet, P., Pinet, F., Bauters, C., Pieske, B., Tomaschitz, A., Mamas, M., Diez, J., Mcdonald, K., Cleland, J. G. F., Rocca, H.-P. B.-L., Heymans, S., Latini, R., Masson, S., Sever, P., Delles, C., Pocock, S., Collier, T., Kuznetsova, T. & Staessen, J. A. 2014. Heart 'omics' in AGEing (HOMAGE): design, research objectives and characteristics of the common database. *Journal of Biomedical Research,* 28**,** pp. 349-359.

Jiang, L., Cai, Z., Wang, D. & Zhang, H. 2012. Improving Tree augmented Naive Bayes for class probability estimation. *Knowledge-Based Systems,* 26**,** pp. 239-245.

Jiang, L., Zhang, H., Cai, Z. & Su, J. 2005. Learning tree augmented naive bayes for ranking. Database Systems for Advanced Applications, Springer, pp. 688-698.

John, G. H. & Langley, P. 1995. Estimating continuous distributions in Bayesian classifiers. Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, 1995. Morgan Kaufmann Publishers Inc., pp. 338-345.

Jong, P., Ahn, S. A., Bangdiwala, S. I. & Rousseau, M. F. 2012. Abstract 11502: Modifying the Seattle Heart Failure Model Improves Survival Prediction in Heart Failure Patients. *Circulation,* 126**,** A11502.

Jonsson, P. & Wohlin, C. 2004. An evaluation of k-nearest neighbour imputation using Likert data. Software Metrics, 2004. Proceedings. 10th International Symposium on, 14-16 Sept, pp.108-118.

Kaisler, S., Armour, F., Espinosa, J. A. & Money, W. 2013. Big Data: Issues and Challenges Moving Forward. *In:* System Sciences (HICSS), 2013 46th Hawaii International Conference on, 7-10 Jan. pp. 995-1004.

Kanj, S., Abdallah, F. & Denoux, T. 2012. Purifying training data to improve performance of multi-label classification algorithms. *In:* Information Fusion (FUSION), 2012 15th International Conference on, 9-12 July, pp. 1784-1791.

Kantardzic, M. 2011. *Data mining: concepts, models, methods, and algorithms,* New Jersey, John Wiley & Sons.

Kapoor, B. & Kleinbart, M. 2012. Building an Integrated Patient Information System for a Healthcare Network. *Journal of Cases on Information Technology (JCIT),* 14**,** pp. 27-41.

Karbing, D., Kjaergaard, S., Smith, B., Espersen, K., Allerod, C., Andreassen, S. & Rees, S. 2007. Variation in the PaO2/FiO2 ratio with FiO2: mathematical and experimental description, and clinical relevance. *Critical Care,* 11**,** R118.

Karlik, B. & Öztoprak, E. 2012. Personalized cancer treatment by using Naive Bayes classifier. *Int J Mach Learn Comput,* 2**,** pp. 339-344.

Kesavaraj, G. & Sukumaran, S. 2013. A study on classification techniques in data mining. *In:* Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on, 4-6 July, pp. 1-7.

Kohavi, R. 1995. The power of decision tables. *Machine Learning: ECML-95.* Springer.

Korkusuz, D., Raharjo, H. & Bergman, B. 2011. Process capability analysis for non-normal distribution with lower specification limit. *In:* Industrial Engineering and Engineering Management (IEEM), 2011 IEEE International Conference on, 6-9 Dec. pp. 1466-1470.

Lalkhen, A. G. & Mccluskey, A. 2008. Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care & Pain,* 8**,** pp. 221-223.

Larose, D. T. 2014. *Discovering knowledge in data: an introduction to data mining,* John Wiley & Sons.

Lee, K., Gray, A. & Kim, H. 2013. Dependence maps, a dimensionality reduction with dependence distance for high-dimensional data. *Data Mining and Knowledge Discovery,* 26**,** pp. 512-532.

Lei, L., Naijun, W. & Peng, L. 2005. Applying sensitivity analysis to missing data in classifiers. Services Systems and Services Management, 2005. Proceedings of ICSSSM '05. 2005 International Conference on, 13-15 June, Vol. 2. pp. 1051-1056

Lejeune, M. A. P. M. 2001. Measuring the impact of data mining on churn management. *Internet Research,* 11**,** pp. 375-387.

Leung, H. & Haykin, S. 1991. The complex backpropagation algorithm. *Signal Processing, IEEE Transactions on,* 39**,** pp. 2101-2104.

Levy, W. C., Mozaffarian, D., Linker, D. T., Sutradhar, S. C., Anker, S. D., Cropp, A. B., Anand, I., Maggioni, A., Burton, P., Sullivan, M. D., Pitt, B., Poole-Wilson, P. A., Mann, D. L. & Packer, M. 2006. The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation,* 113**,** pp. 1424-33.

Li, D., Deogun, J., Spaulding, W. & Shuart, B. 2004. Towards missing data imputation: A study of fuzzy k-means clustering method. *In:* Rough Sets and Current Trends in Computing, Springer, pp. 573-579.

Li, D., Deogun, J., Spaulding, W. & Shuart, B. 2005. Dealing with missing data: Algorithms based on fuzzy set and rough set theories. *Transactions on rough sets IV.* Springer.

Li, D.-C., Liu, C.-W. & Hu, S. C. 2010. A learning method for the class imbalance problem with medical data sets. *Computers in biology and medicine,* 40**,** pp. 509-518.

Li, D. X. 1999. Value at Risk based on the Volatility, Skewness and Kurtosis. *Riskmetrics Group, http://www. riskmetrics. com/kurtovv. html.*

Li, X., ZAÏANE, O. R. & LI, Z. 2006. *Advanced Data Mining and Applications: Second International Conference, ADMA 2006, Xi'an, China, August 14-16, 2006, Proceedings*, Springer.

Liangxiao, J., Zhang, H. & Zhihua, C. 2009. A Novel Bayes Model: Hidden Naive Bayes. *Knowledge and Data Engineering, IEEE Transactions on,* 21**,** pp. 1361-1371.

Liao, Z., Lu, X., Yang, T. & Wang, H. 2009. Missing Data Imputation: A Fuzzy K-means Clustering Algorithm over Sliding Window. Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on, 14-16 Aug, pp. 133-137.

Little, R. J. 1992. Regression with missing X's: a review. *Journal of the American Statistical Association*, 87**,** pp. 1227-1237.

Little, R. J. A. & Rubin, D. B. 1987. *Statistical Analysis With Missing Data*, Wiley.

Longadge, R. & Dongre, S. 2013. Class Imbalance Problem in Data Mining Review. *arXiv preprint arXiv:1305.1707*.

Longadge, R., Dongre, S. & Malik, L. 2013. Class Imbalance Problem in Data Mining Review. *International Journal of Computer Science and Network* 2**,** pp. 83-87.

Lu, Z. & Su, J. 2010. Clinical data management: Current status, challenges, and future directions from industry perspectives. *Open Access J Clin Trials,* 2**,** pp. 93-105.

Luengo, J., García, S. & Herrera, F. 2012. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and information systems,* 32**,** pp. 77-108.

Maimon, O. & Rokach, L. 2008. Data mining with decision trees: theory and applications. USA: World Scientific Publishing.

Mani, S., Pazzani, M. & West, J. 1997. Knowledge discovery from a breast cancer database. *In:* Keravnou, E., Garbay, C., Baud, R. & Wyatt, J. (eds.) *Artificial Intelligence in Medicine.* Springer Berlin Heidelberg.

Marschollek, M., Wolf, K. H., Gietzelt, M., Nemitz, G., Meyer Zu Schwabedissen, H. & Haux, R. 2008. Assessing elderly persons' fall risk using spectral analysis on accelerometric data - a clinical evaluation study. *In:* Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, 20-25 Aug, pp. 3682-3685.

Mathworks, I. 2005. *MATLAB: the language of technical computing. Desktop tools and development environment, version 7*, MathWorks.

Mccallum, A. & Nigam, K. 1998. A comparison of event models for naive bayes text classification. AAAI-98 workshop on learning for text categorization. Citeseer, pp. 41-48.

Mcgregor, C., Catley, C. & James, A. 2012. Variability analysis with analytics applied to physiological data streams from the neonatal intensive care unit. *In:* 25th International Symposium on Computer-Based Medical Systems (CBMS), pp.1-5.

Meila, M. 1999. An accelerated Chow and Liu algorithm: fitting tree distributions to high dimensional sparse data.

Menardi, G. & Torelli, N. 2014. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery,* 28**,** pp. 92-122.

Moon, T. K. 1996. The expectation-maximization algorithm. *Signal Processing Magazine, IEEE,* 13**,** pp. 47-60.

Moore, L. & Kambhampati, C. 2013. The Effect of Features Using Feature Selection for Bayesian Classifier. *In:* 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC),13-16 Oct, pp. 4641-4646.

Moore, L., Kambhampati, C. & Cleland, J. G. F. 2014. Classification of a real live Heart Failure clinical dataset - Is TAN Bayes better than other Bayes? *2014 IEEE International conference on Systems, Man and Cybernetics* San Diego, USA.

Muhammed, L. A. N. 2012 Using data mining technique to diagnosis heart disease. *In:* 2012 International Conference on Statistics in Science, Business, and Engineering (ICSSBE), 10-12 Sept, pp. 1-3.

Musil, C. M., Warner, C. B., Yobas, P. K. & Jones, S. L. 2002. A Comparison of Imputation Techniques for Handling Missing Data. *Western Journal of Nursing Research,* 24**,** pp. 815-829.

Naidu, K., Dhenge, A. & Wankhade, K. 2014. Feature Selection Algorithm for Improving the Performance of Classification: A Survey. *In:* 2014 Fourth International Conference on Communication Systems and Network Technologies (CSNT), , 7-9 April, pp. 468-471.

Nitta, T. 1997. An Extension of the Back-Propagation Algorithm to Complex Numbers. *Neural Networks,* 10**,** pp. 1391-1415.

Noteboom, C. B., Motorny, S. P., Qureshi, S. & Sarnikar, S. 2014. Meaningful Use of Electronic Health Records for Physician Collaboration: A Patient Centered Health Care Perspective. 2014 47th Hawaii International Conference on System Sciences (HICSS), 6-9 Jan, pp. 656-666.

Obenshain, M. K. M. A. T. 2004. Application of Data Mining Techniques to Healthcare Data. *Infection Control and Hospital Epidemiology,* 25**,** pp. 690-695.

Ochiai, M. E., Barretto, A. C., Oliveira, M. T., Munhoz, R. T., Morgado, P. C. & Ramires, J. A. 2005. Uric acid renal excretion and renal insufficiency in decompensated severe heart failure. *European journal of heart failure,* 7**,** pp. 468-474.

Orriols, A., & Bernado-Mansilla E., 2005. The class imbalance problem in learning classifier systems: a preliminary study. *Proceedings of the 7th annual workshop on Genetic and evolutionary computation.* Washington, D.C.: ACM.

Parikh, M. & Webb, S. T. 2012. Cations: potassium, calcium, and magnesium. *Continuing Education in Anaesthesia, Critical Care & Pain*.

Patil, B., Joshi, R. & Toshniwal, D. 2010. Missing Value Imputation Based on K-Mean Clustering with Weighted Distance. *In:* Ranka, S., Banerjee, A., Biswas, K., Dua, S., Mishra, P., Moona, R., Poon, S. H. & Wang, C. L. (eds.) *Contemporary Computing.* Springer Berlin Heidelberg.

Pearson, M. & Cowie, M. R. 2005. *Managing chronic heart failure : learning from best practice : implementing NICE/NCC-CC guidelines on chronic conditions,* London.

Pelckmans, K., De Brabanter, J., Suykens, J. A. & De Moor, B. 2005. Handling missing values in support vector machine classifiers. *Neural Networks,* 18**,** 684-692.

Peng, L., Lei, L. & Naijun, W. 2005. A Quantitative Study of the Effect of Missing Data in Classifiers. The Fifth International Conference on Computer and Information Technology, 2005. CIT 2005. 21-23 Sept, pp. 28-33.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W. & Newcomb, P. 2004. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American journal of epidemiology,* 159**,** pp. 882-890.

Pigott, T. D. 2001. A review of methods for missing data. *Educational research and evaluation,* 7**,** pp. 353-383.

Poolsawad, N., Kambhampati, C. & Cleland, J. 2011. Feature Selection Approaches with Missing Values Handling for Data Mining-A Case Study of Heart Failure Dataset. *World Academy of Science, Engineering and Technology,* 60**,** pp. 828-837.

Poolsawad, N., Kambhampati, C. & Cleland, J. 2014a. Balancing Class for Performance of Classification with a Clinical Dataset. Proceedings of the World Congress on Engineering.

Poolsawad, N., Moore, L., Kambhampati, C. & Cleland, J. F. 2014b. Issues in the mining of heart failure datasets. *International Journal of Automation and Computing,* 11**,** pp. 162-179.

Poolsawad, N., Moore, L., Kambhampati, C. & Cleland, J. G. F. 2012. Handling missing values in data mining - A case study of heart failure dataset. 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 29-31 May, pp. 2934-2938.

Potamias, G. A. & Moustakis, V. S. 2001. Knowledge discovery from distributed clinical data sources: the era for internet-based epidemiology. Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE, vol.4.pp. 3638-3641

Powers, D. 2007. Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation (Tech. Rep.). *Adelaide, Australia*.

Powers, D. M. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Pérez, A., Dennis, R. J., Gil, J. F. A., Rondón, M. A. & López, A. 2002. Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. *Statistics in Medicine,* 21**,** pp. 3885-3896.

Pérez, A., Larrañaga, P. & Inza, I. 2009. Bayesian classifiers based on kernel density estimation: Flexible classifiers. *International Journal of Approximate Reasoning,* 50**,** pp. 341-362.

Quinlan, J. R. 1986. Induction of decision trees. *Machine learning,* 1**,** pp. 81-106.

Quinlan, J. R. 2014. *C4. 5: programs for machine learning,* Elsevier.

Ramanan, A., Suppharangsan, S. & Niranjan, M. 2007. Unbalanced Decision Trees for multi-class classification. International Conference on Industrial and Information Systems, 2007. ICIIS 2007., 9-11 Aug, pp. 291-294.

Sarkar, M. & LEONG, T. Y. Fuzzy K-means clustering with missing values. Proceedings of the AMIA Symposium, 2001. American Medical Informatics Association, 588.

SAS Data Mining and the Case for Sampling - A SAS Institute Best Practices Paper Solving Business Problems Using SAS & reg. *Enterprise Miner & trade; Software. SAS Institute Inc.*

Schrier, R. W. 2008. Blood Urea Nitrogen and Serum Creatinine Not Married in Heart Failure. *Circulation: Heart Failure,* 1**,** pp. 2-5.

Sematech, N. 2006. Engineering statistics handbook. NIST SEMATECH.

Shahriar, M. S. & Anam, S. 2008. Quality Data for Data Mining and Data Mining for Quality Data: A Constraint Based Approach in XML. Second International Conference on Future Generation Communication and Networking Symposia, 2008. FGCNS '08. 13-15 Dec, pp. 46-49.

Shelton, R. J., Clark, A. L., Kaye, G. C. & Cleland, J. G. 2010. The atrial fibrillation paradox of heart failure. *Congest Heart Fail,* 16**,** pp. 3-9.

Shi, H. & Liu, Y. 2011. Naïve Bayes vs. Support Vector Machine: Resilience to Missing Data. *In:* Deng, H., Miao, D., Lei, J. & Wang, F. (eds.) *Artificial Intelligence and Computational Intelligence.* Springer Berlin Heidelberg.

Shi, H.B. & Huang, H.K. 2002. Learning tree-augmented naive Bayesian network by reduced space requirements. *In:* IEEE Proceedings 2002 International Conference on Machine Learning and Cybernetics, pp. 232-1236.

Shuo, W. & Xin, Y. 2012. Multiclass Imbalance Problems: Analysis and Potential Solutions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* 42**,** pp. 1119-1130.

Silva, J. D. A. & Hruschka, E. R. 2013. An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks. *Data & Knowledge Engineering,* 84**,** pp. 47-58.

Silva, L. M., De Sá, J. M. & Alexandre, L. A. 2008. Data classification with multilayer perceptrons using a generalized error function. *Neural Networks,* 21**,** pp. 1302-1310.

Sinha, A. & Gupta, S. 2008. Fast Estimation of Nonparametric Kernel Density Through PDDP, and its Application in Texture Synthesis. Proceedings of the 2008 International conference on Visions of Computer Science: BCS International Academic Conference., 2008. pp. 225-236.

Stehman, S. V. 1997. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment,* 62**,** 77-89.

Subasi, A., Yilmaz, M. & Ozcalik, H. R. 2006. Classification of EMG signals using wavelet neural network. *Journal of Neuroscience Methods,* 156**,** pp. 360-367.

Tillander, A. 2012. Effect of data discretization on the classification accuracy in a high-dimensional framework. *International Journal of Intelligent Systems,* 27**,** pp. 355-374.

Tillander, A. 2013. Classification models for high-dimensional data with sparsity patterns.

Tsuji, H., Venditti, F. J., Manders, E. S., Evans, J. C., Larson, M. G., Feldman, C. L. & Levy, D. 1994. Reduced heart rate variability and mortality risk in an elderly cohort. The Framingham Heart Study. *Circulation,* 90**,** pp. 878-83.

Vaughan, T., Weber, R., Tipton, W. & Nelson, H. 1989. Comparison of PEFR and FEV1 in patients with varying degrees of airway obstruction. Effect of modest altitude. *CHEST Journal,* 95**,** pp. 558-562.

Vaughn, M. 1996. Interpretation and knowledge discovery from the multilayer perceptron network: Opening the black box. *Neural Computing & Applications,* 4**,** pp. 7282.

Von Hippel, P. T. 2005. Mean, median, and skew: Correcting a textbook rule. *Journal of Statistics Education,* 13**,** n2.

Wang, S.C. 2003. Artificial neural network. *Interdisciplinary Computing in Java Programming.* Springer.

Webb, G. I., Boughton, J. & Wang, Z. 2002. Averaged one-dependence estimators: preliminary results. *In:* Proceedings of the Australasian Data Mining Workshop.

Webb, G. I., Boughton, J. R. & Wang, Z. 2005. Not So Naive Bayes: Aggregating One-Dependence Estimators. *Machine learning,* 58**,** pp. 5-24.

Weiss, S. M. 1998. *Predictive data mining: a practical guide*, Morgan Kaufmann.

Weitschek, E., Felici, G. & Bertolazzi, P. 2013 Clinical Data Mining: Problems, Pitfalls and Solutions. 24th International Workshop on Database and Expert Systems Applications (DEXA), 2013, 26-30 Aug, pp. 90-94.

Wielinga, B. J., Schreiber, A. T. & Breuker, J. A. 1992. KADS: a modelling approach to knowledge engineering. *Knowledge Acquisition,* 4**,** pp. 5-53.

Wirth, R. & Hipp, J. 2000. CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000. Citeseer, pp. 29-39.

Witten, I. H. & Frank, E. 2005. *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Xiaoyan, W., Zhenyu, W. & Kang, L. 2008. Classification and Identification of Differential Gene Expression for Microarray Data: Improvement of the Random Forest Method. The 2nd International Conference on Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. 16-18 May, pp. 763-766.

Xiaoyuan, J., Chao, L., Min, L., Yongfang, Y., Zhang, D. & Jingyu, Y. 2011. Class-imbalance learning based discriminant analysis. 2011 First Asian Conference on Pattern Recognition (ACPR), 28 Nov, pp. 545-549.

Xue-Min, M., Chuan-Xi, C. & Bing-Yu, S. 2011. Comparative research on methods of dimensionality reduction in high-dimension medical data. 2011 Fourth International Workshop on Advanced Computational Intelligence (IWACI), 19-21 Oct, pp. 586-589.

Yan, Z.Y., Xu, C.F. & Pan, Y.H. 2011. Improving naive Bayes classifier by dividing its decision regions. *Journal of Zhejiang University SCIENCE C,* 12**,** pp. 647-657.

Yang, B., Janssens, D., Ruan, D., Cools, M., Bellemans, T. & Wets, G. 2012. A Data Imputation Method with Support Vector Machines for Activity-Based Transportation Models. *In:* WANG, Y. & LI, T. (eds.) *Foundations of Intelligent Systems.* Springer Berlin Heidelberg.

Zamora, E., Lupón, J., Urrutia, A., González, B., Mas, D., Díez, C., Altimir, S. & Valle, V. 2007. Prognostic significance of creatinine clearance rate in patients with heart failure and normal serum creatinine. *Revista Española de Cardiología*, 60**,** pp. 1315-1318.

Zhang, H. 2004. The optimality of naive Bayes. *AA,* 1**,** 3.

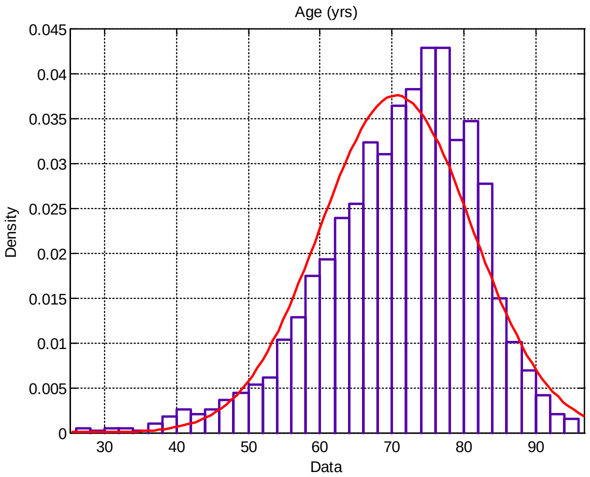
Zhang, S. 2012. Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software,* 85**,** pp. 2541-2552.

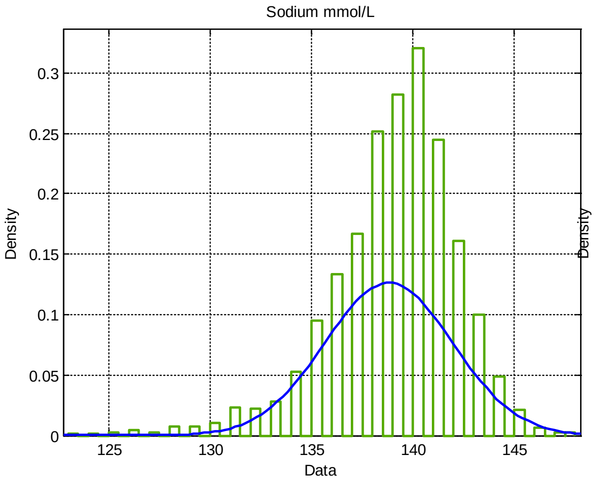
Zhang, Y., Kambhampati, C., Davis, D. N., Goode, K. & cleland, J. G. 2012. A comparative study of missing value imputation with multiclass classification for clinical heart failure data. 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 2840-2844.

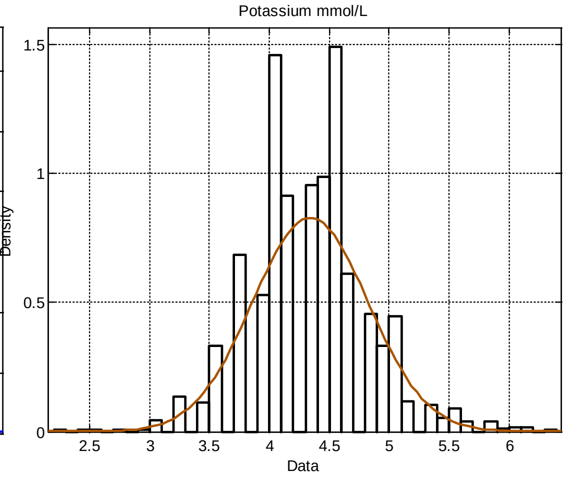
Zhanyu, M. & Leijon, A. 2011. Bayesian Estimation of Beta Mixture Models with Variational Inference. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 33**,** pp. 2160-2173.

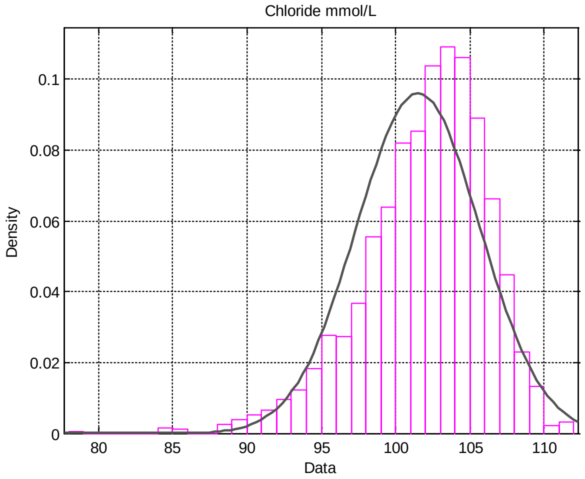
Žalik, K. R. 2008. An efficient k′-means clustering algorithm. *Pattern Recognition Letters,* 29**,** pp. 1385-1391.

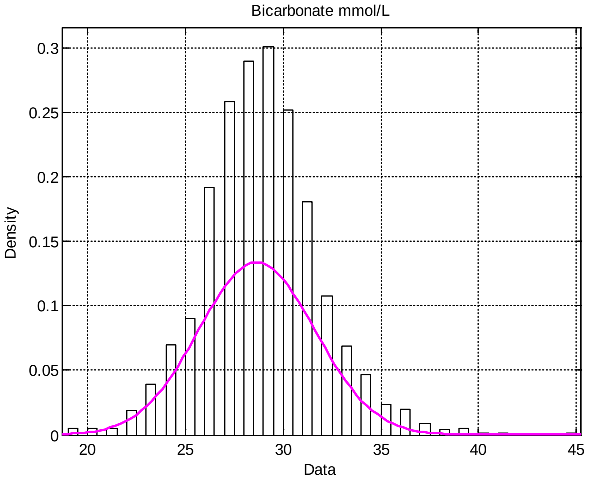
Appendix I Distributions of the raw clinical data 60 variables

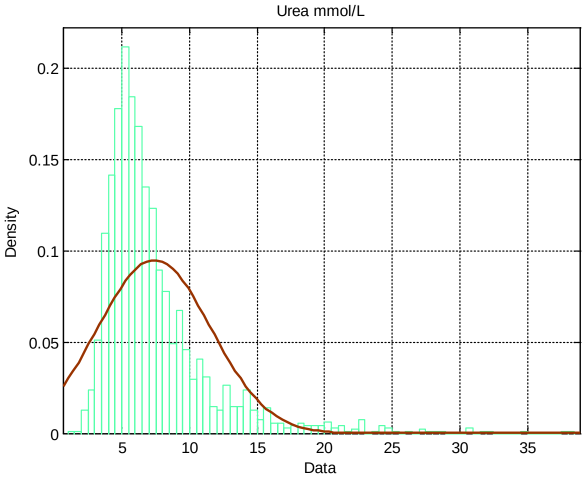


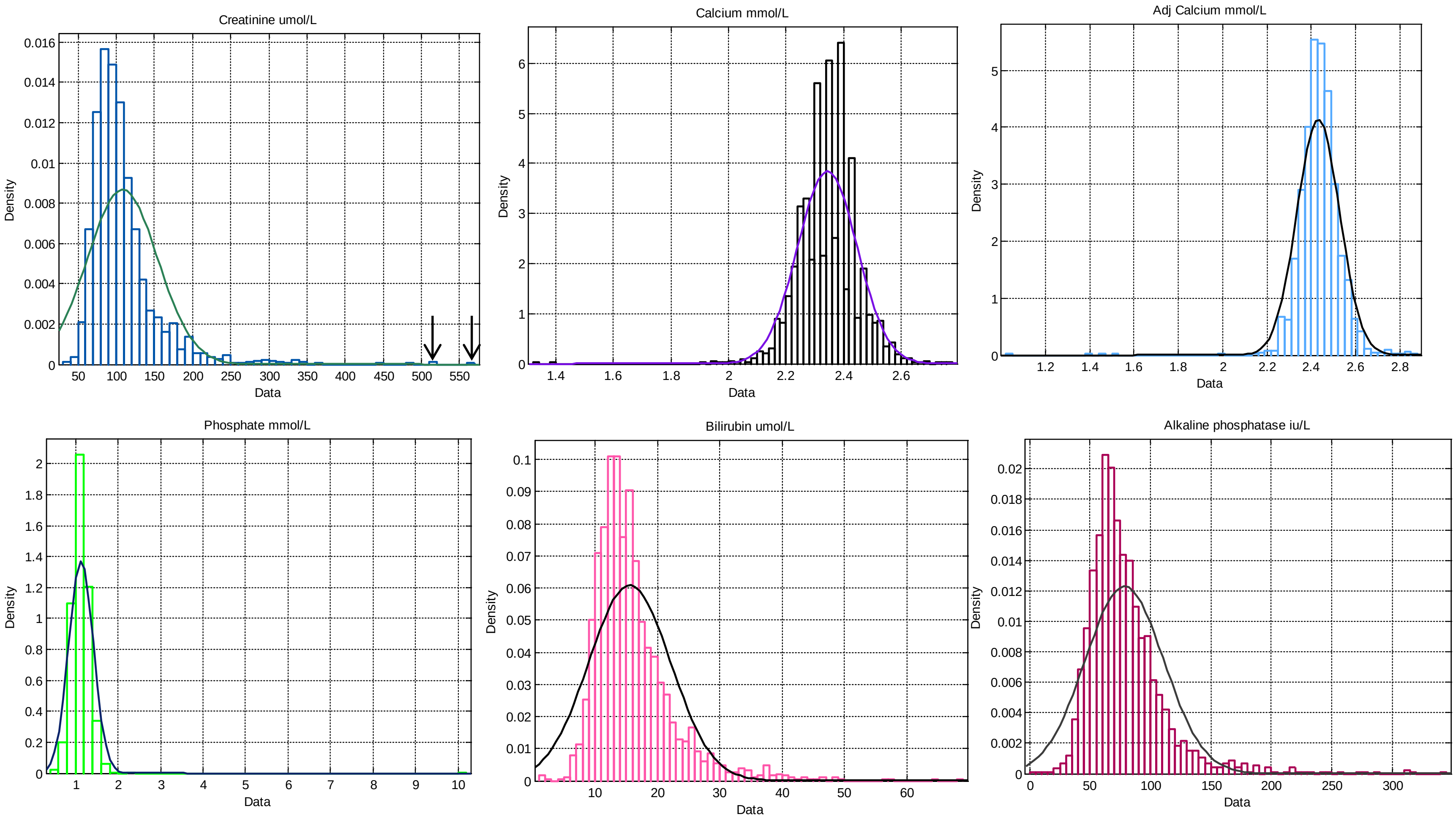


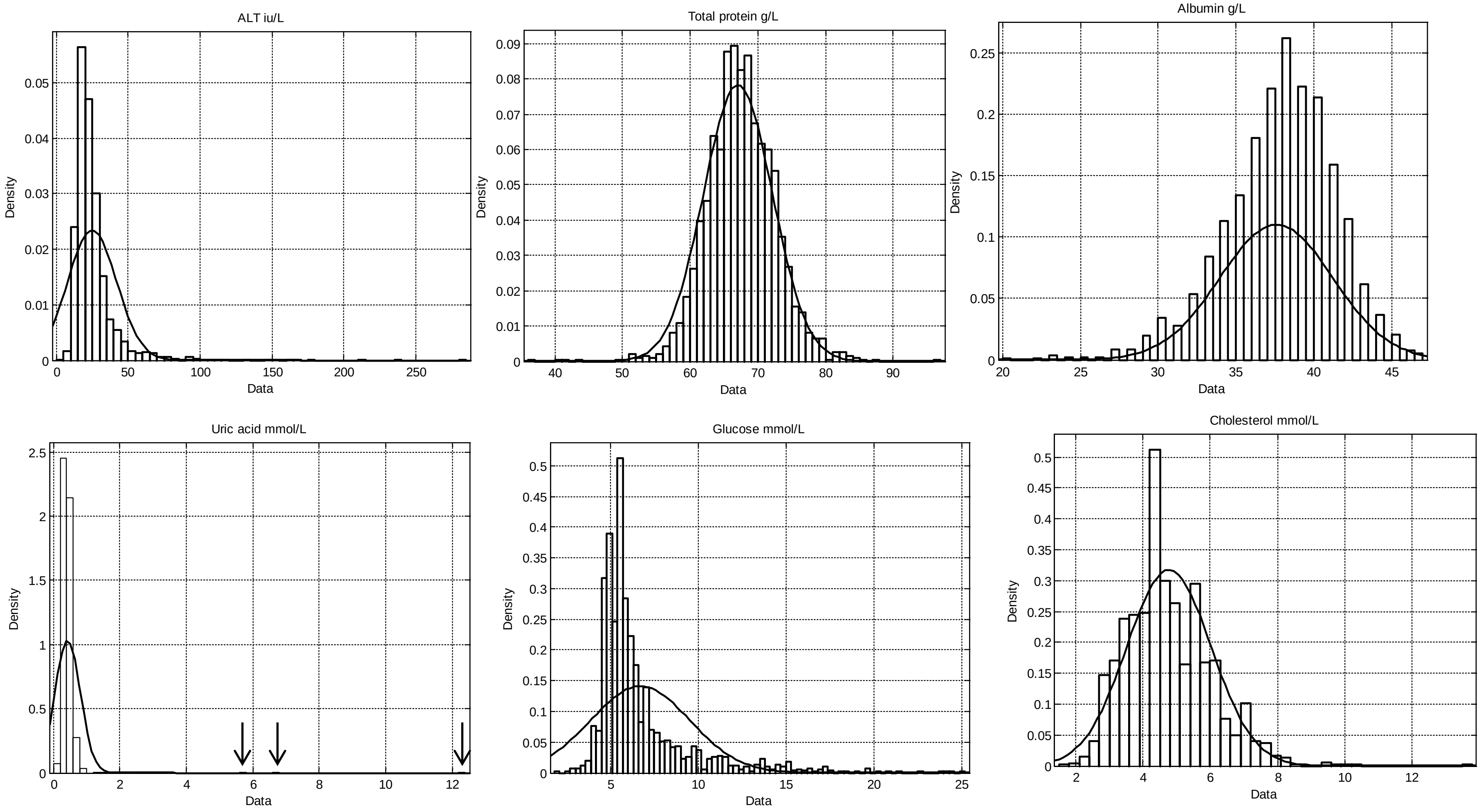


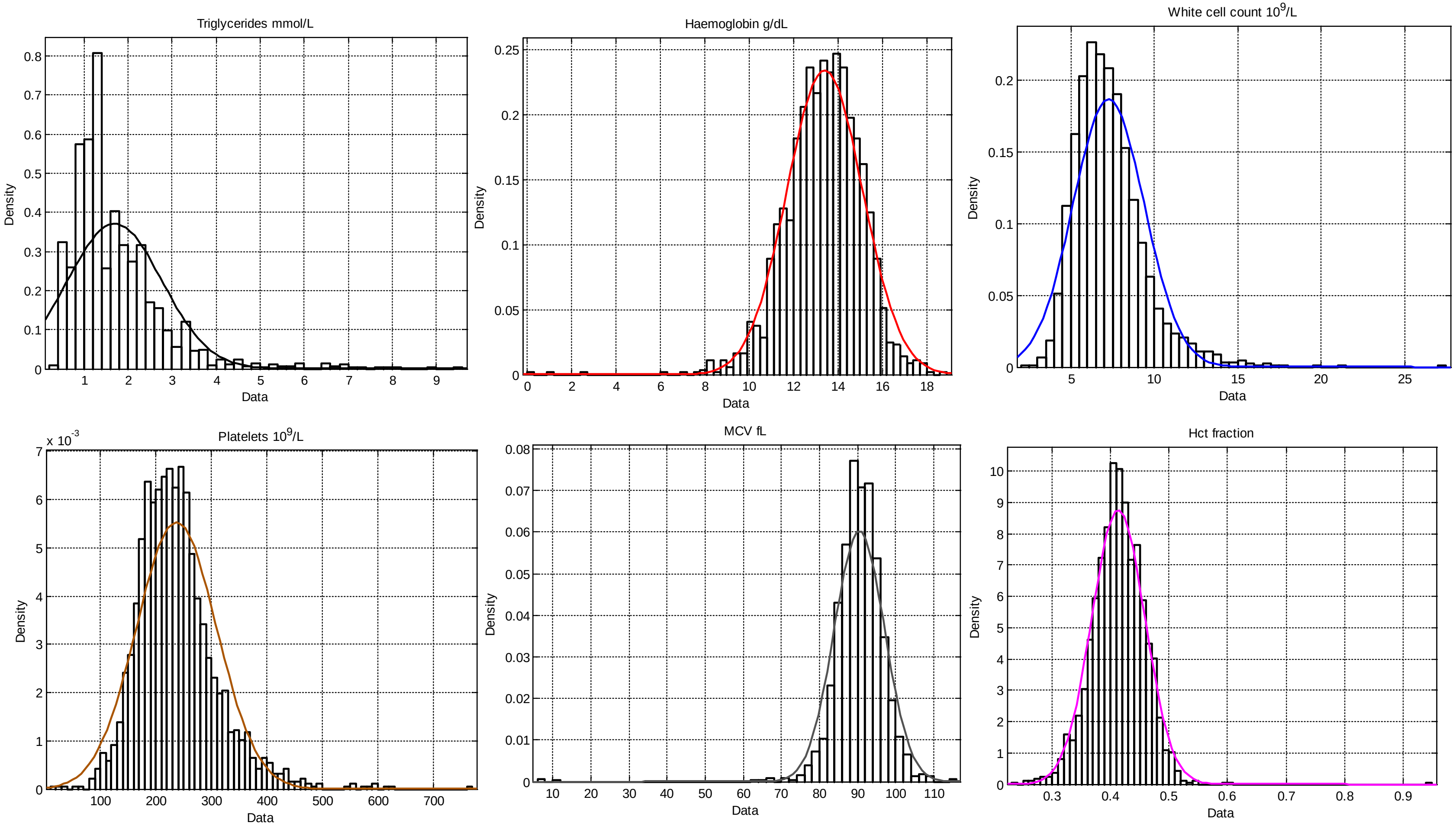


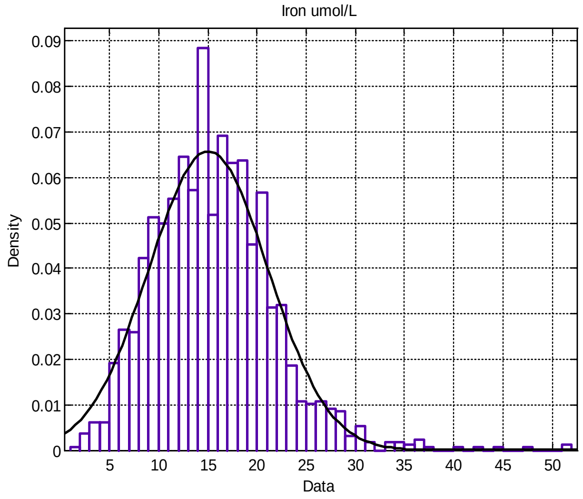


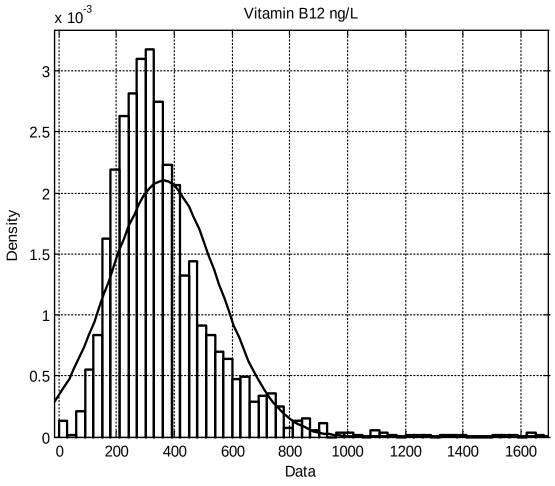


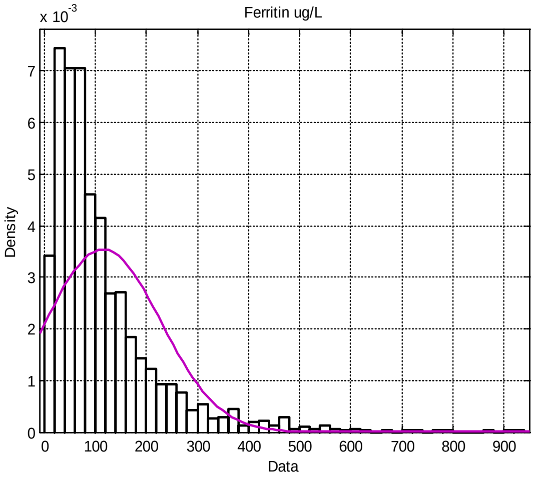

Creatinine umol/L
Density
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
50
100
150
200
250
300
350
400
450
500
550
Data
Calcium mmol/L
Density
6
5
4
3
2
1
0
1.4
1.6
1.8
2
2.2
2.4
2.6
Data
Adj Calcium mmol/L
Density
5
4
3
2
1
0
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
Data
Phosphate mmol/L
Density
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
7
8
9
10
Data
Bilirubin umol/L
Density
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
10
20
30
40
50
60
Data
Alkaline phosphatase iu/L
Density
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
50
100
150
200
250
300
Data

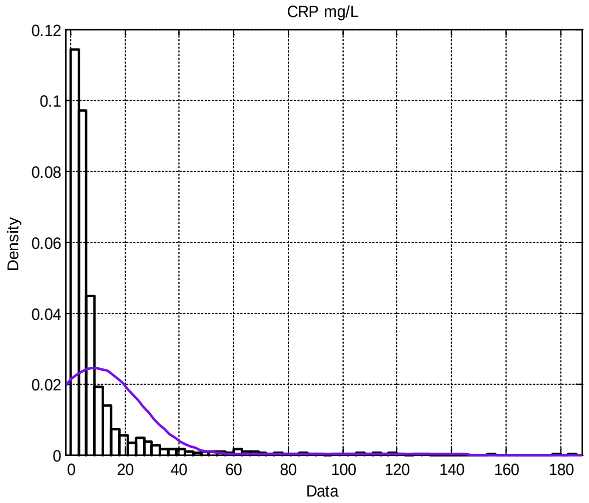

ALT iu/L
Density
0.05
0.04
0.03
0.02
0.01
0
0
50
100
150
200
250
Data
Total protein g/L
Density
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
40
50
60
70
80
90
Data
Albumin g/L
Density
0.25
0.2
0.15
0.1
0.05
0
20
25
30
35
40
45
Data
Uric acid mmol/L
Density
2.5
2
1.5
1
0.5
0
0
2
4
6
8
10
12
Data
Glucose mmol/L
Density
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
5
10
15
20
25
Data
Cholesterol mmol/L
Density
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
2
4
6
8
10
12
Data

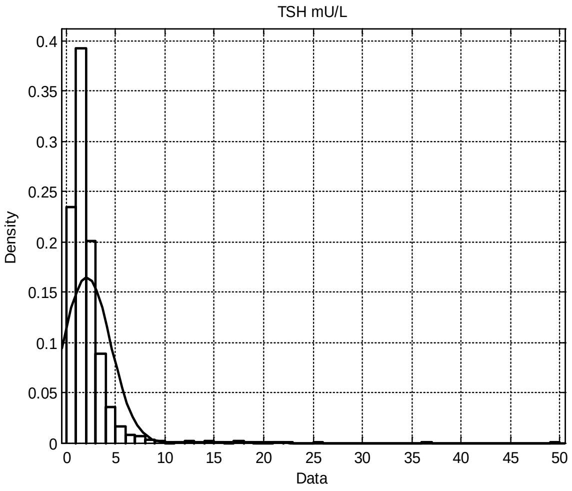

Triglycerides mmol/L
Density
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
1 2 3 4 5 6 7 8 9
Data
Haemoglobin g/dL
Density
0 0.05 0.1 0.15 0.2 0.25
0 2 4 6 8 10 12 14 16 18
Data
White cell count 10^9/L
Density
0 0.05 0.1 0.15 0.2
5 10 15 20 25
Data
Platelets 10^9/L
x 10^-3
Density
0 1 2 3 4 5 6 7
100 200 300 400 500 600 700
Data
MCV fL
Density
0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
10 20 30 40 50 60 70 80 90 100 110
Data
Hct fraction
Density
0 1 2 3 4 5 6 7 8 9 10
0.3 0.4 0.5 0.6 0.7 0.8 0.9
Data

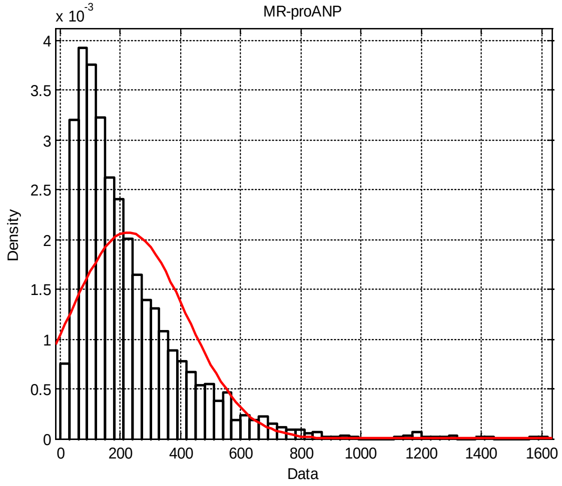

Iron umol/L
Density
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
5
10
15
20
25
30
35
40
45
50
Data

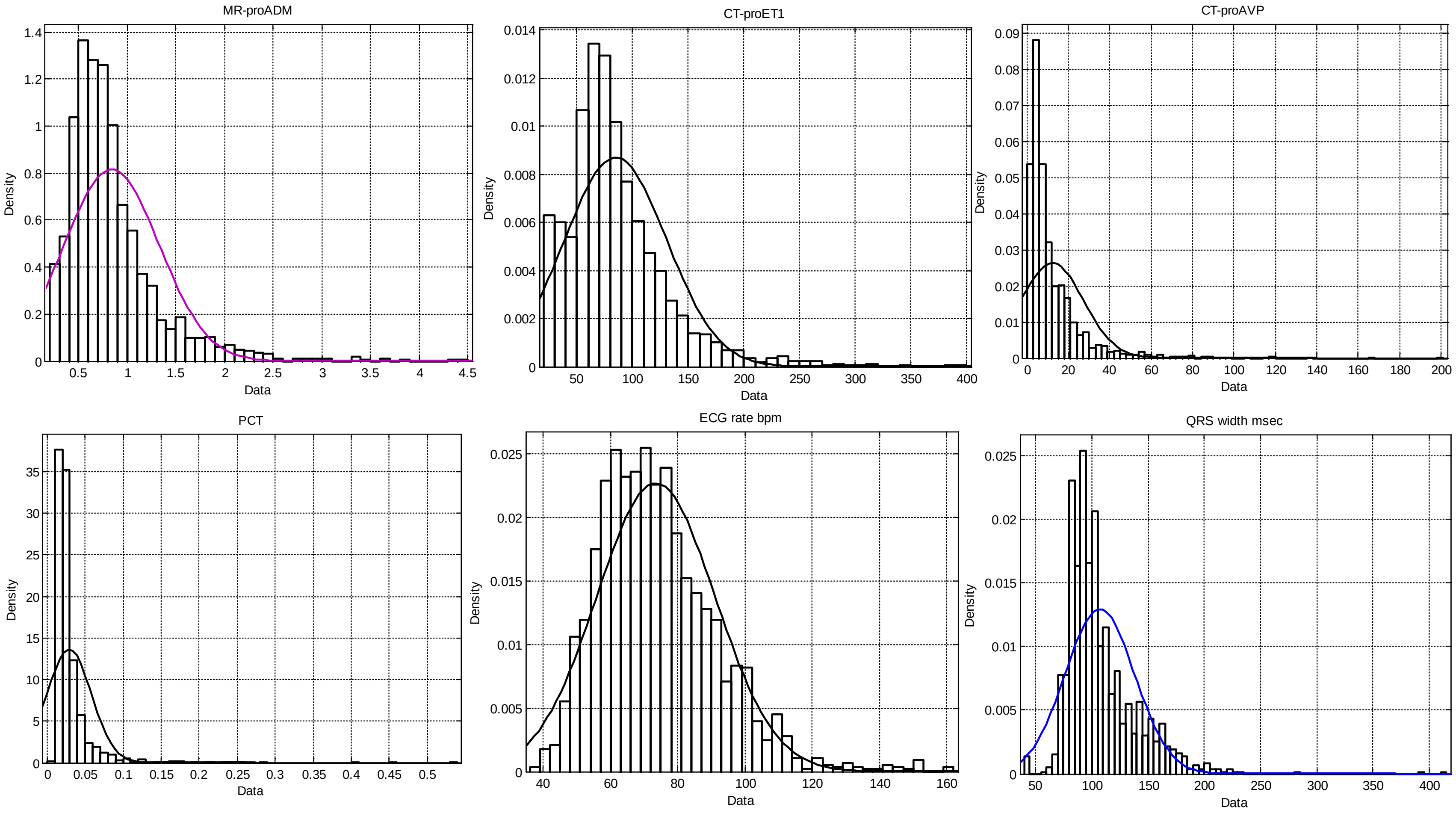

Vitamin B12 ng/L
x 10^-3
Density
3
2.5
2
1.5
1
0.5
0
0
200
400
600
800
1000
1200
1400
1600
Data

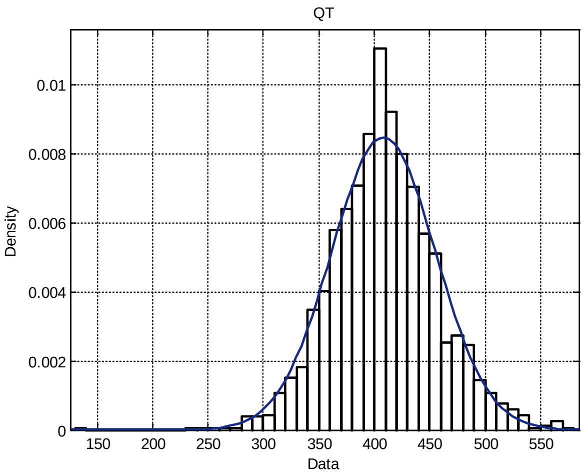

Ferritin ug/L
x 10^-3
Density
7
6
5
4
3
2
1
0
0
100
200
300
400
500
600
700
800
900
Data

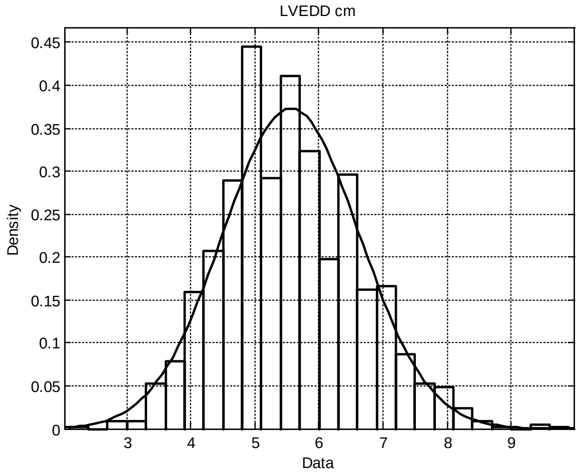

CRP mg/L
Density
0.12
0.1
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
120
140
160
180
Data

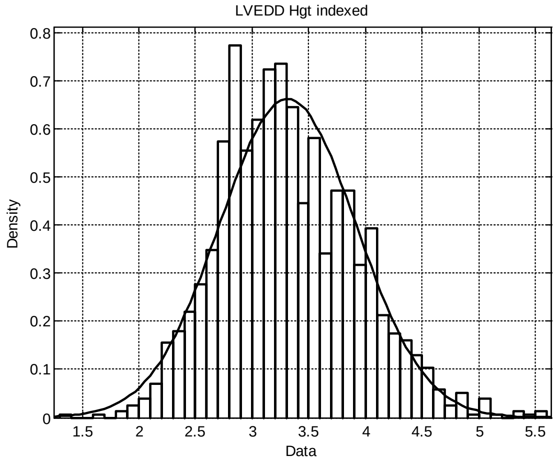

TSH mU/L
Density
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
35
40
45
50
Data

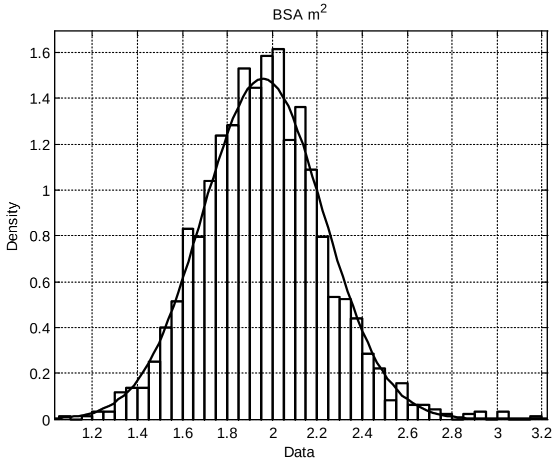

MR-proANP
x 10^-3
Density
4
3.5
3
2.5
2
1.5
1
0.5
0
0
200
400
600
800
1000
1200
1400
1600
Data

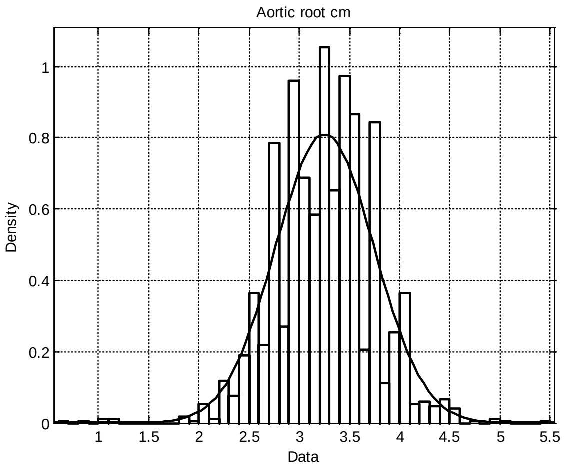

MR-proADM
Density
Data
0 0.2 0.4 0.6 0.8 1 1.2 1.4
0.5 1 1.5 2 2.5 3 3.5 4 4.5
CT-proET1
Density
Data
0 0.002 0.004 0.006 0.008 0.01 0.012 0.014
50 100 150 200 250 300 350 400
CT-proAVP
Density
Data
0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09
0 20 40 60 80 100 120 140 160 180 200
PCT
Density
Data
0 5 10 15 20 25 30 35
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5
ECG rate bpm
Density
Data
0 0.005 0.01 0.015 0.02 0.025
40 60 80 100 120 140 160
QRS width msec
Density
Data
0 0.005 0.01 0.015 0.02 0.025
50 100 150 200 250 300 350 400

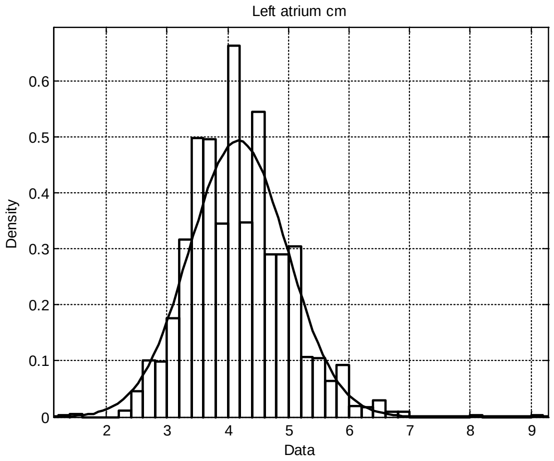

QT

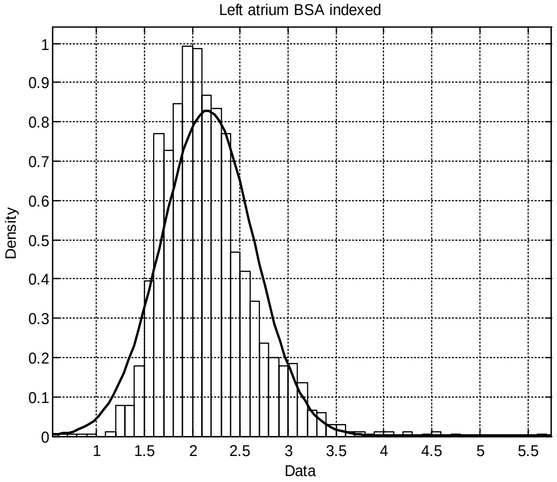

LVEDD cm

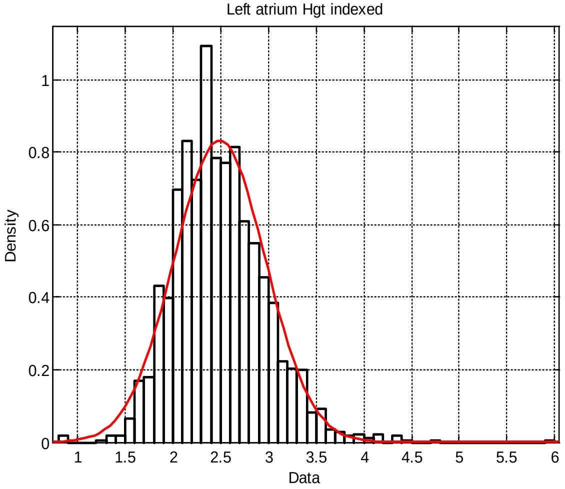

LVEDD Hgt indexed

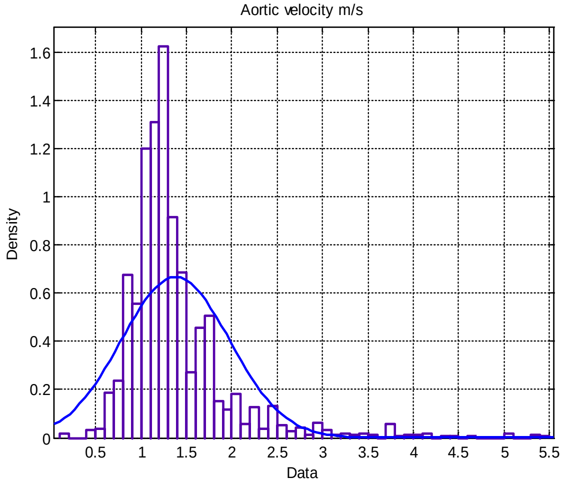

BSA m^2

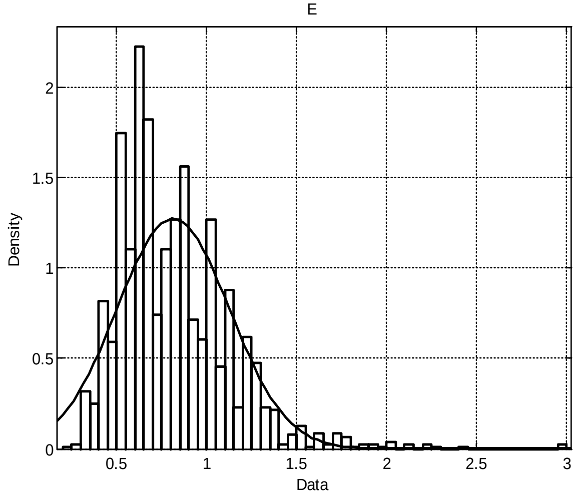

Aortic root cm

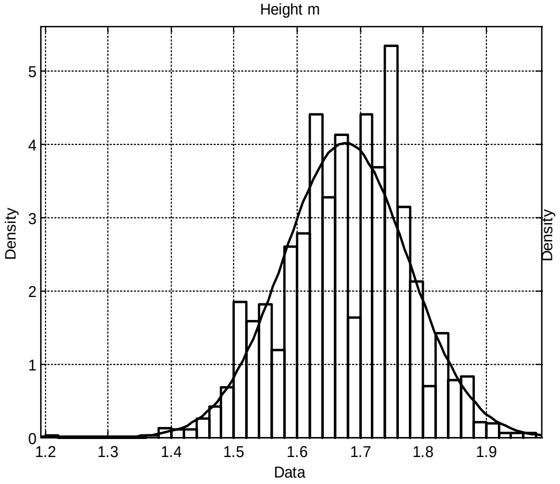

Left atrium cm

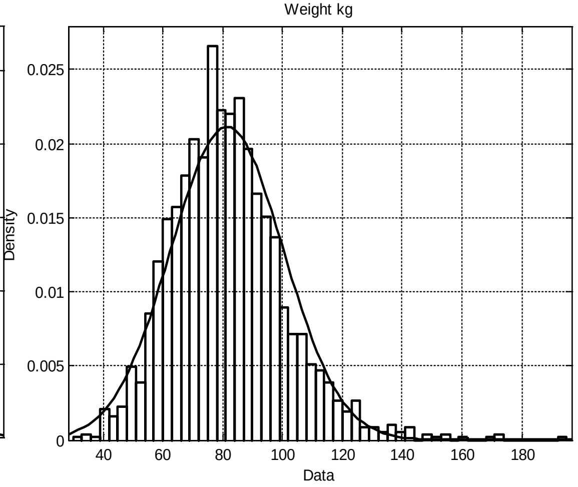

Left atrium BSA indexed
Density
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
Data

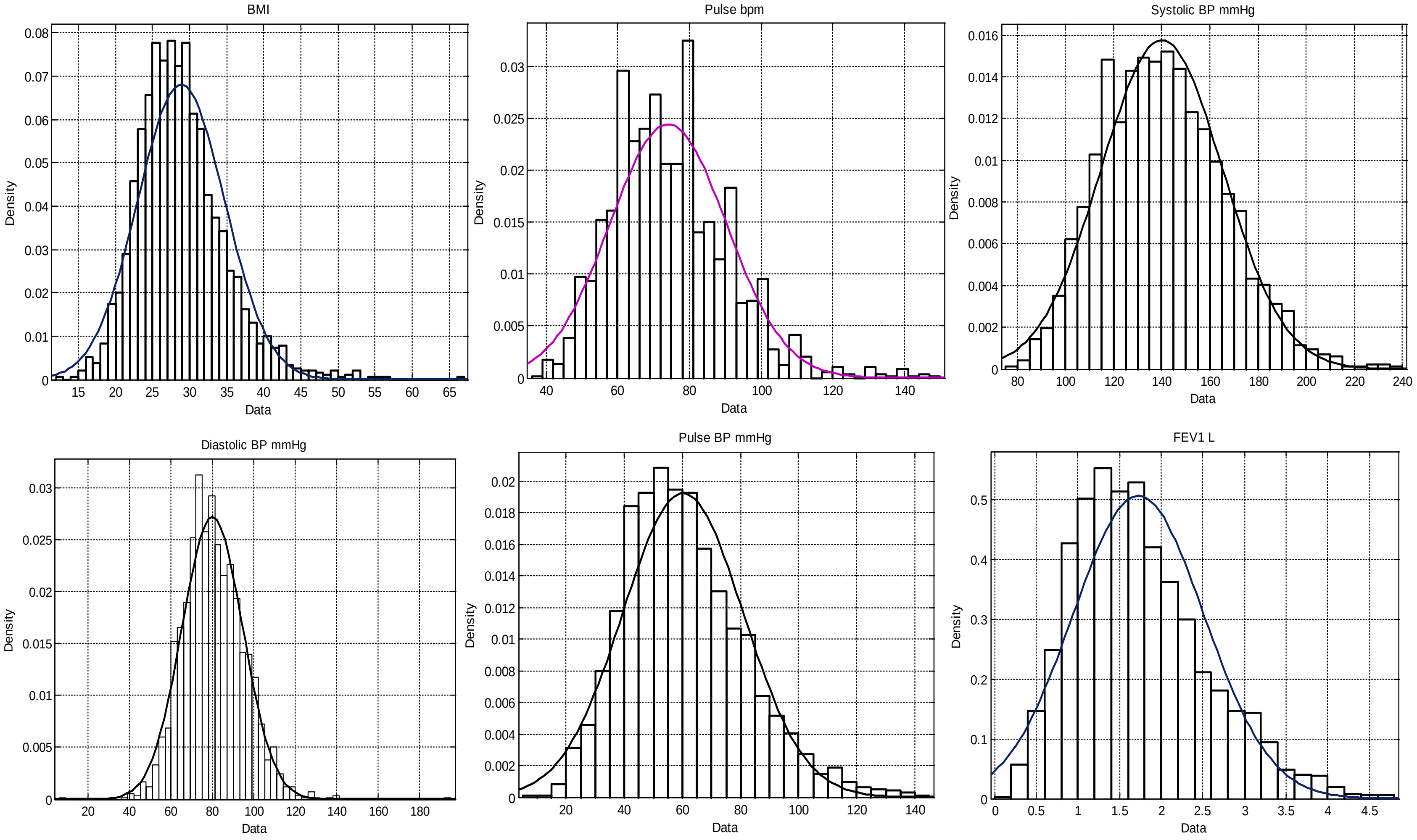

Left atrium Hgt indexed
Density
0
0.2
0.4
0.6
0.8
1
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
Data

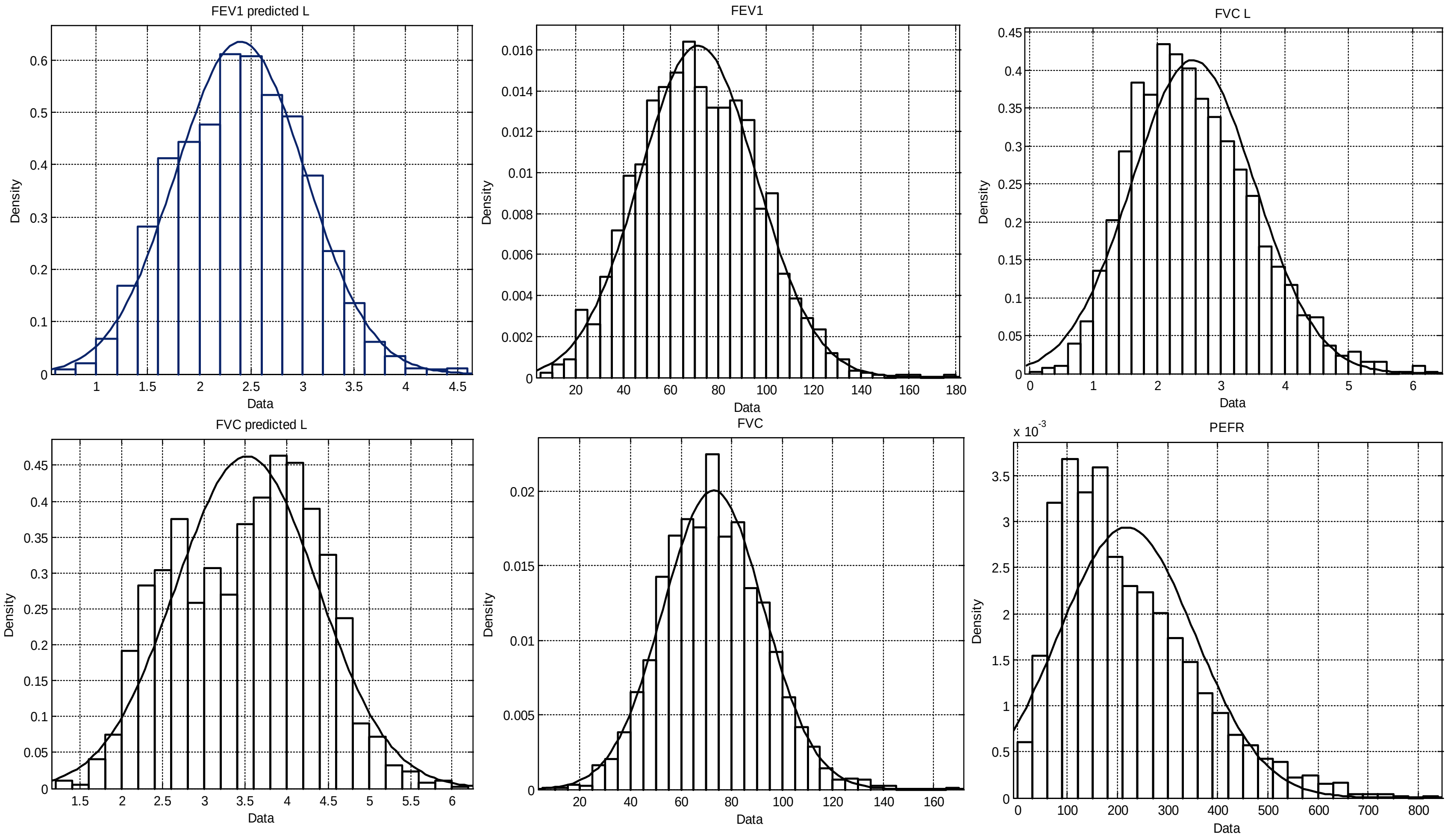

Aortic velocity m/s
Density
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
Data


E
Density
0
0.5
1
1.5
2
0.5
1
1.5
2
2.5
3
Data


Height m
Density
0
1
2
3
4
5
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
Data


Weight kg
Density
0
0.005
0.01
0.015
0.02
0.025
40
60
80
100
120
140
160
180
Data


BMI
Density
Data
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
15 20 25 30 35 40 45 50 55 60 65
Pulse bpm
Density
Data
0.03
0.025
0.02
0.015
0.01
0.005
0
40 60 80 100 120 140
Systolic BP mmHg
Density
Data
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
80 100 120 140 160 180 200 220 240
Diastolic BP mmHg
Density
Data
0.03
0.025
0.02
0.015
0.01
0.005
0
20 40 60 80 100 120 140 160 180
Pulse BP mmHg
Density
Data
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
20 40 60 80 100 120 140
FEV1 L
Density
Data
0.5
0.4
0.3
0.2
0.1
0
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5


FEV1 predicted L
Density
Data
0
0.1
0.2
0.3
0.4
0.5
0.6
1
1.5
2
2.5
3
3.5
4
4.5
FEV1
Density
Data
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014
0.016
20
40
60
80
100
120
140
160
180
FVC L
Density
Data
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0
1
2
3
4
5
6
FVC predicted L
Density
Data
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
FVC
Density
Data
0
0.005
0.01
0.015
0.02
20
40
60
80
100
120
140
160
PEFR
x 10^-3
Density
Data
0
0.5
1
1.5
2
2.5
3
3.5
0
100
200
300
400
500
600
700
800

Appendix II Alive and dead class posterior probabilities and missing data of SVM and EM 79 FP record

No.	**Record no.**	**Alive post**	**Dead post**	**MD**	**True**	**Predicted**
1	**1460**	1	0	0	Dead	Alive
2	1479	1	0	0	Dead	Alive
3	1486	1	0	0	Dead	Alive
4	1493	1	0	0	Dead	Alive
5	**1520**	1	0	0	Dead	Alive
6	1536	1	0	0	Dead	Alive
7	**1540**	1	0	0	Dead	Alive
8	**1543**	1	0	0	Dead	Alive
9	**1544**	1	0	0	Dead	Alive
10	**1558**	1	0	0	Dead	Alive
11	**1591**	1	0	0	Dead	Alive
12	1600	1	0	0	Dead	Alive
13	**1615**	1	0	0	Dead	Alive
14	**1637**	1	0	0	Dead	Alive
15	**1671**	1	0	0	Dead	Alive
16	1686	0.864	0.136	0	Dead	Alive
17	1690	1	0	0	Dead	Alive
18	1694	1	0	0	Dead	Alive
19	**1707**	1	0	0	Dead	Alive
20	1708	1	0	0	Dead	Alive
21	1710	1	0	0	Dead	Alive
22	**1711**	1	0	0	Dead	Alive
23	1712	1	0	0	Dead	Alive
24	**1716**	1	0	0	Dead	Alive
25	1725	1	0	0	Dead	Alive
26	**1729**	1	0	0	Dead	Alive
27	**1759**	1	0	0	Dead	Alive
28	**1768**	1	0	0	Dead	Alive
29	1771	1	0	0	Dead	Alive
30	**1776**	1	0	0	Dead	Alive
31	**1780**	1	0	0	Dead	Alive
32	**1784**	1	0	0	Dead	Alive
33	**1786**	1	0	0	Dead	Alive
34	**1795**	0.886	0.114	0	Dead	Alive
35	1797	1	0	0	Dead	Alive
36	**1800**	1	0	7	Dead	Alive
37	1848	1	0	0	Dead	Alive
38	1850	1	0	0	Dead	Alive
39	1855	1	0	0	Dead	Alive
40	**1856**	1	0	0	Dead	Alive
41	**1858**	1	0	0	Dead	Alive

42	**1859**	1	0	0	Dead	Alive
43	**1861**	1	0	7	Dead	Alive
44	**1863**	1	0	0	Dead	Alive
45	**1865**	0.993	0.007	1	Dead	Alive
46	**1867**	1	0	0	Dead	Alive
47	1868	1	0	0	Dead	Alive
48	**1869**	1	0	0	Dead	Alive
49	1872	1	0	0	Dead	Alive
50	**1873**	1	0	1	Dead	Alive
51	**1874**	1	0	0	Dead	Alive
52	1876	1	0	0	Dead	Alive
53	**1877**	1	0	0	Dead	Alive
54	**1878**	1	0	0	Dead	Alive
55	**1879**	1	0	0	Dead	Alive
56	**1880**	1	0	0	Dead	Alive
57	1883	1	0	1	Dead	Alive
58	**1886**	1	0	0	Dead	Alive
59	1887	1	0	0	Dead	Alive
60	**1888**	1	0	0	Dead	Alive
61	**1890**	1	0	0	Dead	Alive
62	**1892**	1	0	0	Dead	Alive
63	1896	1	0	1	Dead	Alive
64	**1897**	1	0	0	Dead	Alive
65	**1898**	1	0	0	Dead	Alive
66	**1899**	1	0	7	Dead	Alive
67	**1902**	1	0	0	Dead	Alive
68	**1906**	1	0	0	Dead	Alive
69	1910	1	0	0	Dead	Alive
70	**1911**	1	0	0	Dead	Alive
71	**1914**	1	0	0	Dead	Alive
72	1923	1	0	0	Dead	Alive
73	1927	1	0	0	Dead	Alive
74	1929	1	0	1	Dead	Alive
75	1931	1	0	0	Dead	Alive
76	1936	1	0	0	Dead	Alive
77	**1937**	1	0	0	Dead	Alive
78	**1939**	1	0	0	Dead	Alive
79	1943	1	0	0	Dead	Alive

Appendix III Alive and dead class posterior probabilities and missing data of SVM and EM 49 *FP* record

No.	Record no.	Alive post	Dead post	MD	True	Predicted
1	1460	0.826	0.174	0	Dead	Alive
2	1520	0.938	0.062	0	Dead	Alive
3	1540	0.678	0.322	0	Dead	Alive
4	1543	0.963	0.037	0	Dead	Alive
5	1544	0.729	0.271	0	Dead	Alive
6	1558	0.942	0.058	0	Dead	Alive
7	1591	0.87	0.13	0	Dead	Alive
8	1615	0.98	0.02	0	Dead	Alive
9	1637	0.78	0.22	0	Dead	Alive
10	1671	0.514	0.486	0	Dead	Alive
11	1707	0.839	0.161	0	Dead	Alive
12	1711	0.551	0.449	0	Dead	Alive
13	1716	0.745	0.255	0	Dead	Alive
14	1729	0.996	0.004	0	Dead	Alive
15	1759	0.795	0.205	0	Dead	Alive
16	1768	0.73	0.27	0	Dead	Alive
17	1776	0.815	0.185	0	Dead	Alive
18	1780	0.58	0.42	0	Dead	Alive
19	1784	0.841	0.159	0	Dead	Alive
20	1786	0.82	0.18	0	Dead	Alive
21	1795	1	0	0	Dead	Alive
22	1800	0.922	0.078	7	Dead	Alive
23	1856	0.992	0.008	0	Dead	Alive
24	1858	0.988	0.012	0	Dead	Alive
25	1859	0.677	0.323	0	Dead	Alive
26	1861	1	0	7	Dead	Alive
27	1863	1	0	0	Dead	Alive
28	1865	0.586	0.414	1	Dead	Alive
29	1867	0.986	0.014	0	Dead	Alive
30	1869	0.934	0.066	0	Dead	Alive
31	1873	0.98	0.02	1	Dead	Alive
32	1874	0.997	0.003	0	Dead	Alive
33	1877	0.863	0.137	0	Dead	Alive
34	1878	0.84	0.16	0	Dead	Alive
35	1879	0.962	0.038	0	Dead	Alive
36	1880	0.999	0.001	0	Dead	Alive
37	1886	0.535	0.465	0	Dead	Alive
38	1888	0.997	0.003	0	Dead	Alive

39	1890	0.998	0.002	0	Dead	Alive
40	1892	0.698	0.302	0	Dead	Alive
41	1897	0.802	0.198	0	Dead	Alive
42	1898	0.995	0.005	0	Dead	Alive
43	1899	0.521	0.479	7	Dead	Alive
44	1902	0.997	0.003	0	Dead	Alive
45	1906	0.977	0.023	0	Dead	Alive
46	1911	0.987	0.013	0	Dead	Alive
47	1914	0.976	0.024	0	Dead	Alive
48	1937	0.916	0.084	0	Dead	Alive
49	1939	0.847	0.153	0	Dead	Alive

Appendix IV Euclidean distance of the 79 *FP* records of SVM and EM hybrid data

No.	**Record no.**	**SVM & EM Euclidean distance**	
		Dead	Alive
1	1460	1000.64	217.85
2	1479	1053.79	1183.42
3	1486	867.77	355.42
4	1493	944.41	270.54
5	1520	1031.81	397.88
6	1536	831.85	592.05
7	1540	923.87	200.20
8	**1543**	**1162.57**	**343.29**
9	1544	989.82	339.00
10	1558	1102.91	227.14
11	1591	949.05	233.93
12	1600	1034.15	168.65
13	1615	921.10	275.20
14	1637	875.45	222.15
15	1671	858.31	569.25
16	1686	910.45	195.77
17	1690	1013.73	343.31
18	1694	898.40	566.73
19	1707	1158.04	249.20
20	1708	1212.06	1472.88
21	1710	1110.67	1231.11
22	1711	1036.20	419.72
23	1712	917.69	423.97
24	1716	1101.56	273.56
25	1725	1107.94	188.08
26	1729	1110.05	174.01
27	1759	1014.00	144.03
28	1768	1100.99	223.92
29	1771	875.88	494.59
30	1776	905.12	250.45
31	1780	1077.55	944.53
32	1784	1075.99	232.35
33	1786	1021.14	162.66
34	1795	927.56	488.18
35	1797	907.03	308.45
36	1800	881.96	200.01
37	1848	786.13	519.84
38	1850	1063.22	1247.68

39	1855	938.33	318.62
40	1856	963.54	157.17
41	1858	856.44	308.37
42	1859	883.38	421.79
43	1861	947.81	190.16
44	1863	937.13	333.46
45	1865	987.79	170.45
46	1867	1061.38	206.99
47	1868	782.15	353.90
48	1869	1054.52	987.29
49	1872	914.00	753.21
50	1873	1061.45	233.52
51	1874	947.64	167.57
52	1876	1080.09	251.60
53	1877	998.11	1146.66
54	1878	800.10	356.41
55	1879	933.14	175.78
56	1880	971.07	181.32
57	1883	1059.20	801.50
58	**1886**	**762.76**	**436.93**
59	1887	1007.42	179.35
60	1888	1148.98	256.65
61	1890	1076.16	167.70
62	1892	968.43	597.81
63	1896	893.06	562.64
64	1897	1007.28	303.80
65	1898	1040.85	313.59
66	1899	884.50	386.97
67	1902	1013.02	157.08
68	1906	958.96	347.88
69	1910	932.38	307.66
70	1911	1063.06	260.18
71	1914	939.92	578.48
72	1923	965.10	462.61
73	1927	883.51	271.65
74	1929	1008.62	778.66
75	1931	1083.66	645.62
76	1936	891.26	504.94
77	1937	977.81	414.03
78	1939	989.23	263.78
79	1943	800.11	627.22

Appendix V Dead and alive class maximum difference of the PEFR variable

	PEFR		
No.	**Record no.**	**Dead**	**Alive**
1	1460	544.00	62.71
2	1479	703.00	96.29
3	1486	607.00	0.29
4	1493	667.00	60.29
5	1520	697.00	90.29
6	1536	668.00	61.29
7	1540	533.00	73.71
8	**1543**	**786.00**	**179.29**
9	1544	780.00	173.29
10	1558	677.00	70.29
11	1591	614.00	7.29
12	1600	683.00	76.29
13	1615	695.00	88.29
14	1637	535.00	71.71
15	1671	513.00	93.71
16	1686	655.00	48.29
17	1690	763.00	156.29
18	1694	662.00	55.29
19	1707	725.00	118.29
20	1708	611.00	4.29
21	1710	612.00	5.29
22	1711	813.00	206.29
23	1712	765.00	158.29
24	1716	738.00	131.29
25	1725	750.00	143.29
26	1729	710.00	103.29
27	1759	557.00	49.71
28	1768	721.00	114.29
29	1771	641.00	34.29
30	1776	485.00	121.71
31	1780	738.00	131.29
32	1784	539.00	67.71
33	1786	591.00	15.71
34	1795	607.00	0.29
35	1797	614.00	7.29
36	1800	565.28	41.43
37	1848	588.00	18.71
38	1850	534.00	72.71

39	1855	693.00	86.29
40	1856	563.00	43.71
41	1858	454.00	152.71
42	1859	694.00	87.29
43	1861	509.85	96.86
44	1863	314.00	292.71
45	1865	676.55	69.84
46	1867	633.00	26.29
47	1868	391.00	215.71
48	1869	572.00	34.71
49	1872	745.00	138.29
50	1873	791.24	184.53
51	1874	599.00	7.71
52	1876	743.00	136.29
53	1877	569.00	37.71
54	1878	478.00	128.71
55	1879	629.00	22.29
56	1880	561.00	45.71
57	1883	741.02	134.31
58	**1886**	**486.00**	**120.71**
59	1887	720.00	113.29
60	1888	702.00	95.29
61	1890	636.00	29.29
62	1892	734.00	127.29
63	1896	616.81	10.10
64	1897	680.00	73.29
65	1898	530.00	76.71
66	1899	620.64	13.93
67	1902	563.00	43.71
68	1906	735.00	128.29
69	1910	662.00	55.29
70	1911	672.00	65.29
71	1914	642.00	35.29
72	1923	688.00	81.29
73	1927	529.00	77.71
74	1929	750.35	143.64
75	1931	771.00	164.29
76	1936	673.00	66.29
77	1937	665.00	58.29
78	1939	692.00	85.29
79	1943	676.00	69.29

Appendix VI J48 pruned decision tree of the original dataset

```
Urea (mmol/L) <= 9.5
|   FEV1 (L) <= 0.92
|   |   BMI <= 22.838625
|   |   |   CRP (mg/L) <= 10
|   |   |   |   Urea (mmol/L) <= 4.8: Alive (7.78/1.15)
|   |   |   |   Urea (mmol/L) > 4.8
|   |   |   |   |   FEV1 (L) <= 0.42: Alive (2.12/0.01)
|   |   |   |   |   FEV1 (L) > 0.42: Dead (16.92/1.44)
|   |   |   CRP (mg/L) > 10: Dead (15.78/0.06)
|   |   BMI > 22.838625
|   |   |   CRP (mg/L) <= 7.6
|   |   |   |   Aortic Velocity (m/s) <= 1.43
|   |   |   |   |   E <= 0.38: Dead (3.59/0.49)
|   |   |   |   |   E > 0.38
|   |   |   |   |   |   FVC (L) <= 1.33
|   |   |   |   |   |   |   Haemoglobin (g/dL) <= 12.7
|   |   |   |   |   |   |   |   FVC <= 58.620258: Dead (4.31/0.16)
|   |   |   |   |   |   |   |   FVC > 58.620258: Alive (2.43/0.03)
|   |   |   |   |   |   |   Haemoglobin (g/dL) > 12.7: Alive (14.04/0.7)
|   |   |   |   |   |   FVC (L) > 1.33: Alive (40.83/0.33)
|   |   |   |   Aortic Velocity (m/s) > 1.43
|   |   |   |   |   CT-proET1 <= 37: Dead (5.82/0.67)
|   |   |   |   |   CT-proET1 > 37
|   |   |   |   |   |   FEV1 (L) <= 0.65
|   |   |   |   |   |   |   QT <= 414: Dead (5.94/0.1)
|   |   |   |   |   |   |   QT > 414: Alive (6.94/1.9)
|   |   |   |   |   |   FEV1 (L) > 0.65: Alive (17.36/1.96)
|   |   |   CRP (mg/L) > 7.6
|   |   |   |   MR-proANP <= 98
|   |   |   |   |   Uric Acid (mmol/L) <= 0.3: Dead (2.21/0.21)
|   |   |   |   |   Uric Acid (mmol/L) > 0.3: Alive (11.35)
|   |   |   |   MR-proANP > 98
|   |   |   |   |   Height (Exam) (m) <= 1.575: Dead (15.43/1.31)
|   |   |   |   |   Height (Exam) (m) > 1.575
|   |   |   |   |   |   Age (yrs) <= 68: Alive (4.55)
|   |   |   |   |   |   Age (yrs) > 68
|   |   |   |   |   |   |   Height (Exam) (m) <= 1.595: Alive (3.91/0.15)
|   |   |   |   |   |   |   Height (Exam) (m) > 1.595
|   |   |   |   |   |   |   |   Phosphate (mmol/L) <= 1.07: Dead (12.86/1.16)
|   |   |   |   |   |   |   |   Phosphate (mmol/L) > 1.07
|   |   |   |   |   |   |   |   |   FEV1 <= 25.064277: Dead (4.74/0.21)
|   |   |   |   |   |   |   |   |   FEV1 > 25.064277: Alive (7.06/0.84)
|   FEV1 (L) > 0.92
|   |   Systolic BP (mmHg) <= 111
|   |   |   Bilirubin (umol/L) <= 34
|   |   |   |   Albumin (g/L) <= 30
|   |   |   |   |   Platelets (10^9/L) <= 334: Dead (9.25/0.05)
|   |   |   |   |   Platelets (10^9/L) > 334: Alive (4.02/1.0)
|   |   |   |   Albumin (g/L) > 30
|   |   |   |   |   CT-proET1 <= 31
|   |   |   |   |   |   Calcium (mmol/L) <= 2.34: Dead (8.92/0.72)
|   |   |   |   |   |   Calcium (mmol/L) > 2.34
|   |   |   |   |   |   |   TSH (mU/L) <= 1.2: Dead (2.95/0.56)
|   |   |   |   |   |   |   TSH (mU/L) > 1.2: Alive (4.79/0.09)
|   |   |   |   |   CT-proET1 > 31
|   |   |   |   |   |   Bicarbonate (mmol/L) <= 26: Alive (16.53)
|   |   |   |   |   |   Bicarbonate (mmol/L) > 26
|   |   |   |   |   |   |   Left Atrium (Hgt indexed) <= 2.4
|   |   |   |   |   |   |   |   Pulse BP (mmHg) <= 45: Alive (27.85)
|   |   |   |   |   |   |   |   Pulse BP (mmHg) > 45
|   |   |   |   |   |   |   |   |   Albumin (g/L) <= 37: Dead (2.61/0.36)
|   |   |   |   |   |   |   |   |   Albumin (g/L) > 37: Alive (2.75)
|   |   |   |   |   |   |   Left Atrium (Hgt indexed) > 2.4
|   |   |   |   |   |   |   |   Rate (ECG) (bpm) <= 64: Alive (13.63/0.87)
|   |   |   |   |   |   |   |   Rate (ECG) (bpm) > 64
|   |   |   |   |   |   |   |   |   MCV (fL) <= 91.7
|   |   |   |   |   |   |   |   |   |   QT <= 400
|   |   |   |   |   |   |   |   |   |   |   Iron (umol/L) <= 15
|   |   |   |   |   |   |   |   |   |   |   |   Left Atrium (cm) <= 4.33: Alive (3.05/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   Left Atrium (cm) > 4.33: Dead (5.37/0.35)
|   |   |   |   |   |   |   |   |   |   |   Iron (umol/L) > 15: Alive (7.94/1.61)
|   |   |   |   |   |   |   |   |   |   QT > 400: Alive (8.72/0.25)
|   |   |   |   |   |   |   |   |   MCV (fL) > 91.7
|   |   |   |   |   |   |   |   |   |   Cholesterol (mmol/L) <= 4.4: Dead (10.43/1.25)
|   |   |   |   |   |   |   |   |   |   Cholesterol (mmol/L) > 4.4: Alive (2.88/0.94)
|   |   |   Bilirubin (umol/L) > 34: Dead (7.15/0.04)
|   |   Systolic BP (mmHg) > 111
|   |   |   Age (yrs) <= 61: Alive (276.61/9.0)
|   |   |   Age (yrs) > 61
|   |   |   |   CRP (mg/L) <= 3.9
|   |   |   |   |   Haemoglobin (g/dL) <= 11.8
|   |   |   |   |   |   Calcium (mmol/L) <= 2.31
|   |   |   |   |   |   |   Vitamin B12 (ng/L) <= 433
|   |   |   |   |   |   |   |   Triglycerides (mmol/L) <= 1.6: Dead (10.32/1.29)
|   |   |   |   |   |   |   |   Triglycerides (mmol/L) > 1.6: Alive (5.1/1.62)
|   |   |   |   |   |   |   Vitamin B12 (ng/L) > 433: Alive (7.05/0.54)
| | | | | | Calcium (mmol/L) > 2.31
| | | | | | | Age (yrs) <= 86: Alive (19.57/0.31)
| | | | | | | Age (yrs) > 86: Dead (2.5/0.63)
| | | | | Haemoglobin (g/dL) > 11.8
| | | | | | LVEDD (cm) <= 6.91
| | | | | | | Urea (mmol/L) <= 7.3: Alive (339.82/17.21)
| | | | | | | Urea (mmol/L) > 7.3
| | | | | | | | Sodium (mmol/L) <= 137: Alive (13.08/0.05)
| | | | | | | | Sodium (mmol/L) > 137
| | | | | | | | | Urea (mmol/L) <= 7.5: Dead (4.48/1.14)
| | | | | | | | | Urea (mmol/L) > 7.5
| | | | | | | | | | Adj Calcium (mmol/L) <= 2.31: Dead (2.38/0.08)
| | | | | | | | | | Adj Calcium (mmol/L) > 2.31
| | | | | | | | | | | Creatinine (umol/L) <= 139: Alive (35.57/1.14)
| | | | | | | | | | | Creatinine (umol/L) > 139
| | | | | | | | | | | | Sodium (mmol/L) <= 142: Dead (2.94/0.09)
| | | | | | | | | | | | Sodium (mmol/L) > 142: Alive (3.11/0.01)
| | | | | | LVEDD (cm) > 6.91
| | | | | | | Bicarbonate (mmol/L) <= 30
| | | | | | | | Rate (ECG) (bpm) <= 90
| | | | | | | | | CRP (mg/L) <= 1.8
| | | | | | | | | | PCT <= 0.014: Alive (3.68)
| | | | | | | | | | PCT > 0.014
| | | | | | | | | | | Age (yrs) <= 68: Alive (2.38)
| | | | | | | | | | | Age (yrs) > 68: Dead (8.96/1.6)
| | | | | | | | | CRP (mg/L) > 1.8: Alive (10.75/0.1)
| | | | | | | | Rate (ECG) (bpm) > 90: Dead (4.19/0.12)
| | | | | | | Bicarbonate (mmol/L) > 30: Alive (7.25/0.09)
| | | | CRP (mg/L) > 3.9
| | | | | CT-proET1 <= 44
| | | | | | FEV1 <= 42.70618: Dead (5.92/0.03)
| | | | | | FEV1 > 42.70618
| | | | | | | MR-proADM <= 0.644867: Alive (55.18/18.26)
| | | | | | | MR-proADM > 0.644867: Dead (7.34/1.88)
| | | | | CT-proET1 > 44
| | | | | | FVC (L) <= 3.7
| | | | | | | BMI <= 27.274954
| | | | | | | | Cholesterol (mmol/L) <= 3.3: Alive (10.03/0.25)
| | | | | | | | Cholesterol (mmol/L) > 3.3
| | | | | | | | | Adj Calcium (mmol/L) <= 2.38
| | | | | | | | | | Bilirubin (umol/L) <= 14: Dead (16.94/2.92)
| | | | | | | | | | Bilirubin (umol/L) > 14
| | | | | | | | | | | Vitamin B12 (ng/L) <= 373: Alive (5.51/0.07)
| | | | | | | | | | | Vitamin B12 (ng/L) > 373: Dead (2.05/0.05)
|   |   |   |   |   |   |   |   |   Adj Calcium (mmol/L) > 2.38
|   |   |   |   |   |   |   |   |   |   Triglycerides (mmol/L) <= 2.2
|   |   |   |   |   |   |   |   |   |   |   Bilirubin (umol/L) <= 26
|   |   |   |   |   |   |   |   |   |   |   |   Urea (mmol/L) <= 5.2: Alive (31.56/4.16)
|   |   |   |   |   |   |   |   |   |   |   |   Urea (mmol/L) > 5.2
|   |   |   |   |   |   |   |   |   |   |   |   |   Chloride (mmol/L) <= 105
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Hct (fraction) <= 0.403
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Urea (mmol/L) <= 8.7: Alive (16.49/2.59)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   Urea (mmol/L) > 8.7: Dead (2.79/0.02)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Hct (fraction) > 0.403
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   White Cell Count (10^9/L) <= 8.1
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   PCT <= 0.02: Dead (8.0/1.97)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   PCT > 0.02: Alive (9.58/2.12)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   White Cell Count (10^9/L) > 8.1: Dead (7.52/0.05)
|   |   |   |   |   |   |   |   |   |   |   |   |   Chloride (mmol/L) > 105: Alive (5.72/0.06)
|   |   |   |   |   |   |   |   |   |   |   Bilirubin (umol/L) > 26: Dead (3.89/0.06)
|   |   |   |   |   |   |   |   |   |   Triglycerides (mmol/L) > 2.2: Alive (10.78/0.39)
|   |   |   |   |   |   |   BMI > 27.274954: Alive (256.49/42.7)
|   |   |   |   |   |   FVC (L) > 3.7: Alive (32.56/0.3)
Urea (mmol/L) > 9.5
|   PCT <= 0.045
|   |   Diastolic BP (mmHg) <= 98
|   |   |   Iron (umol/L) <= 12
|   |   |   |   Iron (umol/L) <= 5: Alive (4.98/0.42)
|   |   |   |   Iron (umol/L) > 5
|   |   |   |   |   Iron (umol/L) <= 8
|   |   |   |   |   |   LVEDD (Hgt indexed) <= 3.31
|   |   |   |   |   |   |   QT <= 421: Dead (2.3/0.47)
|   |   |   |   |   |   |   QT > 421: Alive (3.17/0.34)
|   |   |   |   |   |   LVEDD (Hgt indexed) > 3.31: Dead (18.4/2.54)
|   |   |   |   |   Iron (umol/L) > 8
|   |   |   |   |   |   Bilirubin (umol/L) <= 9: Alive (5.57/0.03)
|   |   |   |   |   |   Bilirubin (umol/L) > 9
|   |   |   |   |   |   |   CT-proAVP <= 18.5
|   |   |   |   |   |   |   |   Iron (umol/L) <= 9: Alive (4.95/0.51)
|   |   |   |   |   |   |   |   Iron (umol/L) > 9
|   |   |   |   |   |   |   |   |   Iron (umol/L) <= 11
|   |   |   |   |   |   |   |   |   |   White Cell Count (10^9/L) <= 8.2: Alive (11.09/1.71)
|   |   |   |   |   |   |   |   |   |   White Cell Count (10^9/L) > 8.2: Dead (7.63/1.05)
|   |   |   |   |   |   |   |   |   Iron (umol/L) > 11
|   |   |   |   |   |   |   |   |   |   MCV (fL) <= 89.6: Alive (5.42/0.13)
|   |   |   |   |   |   |   |   |   |   MCV (fL) > 89.6
|   |   |   |   |   |   |   |   |   |   |   White Cell Count (10^9/L) <= 7.7: Dead (5.86/0.45)
|   |   |   |   |   |   |   |   |   |   |   White Cell Count (10^9/L) > 7.7: Alive (2.93/0.44)
|   |   |   |   |   |   |   CT-proAVP > 18.5
|   |   |   |   |   |   |   |   Sodium (mmol/L) <= 138
|   |   |   |   |   |   |   |   |   ALT (iu/L) <= 15: Dead (4.13/0.12)
|   |   |   |   |   |   |   |   |   ALT (iu/L) > 15
|   |   |   |   |   |   |   |   |   |   FVC Predicted (L) <= 3.737692: Alive (3.34/0.02)
|   |   |   |   |   |   |   |   |   |   FVC Predicted (L) > 3.737692: Dead (2.54/0.44)
|   |   |   |   |   |   |   |   Sodium (mmol/L) > 138: Dead (16.12/1.05)
|   |   |   Iron (umol/L) > 12
|   |   |   |   Albumin (g/L) <= 36
|   |   |   |   |   Chloride (mmol/L) <= 96: Dead (6.84/0.23)
|   |   |   |   |   Chloride (mmol/L) > 96
|   |   |   |   |   |   Ferritin (ug/L) <= 155
|   |   |   |   |   |   |   TSH (mU/L) <= 2.2: Dead (11.95/2.38)
|   |   |   |   |   |   |   TSH (mU/L) > 2.2: Alive (10.38/3.09)
|   |   |   |   |   |   Ferritin (ug/L) > 155: Alive (12.07/2.1)
|   |   |   |   Albumin (g/L) > 36
|   |   |   |   |   MCV (fL) <= 93.1: Alive (55.65/6.51)
|   |   |   |   |   MCV (fL) > 93.1
|   |   |   |   |   |   Triglycerides (mmol/L) <= 2.7
|   |   |   |   |   |   |   Iron (umol/L) <= 24
|   |   |   |   |   |   |   |   Diastolic BP (mmHg) <= 82: Dead (24.56/6.9)
|   |   |   |   |   |   |   |   Diastolic BP (mmHg) > 82: Alive (3.17)
|   |   |   |   |   |   |   Iron (umol/L) > 24: Alive (3.04/0.28)
|   |   |   |   |   |   Triglycerides (mmol/L) > 2.7: Alive (9.05/0.79)
|   |   Diastolic BP (mmHg) > 98: Alive (15.93/0.49)
|   PCT > 0.045
|   |   Urea (mmol/L) <= 21.2
|   |   |   Rate (ECG) (bpm) <= 54: Alive (8.54/0.28)
|   |   |   Rate (ECG) (bpm) > 54
|   |   |   |   Haemoglobin (g/dL) <= 14.6
|   |   |   |   |   Creatinine (umol/L) <= 132: Dead (13.11/0.05)
|   |   |   |   |   Creatinine (umol/L) > 132
|   |   |   |   |   |   TSH (mU/L) <= 0.41: Alive (4.11/0.1)
|   |   |   |   |   |   TSH (mU/L) > 0.41
|   |   |   |   |   |   |   Adj Calcium (mmol/L) <= 2.45
|   |   |   |   |   |   |   |   Age (yrs) <= 77
|   |   |   |   |   |   |   |   |   Platelets (10^9/L) <= 328: Alive (12.12/1.34)
|   |   |   |   |   |   |   |   |   Platelets (10^9/L) > 328: Dead (2.03/0.02)
|   |   |   |   |   |   |   |   Age (yrs) > 77
|   |   |   |   |   |   |   |   |   Platelets (10^9/L) <= 272: Dead (15.0/1.0)
|   |   |   |   |   |   |   |   |   Platelets (10^9/L) > 272: Alive (3.07/0.07)
|   |   |   |   |   |   |   Adj Calcium (mmol/L) > 2.45: Dead (16.28/1.29)
|   |   |   |   Haemoglobin (g/dL) > 14.6: Alive (4.31/0.01)
|   |   Urea (mmol/L) > 21.2: Dead (21.48/1.12)
```

Appendix VII J48 pruned decision tree of the SVM and EM hybrid dataset

```
SystolicBP(mmHg) <= 233
|   MR-proANP <= 188: Alive (910.0/12.0)
|   MR-proANP > 188
|   |   UricAcid(mmol/L) <= 0.83
|   |   |   LVEDD(HgtIndexed) <= 3.55
|   |   |   |   AlkalinePhophatase(iu/L) <= 135
|   |   |   |   |   Chloride(mmol/L) <= 98
|   |   |   |   |   |   CRP(mg/L) <= 8.8: Alive (42.0/1.0)
|   |   |   |   |   |   CRP(mg/L) > 8.8
|   |   |   |   |   |   |   BSA(m^2) <= 1.741787
|   |   |   |   |   |   |   |   E <= 1.27: Dead (6.0)
|   |   |   |   |   |   |   |   E > 1.27: Alive (2.0)
|   |   |   |   |   |   |   BSA(m^2) > 1.741787: Alive (12.0/1.0)
|   |   |   |   |   Chloride(mmol/L) > 98: Alive (243.0/8.0)
|   |   |   |   AlkalinePhophatase(iu/L) > 135
|   |   |   |   |   Rate(ECG)(bpm) <= 84
|   |   |   |   |   |   Albumin(g/L) <= 39: Dead (6.0)
|   |   |   |   |   |   Albumin(g/L) > 39: Alive (2.0)
|   |   |   |   |   Rate(ECG)(bpm) > 84: Alive (10.0/1.0)
|   |   |   LVEDD(HgtIndexed) > 3.55
|   |   |   |   Albumin(g/L) <= 31
|   |   |   |   |   Calcium(mmol/L) <= 2.28: Dead (8.0)
|   |   |   |   |   Calcium(mmol/L) > 2.28: Alive (3.0)
|   |   |   |   Albumin(g/L) > 31
|   |   |   |   |   Urea(mmol/L) <= 17.6
|   |   |   |   |   |   Triglycerides(mmol/L) <= 1.75
|   |   |   |   |   |   |   MCV(fL) <= 91.6
|   |   |   |   |   |   |   |   Calcium(mmol/L) <= 2.3: Alive (32.0)
|   |   |   |   |   |   |   |   Calcium(mmol/L) > 2.3
|   |   |   |   |   |   |   |   |   Albumin(g/L) <= 38
|   |   |   |   |   |   |   |   |   |   MCV(fL) <= 88.5
|   |   |   |   |   |   |   |   |   |   |   QRSWidth(msec) <= 102: Alive (5.0)
|   |   |   |   |   |   |   |   |   |   |   QRSWidth(msec) > 102
|   |   |   |   |   |   |   |   |   |   |   |   UricAcid(mmol/L) <= 0.38: Alive (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   UricAcid(mmol/L) > 0.38: Dead (7.0)
|   |   |   |   |   |   |   |   |   |   MCV(fL) > 88.5: Alive (11.0)
|   |   |   |   |   |   |   |   |   Albumin(g/L) > 38: Alive (18.0)
|   |   |   |   |   |   |   MCV(fL) > 91.6
|   |   |   |   |   |   |   |   CT-proET1 <= 66.333748
|   |   |   |   |   |   |   |   |   Pulse(Exam)(bpm) <= 54: Alive (2.0)
|  |  |  |  |  |  |  |  |  Pulse(Exam)(bpm) > 54: Dead (12.0/1.0)
|  |  |  |  |  |  |  |  CT-proET1 > 66.333748
|  |  |  |  |  |  |  |  |  LVEDD(cm) <= 6.96
|  |  |  |  |  |  |  |  |  |  CRP(mg/L) <= 9.4: Alive (23.0)
|  |  |  |  |  |  |  |  |  |  CRP(mg/L) > 9.4
|  |  |  |  |  |  |  |  |  |  |  Height(Exam)(m) <= 1.69: Alive (4.0)
|  |  |  |  |  |  |  |  |  |  |  Height(Exam)(m) > 1.69: Dead (2.0)
|  |  |  |  |  |  |  |  |  LVEDD(cm) > 6.96
|  |  |  |  |  |  |  |  |  |  LeftAtrium(BSAIndexed) <= 2.34: Alive (4.0)
|  |  |  |  |  |  |  |  |  |  LeftAtrium(BSAIndexed) > 2.34
|  |  |  |  |  |  |  |  |  |  |  Cholesterol(mmol/L) <= 3.6: Alive (2.0)
|  |  |  |  |  |  |  |  |  |  |  Cholesterol(mmol/L) > 3.6: Dead (7.0)
|  |  |  |  |  |  Triglycerides(mmol/L) > 1.75: Alive (41.0/1.0)
|  |  |  |  |  Urea(mmol/L) > 17.6
|  |  |  |  |  |  TotalProtein(g/L) <= 68
|  |  |  |  |  |  |  MR-proANP <= 494.319106: Alive (3.0/1.0)
|  |  |  |  |  |  |  MR-proANP > 494.319106: Dead (8.0)
|  |  |  |  |  |  TotalProtein(g/L) > 68: Alive (3.0)
|  |  UricAcid(mmol/L) > 0.83: Alive (109.0)
SystolicBP(mmHg) > 233: Dead (404.0)
```

Appendix VIII *FP* Records of the four classifiers

No.	Decision tree	KDE	MLP	Beta distribution
1	1486	1460	1543	1460
2	1493	1479	1600	1479
3	1540	1486	1637	1486
4	1543	1493	1707	1493
5	1637	1520	1729	1520
6	1686	1536	1768	1536
7	1711	1540	1856	1540
8	1716	1543	1858	1543
9	1776	1544	1863	1544
10	1784	1558	1867	1558
11	1850	1591	1876	1591
12	1861	1600	1878	1600
13	1863	1615	1888	1615
14	1865	1637	1890	1637
15	1867	1671	1898	1671
16	1869	1686	1899	1686
17	1874	1690	1902	1690
18	1878	1694	1910	1694
19	1879	1707	1911	1707
20	1880	1708	1927	1708
21	1888	1710	1931	1710
22	1890	1711	1937	1711
23	1898	1712		1712
24	1902	1716		1716
25	1927	1725		1725
26	1939	1729		1729
27		1759		1759
28		1768		1768
29		1771		1771
30		1776		1776
31		1778		1778
32		1780		1780
33		1784		1784
34		1786		1786
35		1795		1795
36		1797		1797
37		1800		1800
38		1848		1848
39		1850		1850

40		1855		1855
41		1856		1856
42		1858		1858
43		1859		1859
44		1861		1861
45		1863		1863
46		1865		1865
47		1867		1867
48		1868		1868
49		1869		1869
50		1872		1872
51		1873		1873
52		1874		1874
53		1876		1876
54		1877		1877
55		1878		1878
56		1879		1879
57		1880		1880
58		1883		1883
59		1886		1886
60		1887		1887
61		1888		1888
62		1890		1890
63		1892		1892
64		1896		1893
65		1897		1896
66		1898		1897
67		1899		1898
68		1902		1899
69		1906		1902
70		1910		1906
71		1911		1910
72		1914		1911
73		1923		1914
74		1927		1923
75		1929		1927
76		1931		1929
77		1936		1931
78		1937		1936
79		1939		1937
80		1943		1939
81				1943