ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

PhD THESIS

Jale BEKTAŞ

USE OF DATA MINING TECHNIQUES TO DETERMINE PRESENCE OF CORONARY ARTERY DISEASE AND DERIVING A RISK SCORE BY EMPLOYING RISK FACTORS

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

ADANA, 2017

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

USE OF DATA MINING TECHNIQUES TO DETERMINE PRESENCE OF CORONARY ARTERY DISEASE AND DERIVING A RISK SCORE BY EMPLOYING RISK FACTORS

Jale BEKTAŞ

PHD THESIS

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

We certify that the thesis titled above was reviewed and approved for the award of degree of the Doctor of Philosophy by the board of jury on ../../2017.

…….....	...…….....	…….....
Assoc. Dr. Turgay İBRİKÇİ	Prof. Dr. S. Ayşe ÖZEL	Assoc. Prof.Dr. Sami ARICA
SUPERVISOR	MEMBER	MEMBER

…….....	……..……
Asst. Prof. Dr. İrem ERSÖZ KAYA	Asst. Prof. Dr. Lütfü SARIBULUT
MEMBER	MEMBER

This PhD Thesis is written at the Department of Institute of Natural and Applied Sciences of Çukurova University.
Registration Number:

Prof. Dr. Mustafa GÖK
Enstitü Müdürü

ABSTRACT

PhD THESIS

USE OF DATA MINING TECHNIQUES TO DETERMINE PRESENCE OF CORONARY ARTERY DISEASE AND DERIVING A RISK SCORE BY EMPLOYING RISK FACTORS

Jale BEKTAŞ

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

Supervisor : Assoc. Prof. Dr. Turgay İBRİKÇİ
Co-Supervisor : Prof. Dr. İ. Türkay ÖZCAN
Year: 2017, Pages: 101
Jury : Assoc. Prof. Dr. Turgay İBRİKÇİ
: Prof. Dr. S. Ayşe ÖZEL
: Assoc. Prof. Dr. Sami ARICA
: Assist. Prof. Dr. İrem ERSÖZ KAYA
: Assist. Prof. Dr. Lütfü SARIBULUT

This study focuses primarily on the problems of collaborative classification with missing data on Coronary Artery Disease (CAD) by applying machine learning algorithms and proposes a risk score prediction system consisting of a 4-classes dataset. Three imputation methods are applied: K-means, multilayer perceptron (MLP), and self-organizing maps (SOMs). The MLP imputation method is obviously the best method among those investigated with the metric values for sensitivity (0.90), and for specificity (0.18).

Dataset imputed with MLP method is employed by transforming into a 4-classes structure. Using the feature selection and the sampling methods with the NN substantially improves the evaluation metrics. The results before the pre-process operations were detected as follows; 72.3% accuracy; after the operations, 84.1% accuracy were achieved with 0.84 sensitivity 0.94 specificity.

This study also presents a hybrid classification procedure that uses Support Vector Machine (SVM) with LR on two distinctive datasets. The results show that the hybrid approach allows developing an efficient algorithm, which solves the problem with all imbalanced dataset training at one time.

Key Words: CAD, Machine Learning Algorithms, Imputation Procedures, Imbalance datasets, Hybrid algorithms

ÖZ

DOKTORA TEZİ

VERİ MADENCİLİĞİ TEKNİKLERİNİN KULLANILARAK KORONER ARTER HASTALIĞININ VARLIĞININ BELİRLENMESİ ve RİSK FAKTÖRLERİNİN KULLANILMASIYLA BİR RİSK SKOR SİSTEMİNİN OLUŞTURULMASI

Jale BEKTAŞ

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

Danışman : Doç. Dr. Turgay İBRİKÇİ
İkinci Danışman: Prof. Dr. İ. Türkay ÖZCAN
Yıl: 2017, Sayfa: 101
Jüri : Doç. Dr. Turgay İBRİKÇİ
: Prof. Dr. S. Ayşe ÖZEL
: Doç. Dr. Sami ARICA
: Yrd. Doç. Dr. İrem ERSÖZ KAYA
: Yrd. Doç. Dr. Lütfü SARIBULUT

Bu çalışma, öncelikle Koroner Arter Hastalığı (KAH) hakkında eksik veriler içeren bir veri seti üzerinde makine öğrenme algoritmaları uygulanarak sınıflandırma sorunlarına odaklanmakta ve dört sınıflı bir veri kümesinden oluşan bir risk puan tahmin sistemini önermektedir. Üç adet veri seti doldurma yöntemi uygulanır: K-ortalamalar, çok katmanlı perseptron (ÇKP), ve öz örgütlenmeli harita (ÖÖH). ÇKP ile test edilmiş ÇKP doldurma metodu, 0.90 duyarlılık ve 0.18 özgüllük ile bu çalışmada araştırılan diğer yöntemler arasındaki en iyi metottur.

ÇKP metoduyla doldurulmuş veri seti dört sınıflı bir yapıya dönüştürülerek ele alınır. YSA yönteminin yüksek seviyede örnekleme ve Relief-f ile birlikte kullanımında önceki sonuçlar %72,3 doğruluk değerine sahipken işlemlerden sonra %84,1 doğruluk, 0.84 duyarlılık ve 0.94 özgüllük değerlerine ulaşılmıştır.

Bu çalışma ayrıca iki farklı veri kümesinde LR ile DVM kullanan bir karma sınıflandırma prosedürü sunmaktadır. Sonuçlar, karma yaklaşımın, problemi bir defada tüm dengesiz veri kümesi eğitimiyle çözen etkili bir algoritma geliştirmeye izin verdiğini göstermektedir.

Anahtar Kelimeler: Koroner Arter Hastalığı, Makine Öğrenmesi Algoritmaları, Kayıp veri doldurma prosedürleri, Dengesiz verisetleri, Hibrid sistemler

ACKNOWLEDGEMENTS

Firstly, I would like to express my great thanks to my supervisor Assoc.Prof.Dr.Turgay İBRİKÇİ for the continuous support of my dissertation study, for his patience, motivation and his valuable time. I would like to express my thanks to my co-supervisor Prof.Dr.İ.Türkay ÖZCAN for the contributions during the data collection phase and for the valuable information on the subject. I would also like to thank my thesis committee members, Prof Dr. Selma Ayşe ÖZEL, Assoc. Prof. Dr. Sami ARICA for their valuable advices during this study.

Finally, I especially want to thank my parents, my husband and my daughter for their continuous support, understanding and encouragement.

CONTENTS **PAGE**

LIST OF TABLES **PAGE**

LIST OF FIGURES **PAGE**

LIST OF SYMBOLS AND ABBREVIATIONS

Nx	:MVs sample
M1	:Complete part of the CAD dataset(*set 1*)
M2	:The part of the CAD dataset has MVs(*set 2*)
C_i	:The i^{th} cluster
c_i	:The centroid of cluster C_i
T	:Target vector
W	:Wight vector of BMN
Θ	:Restraint due to the distance of BMN

1. INTRODUCTION

Cardiovascular disease (CVD) which is a class of disorders affecting the hearth and blood vessels plays major role for incidence of the disease and death rate (mortality) over the world. Studies show that mortality due to cardiovascular disease will increase from 28.9% to 36.3% between 1990 and 2020 (Hennekens, 1998). It is accounted for nearly 600,000 deaths, at the same time the reason for one in every four deaths in the United States (Heron et al., 2004). However, by Coronary artery disease (CAD), nearly 380, 000 individuals died in the U.S. By 2030 nearly the numbers of 23.6 million individuals are predicted to die mainly from heart disease and stroke (WHO, 2013). Early diagnosis is missed in 2% of patients and short period mortality should be high of these patients.

Under the leadership of the Turkish Society of Cardiology, TEKHARF (Turkish Adult Risk Factors for Heart Disease) Study has been carried out since 1990. According to the 12-years follow-up estimations of TEKHARF shows that there are nearly 2.0 million coronary heart patients in Turkey and 160 thousands of individuals died in per year.

TEKHARF study finds that mortality of coronary heart disease for adult men 5.2 per thousand and women 3.2 per thousand for per year. One of every 8 deaths is undetermined, between the other diseases coronary heart disease mortality gets the first place with 42.5 % percentage (Onat et al., 2003).

Because of the extended average life span and development of treatment methods, number of older individuals and patients who are exposed to recurrent cardiovascular events increase. After the diagnosis of coronary heart disease, medical, surgical and invasive treatment methods are applied but these methods require quite a high cost. A significant portion of the adult population are affected in middle age and early old age by this disease and thus financial burden of this situation increases (Kültürsay, 2001). Instead of treatment facilities that are carried on with an extremely high cost, clinical decision support systems (CDSS) become

an important work area for clinicians to enhance health care in the last few decades. CDSS supply improved efficiency, accuracy, benefit-cost control and support the person who is responsible for the clinical decisions with an advice (Berner, 2007). The based on this information the applicability of risk factors of CAD challenge us to improve a CAD diagnosis support system by using different computational learning methods.

1.1. Coronary Artery Disease

The heart provides its own blood circulation from the coronary arteries. Two large coronary arteries are separated from the aorta. These arteries and branches feed all parts of the heart muscle with oxygen-enriched blood represented in Figure 1.1.

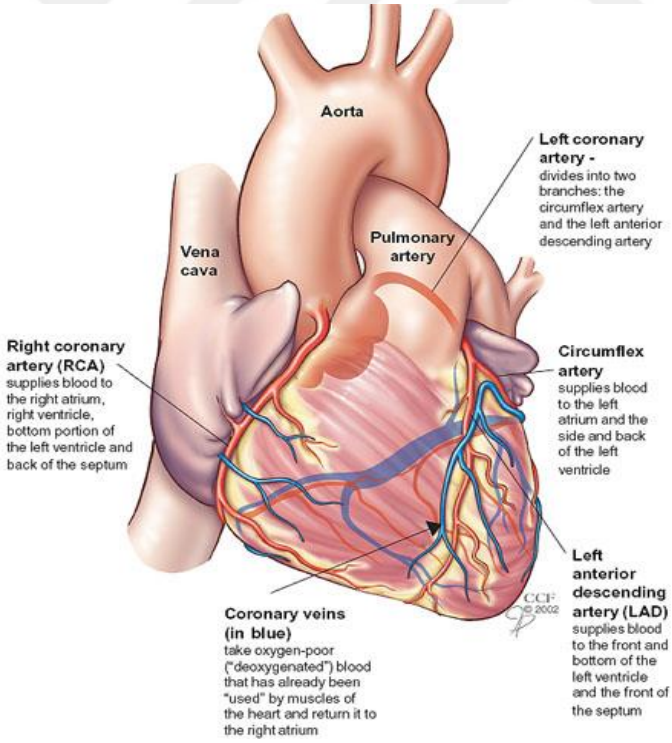

Figure 1.1.Two major coronary arteries and their branches (NLM, 2013)

Cholesterol, calcium, oxygen and nutrients are carried by the blood and accumulate on the wall of the coronary artery. Narrowing of the artery results insufficient blood flow through the vessel and blocks blood supply to the heart muscle. These deposits cause fatty plaque generally over the branch points and the inner walls of the arteries. Narrowing of these arteries as well as of other middle-big arteries such as aorta and cerebral arteries are called Atherosclerosis which causes CAD. Figure 1.2. Shows narrowed areas in coronary arteries.

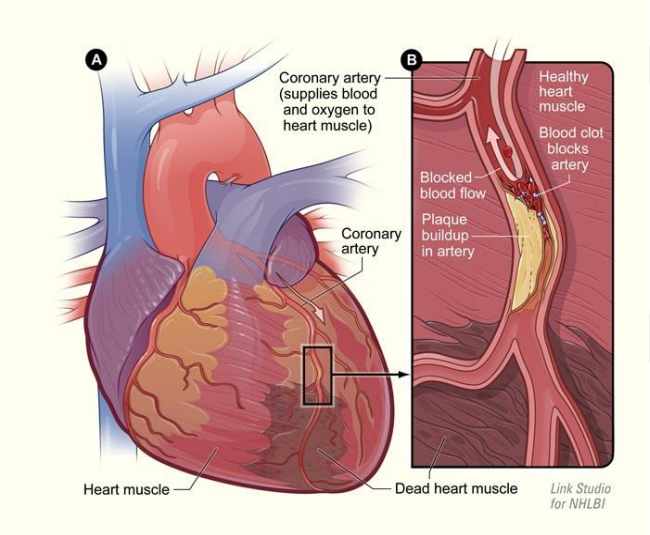

Figure 1.2. A blockage in the coronary artery. Blockage may cause the affected heart muscle to die (NLM, 2013)

Clinical in Atherosclerosis, depends on the content of the plaque. How much plaque richer by lipid, has extremely component and smooth muscle roof is thinner it is more disposed to cause clinical events (Van DER Wal et al., 1999).

Generally, later in life on the branch points of the arteries in some regions thoroughly became evident widely lesions are seen that require some procedural refinements such as stent or bypass.

1.2. Atherosclerotic Risk Factors

Defining risk factors become an important issue in treatment for asymptomatic individuals and also preventing cardiac disease who are exposed to recurrent cardiovascular events. Risk factors classified according to National Cholesterol Education Program (NCEP) ATP III Guidelines published in 2001.

There are many factors which cause a higher risk for CAD can be examined under two headings. These are Lipid risk factors and no lipid risk factors. No lipid risk factors consist of modifiable and no modifiable risk factors. Coronary artery disease risk factors (NCEP, 2002)

1. **Lipid risk factors**

   Low HDL Cholesterol (<40 mg/dl)
   High LDL Cholesterol (≥130 mg/dl)
   Total Cholesterol (≥130 mg/dl)
   Triglyceride

2. **Nonlipid risk factors**

   A. Modifiable risk factors

      a. Hypertension (High blood pressure ≥140/90 mmHg or antihypertensive usage)
      b. Smoking
      c. Diabetes Mellitus (DM)
      d. Obesity
      e. Physical inactivity
      g. Thrombogenic / hemostatic status

B. Nonmodifiable risk factors

a. Age (men ≥45, women ≥55 or early menopause)

b. Gender

c. Family history

Abnormal cholesterol (high total cholesterol and high LDL just the opposite low HDL) increases cardiovascular risk. LDL causes cholesterol deposition on inner walls of the arteries. High LDL plays a role in every term of atherosclerosis thus NCEP determines to reduce LDL cholesterol for lipid-lowering treatment as primary aim (NCEP, 2002). HDL cholesterol protects arteries blocking fatty plaque. When HDL>60 mg/dl, a risk factor is removed from the calculation of risk because high HDL cholesterol reduces the CAD risk. TEKHARF study shows that Total Cholesterol/HDL Cholesterol percentage is the best lipid predictor of coronary heart disease for Turkish people. According to the result of the same study also shows that 2 units' increase increases the CAD and death risk with 68% percentage independently (Onat et al., 2003).

Hypertension is a very important risk factor for CAD. Hypertension is responsible with 35% possibility for all atherosclerotic cardiovascular events. Individuals with hypertensives are under 2-3 times more risk than individuals with normotensives (Hambrecht et al., 2000). Hypertension increases acute myocardial infarcts 2-3 times more. High blood pressure can be controlled by appropriate medication.

Diabetes Mellitus presence is represented as risk equivalent of CAD. DM is an independent risk factor and increases CAD risk increases to 2 times for women and 4 times for men (NCEP, 2002). DM increases cardiovascular diseases with 70% possibility independent from systolic blood pressure, central obesity and dyslipidemia.

Obesity is defined one of the major risk factors by American Heart Association (AHA) (Eckel et al., 1998). Obesity has become a health issue of

which prevalence grows worldwide and reaches epidemic proportions (Consultation, 2000). According to TEKHARF study the prevalence of obesity are 21% over 30 years old for men and 43% for women. In our country body mass index increased 1.26 kg/m2 for women, 1.29 kg/m2 for men in last 10 years. These results show us tend to putting on weight of our society.

Interacting with other risk factors smoking causes the increase of risk to 2-3 times. There is no any evidence for filters to reduce the effect of damage of cigarette (Castelli et al., 1981). Deaths due to cardiac disease are 2.7 times higher for men and 4.7 times higher for women in the case of smoking. Cigarette dropouts are associated with a reduced risk of CAD. This is the most preventable cause of mortality (O'Rourke et al., 2001).

Physical inactivity is an independent risk factor which increases the risk two times for both men and women and decrease cardiovascular function capacity. There is a relationship among cardiovascular death and weekly dose of exercise.

Incidence and prevalence of CAD increases in older age, thus age might be considered the most important risk factor (O'Rourke et al., 2001). Despite early lesions of atherosclerosis occur in childhood death rate of individuals of CAD increase in older age in each decade. From the age of 40 until the age of 60 in the incidence of myocardial infarction has increased more than 5 times (Heron et.al., 2009). People over 45 years for men and 55 years for women assessed as a major risk factor.

In terms of CAD risk, there is a relationship between early detection of the disease and first degree relatives diagnosed with CAD. This risk factor, which consists of genetic predisposition, continues even if other risk factors are absent. The lowest age limit was set at 65 for women and 55 for men. When any individual younger than these limits and diagnosed with CAD, then the risk is assumed that the risk factor is included in the family story. The first-degree relatives with a young individual diagnosed with CAD is assumed as the strongest risk factor (Bassuk and Manson, 2008).

1.3. Missing Value Imputation Problem

Real life datasets often contain missing values and especially this is a common problem in clinical datasets (Allison, 2001). When specific elements of dataset which consist of different types of variables are missing, the data are called incomplete. There are many reasons for the existence of missing values. Two types of databases are available in medical domain (Tsumoto, 2000). The first one is a dataset for which researchers deal with manually data collection issues especially when this is a part of a clinical trial. Another type is large data sets obtained from hospital information systems. This type of database systems are not stored for a special purpose and may not be proper for a specific research such as any clinical decision support systems for diagnosing CAD.

In this study, the CAD dataset is collected manually and thus, a suitable method for solving the missing value imputation problem has been investigated. Machine learning approaches as an alternative to the traditional mean method have been used.

1.4. Imbalanced Dataset Problem

In the field of cardiology, a technology that could predict significant CAD from individuals who are diagnosed with coronary angiography would be helpful. The need to determine the stenosis level and presence of CAD is greater than 50%, leading to new criteria for more specific classification of CAD. After angiography, as for the location of the lesion, one of the three major coronary arteries; Left Anterior Descending (LAD), Left Circumflex (LCX), or Right Coronary Artery (RCA) (stenosis diameter >50% of vessel diameter) is considered as having CAD (Alizadehsani et al., 2013). In case of presence of CAD, the size of the stenosis can be matched with the risk score concept. This situation opens the way to classify the disease with different risk scores. Thus, more than 2 classes cause a dataset with imbalanced structure.

1.5. The Requirement of Hybrid System Design

Traditionally, Support Vector Machines (SVMs) are used in classification and pattern recognition (Maji, 2013), which is also the case in Neural Networks and other learning algorithms. However, a major limitation is that when the training sets are imbalanced, SVM cannot perform satisfying results by the chosen kernel. To overcome this limitation, we propose a hybrid method which uses SVM linear kernel with Logistic Regression (LR) in different manner. The hybrid approach enables us to develop an efficient algorithm, which solves the problem with all imbalanced dataset training at one time. For these reasons, the development of different hybrid algorithms is always a popular research area and over the years, wide spectrum of algorithms and methods has been proposed. To measure the success of hybrid system, distinctive datasets are expected to help at the classification test phase.

As it is stated before, MVs imputation, dealing with imbalanced datasets and hybrid system designs are treated as discrete problems and are solved sequentially. The most suitable output that is obtained as a result of one stage must be used for the other stage. On the other hand, it should be noted that it is a good argument for widen the problem at every step and getting the optimal solutions for the examined problems sequentially in this thesis. Moreover, the hybrid approach could be quite successful when the computational power becomes sufficient to implement this idea in practice. As a result, this thesis's main goal is to develop different MVs imputation procedures to impute datasets and to solve imbalanced dataset problem by using optimal imputed dataset obtained from previous stage. Furthermore, we propose a hybrid approach to use alternatively for classification applications.

2. PREVIOUS RESEARCH

2.1. Computational Methods for MVs Imputation

In recent years, MVs imputation problem has been widely studied by researchers in several clinical datasets such as CAD dataset and in other domains. Due to the nature of the MVs, different number of attributes may consist of different number of blank entries which depends on the specificity of the patient. Thus, there may be not enough robust information recorded for every individual. Existence of missing values usually requires the study to have a preprocessing step. Huge numbers of sample contain missing values for relatively many attributes forms a complexity. Two types of issues are usually encountered with so many missing values when performing data mining (Wang et al., 2010): (1) the loss of efficiency; probable problems in data analysis and processing, and (2) statistical analysis can be biased if incomplete cases are excluded from the analysis.

Three major approaches are usually used to deal with MVs (Farhangfar et al., 2008). The simplest way to deal with missing values is to extract the samples that contain the missing values. This method may be useful when dataset contains relatively small number of samples with missing values. Another approach is to fill in missing values with a new global value that is predetermined. This approach requires encoding the blank entries into a new numerical value. It may be not practical to fill huge number of samples contain missing values for relatively many number of attributes.

There are different types of missing value imputation mechanism that depends on the reason of the lack of MVs (Little et al., 2002). Reasons for MVs are commonly classified as: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). These approaches depend on the distributions in a dataset. In statistical analyses, data-values in a data set are MCAR when the events leading to the disappearance of any data material are independent. This situation is valid when both observable and non-observable

parameters are taken into account and appear entirely random. MAR occurs when the MVs situation is related to a particular variable. When there is a relation between the missing value and the missing cause of the variable and this occurs due to a certain cause, it is expressed as MNAR.

The most preferred traditional methods are deleting case, mean/mode value imputation, and other statistical methods (Little et al., 2002). However, every method may not be suitable for every missing data problem, so it is important to choose appropriate method depending on the nature of the study. Nowadays, fuzzy unordered rule induction algorithm (FURIA) imputation (Rahman et al., 2012) and machine learning (Gajawada et al., 2012; Richman, 2007) is used alternatively as a method for MVs imputation in several clinical and other incomplete datasets. Machine learning algorithms such as multilayer perception (MLP) (Richman et al., 2009), self-organizing maps (SOM) (Jerez et al., 2010; Mingoti et al.,2006), decision tree (DT) and k-nearest neighbors (KNN) (Gajawada et al., 2012) have been used as methods of generating missing values in different areas (Poolsawad et al., 2011). Machine learning algorithms are seen to have better performance when compared with other traditional techniques (Rahman et al., 2013). Comprehensive surveys on the MVs imputation problem are given in Table 2.1. The research presented in this thesis principally inspired from the second row.

Table 2.1. Significant studies on the solution for MVs imputation problem

Year	Authors	Objective	Solutions Method	Dataset
2002	Little et al.	Presents survey on traditional statistical methods for MV imputation	Deleting case Mean/ Mode	An Example Dataset
2006	Mingoti et al.	Machine Learning imputation	SOM, Fuzzy C- means K-means	An Example Dataset
2007	Richman	Multiple imputation by NN	SVM MLP	Clinical dataset
2008	Farhangfar et al.	Gauges Impact of imputation of missing values on classification	C4.5 and Naive-Bayes KNN SVMs	Spl, Tic dataset
2009	Richman et al.	MVs imputation by NN	SVM NN	Example dataset
2010	Jerez et al.	Proposes machine learning techniques led to a significant enhancement of prognosis accuracy	LRA MLP SOM KNN	Real Breast Cancer data
2011	Poolsawad et al.	Propose feature selection and Mvs imputation will enhance THE prediction model.	KNN NN EM	Coronary Dataset
2012	Rahman and Davis	Improves classification performance by imputing dataset	FURIA, Fuzzy Rules, K-Mean	Cardiovascular data
2012	Gajawada et al.	Presents MVs imputation procedure	KNN	Clinical datasets from UCI
2013	Rahman et al.	Propose a method for MV imputation	FURIA K-means	Cardiovascular dataset

- LRA Logistic Regression Analysis, NN Neural Networks, MLP Multi-Layer Perceptron, SVM Support Vector Machines
, SOM Self Organizing Maps, KNN K-Nearest Neighbors, FURIA Fuzzy Unordered Rule Induction Algorithm.

2.2. Comparative Methods on Imbalanced Medical Datasets

The need to determine the stenosis level and presence of CAD is greater than 50%, leading to new criteria for more specific classification of CAD (Alizadehsani et al., 2013). In case of presence of CAD, the size of the stenosis can be matched with the risk score concept. This situation opens the way to classify the disease with different risk scores. In this sense, the purpose must be to determine one of these scores that are determined as 0 to 39 as low-risk, 40 to 69 as moderate, 70 to 85 as high; and >=86 as very high risk (Ergun et al., 2004). As the degree of stenosis increases to four, it becomes inevitable to face the problem of imbalanced data. For CAD dataset, the data shows an imbalance on the status of patients which are classified as “moderate” and “high”.

A dataset is represented as imbalanced if the classification categories are not equally distributed inside. These datasets contain more samples in specific classes from the rest of classes. In this scenario, many classifiers can have good accuracy results over the major classes because the learning model has a natural tendency for these major classes during the training stage. Otherwise, accuracy rates decrease for minority classes.

Many researchers who have studied CAD risk factors by using statistical methods such as LRA have obtained significant results (Larifla et al., 2014; Lu and Yang, 2012). In the past works, linear and nonlinear logistic regression methods were used to classify CAD and predict CAD level using risk factors (Cassar et al., 2009; Nakazato et al., 2012). Traditional and novel risk factors and the stenosis degrees in young patients with CAD were determined by the dependent variable. Here, the patients are diagnosed with carotid artery stenosis using the NN and the LRA (Ergun et al., 2004). In another study, a prediction model for renal artery stenosis was conducted and multiple LRA including with all the binary and categorical variables was performed (Lee et al., 2014). Recently, NN is used for the classification of several diseases; and some of medical diagnosis studies have shown the estimation stability of an NN model. However, NN model enable

development of simple and practical applications in many contexts like CAD diagnosis (Colak et al., 2008). Furthermore, there are studies to bring solutions on class imbalance problem. Some of studies investigate methods of sampling, over-sampling and under-sampling to assess the performance of classification methods such as NN, Radial Basis Functions (RBF), Decision Tree, Support Vector Machines (SVM) (Poolsawad, 2014). Other studies bring solution about class imbalance problem by using sampling methods on clinical datasets. They use over-sampling, SMOTEboost which uses boosting and sampling together, subsampling methods. The main idea of these studies is to increase classification performance and use classification algorithms such as decision tree, SVM, and FURIA (Longadge, 2013; Rahman and Davis 2013). An another study (Natarajan et al., 2013), employs Statistical Relational Learning algorithms for predicting future coronary artery calcification in year 20 for identifying risk factors at the early adult stage (patients between 38 and 50 years old).

An integer risk score to predict the risk of death over a 1-year period after percutaneous coronary intervention was developed (Maluenda et al., 2010). This study is built using LR and bootstrap methods using records of 6,932 consecutive patients. 85 of 738 patients with stable/unstable angina pectoris died at 1 year, and the model predicted 79.3% of those events.

Comprehensive surveys on the class imbalance problem and risk score prediction system are given in Table 2.2. The research presented in this thesis principally inspired from the first, fifth, and ninth row.

Table 2.2. Significant studies for the class imbalance problem and risk score prediction systems

Year	Authors	Objective	Solutions Method	Dataset
2004	Ergun et al.	Classification of two categories/4 categories	LRA, NN	Carotid artery stenosis data
2008	Colak et al.	Two categories classification	NN-MLP	CAD dataset
2009	Cassar and Gersh	Decision support system for CAD or CABG	LRA	MRI dataset
2010	Maluenda et al.	Generate risk score	LR and bootstrap methods	Acute CAD
2012	Nakazato et al.	Measuring relationship between EFVi m using non-contrast CT	Univariable LRA	Cardiovascular risk factors, EFVi with CT
2013	Rahman and Davis	Bring solution about class imbalance problem by using sampling methods	Decision Tree, FURIA, SMOTE, Under-sampling	Cardiovascular Data
2013	Alizadehsani et al.	Prediction the stenosis of arteries	Data mining	CAD dataset
2013	Longadge and Dongre	Study on multiclass imbalance problem	SVM Sampling-SMOTE Feature Selection	Cancer dataset
2013	Natarajan et al.	Predicts future coronary artery calcification	Statistical learning, Decision Tree, SVM	CAD dataset
2014	Lee et al.	A scoring system for predicting significant RAS	Multiple LRA	CAD dataset
2014	Larifla et al.	Generate risk score	Multivariable LRA	CAD dataset
2014	Poolsawad et al.	investigates methods of sampling, over-sampling and under-sampling	MLP, RBF, Decision Tree, SVM, RF	Sample clinical dataset

- LRA Logistic Regression Analysis, NN Neural Networks, MLP Multi-Layer Perceptron, SVM Support Vector Machines, RBF Radial Basis Functions, CABG Coronary Artery Bypass Grafting, MRI Magnetic Resonance Imaging, EFV Epicardial Fat Volume, RAS Renal Artery Stenosis

2.3. Hybrid Classification Approach Using SVM with LR

The use of the kernel concept (Maji, 2013) in many machine learning problems, in which the SVM is the most prominent one, has become common. Traditionally, SVMs are used in classification and pattern recognition, which is also the case in Neural Networks and other learning algorithms. When the SVM model is being formed, it is necessary to optimize some important parameters that may influence its generalization performance (Chapelle, 2002). For this reason, the hybrid classification approach (HCA) has been designed to discover specific features and calculate high-dimensional inputs in an accurate manner. With this approach, the purpose is to perform the classification and automatic variable selection processes in a simultaneous manner. In situations where the training dataset is imbalanced, when it has a top-closed geometrical structure, and where it is not separated with the defined kernel function in a linear manner, it may not show adequate performance. Generally, this problem is resolved by choosing another kernel function in an empirical way.

There are several studies on Kernelized Logistic Regression Method (Elbashir, 2013). In Machine Learning applications, weighting process (Wang et al., 2003) is a popular method with the purpose of using more than one classifier together. Moreover, there are several other studies on the ensemble methods that are used to train imbalanced datasets. However, when weighing the classifiers that have reached different accuracy rates, there have always been uncertainties. Using LR for integrating more than one classifier, finding optimal weights based on statistical modeling theory will relieve us from this problem. HCA must be applied to distinctive datasets such as the Wisconsin Diagnostic Breast Cancer (WDBC) benchmark data from University of California (UCI) repository of machine learning and so, is planned to be tested for only binary datasets for now.

An independent SVM classifier may be operated for each sub-set, which is far from the imbalance form of the whole dataset. In this way, better accuracy estimations may be achieved in the sub-sets when compared with performing all

dataset training (Chang, 2003; Wang et al., 2007). In further stages of this method, the thing that has to be done is converting the results of the classification, which was made with the sub-sets, into one single and collective result that may be decided on. For this purpose, the results of the local classification performed with sub-datasets may be merged with Logistic Regression (LR) technique. Here, the calculation of the HCA will be performed by SVM, and in case the size of the dataset is big, this situation will pose a benefit. In the study, all possible outcomes for all samples must be calculated; however, there must not be any situations that will increase the complexity of the calculation. Comprehensive surveys on the Hybrid system problem are given in Table 2.3. The research presented in this thesis principally inspired from the second and sixth row.

Table 2.3. Significant studies on the design for Hybrid Systems

Year	Authors	Objective	Solutions Method
2002	Chapelle	Uses an ensemble system for kernel selection	Sample Toy data
2003	Chang	Develops an effective hybrid algorithm by SVM with LR	Breast Cancer, Pen digit data
2003	Wang et al.	Use weighting process to assess ensemble classifiers	Credit Card Fraud Data
2007	Wang et al.	Develop an Hybrid SVM that computes regularized solution	Real Microarray data
2013	Maji et al.	Additive kernel applications	Handwritten digits, video retrieval
2013	Elbashir	Proposes a hybrid system which combines (SVMs) and logistic regression (LR) for prediction	BT426, BT547, and BT823 datasets

- SVM Support Vector Machines, LR Logistic Regression

2.4. Motivation of This Thesis

The efficiency of the classification algorithms depends on different factors. Often the preprocessing step on datasets has higher impact on the quality of the solution than using the search algorithm alone. Real datasets usually have MVs and the researchers need to deal with MVs imputation methodology during the preprocessing step. For example, fuzzy unordered rule induction algorithm

(FURIA) imputation (Rahman et al., 2012) and machine learning (Gajawada et al., 2012; Richman, 2007) is used alternatively as a method for MVs imputation in several clinical and other incomplete datasets.

A dataset is represented as imbalanced if the classification categories are not equally distributed inside. These datasets contain more samples in specific classes from the rest of classes. In this scenario, as the numbers of classification categories increase more than two, it becomes inevitable to face the problem of imbalanced data. For CAD dataset, the size of the stenosis is matched with the risk score concept and four risk score are determined to classify CAD stenosis. The dataset showed an imbalance on the status of patients which are classified as "moderate" and "high". Two classification methodologies, NN and LRA are used in this thesis. T-test and Relief-f feature selection methods and over-sampling and stratified sampling are used, so as to demonstrate the performance of classification methods. This second part of the thesis includes implementations of these classification algorithms and their combinations with sampling and feature selection methods for unbalanced CAD dataset.

It is known that the powerful affect in classification of machine learning algorithms, whereas, hybrid systems are preferred to overcome imbalanced training sets. We have to confront the problem of whether a hybrid model can be created by testing existing datasets to overcome this limitation. It is also important to find a solution that can use both machine learning and statistical method. The experience of having worked on the similar issues and methodologies in previous phases of the thesis is increased. In third main part of the thesis answers to parameter selection, model optimization and process flow questions are sought by developing a flexible hybrid procedure.

3. BACKGROUND

3.1. Artificial Neural Networks

Artificial Neural Networks (ANNs) were developed as an information processing system with an inspiration of biological neural networks (White, 1989). ANNs consist of various forms of connections of artificial neural cells each other and are developed by layers. ANNs are used for regression, data classification, pattern recognition, estimation, process control, decision support systems etc. with a learning process. An ANN can learn by examples or by itself.

The most significant characteristics of ANNs are connected neurons, defining weights between connections and transfer function. Every neuron has an activation level. This level identifies inputs which comes to that neuron and sends a signal to other neurons. These signals would be input values for the neurons which came on. Input information is carried on these signals and there is a learning process using these weights, thus network would decide when reaches the optimal weights. Network makes a generalization and with this way ANNs are preferred to solve very large number of problems. Learning process is carried out by either supervised or unsupervised methods. In supervised method, input values correspond to output values and system is constructed to reach desired output by changing weight values, but unsupervised learning categorizes input data into groups without defining the desired output.

The single input neuron forms the basic building block of neural networks. Figure 3.1 shows block neural network architecture with a single-input neuron.

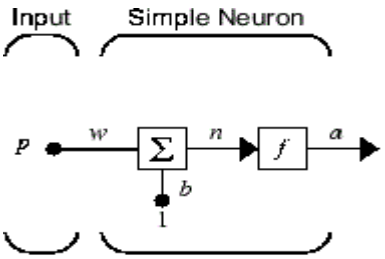

Figure 3.1. The single input neuron for neural networks

$$a = f(wp + b) \tag{3.1.}$$

The product wp is obtained by multiplying the scaler variables weight w *and* input p. To compute the net input n, wp is added to the scalar bias b. Finally, the transfer function f is applied to the net input, which produces the output a. the network architecture is given in Figure 3.2.

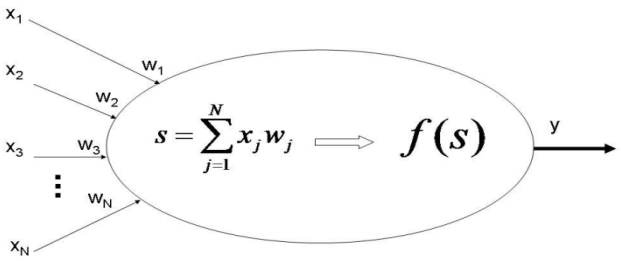

Figure 3.2. Building block of an artificial neuron with vector input (without bias weight)

A neuron with a single N-element input vector is:

$$x_1, x_2 \ldots\ldots x_N \tag{3.2.}$$

are multiplied by weights

$$w_{1,1}, w_{1,2}, \ldots\ldots w_{1,N} \tag{3.3.}$$

the weighted values are summed. The summation is simply expressed as *wx* and finally network produces a single row matrix *w* and the vector *x*.

$$n = w_{1,1}x_1, w_{1,2}x_2, \ldots\ldots w_{1,N}x_N \tag{3.4.}$$

The bias value can be added to the w_x within a matrix form;

$$n = wx + b \tag{3.5.}$$

This n value is used as an input parameter for a transfer function.

3.1.1. Multilayer Perceptron

A multilayer perceptron consists of input layer which have input vector, output layer which have output neurons and one or more than one hidden layer. To solve nonlinear problems hidden layers get importance. The network which has feedforward architecture is shown in Figure 3.3.

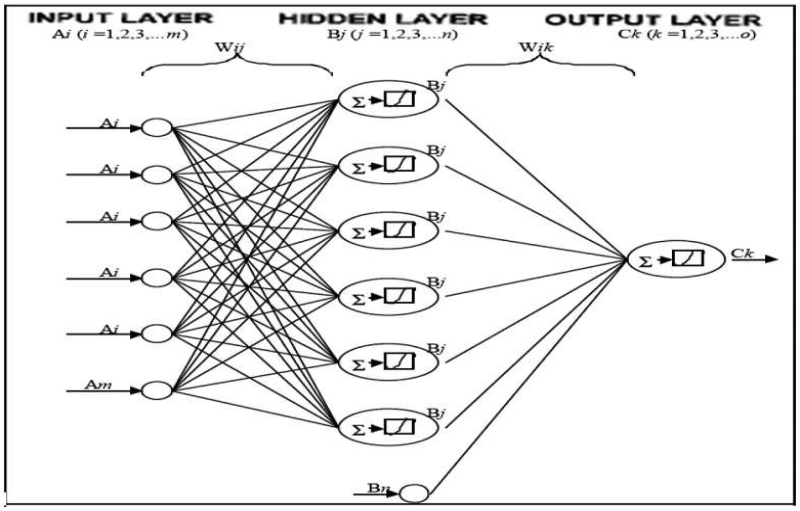

Figure 3.3. Multilayer perceptron network

Learning process starts without preformation about hidden layer number and the use of more neurons prevents generalization. There is a correct proportion between time of learning process and the number of hidden layer. MLP network structure consists of at least two-layer weight values. Weights of the first layer connect the input data. Weights of the second layer connect these hidden neurons to the output units. First, the hidden neuron outputs are computed by giving an n-dimensional input vector x in the form:

$$f(x) = w_0 + \sum_{i=1}^{n} w_i x_i \tag{3.6.}$$

In our notation, the parameters S denote the neuron outputs. N-element input vector $x_1, x_2 \ldots\ldots x_N$ are multiplied by weights $w_{1,1}, w_{1,2}, \ldots\ldots w_{1,N}$. The input layer distributes the values to each of the neurons in the first hidden layer. The following network structure is used to compute the outputs of the first hidden layer. Formula shows the connections between the I-1 hidden layers of NN.

$$h_1 = \begin{pmatrix} h_1^1 \\ h_1^k \\ h_1^{n_1} \end{pmatrix} = \begin{pmatrix} f\left(\sum_{j=1}^{n_0} w_{j,1}^0 x_j\right) \\ f\left(\sum_{j=1}^{n_0} w_{j,k}^0 x_j\right) \\ f\left(\sum_{j=1}^{n_0} w_{j,n_1}^0 x_j\right) \end{pmatrix} \tag{3.7.}$$

The activation function f (.) is generally chosen to be sigmoidal such as logistic sigmoid or the hyperbolic tangent function. Thus, hidden layer outputs are transformed using the most appropriate function.

$$Y=\begin{pmatrix} y_1 \\ y_k \\ y_{n_{N+1}} \end{pmatrix}=\begin{pmatrix} f\left(\sum_{j=1}^{n_N} u_{j,N} w_{j,1}^{N} h_N^j\right) \\ f\left(\sum_{j=1}^{n_N} u_{j,N} w_{j,k}^{N} h_N^j\right) \\ f\left(\sum_{j=1}^{n_N} u_{j,N} w_{j,n_N}^{N} h_N^j\right) \end{pmatrix} \tag{3.8.}$$

Where Y=1,..., n, and n is the total number of network outputs. The choice of f (.) is determined according to the structure of the input data and the features (Bishop, 2006).

Learning process starts without preformation about number of hidden layer and the use of more neurons prevents generalization. When network is being created; the number of hidden layers and neurons are detected with trial and error strategy.

Gradient of the error which is used to adjust the weights can be obtained. Sum of squares of the difference between desired output and target output, E, is defined as:

$$E=\frac{1}{2}\sum_j (y_{dj}-y_j)^2 \tag{3.9.}$$

Here, y_{dj} is the desired output of j^{th} and y_j is the target output. The weights of neurons are calculated according to error. Every w_{ji} weight is summed with Δw_{ji}. Δw_{ji} is calculated according to training algorithm (Chaudhuri at al., 2000; Haykin, 1994).

Different error computation methods should be considered (Bishop, 2006; Duda et al., 2012) depending on the type of output features (continuous, binary, discrete or multi-class). The sum of squares error (SSE) function selection is more common for continuous values.

3.1.2. Self-Organizing Map

A Self-Organizing Map (SOM) is a neural network model made out of an input space and an output layer organized on a map. Each input node fixed to other nodes with weights which consist of coordinates in the map. Figure 3.4 shows a SOM which can accept a 4 –D input vector and can associate this vector to 6X6 map. Each node in output layer has topological coordinates on the map and weight vector has the same dimensions with input vector (Kohonen, 1982).

We define each input value in learning set "n dimensioned X vector" and is shown X_1, X_2, X_3…, X_n. Weight vector of each node on output layer has n dimension and is shown W (W_1, W_2, W_3…,W_n). A SOM does not need a target vector instead of this output layer node which best matches to input layer X and neighboring nodes which has W weight vector organized to the coordinates of network that represents the class where X vector belongs to.

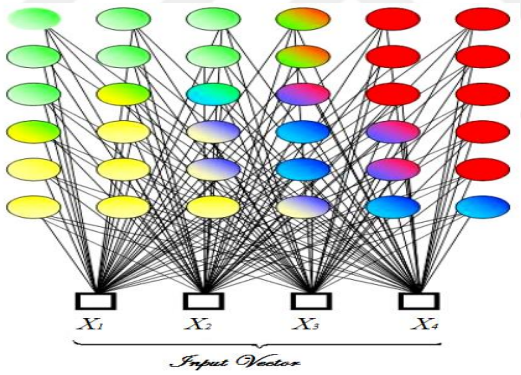

Figure 3.4. SOM network topolized with 6X6 map

The following steps are repeated many times in the learning law.

1) W weight vector in output layer of each node made out of random real numbers in the range [0-1].

2) Take a random X input vector and apply this to the input of SOM network.

3) Similarity measure between X input vector and W weight vector of each node in the output layer defined with Euclidean distance in Eq.3.10.

$$d = \sqrt{\sum_{i=0}^{n}(W_i - X_i)^2} \quad (3.10.)$$

The node which has the minimum distance is defined as best matching node (BMN).

4) According to algorithm W weight vector in the area σ (t) around the neighborhood of each node must match to the X input vector. Thus, neighborhood radius of BMN is calculated. Neighboring regions will vary over time. This is shown in Figure 3.5. BMN illustrated as yellow node and neighboring defined as radius. Neighboring diameter σ (t) is an expression and depends on time. This exponential expression for the neighboring function decreases exponential over time.

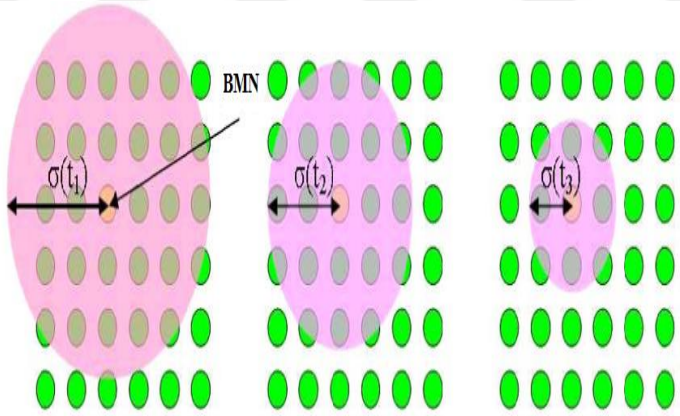

Figure 3.5. Variations of the neighborhood of the BMN with time. Neighborhood radius σ(t) is the function of time

$$\sigma(t) = \sigma_0 e^{-\frac{t}{\theta}} \quad t=1, 2, 3 \ldots \quad (3.11)$$

σ_0 is a scalar value and the half of longest edge of map and θ is a time variant.

5) W weight vector of each node in neighboring is matched to X input vector using Eq.3.12.

$$X(t+1) = X(t) + O(t) \in (t)\big(W(t) - X(t)\big) \quad (3.12)$$

$\in(t)$ is the learning rate and exponentially decreases over time and calculated as in Eq.3.13.

$$\in (t) = \in_0 e^{-\frac{t}{\theta}} \; t=1, 2, 3 \ldots \quad (3.13)$$

O(t) is affected coefficient and varies to the distance of node to BMN and shown in Eq.3.14.

$$O(t) = e^{\frac{distance^2}{2\sigma^2(t)}} \; t=1, 2, 3 \ldots \quad (3.14)$$

From eq.3.12, eq.3.13, e.q.3.14., the difference of every node over W weight vector depends on the distance to the BMN and cycle time. The change of the nodes is greater which are closer to BMN than the change of the nodes which are farther. This is defined as O (t). In the beginning of the self-organizing process $\in (t)$ is high and decreases over time. This mechanism even during in the first loops brings together the vectors which are being associated with each other with topological relations on 2D map. The mapping between samples and nodes performed over time.

3.2. K-Means Clustering

Unsupervised segmentation algorithms, one of them is k-means, are commonly based on clustering approach and preferable because of fast working facility (Garcia-Laencina, 2009). K-means clustering algorithm uses an iterative structure and predefining the cluster numbers, calculates the cluster centers using different distance measure methods such as Euclidean distance, bit-vector distance.

During the clustering process, the main objective is to minimize total intra-cluster variance by using the SSE error function.

$$v = \sum_{i=1}^{k} \sum_{j=1}^{n} \left\| x_j - \mu_i \right\|^2 \qquad (3.15.)$$

Where $\left\| x_j - \mu_i \right\|^2$ is a chosen distance measure between a data point xj and the cluster center, μ_i notates the distance of the data points to the cluster center to which they belong. *k* is positive integer number. The distances between input data and the corresponding cluster center are computed by minimizing the sum of squares.

The algorithm decides the cluster number (i.e. k = 4),

Cluster centers μ_i are randomly located,

The closest cluster center to which each data point x_j belongs is determined,

Compute again the positions of the k centroids, each center finds the centroid of the points it owns,

The process repeats until the cluster centers no longer move and results in the grouping of the data points

Euclidean distance calculation is one of the most used methods for k-means algorithm. An N x N matrix d is calculated. For points with d dimensions, the Euclidean distance *d (p, q)* between two points' *p* and *q* is defined as follows:

$$d(\mathrm{p},q)=\sqrt{\sum_{i=1}^{t}(pi-qi)^2} \tag{3.16.}$$

where p_i and q_i represents the i^{th} dimension values of p_i and q_i respectively.

3.3. Support Vector Machines

Support Vector Machines (SVM) is a classification algorithm based on statistical learning theory. Mathematical algorithms which SVM uses, primarily solves classification problem of linear datasets of two classes. SVM is generalized for classification of multi-class and nonlinear datasets. SVM deals with to find optimal decision function which discriminates two class, in other words determines the most suitable hyper-plane (Vapnik, 2000). SVM has been studied extensively in recent years about medical data analysis and classification (Bernstam at al., 2010; Jayasurya at al., 2009). It is not possible to form a linear hyper-plane for datasets as well as for medical data classification in Figure 3.6.

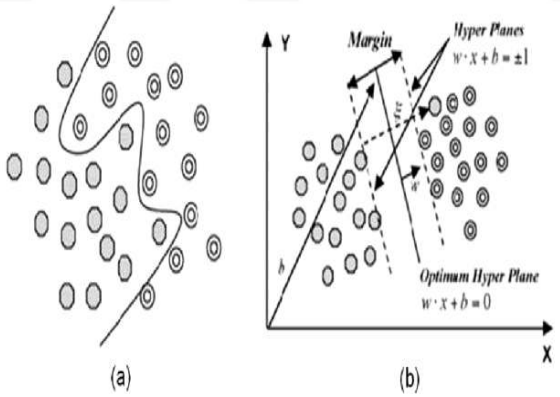

Figure 3.6. (a) Separation of nonlinear datasets. (b) Generalization of the solution introducing slack variable for nonlinear data

Some of training datasets outlies optimal hyperplane so this problem is solved defining an artificial variable ($i\xi$). Balance among maximizing the boundary and minimizing the misclassification is controlled with a C parameter which has

positive values ($0 < C < \infty$). Mathematical illustration of optimization problem for nonlinear datasets is in Formul.28.

$$\min\left[\frac{\|w\|^2}{2}+C.\sum_{i=1}^{r}\xi_i\right] \tag{3.17}$$

According to the equation above restrictions are described as:

$$y_i(w.\varphi(x_i)+b)-1\geq 1-\xi_i \tag{3.18}$$

$\xi_i \geq 0$ and $i=1,\ldots..,N.$ Samples which are not linearly separable are shown in a high dimensional space which is described as feature space. Thus samples are discriminated and hyper plane can be determined between classes.

Mathematically notation of support vector machines is $K(x_i,x_j)=\varphi(x).\varphi(x_j)$ which is identified as kernel functions. With the help of kernel function, it is probable to make nonlinear transformations and kernel function allows samples to be linearly separable in high dimension. As a result of this, decision rule of kernel function with two classes is in Formul.30 (Osuna at al., 1997).

$$f(x)=sign\left(\sum_i \alpha_i y_i \varphi(x).\varphi(x_i)+b\right) \tag{3.19}$$

Kernel function and optimal parameters which belong to this function are defined for classification with SVM. Kernel functions are represented in Table 3.1. Polynomial and radial based functions are easy to understand and more clear among all kernel functions. Whenever mathematical notation is seeming to be easy

the increase the degree of the polynomial may cause the algorithm to be more complicated. This issue not only significantly increases the duration of process but also decreases the classification accuracy.

Table 3.1. Basic Kernel functions which are used in SVM

Kernel function	Mathematically notation	Parameter
Normalized polynomial kernel	$\frac{((x.y)+1)^d}{\sqrt{((x.x)+1)^d((y.y)+1)^d}}$	polynomial degree (d)
Radial basis function kernel	$e^{-\gamma\|(x-x_i)\|^2}$	kernel dimension (γ)
Pearson vii (puk) kernel	$\frac{1}{\left[1+\left(\frac{2\sqrt{\|x-y\|^2\sqrt{2^{(1/\omega)}-1}}}{\sigma}\right)^2\right]^{\omega}}$	Pearson expanse parameters (σ,ω)

3.4. Decision Tree Algorithms

Extracting valuable information from large sets of data is a multidisciplinary field. Despite this fact, in cardiology related studies it is widely used. Decision tree algorithms are basically used to develop practical software because of on many environments. Due to its ease of understanding, no need prior assumptions about the data and ability to process both numerical and categorical data these algorithms are basically can be applied to real problems. Other hand limitations to one output attribute and tree complexity when working numerical datasets are handicaps.

A decision tree is a prediction method which is used in statistics, data mining and machine learning and can be easily used for clinical decision making (Podgorelec et al., 2001).

A decision tree consists of different numbers of nodes, branches and leaves (Rokach, 2008). A tree is generated by splitting the input data set into subsets based on feature value test. The subset in a node has exactly the same value as the target variable. The result causes tree to branch without losing data. If a node has a specific class, then this node is a leaf and branching stop at this point. There are three rules are observed in Figure 3.7.

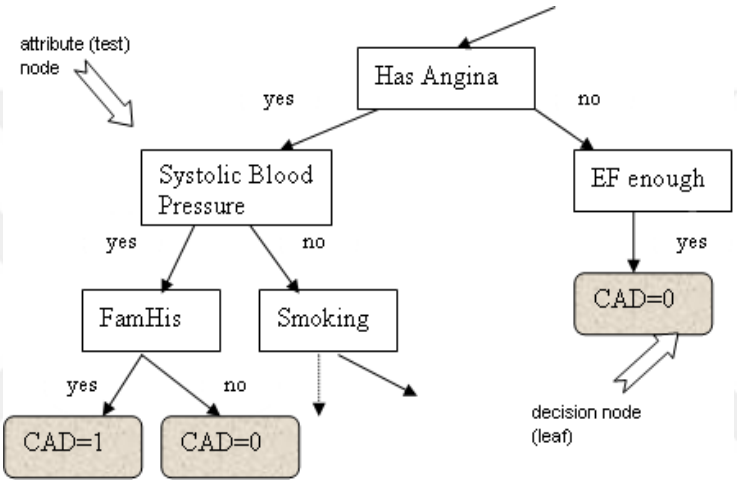

Figure 3.7. Shows a sample tree where a leaf represents a class, root and other nodes are the features of data set

Decision trees are interpreted easily. Rule sequences are placed from root node to leaf.

Rule 1:

If Has Angina= no and EF enough=yes Then CAD=0

Rule 2:

If Has Angina= yes and

If Systolic Blood Pressure=yes and

If FamHis=yes Then CAD=1

Rule 3:

If Has Angina= yes and

If Systolic Blood Pressure=yes and

If FamHis=no Then CAD=0

Decision trees are also described as an integration of mathematical and computational techniques to support the classification and generalization of a given data set. Data set can be described as:

$$(x,Y)=(x_1,x_2,x_3,\ldots\ldots x_k,Y) \tag{3.20}$$

The target variable, Y, contains classification or generalization information. The vector x is organized with the input variables.

3.4.1. Logistic Model Trees

Logistic Model Trees (LMT) is a classification model which consists of the combination of linear logistic regression and tree induction methods (Landwehr et al. 2005). While regression model in LMT is described as classification tree target, feature values are named as regression trees in model consisting of continuous values (Li et al. 2010). This study is set up to logitboost algorithm of splitting criteria.

LMT classification is based on decision trees and it is determined with features which consist of categorical or separate values by a threshold. Tree is generated by using threshold value, smaller than the threshold branched to left, otherwise to the right. Logit Boost method is used to determine the threshold value (Landwehr et al., 2005). Logit Boost uses an ensemble of functions F_K to predict classes *1, . . . , K* using *M* "weak learners".

$$F_{k(x)}=\sum_{m=1}^{K} f_{mk}(x) \tag{3.21}$$

3.4.2. Random Forest

Random Forest (RF) is an ensemble system (Breiman, 2001) that contains two or more classifiers. A number t of bootstrap samples is taken from the training data for the construction of t trees. For the convenience of evaluation, the features are determined randomly and one is chosen for the best division.

3.5. Multivariate Logistic Regression

The Multiple Logistic Regression Analysis (LRA) finds wide area of usage in medical researches (Altman et al., 2009). LRA has the advantage because of the regression-type logic and is more useful than the other analysis methods. Thus, LRA takes an important place in data analysis. At the same time, the mathematical model which is derived from the analysis with LRA is interpreted simpler.

Logistic or logit transformation (Hosmer at al., 2013) $\log\frac{p}{1-p}$ is a basic change of unlimited log p range. Logistic transformation is a linear function of x. The model can be represented as:

$$\log\frac{p(x)}{1-p(x)} = \beta_0 + x\beta \tag{3.22}$$

Solving for p, this gives

$$p(x;b,w) = \frac{e^{\beta_0+x\beta}}{1+e^{\beta_0+x\beta}} = \frac{1}{1+e^{-(\beta_0+x\beta)}} \tag{3.23}$$

Y = 1 when p ≥ 0.5 and Y = 0 when p < 0.5 probabilities should be predicted to minimize misclassification rate. When β_0+xβ non-negative then results scalar 1, and 0 otherwise. Therefore, the decision boundary which solves linear

classification problem of $\beta_0 + x\beta = 0$, which is a point if *x* is one dimensional, a line if it is two dimensional, etc. Boundary regularizations and distances and especially β coefficient determine the probabilities of classes. When these situations are evaluated it is seen that logistic regression is not just a classifier, can make more detailed predictions in different ways unfortunately, can result with wrong fit.

3.6. Feature Selection

There is not essentially a single best classification tool. Furthermore, before the best performing classification results the features are need to be analyzed. When the features are obtained from health care data selection of features becomes more important. Feature selection is a combination of methods which are used in pattern recognition, statistics and data mining (Kim at al., 2003). High dimensionality in clinical data sets that consist of unbalanced classes, redundant features and noisy or missing data is a great problem to deal with (Poolsawad at al., 2011).

Various feature selection methods are applied over datasets, but the choice of technique is dependent on the nature of the final classification result. When classification is preferred by choosing the strong features, the results can be stronger.

3.6.1. Relief-F

Estimation the importance of features is a major problem in a variety of scopes in machine learning (Robnik-Sikonja at al., 2003). Relief algorithms are accepted as successful feature estimators in general. These algorithms have strong ability for structuring conditional connections between features that especially provides earnings in preprocessing step.

Relief-F is an extension of Relief algorithms and has some advantages such as success in datasets which have more than 2 classes and handling noisy datasets. The concept of Relief-F (Kononenko, 1994) is choosing instances randomly and changing the weights of the feature relevance using nearest neighbors. Finally

obtains a feature weighting vector to give more weight to the features that distinguish the instance from neighbors of different classes. The quality of feature W[f] is an approximation of the following difference possibilities.

W[f] = P(different value of f|near instance with different class prediction) –P (different value of f| near instance with same class prediction)

Relief-F method presents some advantages such as noise-tolerance as well as being applicable easily for binary or continuous data. However, the inadequate training samples affect the algorithm not to distinguish redundant features.

3.6.2. Independent t-test Analysis

A t-test is used especially in statistical analysis which determines whether a subset or subsets of scores belong to the same population. The independent sample t-test is used to compare a data set with a pre-averaging distribution. Independent t-test analysis is used for the selection of variables for logistic regression especially in medical diagnosis (Lin at al., 2010; Ergun at al., 2004).

3.7. Sampling Methods

3.7.1. Stratified Sampling

Stratified sampling method produces a random subsample of the original dataset using either sampling without replacement. Stratified random sampling method divide the data set into smaller subsets known as strata. Then, the strata are generated based on categories or characteristics of class feature. Random samples from each class tag are taken in a number proportional to the size of total classes when compared to the data set. Thus, the number of samples in the generated dataset is considered to become same.

3.7.2. Oversampling

An artificial way of rebalancing unbalanced datasets is used by applying sampling methods to solve the problem of uneven distribution of classes. Thus, each class has sufficient sample for classifiers and better trained classifier algorithm offers a better success rate. Oversampling is one of the sampling methods which duplicates in the minority class examples randomly and makes the dataset to increase with much more samples. Moreover, this method helps to achieve balancing class distribution by replication minority class sample. Sampling method is visualized in Figure 3.8.

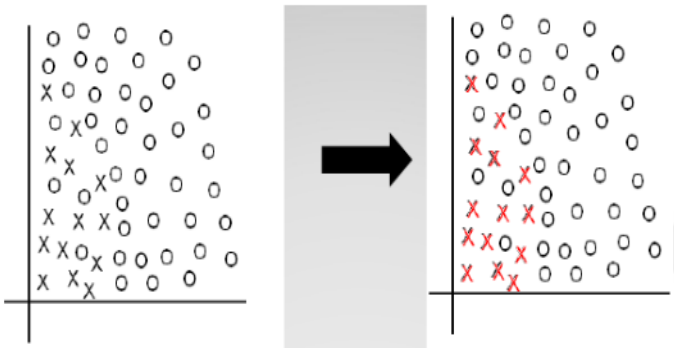

Figure 3.8. Replicate the minority class samples

3.8. Normalized Root Mean Square Error (NRMSE)

Non-dimensional forms of the RMSE are useful because often one wants to compare RMSE with different units. There are two approaches: normalize the RMSE to the range of the observed data, or normalize to the mean of the observed data.

$$NRMSE = 1 - \frac{RMSE}{X_{obs,max} - X_{obs,min}} \tag{3.24}$$

4. DATA SET DESCRIPTION

The dataset included 459 patients who presented to the department of cardiology with the suspect of CAD and was approved by Mersin University, Mersin, Turkey Research and Training Hospital. In this Chapter, the original characteristics of the dataset are presented. The statistical properties of the final state which is selected for use at other stages after the imputation are given in the Table 4.2.

4.1. Cardiovascular Dataset

Angiography is applied as a standard procedure to determine the stenosis. It is necessary to determine stenosis size in more than 50% of cases, resulting in the need for new criteria to classify coronary artery disease more specifically. After angiography has been used to determine the location of the lesion, left main coronary artery disease (stenosis diameter > 50% of vessel diameter) is considered to indicate the presence of CAD.

4.1.1. Prognostic features

The present study categorizes 22 clinical comorbidities, cardiac status, medical history features into categorical-type (Table 4.1) and continuous-type format (Table 4.2). Demographic features consisted of age and gender. The comorbidity features consisted of smoking, renal failure, chronic obstructive pulmonary disease (COPD), asthma, diabetes mellitus and hyperlipidemia. Cardiac medical history consisted of hypertension and family history of CAD.

4.1.2. Class feature and distribution of MVs

The cardiovascular dataset has 314 cases of CAD and 145 cases without CAD with 22 features for each case. Of the total number of cases in the dataset, 277 are completed with all the features and 182 have at least one MV. The

complete part of the dataset consists of 210 CAD and 67 without CAD samples. We observed that 182 out of 459 cases have 4.8% to 33.3% MVs in their features. The CAD dataset contains 11 continuous-type of features which has MV at random drop-outs.

Table 4.1. Clinical characteristics for categorical features of individuals relevant to diagnosis

Feature Label	Variable Type	Missingness(%)
Gender (İndividual's gender)	Qualitative/categorical	-
DM (Diabetes Mellitus)	Qualitative/categorical	-
FH (Family History)	Qualitative/categorical	-
HTN (Hypertension)	Qualitative/categorical	-
Smoke	Qualitative/categorical	-
Hyp (Hyperlipidemia)	Qualitative/categorical	-
Asthma	Qualitative/categorical	-
Renal Failure	Qualitative/categorical	-
COPD	Qualitative/categorical	-

4.1.3. Statistical Measurements of the Imputed CAD Dataset

After imputation procedure, a complete version of the CAD dataset is obtained and used at other stages of the thesis. In case of presence of CAD, the size of the stenosis can be matched with the risk score concept. This situation opens the way to classify the disease with different risk scores. In this study, the purpose is to determine one of these scores that are determined as 0 to 39 as low-risk, 40 to 69 as moderate, 70 to 85 as high; and >=86 as very high risk. As for the dataset, 182 individuals were of “moderate” (out of 459 records; 39.65% of the total records) and also 164 individuals were defined as “high” (out of 459 records; 35.73% of the

total records). The dataset shows an imbalance on status of “low” or “very high”. According to the risk assessments, only 43 individuals were of “moderate” (out of 459 records; 9.36% of the total records) and also 70 individuals are defined as “high” (out of 459 records; 15.25% of the total records) (Table 4.3).

Table 4.2. Clinical characteristics of individuals relevant to diagnosis. Missingness percentages are shown for numeric features which have MVs

Feature Label	Variable Type	Missingness(%)
Age (İndividual's age)	Quantitative/numeric	-
LDL (Low-Density Lip.) (mg/dl)	Quantitative/numeric	**13.07**
HLD (High-Density Lip.) (mg/dl)	Quantitative/numeric	**12.41**
TG (Triglyceride) (mg/dl)	Quantitative/numeric	**12.41**
TC (Total Cholesterol) (mg/dl)	Quantitative/numeric	**13.72**
Urea(mg/dl)	Quantitative/numeric	**16.78**
Cr (Creatine) (mg/dl)	Quantitative/numeric	**0.6**
RDW (red cell distribution)(mg)	Quantitative/numeric	**3.7**
MCV (mean corpus. volume)(mg)	Quantitative/numeric	**3.7**
EF (Electrocardiographic result)(%)	Quantitative/numeric	**18**
FBS(Fasting Blood Sugar) (mg/dl)	Quantitative/numeric	**5.22**

Table 4.3. Statistical measurements of variables relevant to diagnosis model

Variable Type	Input Variables	Mean	Max.	Min.	Stand. Dev.
Demographic	Gender(İndividual's gender)	0.34	-	-	0.47
	Age(İndividual's age)	58.69	88	25	12.05
Comorbidity	Hyperlipidemia(Hyp)	0.27	-	-	0.45
	DM(Diabetes Mellitus)	0.34	-	-	0.48
	COPD	0.05	-	-	0.22
	Smoke	0.31	-	-	0.46
	Renal Failure	0.05	-	-	0.21
	Asthma	0.05	-	-	0.21
Medical History	Family History(FH)	0.17	-	-	0.38
	Hypertension(Htn)	0.53	-	-	0.5
Laboratory and EF	Low-Density Lipoprotein (LDL) (mg/dl)	111.02	1200	25.09	63.29
	High-Density Lipoprotein (HDL) (mg/dl)	42.93	97.2	21	12.34
	Urea(mg/dl)	36.88	238.9	6	21.38
	Hemoglobin(Hb) (g/dl)	13.33	26.4	7.7	1.95
	Creatinine(CR) (mg/dl)	0.94	10.8	0.41	0.93
	Triglyceride(TG)(mg/dl)	185.77	1423	40	144.25
	Fasting Blood Sugar(FBS)(mg/dl)	138.05	500	11.59	71.61
	Red cell distribution width (RDW)(mg)	14.07	84.6	12.2	3.74
	Mean corpuscular volume (MCV) (mg)	86.08	108.6	13.7	6.9
	Total(TC) (mg/dl)	184.42	431.2	92.8	44.21
	Electrocardiographic result	52.22	70.46	10	10.82

5. PROPOSED ALGORITHMS FOR MVS IMPUTATION

ANN, K-means methods are explained in Chapter 3. In this chapter, the MVs imputation implementations of ANN that includes multilayer networks and SOM and K-means, with the proposed initial parameters are presented.

5.1. Proposed MLP Model for MVs Imputation

An MLP network can undergo a learning process so as to predict MVs. Features comprising MVs in the training process are used as output while the remaining features are used as input (Poolsawad et al., 2011). The dataset is divided in two groups. It is defined as one group which includes samples containing MVs in features and the samples which have no MVs are placed in another group. Start point of missing values for each patient depends on number of attributes which have blank entries. Dataset is ordered according to the number of attributes which have less to more MVs. Complete dataset is trained and then incomplete data is presented to the network to predict the features which have MVs. All the attributes which have MVs are given to the model respectively. Mean squared error function is considered within a maximum likelihood approach and conjugate gradient method is used as the optimization technique for training the MLP network (Bishop, 2006). At the end of the process, training dataset and imputed dataset are combined to make a final dataset to be fed to selected classifiers for classification. Imputation process with MLP can be described as follows.

- **In Step 1:** Given a dataset consist of 22 attributes (columns) one of them indicates class label and 459 samples. Determine which features and samples contain missing data.

- **In Step 2:** Divide the attributes that do not contain any missing data *(set 1)* and with the ones that have missing data *(set 2).* Arrange *set* 2 in the order of their missing values.
- **In Step 3:** Choose one target attribute in *set 1* that corresponds to attribute which has MVs in *set 2* and use remaining attributes as input values in *set 1*. Then construct a MLP network and train this network using *set 1*. Give attributes except target attribute as input values in *set 2* to the network. Network produces an output vector in which every value corresponds to a blank entry of this attribute in *set 2*.
- **In Step 4:** Repeat *Step3* for each attribute which has MVs in *set 2*.
- **In Step 5:** Merge every imputed sub sets from *Step3*. Then merge these subsets with *set 1*.

Algorithm 1: Multilayer Networks Imputation

Procedure **MLP_Imp (params)**

```
Input
m1,m2
Output
m2
begin
  Evaluate feature index Fx = find( empty( m2 ) )
  Set parameters( hidden_layer_size = 11, train_ratio = 70/100,
   Val_ratio= 15/100, Test_ratio = 15/100 )
  For i = 1 to col no of Fx
    T_i = m1_i;
    m1 = m1 – T_i
    Train MLP network by T_i, m1
    In_i = m2 - Fx_i
    Fx_i = Test MLP network by In_i
    m2 = m2 + Fx_i
  next i
end
```

5.2. Proposed Self-Organising Map Model for MVs Imputation

According to the CAD dataset, examining the relationships between features, MVs are estimated by excluding the features which have MVs. To this end, concerning the size of the map, a square map is used to get rid of border effects that may develop depending on the shape (Fessant at al., 2002). We chose a self-organizing map which has 8X8 dimensions by testing a whole range of values between 4X4 and 20X20. We noticed that a smaller size does not supply obtaining the data distribution satisfactorily, whereas a larger size results shows the model is strongly fitted to training data.

Complete dataset is trained and then incomplete data is given respectively to the model to predict the features which have MVs. All the features which have MVs are given to the model respectively. When an incomplete sample is presented to the SOM network, samples of missing entries during BMN selection are not considered. The incomplete samples are imputed with the average feature values of the BMN.

Imputation process used in our dataset with SOM can be described shortly as follows.

- **In Step 1:** Given a data set consist of 22 features one of them indicates class label and 459 samples. Identify which samples and features that have missing data.
- **In Step 2:** Separate the features that do not contain any missing data (set 1) and with the ones that have MVs (set 2). Arrange *set 2* in the order of their MVs.
- **In Step 3:** Design an 8X8 SOM network using set 1 and find neighboring relations between nodes in the map. For each sample in *set 2*, take a sample and find the BMN by minimizing the distance between the sample and nodes using Euclidean Distance. Missing dimensions are

excluded from the distance calculation. Select the activation group which is created by neighbors of BMN in the map.

- **In Step 4:** Replace value for a missing item based on the weights of the activation group which is generated by neighbors of BMN.
- **In Step 5:** Merge the imputed sample from previous step with *set 1*. Then use Step 3 until there is no sample in *set1*. The algorithm of the MV imputation process is as follows:

Algorithm 2:SOM Imputation

Procedure **SOM_Imp (params)**

Input
m1,m2
Output
m1
begin
  for i = 1 to row no in m2
Evaluate feature index Fx = find(empty(Nx_i))
    Set parameters
      epochs = 1000
      performFcn = SSE
      mapdim = 8 X 8
    Evaluate $_{train}C_i$, $_{test}C_i$
    for j = 1 to col no of $_{test}C_i$
      for all $w_{ij} \in W$ $d_{ij} = \|Nx_i - w_{ij}\|$
      Evaluate weight values of activation group of BMU
      $W_{(t+1)} = W_{(t)} + \Theta_{(t)}(Dij_{(t)} - W_{(t)})$
      Impute Nx_{ij} *with mean values of* $W_{(t+1)}$
    next j
    $m1 = m1 + Nx_i$
  next i
end

5.3. Proposed K-means Model for MVs Imputation

The dataset is separated into two groups. The samples having MVs in their attributes are in one group and the samples which have no MVs are placed in another group. Start point of missing values for each patient depends on number of attributes which have blank entries. Dataset is ordered according to the number of attributes which have less to more MVs. K-means clustering is applied on complete instances set to obtain clusters. Number of clusters is assigned considering similarity between cluster centers.

For the CAD dataset cluster center number is identified as 6 before clustering. Our experiments on complete dataset shows that when k is greater than cluster size cluster center values differs in more wide range. Thus, the probability to be included of incomplete samples to wrong cluster decreases. At the same time this situation indicates the increase of sensitivity.

The incomplete instance which is first in the list is taken and the nearest cluster to this sample is found. Nearest neighbors of the incomplete instance in the cluster nearest to it is found where k is set to cluster size. The MVs in incomplete instance are imputed with the values of nearest cluster to this instance. Then the imputed incomplete instance is moved to complete instance set which is called first group and assigned to the cluster used for imputing that instance. The cluster centers are updated with new recruits. The imputed instance is stored in complete instance set and will be one of the members of all further imputations of other incomplete instances. Imputation process used in our dataset with K-means can be described shortly as follows.

- **In Step 1:** Given a data set consist of 22 attributes (columns) one of them indicates class label and 459 samples (rows). Determine which features and samples contain missing data.

- **In Step 2:** Divide the features that do not contain any MVs (*set 1*) and with the featues that have MVs (*set 2*). Sort *set 2* in the ascending order of their missing value range.
- **In Step 3:** Construct K-means algorithm using set 1 and find cluster centers given 6 cluster numbers. For each row in *set 2*, take a sample and use Euclidean Distance to find the nearest cluster to this sample. Include this sample to this cluster and impute the blank entries (the numbers of blank entries and which column they belong to differ sample to sample) with cluster values to where this instance is included.
- **In Step 4:** Merge the imputed sample from previous step with *set 1*. Then construct Step 3 until there is no sample in *set 2*.

The algorithm of the MV imputation process is as follows:

Algorithm 3:K-means Imputation

Procedure **K_means_Imp (params)**

Input
m1,m2
Output
m1
begin
  for i = 1 to row no in m2
    evaluate feature index $F_X = fmd(empty(Nx_i))$
    Fy=~Fx
    evaluate C_i, c_i *of the m1 by K-means*
    index = 0
    for j = 1 to column no of Fx
      mindist = 0

… ***Continued***

```
13          for k = 1 to 6
14             dist_k = ||Fy_i – C_i||
15             if mindist_k > dist_k then
16                mindist_k = dist_k
17                   index = k;
18             end if
19          next k
20          impute Fx with mean values of C_index
21       next j
22       m1 = m1 + (Fx_i + Fy_i)
23    next i
24 end
```

5. PROPOSED ALGORITHMS FOR MVS IMPUTATION Jale BEKTAŞ

6. RISK SCORE PREDICTION PROCEDURE

Acute chest pain is one of the frequent causes of presentation to emergency department; however, only 15-20% these presentations of chest pain really have (Pope et al., 2000). In addition, it becomes difficult to diagnose the ones which do not have specific symptoms or electrocardiographic signs. Coronary Artery Diagnosis (CAD) is missed in 2% of patients and short period mortality should be high of these patients. Furthermore, there are many factors which cause higher risk for CAD and can be classified according to the National Cholesterol Education Program (NCEP) ATP III Guidelines.

Angiography is used as standard procedure to determine the CAD and different levels of stenosis. The need to determine the stenosis level and presence of CAD is greater than 50%, leading to new criteria for more specific classification of CAD. After angiography, as for the location of the lesion, one of the three major coronary arteries; Left Anterior Descending (LAD), Left Circumflex (LCX), or Right Coronary Artery (RCA) (stenosis diameter > 50% of vessel diameter) are considered as having CAD (Alizadehsani et al., 2013; Alizadehsani et al., 2016). In case of presence of CAD, the size of the stenosis can be matched with the risk score concept. This situation opens the way to classify the disease with different risk scores (Compare et al., 2013). In accordance with this purpose, one of these scores that has prevalence 0 to 39 as low-risk, 40 to 69 as moderate, 70 to 85 as high; and >= 86 as very high risk are determined to classify CAD stenosis (Ergun et al., 2004). As the degree of stenosis increases to four, it becomes inevitable to face the problem of imbalanced data. For CAD dataset, the data shows an imbalance on the status of patients which are classified as “moderate” and “high”.

This chapter presents the Risk Score Prediction System on imbalance CAD dataset. The methodology of the system is presented and the steps of procedure are examined sequentially. The feature selection approaches and sampling methods

which are used to improve the classification results of the imbalanced dataset are mentioned.

6.1. Representation

A multilayer perceptron consists of input layer which have input vector, output layer which have output neurons and one or more than one hidden layer. To solve nonlinear problems hidden layers get importance and thus, network as shown in Figure 6.1 is transformed to a multilayered structure.

Learning process starts without pre-information about hidden layer count and the use of more neurons prevents generalization. When the number of hidden layer increases the time of learning process increases. The choice of f (.) is determined according to the structure of the data and the variables. In this chapter, the activation function f (.) is chosen to be logistic sigmoid function.

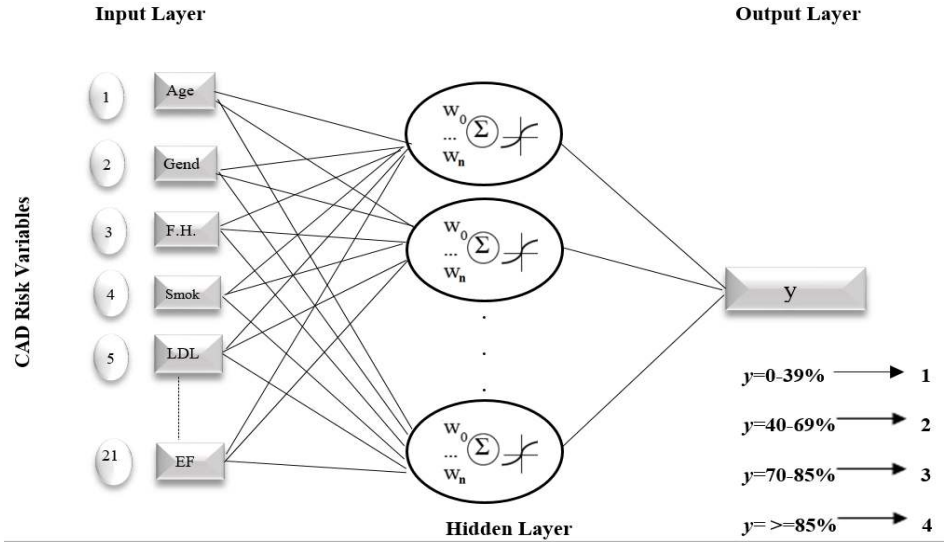

Figure 6.1. Neural architecture of the system used to predict the lesion degree

To determine the risk score and analyze the data, the risk factors are identified with an integer coefficient. The integers should be selected in proportion

to the estimated continuous coefficients of the logistic model. Training dataset was determined with the four risk score numbers: 1) low, 0 to 39; 2) moderate,40 to 69; 3) high, 70 to 85; 4) very high,>85.

NN can be used to predict a risk score which is identified as dependent variable using the remaining complete features as inputs. Mean squared error function is considered within a maximum likelihood approach, and the Conjugate Gradient Method is used as the optimization method for training the NN. The dataset is divided into 10 sub-sets in which the accuracy was evaluated by dividing data into training and test sub-sets for each set. When learning ends, the network is evaluated from the test data set and the test dataset is presented to the network to predict the dependent variable. The prediction process can be described shortly as follows.

- **In Step 1:** Divide the dataset into 10 sets.
- **In Step 2:** Take 75% for training and 25% for test.
- **In Step 3:** Each set is divided into two subsets, namely *set 1*, and *set 2*. Unlike *set 2, set 1* has the target feature.
- **In Step 4:** NN is organized and trained using *set 1*. The other features are given as the input vector in the network by excluding the target vector in *set 2*. The network produces an output vector in which every sample corresponds to a target value in *set 2*.
- **In Step 5:** Choose one of the risk categories considering the risk score which was evaluated by the network.
- **In Step 6:** Results of the output vector produced by the network and the target vector values are evaluated.
- **In Step 7:** Repeat *(ii)* 10 times for each subset.

Algorithm 1: Multilayer Networks Prediction

Procedure **MLPNN_Pred (params)**

```
01  Input
02  m1,m2
03  Output
04  m2
05  begin
06    Evaluate feature index Fx = find( empty( m2 ) )
07    Set parameters( hidden_layer_size = 11,
                      train_ratio = 70/100,
                      Val_ratio= 15/100,
                      Test_ratio = 15/100 )
08    For i = 1 to col no of Fx
09      T_i = m1_i;
10      m1 = m1 – T_i
11      Train MLP network by T_i, m1
12      In_i = m2 - Fx_i
13      Fx_i = Test MLP network by In_i
14      m2 = m2 + Fx_i
15    next i
16  end
```

7. HYBRID CLASSIFICATION APPROACH (HCA)

In many practical applications, it is difficult to reach a good estimation ratio in case the distribution of the dataset is not homogenous. In modern data collection techniques and in many classification applications, it is common to work with big and imbalanced datasets like healthcare services, text classifications, etc. The number of the negative examples may mostly double the positive examples. On the other hand, we do not have adequate idea on what the ratio of positive examples to negative examples should be.

Let us assume that the number of the negative examples is more than the positive samples, or the situation may be just the opposite. In such a situation, the class that has less samples may be defined as set1, and the class that has more samples may be separated into smaller datasets with clustering algorithm (e.g. subset1, subset2 …). In the next step, independent classifiers that consist of the integration of set1 and the subsets are formed. Then, the LR Model has been used to integrate and convert these independent classifiers into a result that may be decided collectively on learning. The empirical results have shown that the hybrid method may be used in coping with imbalanced training datasets classification problem.

This hybrid method has been tested on two datasets. Datasets were taken from UCI Machine Learning Archives. The WDBC consists of 699 samples and 9 independent features that include continuous data, and one dependent feature includes class info. Other one is Dermatology dataset which consists of 366 samples, 32 independent features and one dependent feature. The features include categorical and continuous type data.

This chapter presents the HCA on two benchmark datasets and imbalance CAD dataset. The methodology of the system is presented and the steps of procedure are shown.

7.1. Using SVM for HCA

The important thing in SVM is the finding of the hyperplane that may separate the training samples in R^N. Training samples are considered to be preassigned to two classes A or B.

We consider having *n* training samples. For each x_i, $i = 1,..., n$ is a member of class A or class B. We look for a separating rule of the form.

$$\omega^T x \geq \gamma, \tag{7.1.}$$

where $x \in R^N$ and denotes the feature vector while $\omega \in R^N$, and $\gamma \in \{-1,1\}$ denotes the class labels of samples. To obtain ω and γ we solve the following linear support vector machine optimization problem:

$$minimize \frac{1}{2}\|\omega\|^2 + Ce^T y, \tag{7.2.}$$

where $\omega \in IR^N$, $\gamma \in IR^N$, $y \in IR^N$

The optimal ω and γ of this problem yields the separating hyperplane. The data points that are defined as the support vector and that are the closest to the Hyperplane xj are found. The signed distances are calculated with the following formula in the ω vector.

, $$\omega^T = \left[\left(\sum_j \alpha_j x_j\right)^T b\right] \tag{7.3.}$$

where *b* value shows bias.

7.2. Logistic Regression (LR)

By giving an input vector x_i ϵ 〖R〗^N and the output values γ∈{0, 1} , LR may be fit with the likelihood principle that may estimate the probability of the outcome. This probability will be p, $\gamma_i = 1$, or *1-p* if $\gamma_i = 0$.

$$L(\emptyset) = \prod_{i=1}^{n} (p_i)^{\gamma_i} (1 - p_i)^{(1-\gamma_i)} \quad (7.4.)$$

It is considered mathematically more practical to calculate log of equation. The log-likelihood can be identified as follows:

$$lnL(\emptyset) = \sum_{i=1}^{n} \left(y_i ln p_i + (1 - \gamma_i) ln(1 - p_i) \right) \quad (7.5.)$$

$L(\emptyset)$ is maximized by the change of $\emptyset$ which is called the maximum likelihood estimate. (DEV) or with a known name loss function is the negative log-likelihood and calculated in following equation.

$$DEV = -2lnL(\emptyset) \quad (7.6.)$$

7.3. Representation

Labelled A and B are assumed to represent 2 classes in the dataset. In an imbalanced structure in which Class A has more samples than Class B in the dataset, Class A may be separated into n number subsets $(A_1, A_2, \dots A_n)$ by using clustering algorithm. The separation number *n* may be dependent on the $|A|/|B|$ ratio. In this study, the testing has been made with different *n* values in such a manner in which the $|A|/|B|$ ratio must not be too far from 1. Each I is merged with the samples in the *i =1, 2,…, n* subset with Class B, and presented to the SVM classifiers as $A_i \cup B$. If we define each classifier as $\aleph_i$, we will have *n* number classifiers in total. For this reason, each time when the algorithm is run, each

classifier will form a hyperplane H_i for $\aleph_i$. In the next step, the signed distances d_k to these hyperplanes must be calculated for each k sample in the dataset, and a vector with n dimension $d = (d_{k,1}, \ldots d_{k,n})$ must be obtained. The D vector is prepared to be presented to the LR Model and will weigh the responses on SVM models in the study and then estimation probability is calculated.

The procedure starts with the separation of class samples, which are many in number, with the k-means clustering method. The k parameter has been given as 2, 3 and 4, respectively; and the results are evaluated separately. The SVM classifiers have been structured by using the merged sub-sets with the same number of k defined during clustering. Linear Kernel has been used for the SVM models that are trained and tested with cross validation, and the classifiers have been merged with the LR analysis. The selection of statistical model for LR will influence the performance of end-classifier at a significant level. During the model selection, in order to facilitate the interpretation and to obtain a structure that is far from being over-fit, the estimated coefficient values are examined. Then, the selected variables are fit to LR by using the binomial analysis method. Possible connections are created between each pair of the variable. In the hybrid model, the outcome of the SVMs is $d = (d_{k,1}, \ldots d_{k,n})$ vector. The linear equation obtained by the LR with this vector will be $i, i=1,2,\ldots,n$ and $b_0+b_1d_1+b_2d_2+b_3d_3, + \ldots + b_nd_n$. Here, B is the parameter vector. The whole framework in the hybrid algorithm is given in Figure 7.1.

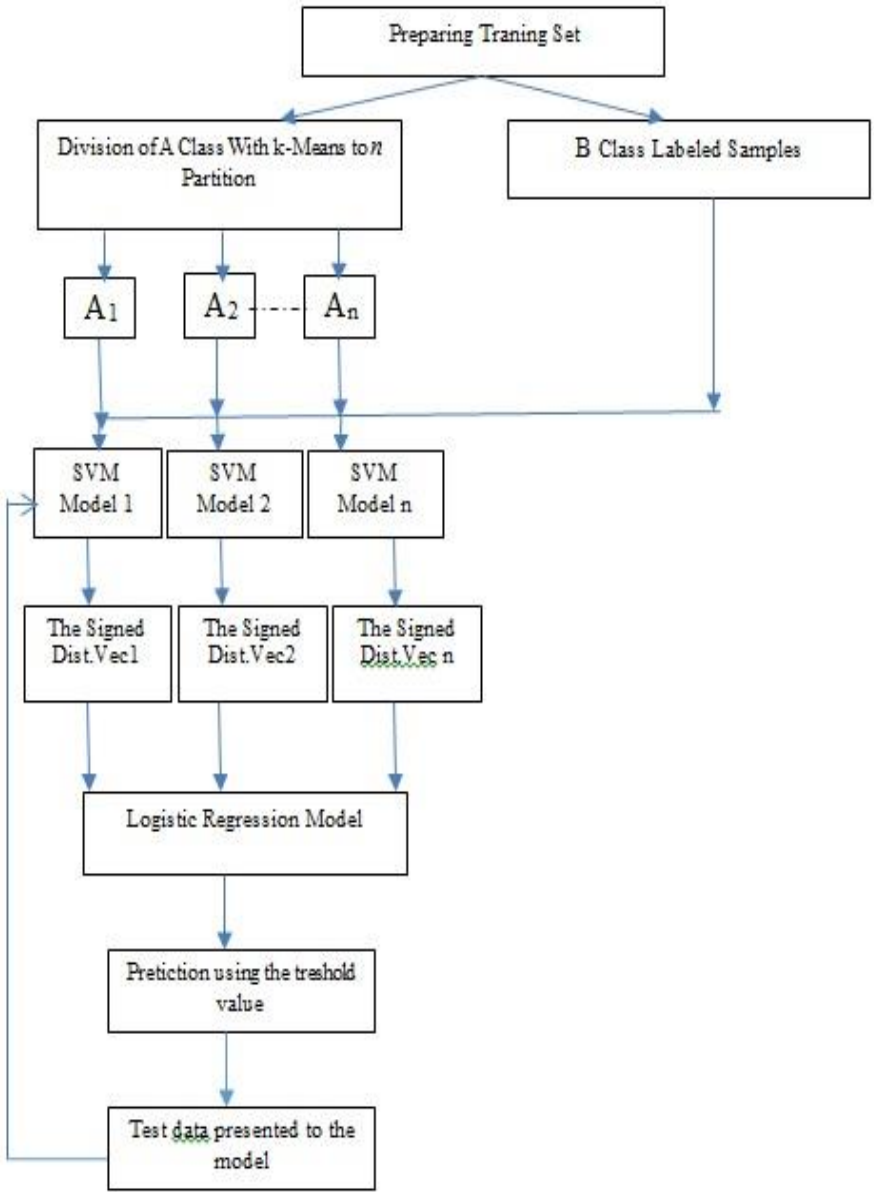

Figure 7.1. Framework of the hybrid algorithm

The algorithm used to carry out the HCA is as follows:

Algorithm 1: Hybrid Classification Approach

Procedure **HCA (params)**

01 ***Input***
02 *mpos, mneg, k*
03 ***Output***
04 *Outreal, SvmStruct*
05 ***begin***
06 *Mfull=mpos+mneg*
07 *evaluate C_i, c_i of the mfull by K-means*
08 *index = 0*
09 *mindist = 0*
10 ***for*** *i = 1 to k*
11 *$dist_i = ||mpos_i - C_i||$*
12 ***if*** *$mindist_i > dist_i$* ***then***
13 *$Mindist_i = dist_i$*
14 *index = i;*
15 ***end if***
16 *$mfull=mneg+mpos_{index}$*
17 *Classify mfull and compute signed distances of the samples by using separating hyperplane*
18 *Sv=SvmStruct.SupportVectors*
19 *alphahat=SvmStruct.Alpha*
20 *bias=SvmStruct.Bias*
21 *kfun=SvmStruct.Kernelfunction*
22 ***for*** *j = 1 to row no of mfull*
23 *$dist_{i,j} = sign||mfull_j * kfun|| * alphahat + bias$*
.. ***Continued***

```
24  Next j
25  Evaluate LR function and compute output vector Out real
26  Bo=fitinfo of LR function
27  for j = 1 to row no of dist
28  Poj=0
29  for n=1 to k
30  Po_n=Po_n+(dist_j,n Bo_n)
31  end k
32  end j
33  T=Nanmean(Po)
34  If Po>=T
35  Outreal=[Outreal 1]
36  Else
37  Outreal=[Outreal 0]
38  end
49  End k
40  end
```

8. RESULTS

The implementations for the proposed algorithms are written in Matlab Environments. On the other hand, SPSS tool library is used for Logistic Regression Analysis. Machine learning algorithms are used to impute MVs in cardiovascular dataset. At the end of the process, training dataset and imputed dataset are combined to make a final dataset to be fed to selected classifiers for classification. SVM, MLP, Random Forest and Logistic Model Trees are used to obtain performance results. The performances are compared with the most commonly used mean imputation method. This experiments are given in Section 8.1.

In the next process, various multivariate logistic regression combinations, preprocessed with feature selection and sampling methods, are performed on 4 classed imputed CAD dataset. The implementations and the classification results are presented with details in Section 8.2. A new interface is designed to predict a risk score from four outputs given in Section 8.3. In HCA's experiments, two distinctive datasets for benchmark classification and the same imputed CAD dataset are used. Section 8.4 shows implementations of the novel HCA algorithm for the benchmark and the CAD data.

8.1. The Impact of Imputation Procedures on the Performance of Classifiers

It is important to assess the classifier for clinical data analysis based on how well the classifier is performing to predict patients diagnosed with CAD. As seen over results the dataset shows an imbalance on status of patients. According to the highest accuracy and sensitivity results only 25 samples are of NCAD out of 459 records (5.45% of the total records). The classifier can provide very high accuracy when the NCAD class is classified substantially correct because of limited sample number. As for analysis it is more important to evaluate sensitivity and specificity than accuracy to compare the classification outcome. Performance

results were given in following tables as for the confusion matrixes, ACC, Sen, Spec, MAE, RMS values respectively placed in Table 8.1, Table 8.2, Table 8.3.

Table 8.1. Different Classification Methods for MVs with mean imputation approach

	Confusion Matrix			**ACC**	**SEN**	**SPEC**	**MAE**	**RMS**
MLP	Status	Classified CAD	Classified NCAD	87.58	0.884	0.145	0.146	0.337
	CAD	296	18					
	NCAD	39	106					
SVM	Status	Classified CAD	Classified NCAD	81.69	0.814	0.172	0.18	0.43
	CAD	298	16					
	NCAD	68	77					
LMT	Status	Classified CAD	Classified NCAD	84.80	0.882	0.225	0.18	0.37
	CAD	283	31					
	NCAD	38	107					
RF	Status	Classified CAD	Classified NCAD	87.36	0.864	0.093	0.209	0.303
	CAD	304	10					
	NCAD	48	97					

Table 8.2. Different Classification Methods for MVs with K-Means imputation approach

	Confusion Matrix			ACC	SEN	SPEC	MAE	RMS
MLP	Status	Classified CAD	Classified NCAD	85.18	0.884	0.223	0.158	0.356
	CAD	283	31					
	NCAD	37	108					
SVM	Status	Classified CAD	Classified NCAD	83.66	0.848	0.198	0.163	0.4
	CAD	291	23					
	NCAD	52	93					
LMT	Status	Classified CAD	Classified NCAD	86.92	0.896	0.194	0.16	0.36
	CAD	287	27					
	NCAD	33	112					
RF	Status	Classified CAD	Classified NCAD	87.36	0.865	0.1	0.21	0.31
	CAD	303	11					
	NCAD	47	98					

It is shown that the size of parameter k affects the result when the K-means imputation process is applied. When k is larger than the size of the cluster, the cluster center values can be differentiated over a larger distance. Thus, the probability of incomplete cases being placed into the wrong cluster decreases, resulting in increased sensitivity, so the number of clusters k is chosen as 6.

Table 8.3. Different Clustering Methods for MVs with MLP Imputation approach

	Confusion Matrix			ACC	SEN	SPE	MAE	RMS
MLP	Status	Classified CAD	Classified NCAD	87.5	0.9	0.18	0.14	0.33
	CAD	289	25					
	NCAD	32	113					
SVM	Status	Classified CAD	Classified NCAD	85.62	0.85	0.15	0.14	0.38
	CAD	297	17					
	NCAD	49	96					
LMT	Status	Classified CAD	Classified NCAD	85.3	0.88	0.19	0.16	0.36
	CAD	288	26					
	NCAD	39	106					
RF	Status	Classified CAD	Classified NCAD	86.27	0.84	0.09	0.21	0.31
	CAD	305	9					
	NCAD	54	91					

On the other hand, when the SOM imputation process is operated, the relationship between features of dataset is examined and MVs are predicted one by one by excluding other features including MVs. In the course of this prediction, it is observed that the size of the map affects the result. In order to eliminate limit restrictions, a square map that can be created based on the shape is used (Fessant, 2002). 8 × 8 SOM network is selected by testing maps of various dimensions ranging between 4 × 4 and 20 × 20. While dimensions smaller than 8 × 8 do not present the data distribution satisfactorily, larger models are observed to fit the training data. Figure 8.1 illustrates ROC curves for the classification performances based on four imputation procedures. Moreover, the MVs-deleted dataset condition was used as a baseline of performance.

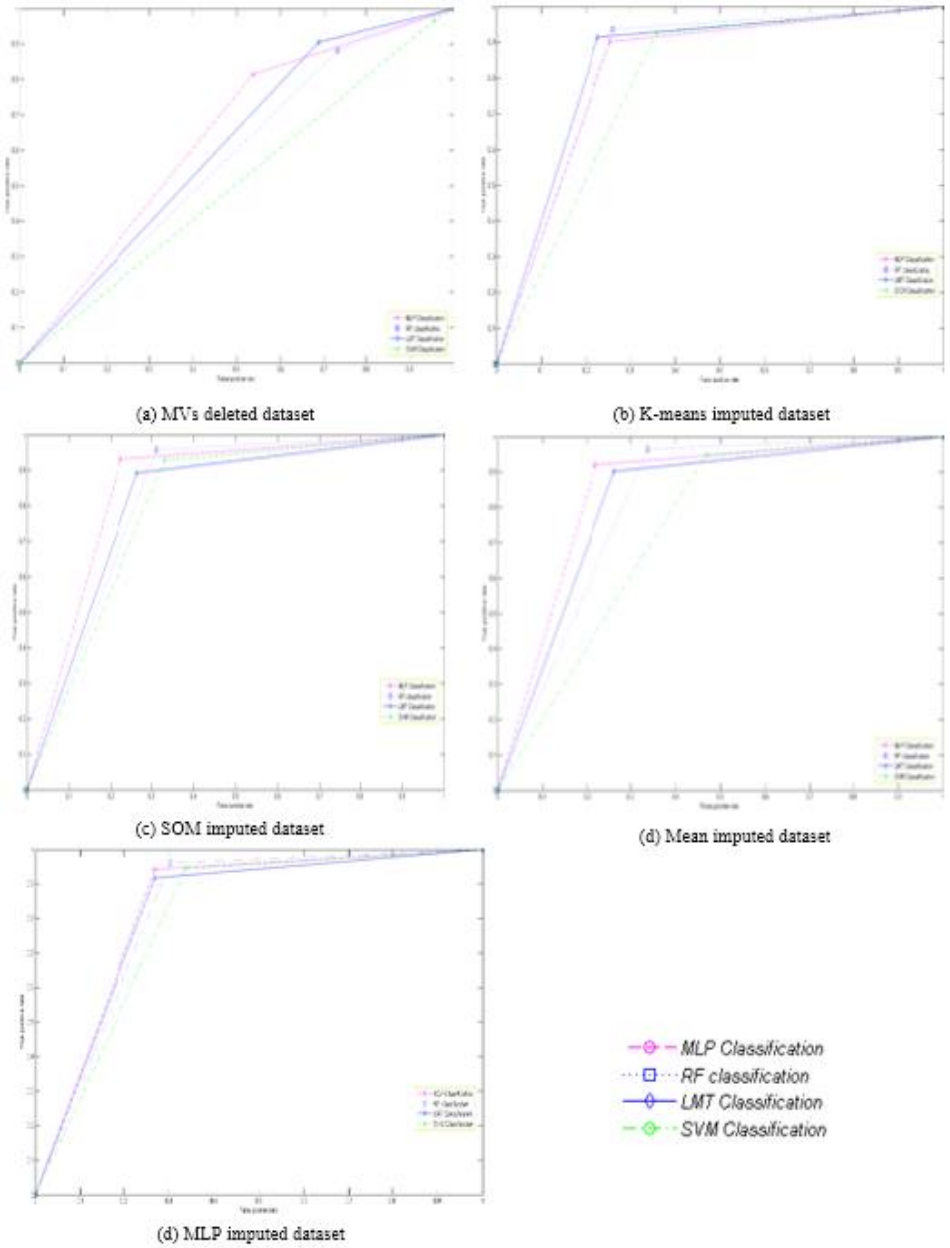

Figure 8.1. ROCs comparisons of classifiers based on imputed dataset

NRMSE that measures the normalized difference between predicted and observed values is adopted as one of the performance indicators in this study. SOM imputation method and tested by the MLP classifier yielded the best results with the RMSE value of 0.245 overall and MVs - deleted condition resulted the worst. NRMSE results were presented in Table 8.4.

Table 8.4. NRMSE values for each imputation-classification method pairs

	SOM Imp	**MLP Imp**	**K-means Imp**	**Mean Imp**	**MVs Del**
LMT	0.141	0.169	0.213	0.154	-0.375
MLP	***0.245***	0.206	0.162	0.231	-0.197
RF	0.170	0.189	0.205	0.145	-0.427
SVM	0.106	0.120	0.070	-0.064	***-1.714***

For the original part of dataset with 277 samples, LMT achieved an accuracy ratio of 76.17% and the sensitivity value of 0.51, which was the top value for this dataset. This accuracy value remains under 7.49 points obtained by K-means imputation method and tested by the SVM classifier which is the lowest performance between imputations – classifications pairs.

The results indicated 87.58% accuracy on two datasets that are imputed using both the mean method and the MLP imputation method and tested by the MLP classifier. In addition, this dataset has the highest sensitivity, with a value of 0.9, as seen in Table 8.5. It is observed that the lowest accuracy (81.69%) and the lowest sensitivity (0.82) were obtained in the tests conducted using the SVM classifier. The classification performance on datasets obtained as a result of imputation methods is presented in Table 8.5 and Table 8.6.

Table 8.5. Classification results from different type of MV replacement methods

		ACC	Sen	Spec	F_{meas}	Prec
Mean	**SVM**	81.69	0.82	0.17	0.81	0.81
Imp.	**RF**	86.92	0.86	0.09	0.875	0.877
	MLP	87.58	0.88	0.15	0.875	0.875
	LMT	84.8	0.88	0.23	0.849	0.848
K-means	**SVM**	83.66	0.85	0.19	0.835	0.83
Imp.	**RF**	87.36	0.87	0.1	0.87	0.87
	MLP	85.18	0.88	0.22	0.85	0.85
	LMT	86.92	0.89	0.19	0.868	0.868
MLP	**SVM**	85.62	0.86	0.15	0.856	0.856
Imp.	**RF**	86.27	0.85	0.09	0.866	0.869
	MLP	87.58	***0.9***	***0.18***	0.875	0.875
	LMT	85.38	0.88	0.19	0.857	0.856
SOM	**SVM**	84.31	0.85	0.17	0.842	0.843
Imp.	**RF**	87.14	0.87	0.11	0.861	0.87
	MLP	***88.23***	0.89	0.1	***0.879***	***0.881***
	LMT	84.09	0.87	0.22	0.84	0.84

Table 8.6. Classification results in the case of deleting MVs

		ACC	Sen	Spec	F_{meas}	Prec
Deleting	**SVM**	74.76	0.51	0.19	0.74	0.73
MV Cases	**RF**	75.81	0.5	0.21	0.72	0.72
	MLP	72.92	0.44	0.17	0.73	0.73
	LMT	76.17	0.51	0.19	0.74	0.73

When K-means imputation method is considered, the sensitivity results of the LMT classifier have the highest values among the classifiers. For the dataset prepared by the SOM imputation method, MLP achieved an accuracy ratio of 88.23%, which was the top value among the imputation–classification method pairs.

It is significant to evaluate the prediction performance of classifiers for CAD from different perspectives in clinical data analysis. Performance results based on evaluation metrics are presented in Figure 8.2.

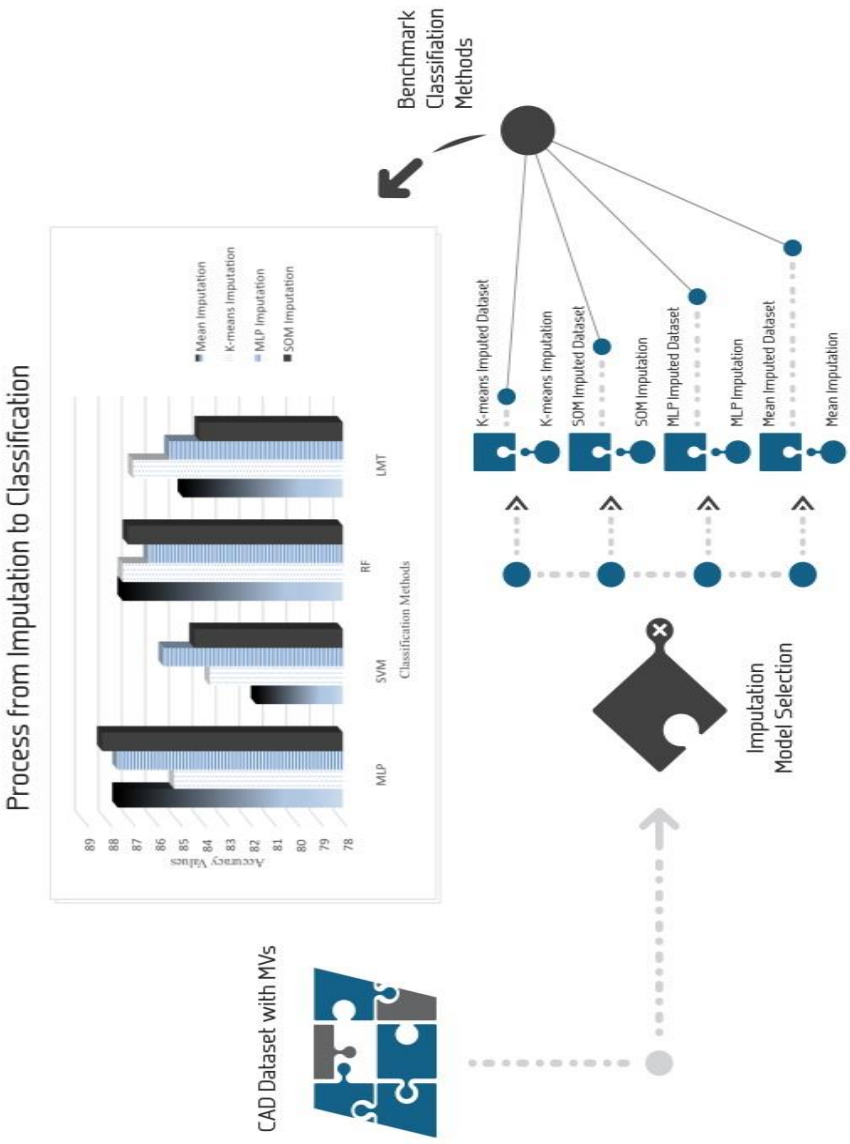

Figure 8.2. Accuracy results of classification methods for mean, K-means, SOM, and MLP imputed datasets

ROC field is calculated for all imputation–classification method pairs and illustrated in Figure 8.3. The sensitivity value of 0.9 and specificity value of 0.18 tested by the MLP classifier are marked in Figure 8.3. This combination performed better than machine learning and mean method with regard to both MV imputation and classification. Moreover, other metrics were evaluated for all imputation–classification pairs and SOM imputation method has the best values obtained by MLP, with the accuracy of 88.23%, precision of 0.88, and F-measure of 0.879. Other metric results are presented in Figure 8.4.

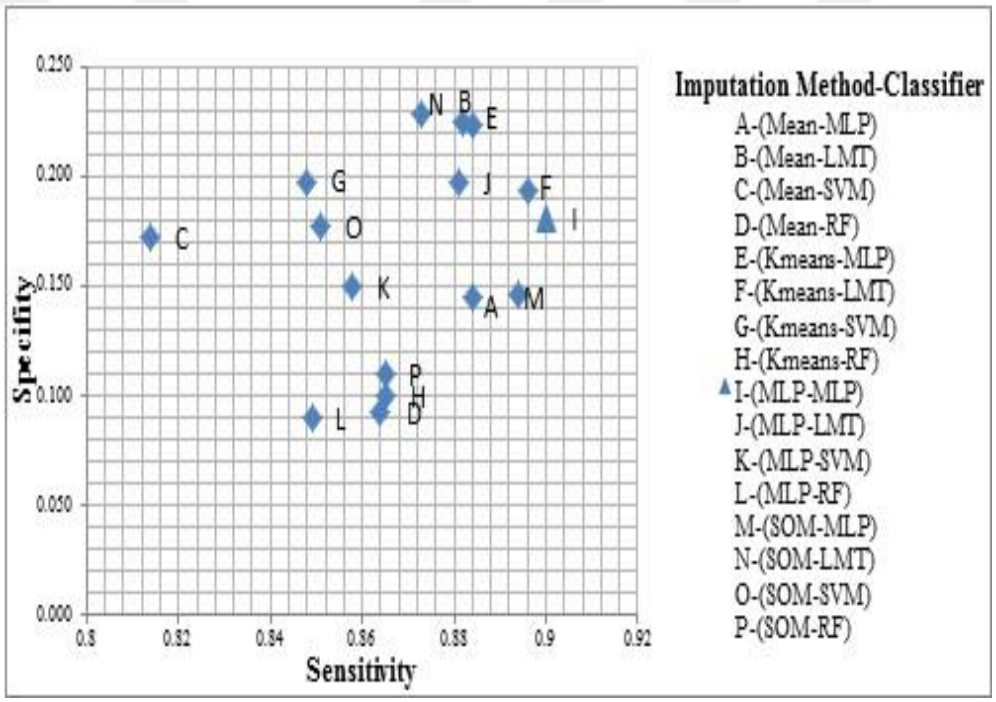

Figure 8.3. ROCs of classification results of mean, K-means, SOM, and MLP imputed datasets in turn with MLP, LMT, SVM, and RF classifiers

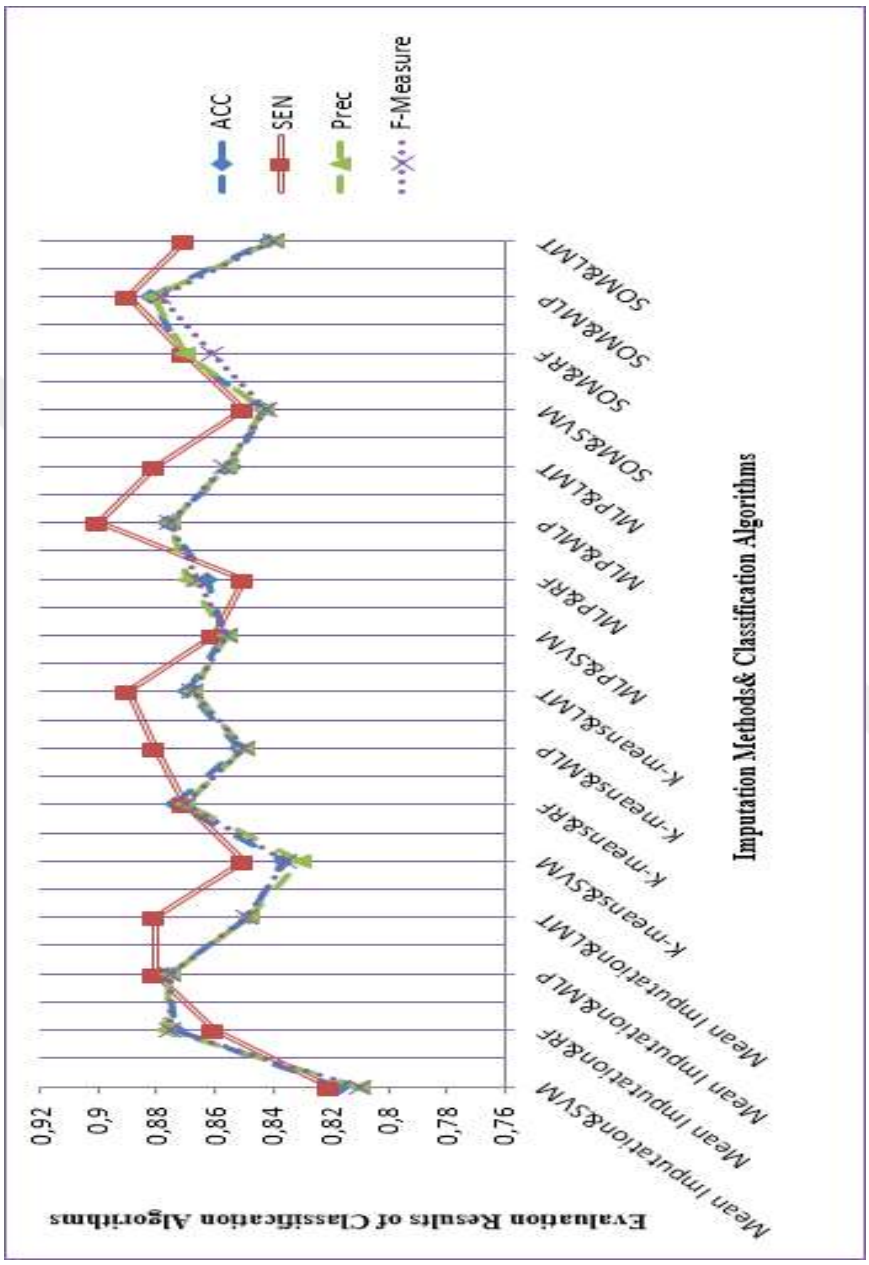

Figure 8.4. Evaluation metrics of classification results of mean, K-means, SOM and MLP imputed datasets

Moreover, according to the average values of the classification results on the basis of the imputation method, it is found that the dataset imputed by MLP is more stable than other imputed datasets and increases the performance of all classifiers in the study. The results are presented in Figure 8.5.

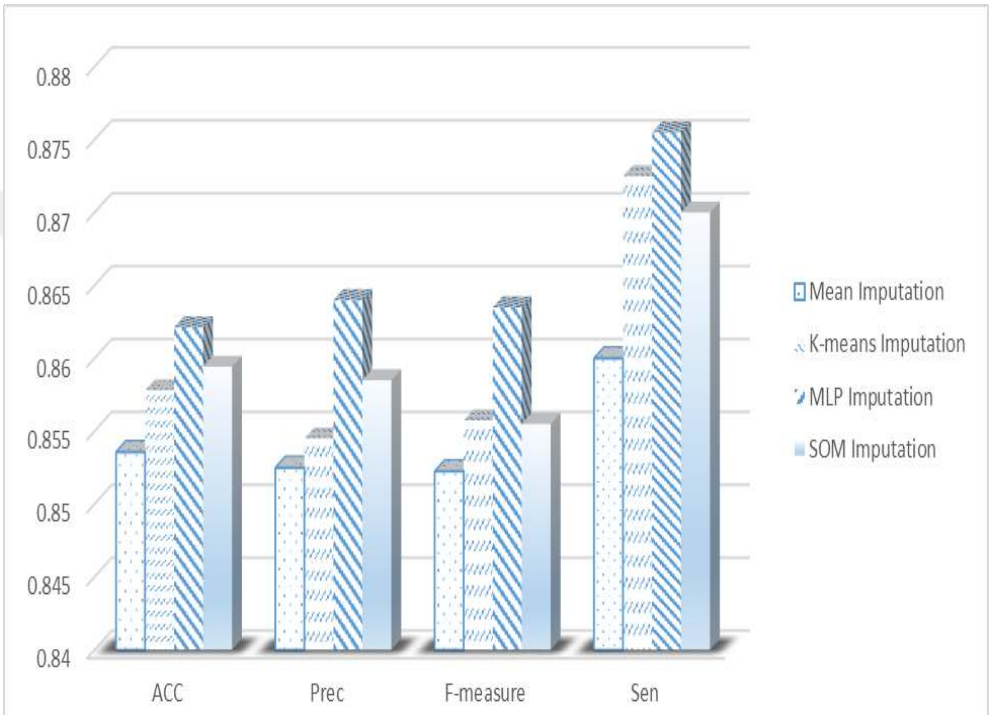

Figure 8.5. Evaluation metrics of average classification results of mean, K-means, MLP, and SOM imputed datasets with classifiers

8.2. Classification Results of Imbalanced 4-Classed Cardiovascular Data Using Feature Selection and Sampling Methods

8.2.1. Logistic Regression Analysis With Feature Selection and Resampling Methods

Independent t-test analysis is used to select candidate variables out of 21 variables. The threshold value has been identified as $p < 0.05$ to find the significant variables that may statistically distinguish the samples that are CAD of not CAD. Wald value is an assessment measurement in LRA. Wald is accepted as important

when bigger than two. Probability value is also calculated. Furthermore, the effect of variable selection on the classification performances is considered in each step.

After variable selection, 10 variables are defined to form the LRA model; and the accuracy increased to 77.1% (0.6 sensitivity and 0.81 specificity) from 72.8% (0.6 sensitivity and 0.81 specificity).

LRA equation is:

$$u = -1{,}908 - 0.036X_{age} + 0.021\mathrm{X}_{HDL} - 0.007\mathrm{X}_{FBS} + 0.004\mathrm{X}_{total} + 0.054\mathrm{X}_{Ef} - 0.326\mathrm{X}_{gender} - 0.184\mathrm{X}_{DM} + 0.332\mathrm{X}_{Hyp} + 0.817\mathrm{X}_{smoke} - 0.376\mathrm{X}_{asthma}.$$

Dependent variable is integrated into four different classes. The classification performance decreases when the number of classes increases from 2 to 4. Logistic model produces 64.5% accuracy with 0.66 sensitivity and 0.87 specificity.

Resampling method was applied and the accuracy of 77.3% with the sensitivity of 0.79 and the specificity of 0.75 are taken over the resampled dataset after logistic model construction for two classes. LRA is performed for four classes by using the same variables that are used above. Logistic model produces 61.4% accuracy that is lower than before i.e. the resampling application.

Another perspective is to exclude the feature that includes two category information (CAD or without CAD). Then, this feature is added to the dataset, and the t-test analysis is applied to choose new candidate features. Fifteen new variables were obtained according to the t-test analysis results. So, LRA equation is:

$$u = 3.878 - 0.039X_{age} + 0.022X_{HDL} - 0.005X_{FBS} + 0.004X_{total} + 0.038X_{Ef} + 0.029X_{hb} - 0.288X_{cr} + 0.020X_{rdw} + 0.030X_{mcv} - 30034X_{gender} - 30740X_{dm} - 11092X_{hyp} - 4.698X_{smoke} - 7.668X_{KOAH} + 2.492X_{FH}.$$

Multivariate LRA is applied to classify the resampled dataset with 2 classes and 79.1% accuracy is obtained with 0.79 sensitivity and 0.79 specificity. However, the logistic model produces 61.4% accuracy when it is applied to 4 classes.

The resampling method is applied using 4 class information. Then, the Relief-f method is used for eliminating the features. 12 features (gender, age, hyp, ef, dm, hdl, smoke, fh, total, fbs, hb, hl) are selected for further operations. According to the new features, the LRA model gives the best results with 69.2%.

The oversampling, and then the t-test method is used, and finally 9 variables are defined. The LRA classification gives an accuracy of 82.0% with the 0.98 sensitivity and the 0.42 specificity on the oversampled dataset with 2 classes and new features. According to the new features, LRA model gives the poorest results with 55.1% over 4 classes.

$$u = 0{,}474 - 0.058X_{age} + 0.034X_{HDL} - 0.009X_{FBS} + 0.064X_{Ef} + 0.008X_{LDL} + 14.36X_{DM} - 3.48X_{Hyp} - 1.03X_{smoke} - 2.45X_{gender}.$$

The results of three t-test analyses are shown in Table 8.7. As shown in this part, the information of which the features are related for each three tests are identified. Effective seven features are chosen to take into three LRA models. Three irrelevant features TG, urea and renal failure are not used for classification.

Table 8.7. Features and coefficients involved into Logistic regression model after t-test analyses

Feature name	t-test	Sampling & t-test	Oversamp. & t-test	Coefficient
Gender	☑	☑	☑	-10012.3 ± 14157.5
Age	☑	☑	☑	-0.044 ± 0.009
Hyperlipidemi	☑	☑	☑	-3698.38 ± 5228
DM	☑	☑		10241.9 ± 14494
COPD		☑		−7.668
Smoke	☑	☑	☑	-1.637 ± 2.29
Asthma	☑			−0.376
Hypertension			☑	-3698.38 ± 5228. 08
LDL				0.008
HDL	☑	☑	☑	0.026 ± 0.007
HB		☑		0.029
CR		☑		−0.288
FBS	☑	☑	☑	0.007 ± 0.002
RDW		☑		0.02

Table 8.7. Features and coefficients involved into Logistic regression model after t-test analyses (Continued)

MCV		☑		0.03
TC	☑	☑		0.004
EF	☑	☑	☑	0.052 ± 0.013
Family History		☑		2.49
Renal Failure	☒	☒	☒	
Urea	☒	☒	☒	
TG	☒	☒	☒	

8.2.2. Performance comparisons of feature selection and sampling methods

The performance measures for two methods are presented in Table 8.8, and the results of the feature selection and the sampling methods are discussed. The experiments on original CAD dataset are placed in the first part, resampling method and feature selection effects on the classification algorithms are given in the second part; and finally, the oversampling and the feature selection method results are given in the third part.

For the resampling dataset, the accuracy and specificity of all the methods have increased because of the resampling process. Moreover, using Relief-f with LRA achieved higher accuracy than LRA with the t-test, which is the best result among the LRA experiments. Also, the relative order of the classification methods according to the evaluation metrics (accuracy, F-measure, precision, sensitivity, and specificity) are similar in the second part and third part of Table 8.8 when considering the NN results.

Table 8.8. Evaluation results of different processes of CAD data set with 4 classes

	Original CAD Dataset			Resampled CAD dataset				Oversampled CAD dataset		
	NN	NN& Relief-f	Log.R. & t-test	Relief-f				Relief-f		
				Log. Reg &t-test	NN	Log.R	NN	Log.R.	NN	NN
ACC %	72.3	71.2	64.5	61.44	80.8	69.28	83.4	55.1	83.1	84.1
Weight. F-Meas	0.72	0.70	0.620	0.529	0.758	0.638	0.800	0.546	0.828	0.839
Prec	0.751	0.718	0.632	0.602	0.770	0.683	0.837	0.545	0.828	0.840
Sen	0.701	0.698	0.660	0.513	0.750	0.613	0.777	0.550	0.831	0.840
Spec	0.870	0.865	0.870	0.809	0.920	0.858	0.931	0.786	0.936	0.941

Oversampled CAD dataset with 729 samples is experimented; oversampling has increased the accuracy and the other results in comparison to the first and also second part, except for the LRA method. The LRA eliminates the repetitions in the dataset, and works with a subpopulation extracting redundant samples. It shows that selecting a subset could mislead the LRA classifier into false predictions. The highest accuracy in this case is obtained by NN and Relief-f feature selection method, which is 84.12%.

The chart that consists of the evaluation metrics versus percentages are given in Figure 8.6. Each line corresponds to one process step for the classification methods. The results are ordered from the highest performance to the lowest. The highest performance is obtained by the NN with oversampled dataset and the Relief-f feature selection method; and afterwards, again, NN with resampled dataset and Relief-f feature selection, NN with oversampled dataset, NN with resampled dataset, respectively. The lowest four performances belong to the LRA despite all the enhancing operations.

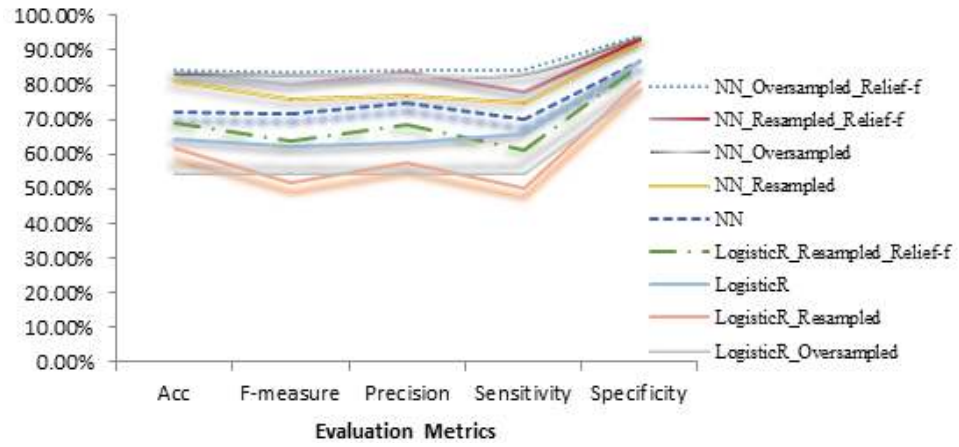

Figure 8.6. Diagram of evaluation metrics versus percentages. Each line corresponds to one process

Considering previous research, there is no relationship between classification performance and data structure, and the results of classifiers vary according to data sets in different data structures. In classifying the two-class CAD

data, the ANN method produces very good results compared to LR and RBF (Lin et al., 2010). Moreover, as the number of class increases to four, sampling (Poolsawad et al., 2014; Kurt et al., 2008) and feature selection methods are required to evaluate the final performance of the classification algorithms. The results show how the right choice for methodology has been made in this regard. Flow chart and the sequence of the processes are given in Figure 8.7.

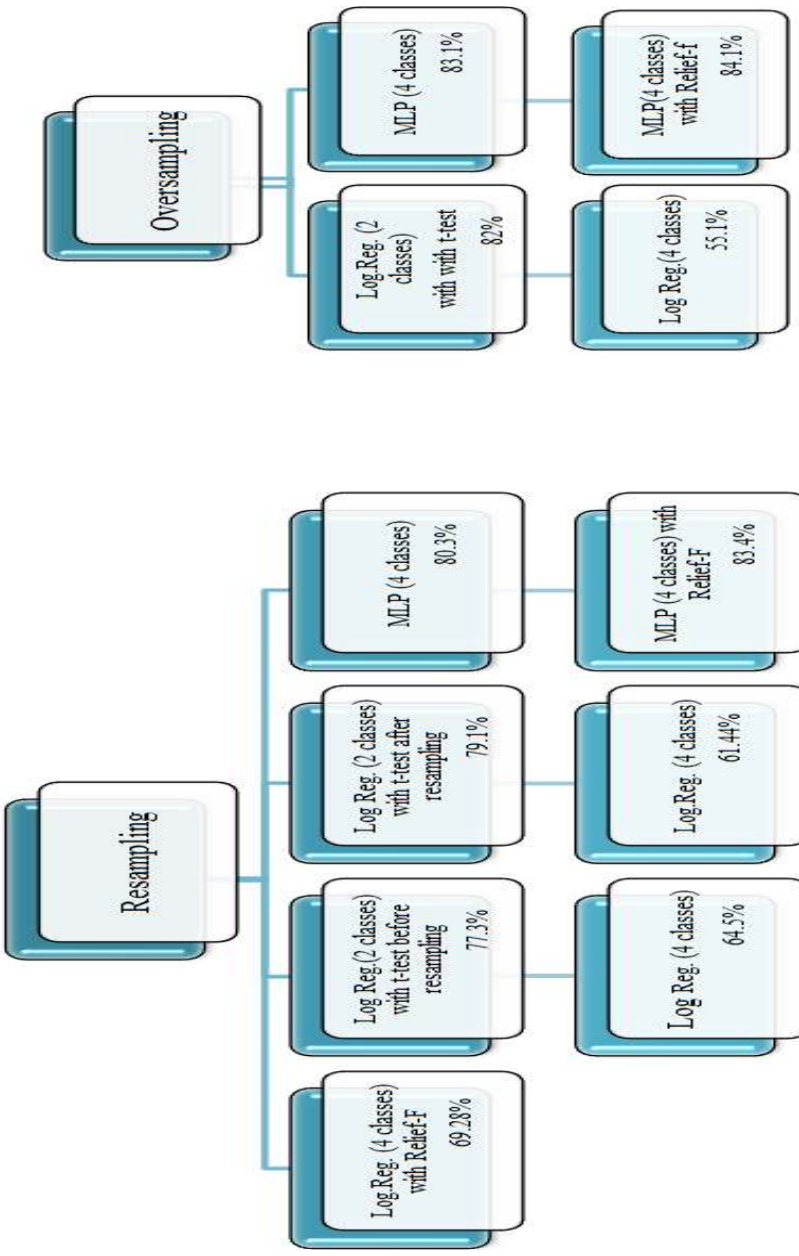

Figure 8.7. Effects of sampling processes on accuracy performance

8.3. Risk Score Prediction Process

Consequent 343 samples are used for training and 116 samples are used for testing, totally 459 samples. As for the dataset 182 individuals are of “moderate” out of 459 records (39.65% of the total records) and also 164 individuals are defined as “high” out of 459 records (35.73% of the total records). Dataset shows an imbalance on status of “low” or “very high”. According to the risk assessments only 43 individuals are of “moderate” out of 459 records (9.36% of the total records) and also 70 individuals are defined as “high” out of 459 records (15.25% of the total records).

At this stage of the study, it is important to use the confusion matrix to predict how well the classifier make right predictions one of the four classes and capture the deviations. Among four classes performances are evaluated regarding the confusion matrix in Table 8.9. Table 8.10 that consists of measurements would contain derivations from four situations combined.

Table 8.9. Confusion matrix for 116 individuals

		Class			
		Low	Moderate	High	Very high
Actual Class	Low	46	0	0	0
	Moderate	0	6	4	1
	High	0	0	11	7
	Very high	0	0	13	28

Table 8.10. Measurements from four confusion matrixes are integrated

Spec	***Sen***	***ACC***	***Fmeasure***
0.93	*0.71*	*0.88*	*0.73*

Regarding to the measurements in Table 8.10, MLP network was saved and a GUI is implemented using Matlab Environments and a prediction system was constructed which can be used for every sample. Blank GUI is given in following Figure 8.8 and figures are placed consequently for different input parameters. Final results are given in Figure 8.9, and Figure 8.10 produced by MLP network.

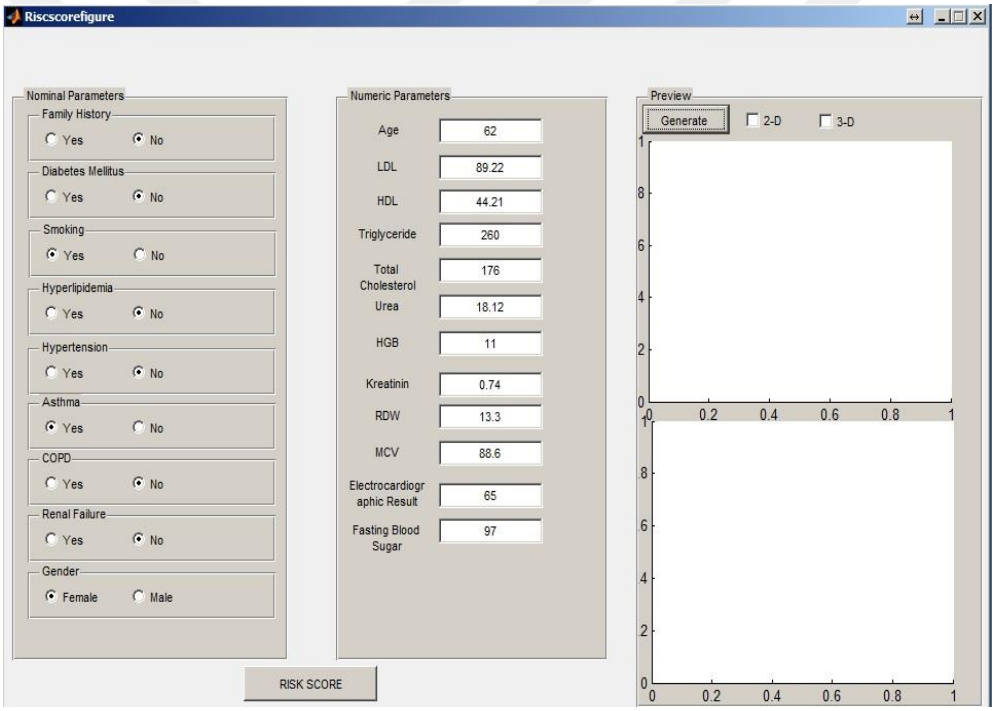

Figure 8.8. Blank GUI developed for risk score prediction system

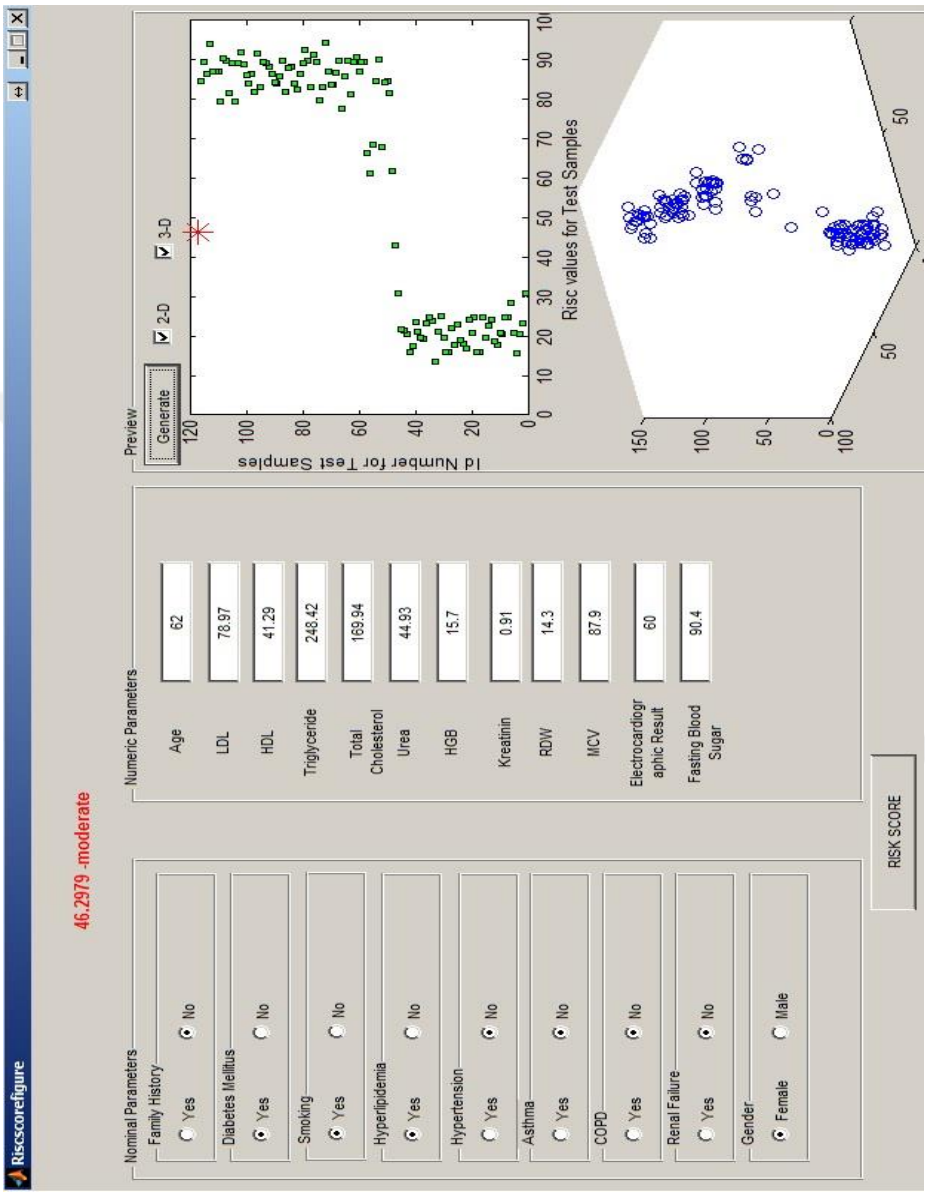

Figure 8.9. System prediction for given input parameters

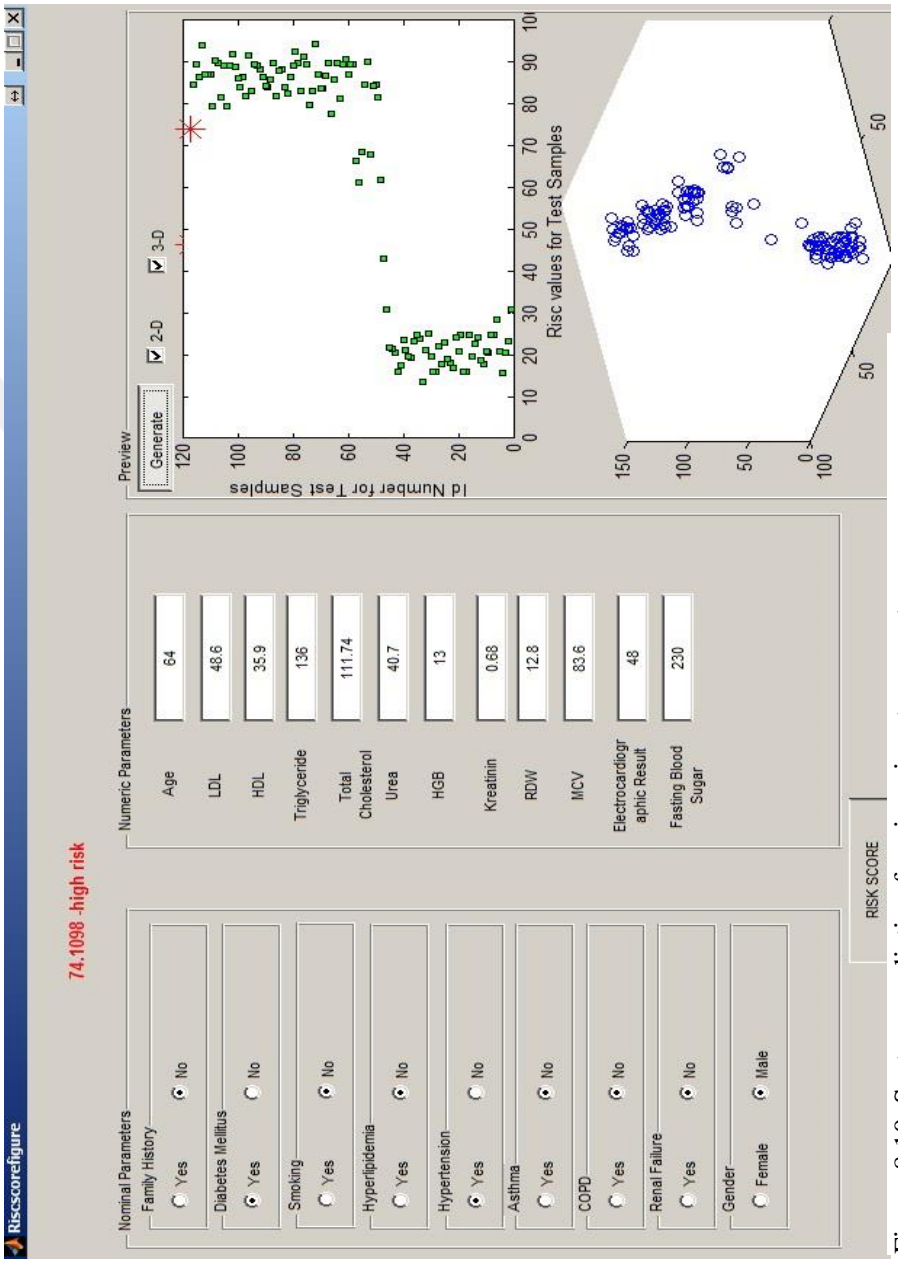

Figure 8.10. System prediction for given input parameters

8.4. HCA Results on Three Distinctive Datasets

The WDBC dataset has been presented to the classifiers as raw without any pre-processing steps. However, a transformation process was performed on the Dermatology dataset. Class distribution of dermatology dataset consists of 5 different values. It is not intervened to class number #1 that corresponds to 112 samples and other 254 samples are associated with unique class that is labelled with class number #2.

Clustering results are presented for each k = 2, 3, 4 and for negative samples. WDBC dataset is clustered and input data distribution for two features versus cluster number is given in Figure 8.11.

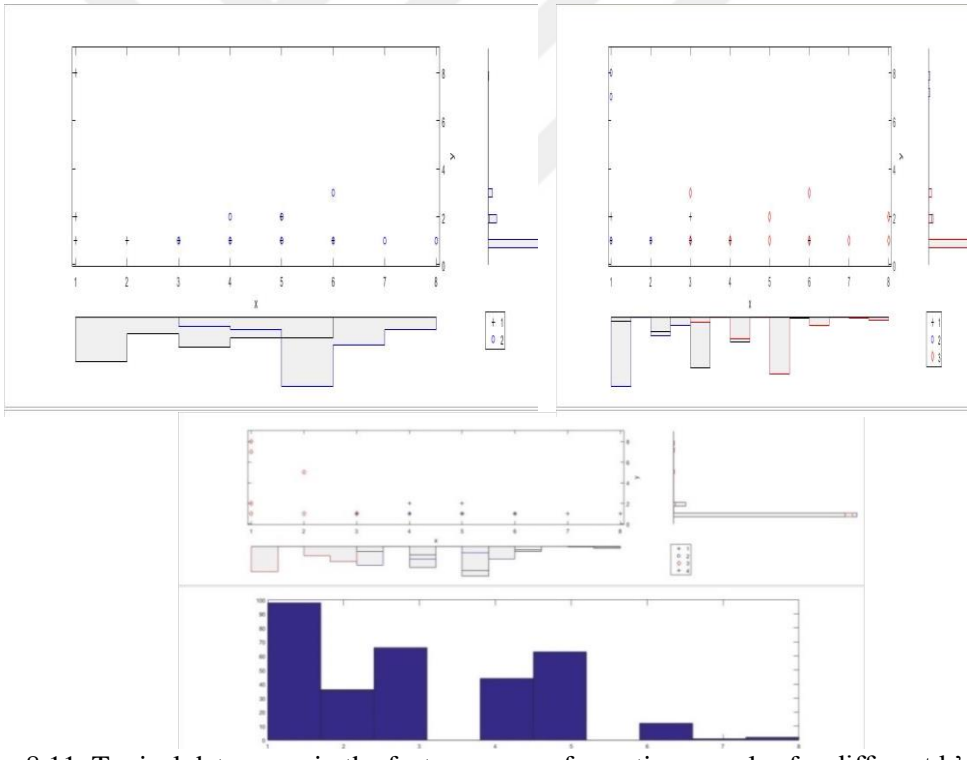

Figure 8.11. Typical data maps in the feature space of negative samples for different k's. Histograms are given for k = 2, k = 3, k = 4, respectively

HCA algorithm is executed with different cluster numbers, 210 samples are tested for WDBC dataset and also 110 samples for Dermatology dataset. Each negative subset is merged with positive samples and presented to the SVM classifiers as $Xn_i \cup X_p$. If each classifier is defined as $\aleph_i$, then n number classifiers in total are obtained. The margin distance around the hyper plane is calculated as the mean distance of the support vectors. The mean distance is accepted as threshold value.

In the test dataset, prediction probabilities for all testing samples are computed using logistic regression model. Testing sample is accepted as positive, when probability value is larger than the threshold value, otherwise the sample is labelled as negative. False positive and false negative rates are taken into consideration and classifier performances are given in Table 8.11.

Table 8.11. Hyb.Proc. = SVM Linear + Logistic regression, SVM L = SVM classifier with linear kernel, SVM R = SVM classifier with radial base kernel function for different k's

Method		**k**	**TP**	**FN**	**FP**	**TN**	**ACC**	**Sen**	**Spe**
Hyb. Proc	**WDBC**	2	66	1	8	135	0.96	0.89	0.99
		3	74	5	0	131	0.98	1	0.96
		4	69	1	5	135	0.97	0.93	0.99
SVM L.			66	1	7	136	0.96	0.95	0.98
SVM RBF			62	32	5	111	0.84	0.75	1
Hyb. Proc	**Dermatolog**	2	23	1	12	74	0.88	0.66	0.98
		3	28	8	8	66	0.85	0.79	0.89
		4	34	2	4	70	0.94	0.89	0.97

The results are compared between Hybrid, SVM Linear and SVM RBF for WDBC dataset. It is found that Hybrid has more balanced false positive and false negative rates. It is observed to have an impact for different clusters numbers (k = 2, 3, 4) of the Hybrid procedure, the best balanced results were taken for k=3. However, Hybrid procedure evaluation metrics are better than using SVM alone with linear or RBF kernel.

According to the Table 8.11 for Dermatology dataset, the most effective results are taken for k=4. Depending on the natural structure of this dataset it is known that class number #2 has the knowledge to carry information for 4 different sample groups. It is likely that the best evaluation performance is obtained for k = 4. This suggests that subset separation value, which delivers the best result at different values for both datasets, can play an important role in the results of Hybrid procedure.

Hybrid procedure is executed for the Dermatology dataset and different k values were evaluated as well as WDBC. ROC curves are integrated for different k's of Hybrid procedure and given in Figure 8.12 and Figure 8.13 respectively.

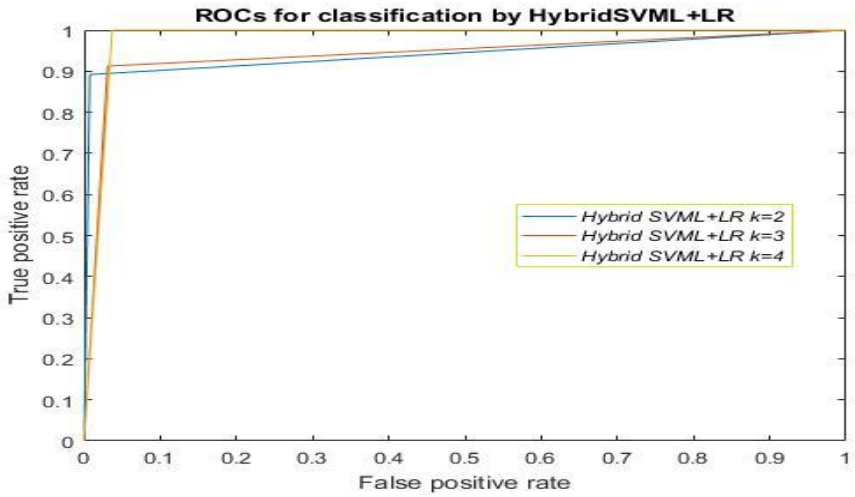

Figure 8.12. ROC curves for k = 2, 3, 4 of Hybrid procedure for WDBC dataset

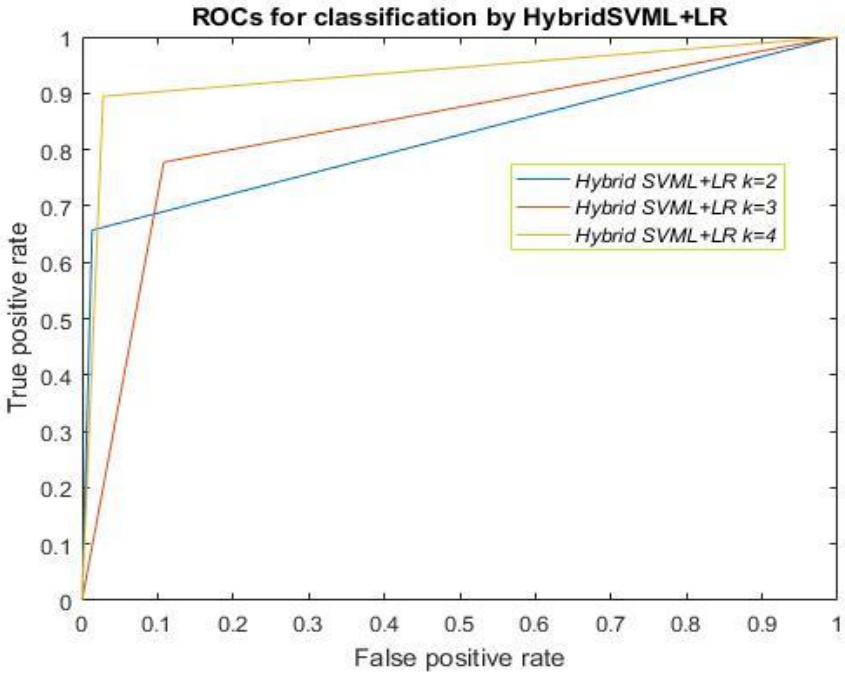

Figure 8.13. ROC curves for k = 2, 3, 4 of Hybrid procedure for Dermatology dataset

Clustering results are presented for each k = 2, 3, 4 and for positive samples. When training dataset SVM supports mapping the predictor data using kernel functions Therefore, iteration number is exceeded to 500,000 for CAD dataset to avoid maximum convergence obstacle. Experiments showed the iteration number is needed to be increased for nonlinear datasets. Another parameter is C for the soft margin. C can be a scalar, or a vector of the same length as the training data and increasing C value might decrease the number of support vectors, but also might increase training time. Experiments also showed to decrease the number of support vectors is needed to increase the performance. In this study 50 for C value is chosen.

After the parameters are edited, each positive subset is merged with negative samples and presented to the SVM classifiers as $Xn_i \cup X_p$. If each classifier is defined as $\aleph_i$, then n number classifiers in total are obtained. The

margin distance around the hyper plane is calculated as the mean distance of the support vectors. The mean distance is accepted as threshold value.

In the test dataset, prediction probabilities for all testing samples are calculated using logistic regression model. Testing sample is accepted as positive, when probability value is larger than the threshold value, otherwise the sample is labelled as negative. To evaluate Hybrid method, classifier performances are given in Table 8.12.

Table 8.12. Hybrid Proc. = SVM Linear + Logistic regression with linear kernel function for different k's of CAD dataset

Method	**k**	ACC	Sen	Spe	F-measure	AUC	NRMSE
Hybrid	2	0.82	0.81	0.82	0.76	0.81	0.124
Proc.	3	0.83	0.77	0.86	0.79	0.82	0.138
	4	0.89	0.85	0.92	0.85	0.88	0.320

It is observed to have an impact for different clusters numbers (k = 2, 3, 4) of the Hybrid procedure, the best balanced results were taken for k=4. However, as the number of k increases, it evaluation metrics gradually heals. ROC curves were given in Figure 8.14.

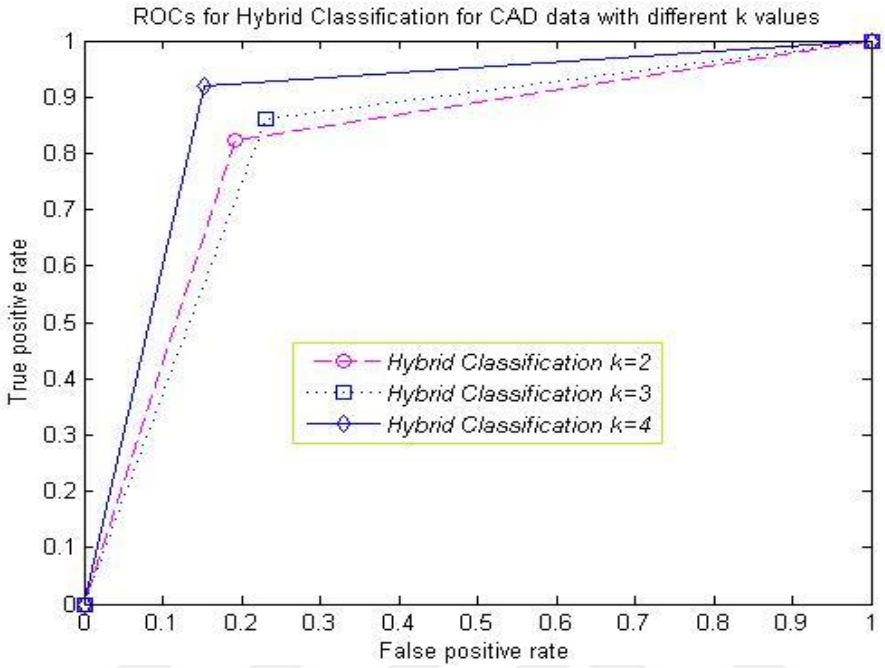

Figure 8.14. ROC curves for k = 2, 3, 4 of Hybrid procedure for CAD dataset

The Hybrid procedure uses a clustering algorithm such as k-means to separate class samples into smaller subsets which is moreover. The testing is made with different partition values in such a manner, in which the ratio of classes to each other must not be too far from 1. In this study, clusters numbers (k = 2, 3, 4) are chosen respectively. In the next step, independent classifiers which use the integration of the other class samples and one of the subsets are utilized. Then, the LR Model is used to integrate and convert these independent classifiers into a result that may be decided collectively on. The hybrid procedure is used to measure the evaluation metrics to display the success of the model. Linear kernel for SVM and binomial LR are used in order not to increase the complexity of the model. This approach enables us to develop an efficient algorithm, which solves the problem with all dataset training at one time.

Depending on the natural structure of the datasets the cluster number may differ as for the knowledge to carry information. It is likely that the best evaluation performance is obtained for k = 4. .

9. CONCLUSION AND FUTURE RESEARCH

This study presents three imputation procedures, which are applied on the CAD dataset consisting of 459 successive cases so as to solve MVs problem. MVs are imputed with these procedures, that are constructed by MLP, SOM, and K-means algorithms and complete datasets are produced at the end of the processes. The CAD classification process is performed by using four different classifiers, and the prediction performance of the classifiers is gauged. When the MLP classifier is used on the dataset to which the MLP imputation algorithm is applied, it produces better results than the other methods, with a sensitivit0y of 0.9 and a specificity value of 0.18. According to the average classification results, the MLP imputation method produces more stable results than all the methods operated with different types of classifiers, giving a sensitivity of 0.873 and specificity value of 0.179. It is expected that a proposed algorithm can generate even better results compared to the previous methods, such as the mean method, which is widely used by researchers and has been used in this study as a reference. Statistical mean method ranks fourth among the algorithm applications. However, when different types of classifiers are considered for all other cases, it should be noted that application of the imputation technique by using the MLP algorithm is more effective than when other algorithms are used according to the average metric values.

Another point to note is that as the degree of stenosis increases to four for CAD diagnosis, this situation creates four class label, it becomes inevitable to face the problem of imbalanced data. Effective features are examined on the CAD dataset by the t-test analyses and Relief-f methods. The effective seven features are identified to take into all three LRA models. On the other hand, the Relief-f method is used for eliminating the features. 12 features (gender, age, hyp, ef, dm, hdl, smoke,fh, total, fbs, hgb, hl) are selected for further operations. According to the new features, the LRA model gives the best accuracy of 69.2%. Significant increase is seen on the LRA performance with the Relief-f feature selection method

rather than the t-test feature selection method. Moreover, the resampling and oversampling techniques were applied to increase the performance of the both LRA and MLP classification; and remarkable accuracy values were achieved. These values consisted of 84.1% of accuracy, 0.839 of F-measure, 0.84 of precision, 0.84 of sensitivity, and 0.94 of specificity were obtained by NN with the oversampled dataset and the Relief-f feature selection method. The sampling methods used in this thesis were successful in increasing all the evaluation metrics except LRA.

To improve the classification effect of imbalanced datasets, we propose HCA which utilizes the SVM with LR up to the number of partitions selected during clustering is used. HCA uses a clustering algorithm such as k-means to separate class samples into smaller subsets for classification with SVM which is moreover. Then, the LR model is used to integrate and convert these independent classifiers into a result that may be decided collectively on. Linear kernel for SVM and binomial LR are used in order not to increase the complexity of the model. This approach enables us to develop an efficient algorithm, which solves the problem with all dataset training at one time.

It can be concluded that, using the feature selection and the sampling methods with the NN substantially improves the prediction accuracy as well as the other metrics when working with imbalanced data such as the CAD dataset. The proposed MVs algorithms and sampling methods will be tested on different real healthcare datasets which have imbalanced characteristic to indicate the power of the proposed algorithms in the future. Also, better accuracy estimations are achieved with promising results on classification with HCA when compared with SVM that uses RBF and linear kernel. In this way, it can be concluded that HCA can provide a good starting point in the development of different hybrid algorithms in the future. Depending on the natural structure of the datasets the number of partitions can be found automatically by different clustering algorithms in the future.

REFERENCES

Allison, P. D., 2001. Missing data. Sage Publications, Thousand Oaks, CA, US.

Alizadehsani, R., Habibi, J., and Hosseini, M., 2013. A data mining approach for diagnosis of coronary artery disease. Computer Methods and Programs in Biomedicine, 111(1): 52-61.

Alizadehsani, R., Zangooei, M., Hosseini, H., et al., 2016. Coronary artery disease detection using computational intelligence methods. Knowledge-Based Systems, 109: 187-197.

Altman, D.G., Vergouwe, Y., Royston, P., and Moons, K. G., 2009. Prognosis and prognostic research: validating a prognostic model. Bmj, 338: b605.

Barbero, A., Takeda, A., and López, J., 2015. Geometric Intuition and Algorithms for Eν–SVM. Journal of Machine Learning Research, 16(1): 323-369.

Bassuk, S. S., and Manson, J. E., 2008. Lifestyle and risk of cardiovascular disease and type 2 diabetes in women: A review of the epidemiologic evidence. American Journal of Lifestyle Medicine, 2(3): 191-213.

Berner, Eta S., 2007. Clinical Decision Support Systems. Springer Science, 233.

Bernstam, E. V., Herskovic, J. R., and Reeder. P., 2010. Oncology research using electronic medical record data. In: J Clin Oncol, 28(15):16501.

Bishop, C.M., 2006. Pattern recognition and machine learning. Springer, New York.

Breiman, L., 2001. Random forests. Machine learning, 45: 5-32.

Cassar, A., Holmes, D.R., Rihal, C.S., and Gersh, B.J., 2009. Chronic coronary artery disease: diagnosis and management. In: Mayo Clinic Proceedings, Elsevier, 84(12):1130-1146.

Castelli, W. P., Dawber, T., and Feinleib, M., 1981. The filter cigarette and coronary heart disease: The Framingham study, The Lancet, 318(8238): 109-113.

Chang, Y.I., 2003. Boosting SVM classifiers with logistic regression, See "www. stat. sinica. edu. tw/library/c_tec_rep/2003-03. Pdf".

Chapelle, O., 2002. Choosing multiple parameters for support vector machines, Machine learning, 46(1): 131-159.

Chaudhuri, B.B., Bhattacharya, U., 2000. Efficient Training and Improved Performance of Multilayer Perceptron in Pattern Classification, Neurocomputing, 34:11-27.

Colak, M.C., Colak, C., Kocaturk, H., et al., 2008. Predicting coronary artery disease using different artificial neural network models. Anadolu kardiyoloji dergisi: AKD= the Anatolian journal of cardiology, 8: 249-254.

Compare, A., Grossi, E., Buscema, M., Zarbo, M., et al, 2013. Combining Personality Traits with Traditional Risk Factors for Coronary Stenosis: An Artificial Neural Networks Solution in Patients with Computed Tomography Detected Coronary Artery Disease, Cardiovascular psychiatry and neurology, 2013.

Consultation, 2000. W. H. O. Obesity: preventing and managing the global epidemic. World Health Organization technical report series, 160: 2581-9.

Duda, R.O., hart, P.E., and stork, D.G., 2012. Pattern classification. John wiley & sons.

Eckel, R. H., Krauss, R. M., and Ronald, M., 1998. American Heart Association call to action: obesity as a major risk factor for coronary heart disease. Circulation, 97(21): 2099-2100.

Elbashir, M.K., 2013. Predicting beta-turns in proteins using support vector machines with fractional polynomials. Proteome science, 11(1):1.

Ergun U., Serhatlioglu, S., Hardalac, F. et al., 2004. Classification of carotid artery stenosis of patients with diabetes by neural network and logistic regression. Computers in biology and medicine, 34: 389-405.

Farhangfar, A., Kurgan, L., and Dy, J., 2008. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition, 41(12): 3692-3705.

Fessant, F., and Midenet, S., 2002. Self-organising map for data imputation and correction in surveys. Neural Computing & Applications, 10(4): 300-310.

Gajawada, S., and Toshniwal, D., 2012. Missing Value Imputation Method Based on Clustering and Nearest Neighbours. International Journal of Future Computer and Communication, 1(2): 206-208.

Garcia-laencina, P.J., Sancho-Gomez, J. L., Figueiras-Vidal, A. R., and Verleysen, M., 2009. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing, 72(7): 1483-1493.

Hambrecht, R., Wolf, A., and Gielen, S., 2000. Effect of exercise on coronary endothelial function in patients with coronary artery disease. New England Journal of Medicine, 342(7): 454-460.

Haykin, S., 1994. Neural Networks: A Comprehensive Foundation, New York, Macmillan College Publishing Company Inc. Heart Disease Health Center. Coronary Artery Disease Directory. http://www.webmd.com/heart-disease/coronary-artery-disease-directory (Access Date: April 3 2013)

Hennekens, C. H., 1998. Increasing Burden of Cardiovascular Disease Current Knowledge and Future Directions for Research on Risk Factors. Circulation, 97(11): 1095-1102.

Heron, M., Hoyert, D. L., and Murphy, S. L., 2009. National vital statistics reports. National Vital Statistics Reports, 57:14.

Hosmer J.R., David, W., Lemeshow, S., 2013. Sturdivant, Rodney X.Applied logistic regression.Wiley.http://www.who.int/mediacentre/factsheets/fs317/en/index.html(Access Date: September 3 2013).

Jayasurya, K., Fung, G., and Yu, S., 2010. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. Medical physics, 37: 1401.

Jerez, J.M., Molina, I., Garcia-Laencina, P. J., Alba, E., Ribelles, N., Martin, M., and Franco, L., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial intelligence in medicine, 50(2): 105-115.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biological cybernetics, 43(1): 59-69.

Kononenko, i., 1994. Estimating attributes: analysis and extensions of relief. In: machine learning: ecml-94. Springer berlin heidelberg. P. 171-182.berkson, j. Application of the logistic function to bio-assay. Journal of the american statistical association, 39(227): 357-365.

Kurt, I., Ture, M., and Kurum, A. T., 2008. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Systems with Applications, 34.1: 366-374.

Kültürsay, H., 2001. Koroner Kalp Hastalığı Primer ve Sekonder Korunma.Argos İletişim Hizmetleri Reklamcılık ve Ticaret Anonim Şirketi. 101-190.

Landwehr, N., Hall, M., and Frank, E., 2005. Logistic model trees. Machine Learning, 59(1): 161–205.

Larifla, L., Armand, C., Velayoudom-Cephise, F.L. et al., 2014. Distribution of coronary artery disease severity and risk factors in Afro-Caribbeans. Archives of cardiovascular diseases, 107:212-218.

Lee, Y., Shin, J.H., Park, H.C., Kim, S.G., et al., 2014. A prediction model for renal artery stenosis using carotid ultrasonography measurements in patients undergoing coronary angiography. BMC nephrology, 15: 1- 60.

Li, H., Sun, J., and Wu, J., 2010. Predicting business failure using classification and regression tree: An empirical comparison with popular classical

statistical methods and top classification mining methods. Expert Systems with Applications. 37: 5895-5904.

Lin, J. H., Haug, P. J., 2006. Data preparation framework for preprocessing clinical data in data mining. In AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2006:489.

Lin, C.C., Bai, Y. M., Chen, J. Y., Hwang, et al., 2010. Easy and low-cost identification of metabolic syndrome in patients treated with second-generation antipsychotics: artificial neural network and logistic regression models. Journal of Clinical Psychiatry, 71(3): 225.

Little, R.J., and Rubin, D.B., 2002. Statistical Analysis with missing data. 2nd ed. New York, Wiley.

Longadge, R., and Dongre, S., 2013. Class imbalance problem in data mining review. arXiv preprint arXiv.1305:1707.

Lu, M., and Yang, W., 2012. Multivariate logistic regression analysis of complex survey data with to BRFSS data. Journal of Data Science, 10:157-173.

Maji, S., Berg, A. C., and Malik, J., 2013. Efficient classification for additive kernel SVMs. IEEE transactions on pattern analysis and machine intelligence, 35(1):66-77.

Maluenda, G., Delhaye, C., and Gaglia J.R., 2010. A novel percutaneous coronary intervention risk score to predict one-year mortalitiy, The American journal of cardiology, 106(5):641-645.

Mingoti, S.A., and Lima, J.O., 2006. Comparing SOM neural network with Fuzzy c means K means and traditional hierarchical clustering algorithms. European Journal of Operational Research, 174(3): 1742-1759.

Nakazato, R., Dey, D., Cheng, V.Y., et al., 2012. Epicardial fat volume and concurrent presence of both myocardial ischemia and obstructive coronary artery disease. Atherosclerosis, 221(2): 422-426.

Natarajan, Kersting, K., and Ip, E., 2013. Early Prediction of Coronary Artery Calcification Levels Using Machine Learning. In: Twenty-Fifth IAAI Conference.

National cholesterol education program national heart, lung, and blood institute, 2002. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) Final Report. National Institutes of Health, NIH Publication No. 02- 5215.

Nlm, S. National Library of Medicine. National Institutes of Health. http://www.nlm.nih.gov/ (Access Date: September 3 2013).

Onat, A., Sansoy, V., and Soydan, İ., 2003.TEKHARF; Oniki Yıllık İzleme Deneyimine Göre Türk Erişkinlerinde Kalp Sağlığı. Argos İletişim Hizmetleri Reklamcılık ve Ticaret Anonim Şirketi.

O'rourke, R.A., Valentin Fuster, V., and Alexander, R.W., 2001. Hurst's the heart manual of cardiology. McGraw-Hill Medical, p.1065-1109.

Osuna, E., Freund, R., and Girosi, F., 1997. Support vector machines: Training and applications.

Pampel, F. C., 2000. Logistic regression: A primer (quantitative applications in the social sciences). SAGE Publications, Thousand Oaks, CA.

Podgorelec, V., Kokol, P., and Stiglic, B., 2002. Decision trees: an overview and their use in medicine. Journal of medical systems, 26(5): 445-463.

Poolsawad, N., Kambhampati, C., and Cleland, J. G. F., 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, 828-837.

Poolsawad, N., Kambhampati, C., and Cleland, J. G. F., 2014. Balancing class for performance of classification with a clinical dataset. In Proceedings of the World Congress on Engineering, Vol. 1.

Pope, J.H., Aufderheide, T.P., RuthazeR, R., et al., 2000. Missed diagnoses of acute cardiac ischemia in the emergency department. N Eng J Med, 342(16):1163 - 1170.

Rahman, M.M., and Davis, D.N., and Darryl N., 2012. Fuzzy unordered rules induction algorithm used as missing value imputation methods for K-Mean clustering on real cardiovascular data, Lect Notes Eng Comput Sci, 2197(1):391--4.

Rahman, M. M., and Davis, D. N., 2013. Addressing the class imbalance problem in medical datasets. International Journal of Machine Learning and Computing, 3(2): 224.

Ramirez, J., Gorriz, J. M., Segovia, F., Chaves, R., et al., 2010. Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification. Neuroscience letters, 472: 99-103.

Richman, M.B., Trafalis, T.B., and Adrianto, I., 2007. Multiple imputation through machine learning algorithms. In: Fifth conference on artificial intelligence applications to environmental science.

Richman, M.B., Trafalis, T.B., and Adrianto, I., 2009. Missing data imputation through machine learning algorithms. In: Artificial Intelligence Methods in the Environmental Sciences. Springer Netherlands. p. 153-169.

Robnik-Sikonja, M., and Kononenko, I., 2003. Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53(1-2): 23-69.

Rokach, L., 2008. Data mining with decision trees: theory and applications. World Scientific.

Tsumoto, S., 2000. Problems with mining medical data. In: Computer Software and Applications Conference. COMPSAC 2000. The 24th Annual International. IEEE, 467-468.

Van der wal, A. C., and Becker, A. E., 1999. Atherosclerotic plaque rupture–pathologic basis of plaque stability and instability. Cardiovascular research, 41(2): 334-344.

Vapnik, V.N., 2000. The Nature of Statistical Learning Theory, 2. Edition, Springer-Verlag, New York.

White, H., 1989. Learning in artificial neural networks: A statistical perspective. Neural computation, 1(4): 425-464.

Wang, H., Fan, W., Yu, P. S., and Han, J., 2003. Mining concept-drifting data streams using ensemble classifiers. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (ACM, 226-235.DOI= http://doi.acm.org/10.1145/956750.956778.)

Wang, L., Zhu, J., and Zou, H., 2007. Hybrid huberized support vector machines for microarray classification. In Proceedings of the 24th international conference on Machine learning (Corvalis, Oregon, USA — (June.2007), 20 – 24. DOI= http://dl.acm.org/citation.cfm?doid=1273496.1273620)

Wang, H., and Wang, S., 2010. Mining incomplete survey data through classification. Knowledge and information systems, 24(2): 221-233.

WHO, World Health Organization, 2013.

CURRICULUM VITAE

Jale BEKTAŞ was born in Adana, Turkey, in 1979. She received a B.S. degree from the Department of Computer Engineering from Mersin University, Turkey, in 2002. She earned a M.S. degree from the Department of Electrical and Electronics Engineering at Çukurova University in Adana, Turkey, in 2009.

She worked as software engineer at the Norsan Computer Systems Import and Export Co. from 2002 to 2006. She developed and managed chambers of commerce and industry automation which has been using at 16 chambers of commerce and industry since 2003. Since 2006, she has been working as the lecturer at School of Applied Technology and Management Department of Mersin University, Erdemli, Mersin, Turkey.

Her research interests are in machine learning, software engineering and data mining.