A STUDY OF A GROUP OF HEALTHCARE DATASETS IN DATA MINING DOMAIN

by

FİKRİ MERT KURUM

B.S., Computer Science and Engineering, Işık University, 2007

Submitted to the Institute for Graduate Studies in

Science and Engineering in partial fulfilment of

the requirements for the degree of

Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2013

A STUDY OF A GROUP OF HEALTHCARE DATASETS IN DATA MINING DOMAIN

APPROVED BY:

Prof. Fikret Gürgen . . . . . .

(Thesis Supervisor)

Assist. Prof. Arzucan Özgür . . . . . .

Assoc. Prof. Cengizhan Öztürk . . . . . .

DATE OF APPROVAL: 23.01.2013

ACKNOWLEDGEMENTS

I would like to thank my thesis supervisor Prof. Fikret Gürgen, for his support throughout this study with not only the subject itself but also many other common interests of ours. With his significant contribution, I was able to finish this thesis work.

I further would like to thank to my family and friends for their continuous encouragement and support.

Last but not least, a special thanks to my dear friends, Mr. Ömer Uzun for their encouragement and concern to my study. There is a tremendous sense of support and guidance in completing this study.

ABSTRACT

A STUDY OF A GROUP OF HEALTHCARE DATASETS IN DATA MINING DOMAIN

This research is to search for alternatives to the resolution of complex medical diagnosis where human knowledge should be apprehended in a general fashion. Successful application examples show that human diagnostic capabilities are significantly worse than the neural diagnostic system.

The study presents the particular case of analysis of eleven datasets containing data associated to several Healthcare datasets. The datasets are analyzed in various Healthcare domains to target different Medical areas. Paradigm of artificial neural networks is shortly introduced and the main problems of medical data base and the basic approaches for training and testing a network by medical data are described. There are eight algorithms used in this study, which are DT, SVM, RBF, MLP, k-NN, Naïve Bayes, Bayes Net and Logistic Regression. These eight algorithms have been performed with using 10-fold cross validation and train/test split over the eleven datasets. It's also examined what is the effect of Principal Component Analysis inside the research.

The performance metrics that are focused in this thesis are Percent Correct, True Positive Rate, False Positive Rate, Precision, Recall, F-Measure, AUC and Error Rates.

As this is a benchmarking study for different classifiers and datasets, a special benchmarking criterion has been created for the evaluation of the thesis.

ÖZET

BİR GRUP SAĞLIK VERİSİ ÜZERİNE VERİ MADENCİLİĞİ ÇALIŞMALARI

Bu araştırma farklı tıbbi tanılar için insan bilgisinin yanında, veri madenciliği üzerinde farklı alternatifler aramak içindir. Yapılan birçok başarılı uygulama örnekleri içerisinde insan teşhis yeteneklerinin sonuçlarının nöral teşhis sistemi sonuçlarına göre daha kötü sonuçlar verdiği göstermektedir.

Çalışma sağlık alanındaki Onbir adet verinin veri madenciliği analizi ile ilgilidir.Çalışma içerisinde farklı Tıbbi alanlar hedeflenerek Veri Madenciliği analizlerinin çeşitli Sağlık verilerine yaptığı yorumlar incelenmiştir. Yapay sinir ağlarının sağlık alanındaki çalışmaları kısaca tanıtıldı, tıbbi veri tabanı ve eğitim, ve tıbbi verilerin bir sinir ağı test edilmesine yönelik temel yaklaşımların üzerinde duruldu. Birkaç test yapılandırmaları, algoritmalar için en iyi ayarı belirlemek için test edilmektedir. Bu çalışmada kullanılan Sekiz adet veri madenciliği algoritmaları bulunmaktadır. Detayları ise şu şekildedir: DT, SVM, RBF, MLP, k-NN, Naive Bayes, Bayes Net ve Lojistik Regresyon. Belirtilen Sekiz adet algorithma “10-fold cross validation” ve “train/test split” değerlendirmeleri göz önüne alınarak Onbir adet veri üzerinde değerlendirilmiştir. Bununla beraber PCA için de ayrı bir değerlendirme gerçekleştirilerek araştırma içerisindeki farklılıkları gösterilmiştir.

Tez içerisinde odaklanılan performans metrikleri ise şu şekildedir: Percent Correct, True Positive Rate, False Positive Rate, Precision, Recall, F-Measure, AUC ve Error Rates.

Bu tez farklı algoritmalar ve veriler üzerien bir kıyaslama çalışması olduğundan dolayı, tez çalışmasının değerlendirilmesi için özel bir kıyaslama ölçütü oluşturulmuştur.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS / ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
AI	Artificial Intelligence
AUC	Area under the curve
BN	Bayesian networks
CV	Cross Validation
CVD	Cardiovascular Disease
DT	Decision Tree
ERNN	Extended Recurrent Neural Network
FIR	Finite Impulse Response
FK-NN	Fuzzy and Nearest Neighbor Algorithm
K-NN	K Nearest Neighbors
LDA	Linear Discrimination Analysis
LR	Logistic Regression
MDM	Minimum Distance-to-Mean
MLBPNN	Multilayer BackPropagation Neural Network
MLC	Maximum Likelihood Classification
MLP	Multilayer Perceptron
NB	Naive Bayes
NN	Neural Networks
PCA	Principal Component Analysis

PWV	Pulse Wave Velocity
RBF	Radial Basis Function
SEER	Surveillance Epidemiology and End Results
SEP	Score for Expression Profile
SVM	Support Vector Machine
TN	True negative
TP	True positive
TPR	True positive rate

1. INTRODUCTION

The major problem in healthcare field is to diagnose disease. Human being make mistake and because of its limitation, diagnosis would give the major issue of human expertise. One of the most important problems of medical diagnosis, in general, is the subjectivity of the analysts. It can be noted, in particular in pattern recognition activities that the experience of the professional is closely related to the final diagnosis. This is due to the fact that the result does not depend on a systematized solution but on the interpretation of the patient's signal (Lanzarini, 2000).

The health of population, which is based primarily on the result of medical research, has a strong impact upon all human activities. Among the most important medical aspects are considered the good interpretation of data and setting the diagnosis. But medical decision making becomes a very hard activity because the human experts, who have to make decisions, can hardly process the huge amounts of data. So they need a tool that should be able to help them to make a good decision. They could use some expert systems or artificial neural networks, which are part of artificial intelligence. Doctors use a combination of a patient's case history and current symptoms to reach a health diagnosis when a patient is ill. In order to recognize the combination of symptoms and history that points to a particular disease, the doctor's brain accesses memory of previous patients, as well as information that has been learned from books or other doctors. A neural network has the ability to mimic this type of decision-making process, and use a knowledge base of information, and a training set of practice cases, to learn to diagnose diseases (Dehariya, 2011).

1.1. Research Overview

It is possible to identify certain diseases and problems with using different classification techniques analyzed in this thesis. This analysis can be used by any medical

field practices by doctors, specialists or even researchers. To do that, it's important to look at diagnosis results and target classes if there is an infection or not.

Medical databases as there are examples in other domains such as banking and finance, has large amount of data about patients, medical conditions and diseases medical details. The relationship and patterns within this data could provide new medical knowledge and the analysis in medical field could open up new treatment technique coming from historical information and analysis and information from different inputs.

In the research, there are 11 sets of data in medical field that have been analyzed in order to help decision making process. This medical datasets are coming from different resources and fortunately there are datasets analyzed which is coming from a local entity and have not been analyzed in Machine Learning Domain before.

There are also different classification methods applied in this research to focus different classification techniques advantages comparing other methods in the field. Percent Correct is the number one objective; however there are various machine learning objectives such as Percent Correct, True Positive Rate, False Positive Rate, Precision, Recall, F-Measure and AUC.

1.2. Proposed Method

Before utilizing a model pronounced by a classification algorithm, it is assessed with respect to some criterion. The model will probably result in some errors therefore the data miner should take it into account while selecting a model (Cios, 2007). Accuracy, which is the percentage of instances that are correctly classified by the model, is the most commonly used decision criteria for the model assessments (Han, 2005).

However, there is also other criterion used to compare and evaluate the models. Berson defines the assessment concept as accuracy, explanation and integration abilities (Berson, 1999). Rokach introduces the comparison criterion as the generalization error of the model, the computation complexity that is the amount of CPU consumed by inducer, the comprehensibility that is the ability to understand the model, the scalability that is the ability to run efficiently on larger databases, the robustness that is the ability to handle

missing or noisy data, the stability that is the ability to produce repeatable results on different datasets and lastly the interestingness that is ability to generate valid new knowledge (Maimon, 2005).

For all algorithms, splitting the data into train and test splits has been selected as the validation method. 66% of the data has been set as the training part and the rest has been set as the testing part.

Also 10-fold cross validation has been implemented on the same datasets for all algorithms.

For Feature Extraction, Principal Component Analysis (PCA) is used.

There are also eight Classification algorithms used in the thesis, which are J48 – Decision Tree, Support Vector Machine, Radial Basis Function, Multilayer Perceptron, k-Nearest Neighbors, Naïve Bayes, Bayes Net and Logistic Regression.

1.3. Research Questions

To summarize the general objectives of this study, there are four research questions focused in the thesis:

Research Question 1 (RQ1):

Does implementing the same classification algorithms on multiple datasets and with different implementation techniques result in significantly different performance indicators?

Research Question 2 (RQ2):

How is the evaluation of the training performance in terms of Percent Correct, True Positive Rate, False Positive Rate, Precision, Recall, F-Measure, AUC and Error Rate for different datasets and algorithms?

Research Question 3 (RQ3):

What is the best implementation within this experiment?

Research Question 4 (RQ4):

Do Probabilistic classifiers have any advantages comparing other classifiers used in this thesis?

1.4. Thesis Outline

Following chapters describe thesis study in detail.

Chapter 2 talks about Literature Review and Application of Classification Algorithm Techniques applied within the Healthcare domain of this study.

Chapter 3 includes Methodology of the Thesis with focusing into the Data Selection, Data Analysis and Preprocessing, Data Reduction and Transformation, Feature Selection and Extraction, Machine Learning Process, and Evaluation and Knowledge Utilization

Chapter 4 focuses into Feature Selection and Feature Extraction, and Classification techniques J48 – Decision Tree, Support Vector Machine, Radial Basis Function, Multilayer Perceptron, k-Nearest Neighbors, Naïve Bayes, Bayes Net and Logistic Regression in detail context.

Chapter 5 describes experiments on datasets and their results among performance metrics.

Chapter 6 shares detailed analysis

Chapter 7 is the conclusion of the Thesis with the project report where compares several classifiers techniques in attempt to measure optimal classifier using difference type of datasets in terms of Four Questions listed in Section 1.3.

2. LITERATURE REVIEW AND RECENT STUDIES WITHIN THE DOMAIN

In literature, data mining community has been very interested in comparing classification type of algorithms but they usually compare the classifiers on a single dataset or they compare only a few of the algorithms not including the recent ones.

It is not easy to find empirical results of how classifiers perform on different multiple datasets; therefore the basic concern about this study is the repetitive algorithm implementations on multiple datasets thus some idea about the effects of dataset characteristics on the performance can be derived from the study. The same concern is also valid for the complexity comparison that is the consumed CPU time by each classifier.

Knowledge discovery process of data mining projects include the data pre-processing stage and recommends steps such as data cleaning, reductions, discretisations or component analysis if necessary. This study aims to find out if those data pre-processing activities have any effect on the classifier accuracy or model development time. Lastly the study aims to figure how robust the selected classifiers are. In order to understand their robustness, iterative implementations are done before and after cleaning noise in the datasets.

2.1. Literature Review

(Brause, 2001) and (Dehariya, 2011) highlighted that almost all the physicians are confronted during their formation by the task of learning to diagnose. Here, they have to solve the problem of deducing certain diseases or formulating a treatment based on more or less specified observations and knowledge. For this task, certain basic difficulties have to be taken into account:

- The basis for a valid diagnosis, a sufficient number of experienced cases, is reached only in the middle of a physician's career and is therefore not yet present at the end of the academic formation. (Fuster, 2001)
- This is especially true for rare or new diseases where also experienced physicians are in the same situation as newcomers. (Fuster, 2001)
- Principally, humans do not resemble statistic computers but pattern recognition systems. Humans can recognize patterns or objects very easily but fail when probabilities have to be assigned to observations. (Zadeh, 1969)
- The quality of diagnosis is totally depends on the physician talent as well as his/her experiences.(Fuster, 2001)
- Emotional problems and fatigue degrade the doctor's performance. (Fuster, 2001)
- The training procedure of doctors, in particular specialists, is a lengthily and expensive one. So even in developed countries we may feel the lack of MDs. (Fuster, 2001)
- Medical science is one of the most rapidly growing and changing fields of science. New results disqualify the older treats, new cures and new drugs are introduced day by day. Even unknown diseases turn up every now and then. So a physician should always try hard to keep his/herself up to date. (Pomi, 2006)

Regarding problems above and also many others, the question would be how computers could help in medical diagnosis. Since decades ago, computers have been employed widely in the medical sector. From local and global patient and medicine databases to emergency networks, or as digital archives, computers have served well in the medical sector. Meanwhile, in the case of medical diagnosis, regarding the complexity of the task, it has not been realistic yet to expect a fully automatic, computer-based, medical diagnosis system. However, recent advances in the field of intelligent systems are going to materialize a wider usage of computers, armed with AI techniques, in that application. A computer system never gets tired or bored, can be updated easily in a matter of seconds, and is rather cheap and can be easily distributed. Again, a good percentage of visitors of a clinic are not sick or at least their problem is not serious, if an intelligent diagnosis system

can refine that percentage, it will set the doctors free to focus on nuclear and more serious cases.

Datamining for the discovery of knowledge in databases have been generating a significant amount of interest in research (Ramaswamy, 2000). Other cases by (Brause, 2001) also give an example of a study in the year 1971, showed these basic facts in the medical area. This study had shown that human have many limitations in diagnosis. The results of this experiment show that best human diagnosis (most experienced physician) give 79.7%, computer with expert database give 82.2% and computer with 600 patient data give 91.1%.

(Ultsch, 1995) used the capability of Neural Networks to diagnose acidosis diseases by using knowledge based system in their hybrid system. The data set consists 11 attributes originating from the blood analysis. Several classification methods were used to explain this data. The Neural Network together with the UMatrix method was able to classify the data into the subcategories healthy, lacacidemia, metabolical acidosis, respiratory acidosis and one patient with cerebral deficiency.

(Florea, 2004) compared the feature selection and classification techniques for medical images modality categorization. Here, they investigated the automatic categorization of medical images according to their corresponding modalities; using five classifiers. The classifiers are Multilayer Perceptron (MLP), Support Vector Machine (SVM), Logistic Model Trees (LMT), Random Forest and K Nearest Neighbors. Moreover, the experiments explained in their study used WEKA with medical images database. It was found that the best classifications results were obtained with MLP, resulting nearly 90% of precision, but at high time-cost. The SVM classifier was reported to be the poorest ones in precision, with all the considered vectors. MLP and LMT took several hours of computation time, but obtained better classification results.

Making prognosis for patients with congestive heart failure is difficult due to the complex nature of this multisystem disease. No single criterion helps to identify patients at risk, and a combination of several prognostic parameters is recommended (Cowburn, 1998). Neural Networks are associative self learning techniques with the ability to identify multidimensional relationships and perform pattern recognition in non-linear domains. (Kononenko, 1993) conducted a study test in four medical diagnostic problems used

Inductive and Bayesian Learning in Medical Diagnosis. The results showed that Naïve Bayesian Classifier, despite its naivety achieved better classification accuracy. In another research by (Koutroumbas, 2001) that considered other application methods are: Decision Tree Induction, Boosted Decision Trees, Naïve Bayesian Classifier, and Radial Basis Function Neural Networks. In their study, they examined which method that is suitable for specific medical diagnosis task. The study showed that AdaBoost Method had high performance to diagnose cytology by combining different classifiers.

(Zhou, 2002) named his an automatic pathological diagnosis procedure named Neural Ensemble Based Detection (NED). It is proposed and realized in early stage Lung Cancer Diagnosis System (LCDS). NED utilizes an Artificial Neural Network ensemble to identify cancer cells in the images of specimens of needle biopsies obtained from the bodies of the subjects to be diagnosed. A fast adaptive neural network classifier had been used to identify the lung cancer cell. In their study, stated that a fast adaptive neural classifier that performs one-pass incremental learning with fast speed and high accuracy and does not require the user manually set up the number of hidden units.

Accurate risk stratification of heart failure patients is critical to improve management and outcomes. Heart failure is a complex multisystem disease in which several predictors are categorical. Neural Network Models have successfully been applied to several medical classification problems. By using a Simple Neural Network, (Atienza, 2000) assessed one-year prognosis in 132 patients, consecutively admitted with heart failure, by classifying them in three groups: death, readmission and one-year event-free survival. Given the small number of cases, the Neural Network model was trained using a resampling method. They indentified relevant predictors using the Automatic Relevance Determination (ARD) method, and estimated their mean affect on 3 different outcomes.

Fan *et al.* had used data mining techniques in predicting breast cancer recurrence. This technique has been chosen by researchers due to the fact that it had proved its efficiency in medical field. For this research, a large data was taken from SEER (Surveillance Epidemiology and End Results) to provide good resources when analyzing. Five data mining algorithms were investigated and were used to predict the breast cancer recurrence based on SEER. SPSS is thus used to prepare and classify data. The results have shown a promising outcome of predicting breast cancer through the use of data mining

techniques. The researchers have provided evidence in this study that the data mining techniques could also be use to predict other high recurrence cancer data.

Gustavo *et al.* have chosen 32 patients to carry out research in ways to reduce amount of monitoring and to improve clinical outcomes of kidney transplant patients by using neural networks. In this study, predicting blood level and the cyclosporine A (CyA) dose drug is important in determining the effect it might caused if taken without right prescription. At the initial test, a small dosage of (CyA) was given to patients within the aim range. Meantime, blood sample test was also taken. In this case, two models were used to estimate the blood concentration. Three models were chosen MLP, The Finite Impulse Response – FIR Neural Networks, and Recurrent Networks to predict. Since the workload is high due to its computerized work, researchers have decided to perform analyzing and prediction in MATLAB. After much analyzing, the prediction result showed no systematic bias. From the research, dynamic neural network has provided an excellent result and the FIR is known to be a great tool for resolving this particular problem.

According to (Tian, 2010), in order to predict health condition of gears, recurrent neural network is needed in this field. The researchers later propose the use of extended recurrent neural network (ERNN) which has two layers knows as Elman and Jordan context layer. With the introduction of Elman context layer in ERNN, it will boost its capability in modeling nonlinear and also in dynamic systems. The proposed ERNN is only suitable for time series prediction and researchers are hoping ERNN will perform better than all other neural network. All training and prediction was done 10 times in order to gain average result performed with the help of ERNN model. After much comparing and investigating, FCRNN model produced a good prediction. Thus researchers concluded that these two models are able to produce acceptable good gearbox health condition in predicting.

According to Siamak, MLP Neural Network is known for its uncomplicated and popularity. Thus to reduce sizes of network, it should be applied in MLP. Therefore, this technique is used to run test on Human Chromosome Classification. In this study, MLP has three layers and one of them is hidden layer and to improve the classification precision, at least one hidden layer should be added to network. Researcher has used Sommon algorithm to reduce the input and output of the network and will improve efficiency of MLP Neural Network. To train the network, 304 samples of Copenhagen data set were

used. The result shows an accuracy of more than 80%, which is an excellent result. Thus the researcher believed that with his study, this idea of using the technique can be applied in any Neural Network Classification.

(Seker, 2001) aimed to develop intelligent methods which are able to diagnose breast cancer accurately. For this study, the researchers are using Logistic Regression (LR) method, Multilayer BackPropagation Neural Network (MLBPNN) method, fuzzy and nearest neighbor algorithm (FK-NN) method, a fuzzy measurement and the leave-one-out error method to predict the patients with breast cancer. As for the data, researchers have collected data from 100 women who were diagnosed with this breast cancer. Results varied according to the different methods used to accurately predict the breast cancer. Though varies, FK-NN method showed a very promising outcome. Researchers thus agreed that FK-NN and Fuzzy measurement are the finest methods in determining and predicting vital prognosis factor.

(Zhang, 2003) used expression profile to predict medical outcome of breast cancer. This expression profile is good in detecting differences of tumor and normal breast tissues in patients who suffer from breast cancer. Hierarchical clustering is an additional method used to make lists of genes and prediction. The researchers have designed Score for Expression Profile (SEP) in order to computer expression profile from a collection of genes. Only 48 out of 85 samples managed to be classified into groups. Analysis has shown that gene will decrease and increase based on the chance of recurrence, which is either positive or negative. The outcome and calculation of the partial correlation coefficient of gene expression was done through a modified partial correlation method. Researchers are still continuing their research in building a model called “multivariate logistic regression”.

(Hafner, 2007) compared classification techniques such as K-NN, SVM and NN classifiers to classify endoscope imagery, where it used co-occurrence histograms as the features. In their research, the researchers found that co-occurrence histograms may improve the classification accuracy of simple 1D color histograms and this only the K-NN classification technique can perform such result. However, for the classification of SVM and NN classifiers have turned out to be non-competitive and both of the techniques do not improve the classification result of 1D color histograms.

In an experiment with a large medical dataset (22,000 instances, 32 attributes, and 60 classes), analysts' found that people can understand large systems of rules with exceptions more readily than equivalent systems of regular rules because that is the way they think about the complex medical diagnoses that are involved. Richards and Compton describe their role as an alternative to classic knowledge engineering (Richards, 1998). People analyzing medical databases have noticed that cases may, in some circumstances, be diagnosable simply from the tests that a doctor decides to make regardless of the outcome of the tests. Then a record of which values are "missing" is all that is needed for a complete diagnosis – the actual values can be ignored completely (Witten, 2011). From this result, we can see that humans cannot ad hoc analyze complex data without errors.

2.2. Recent Studies within the Domain

The study based on the work done by (Keogh, 1999) that used the Naïve Bayes Classifier on two methods for finding the set of augmenting arcs, a greedy hill-climbing search, and a novel, more computationally efficient algorithm that the call SuperParent and compare these methods to TAN; a state of art distribution-based approach to finding the augmenting arcs. The dataset used in their study obtained from the UCI repository. The results show that HSC approach is usually more accurate than TAN approach. Empirical comparisons of supervised machine learning techniques in bioinformatics were conducted by (Tan, 2003) in finding the best algorithm suitable for their dataset. Four datasets obtained from UCI were tested. However, the algorithms used in this study were Decision Trees, One Rule, Decision Rule, Naïve Bayes, Instance Based, SVM, Neural Networks, Stacking, Bagging and Boosting. The learning methods perform better then bagging. The Decision Trees, One Rule and Decision Rules performed better than Naïve Bayes learning.

(Larsen, 2002) in their study used a data mining algorithm C4.5, which is a well-known decision-tree induction algorithm. In order to evaluate the effectiveness of the new attributes constructed by Genetic Algorithm (GA), they compare the performance of C4.5 with and without the new attributes constructed by the GA across several data sets in terms of classification error rate. The approach was tested in seven public domain datasets obtained from the UCI Repository. Furthermore, (Forman, 2004) compared the

classification using less training to advance the state of meta-knowledge about selecting which learning models to apply in which situations using Support Vector Machines, Logistic Regression, Naïve Bayes and Multinomial Naïve Bayes. The data were obtained from UCI and the feature selection should not be decoupled from the model selection task, as different combinations are best for regions of learning surface.

According to (Gonçalves, 2004) have been deployed techniques to automatically classify these juridical documents is proposed. Support Vector Machines are used as learning algorithm and the obtained results are presented and compared with other approaches, such as C4.5 and Naïve Bayes. The results showed its quite clear that Naïve Bayes classifier performs quite worse than the other two classifiers. However, it's important to point out that the temporal complexity of C4.5 is much higher than SVM algorithms and the worst SVM classification models remain bad classification models in C4.5.

(Auer, 2005) used eight datasets with binary classification tasks and few missing values in the data from the UCI machine learning repository Wisconsin breast-cancer, King-Rook vs. King-Pawn Chess Endgames, German Numerical Credit Data, Pima Indian Diabetes, Cleveland heart disease, Ionosphere, Thyroid disease records and Sonar. The experiment compared the results between p-delta rules with the implementation in WEKA of Multilayer Perceptrons with back propagation, the decision tree algorithm C4.5, Support Vector Machines and MADALINE. The results show that the performance of the p-delta rule is comparable with that of other classification algorithms. A comparative study of Decision Tree and Support Vector Machine to classify gene sequence was conducted by (Yuan, 2003). The results showed that bagged and boosted decision trees, SVMs and GS-SVM generated better classifiers than traditional decision tree. (Pal, 2003) used three classification algorithms, which are maximum-likelihood, Multilayer Back Propagation, Neural Network, Support Vector Machine classifier using multi and hyper-spectral data sets. The results showed that SVM classifier outperformed Multilayer and Neural Network classifiers in terms of classification accuracy with both data sets. The level of classification accuracy achieved with the Support Vector classifier was better than both Multilayer and Neural Network classifiers when they are used with small number of training data. Many statistical classifiers are based on some approximation to the ideal Bayesian Classifier. According to (German, 1999) have been used the specific classifiers are the Minimum

Distance-to-Mean (MDM), Maximum Likelihood Classification (MLC) classifiers, Linear Discrimination Analysis (LDA), Neural Network Based Multilayer Perceptron (MLP) and the Decision Tree used is popular C4.5 classifier. The aim of this study is to compare these classifiers based on terms of their learning ability, generalization ability and speed. The results of this study showed that the generalization ability of MLP is a consistently superior classifier. By contrast, the popular MLC classifier is not significantly faster. The MDM classifier has a significant benefit in terms of speed, the LDA classifier is a better choice; it outperforms the MDM and even the MLC on the more complex dataset. The Decision Tree classifier is the best all-round classifiers it is as fast as LDA.

A comparison of stacking with meta Decision Trees to bagging, boosting and stacking with other methods, work have been done by (Zenko, 2001). This study has been compared classifier ensembles combined with MDTs to bagged and boosted Decision Trees and to classifier ensembles combined with other methods are voting and stacking with three different meta-level classifiers are ordinary Decision Trees, Naïve Bayes and Multi-response Linear Regression (MLR). The results showed that the performance of stacking with MDTs the relative improvement in accuracy by stacking with MDTs as compared to bagging, boosting, voting and stacking with J4.8, Naïve Bayes and MLR. Stacking with Naïve Bayes performs poorly. Stacking with MLR slightly outperforms stacking with MDTs.

(Alty, 2003) used to predict Cardiovascular Disease (CV); to predicting patients' CVD, the researcher used a simple measurement of a patients' volume pulse measured at the finger-tip (Digital Volume Pulse) where it used an infra-red light absorption detector placed on the index finger. Suitable features are extracted from the waveform and according to researcher SVM classifier has been found to make accurate prediction of high or low arterial stiffness as indicated by the Aortal Pulse Wave Velocity (PWV).

(Ali, 2010) have chosen SVM algorithms to make sure of its possibility to predict breast cancer survival chances and also due to its great tools in analyzing where complex nonlinear interactions is found. In this study, researchers have analyzed both of SVM classifiers with three models known as Radial Basis Function (RBF), Polynomial and Sigmoid SEER breast cancer data set. While performing, researchers have found out that out of three different models, decision trees proved to be one of the best classifier. In this case, some 162,500 records of breast cancer dataset were selected and used to run test. The

overall result showed a best performance with an excellent accuracy through RBF kernel. Researchers have provided some outlines for future work which will optimize all the research that they made.

(Garcia-Orellana, 2007) used SVM and MLP to compare the mammographic CAD. The researchers used to detect and diagnosis of micro calcification clusters in mammograms. The result shows that SVM and MLP have the same performance in classifying mammographic CAD task.

(Glotsos, 2003) used SVM and DT as their tools to classify brain tumor astrocytomas (ASTs). There are two levels hierarchical of DT, where the first level was concerned with the detection of low versus high-grade tumors and for the second level the detection of less aggressive as opposed to highly aggressive tumors. SVM classifier training was based on the leave one out method. SVM outputs were compared with Bayesian Classifier and Probabilistic Neural Network and the result shows that SVM come out with a good performance event with the limited training samples were verified in this research.

(Sewak, 2007) used SVM approach to classify breast cancer, which the dataset was provided by the University of Wisconsin Hospital (WDBC). The classification SVM with Linear, Polynomial and RBF kernel functions were trained using a fraction of WDBC dataset as a training set. SVM has successfully classified the testing data and give 100 percent for tumor prediction accuracy.

(Zhang, 2009) compared study of ensemble learning approaches in the classification of breast cancer metastasis. The researchers combined the gene expression profiles and protein interaction networks by selecting a small number of sub networks as disease markers and used for the classification of metastasis. Researchers compared three ensemble learning approaches such as AdaBoost, LogitBoost and Random Forest with used classifiers Logistic Regression and Support Vector Machine to classify Breast Cancer Metastasis. As the result, the ensemble learning methods can perform a good result over the other two methods and it shows these ensemble learning approaches with sub networks makers can be more suitable in handling the classification problem such as Breast Cancer Metastasis.

Sandhya *et al.* have based on the most influencing risk factor through the usage of various methods in order to categorize Alzheimer's and Parkinson's disease. A data set of

487 patients' records was obtained. In this study, researcher has only focused on the major risk issue, which is related to both of the diseases. The Objectives of the study were: to classify the diseases using ML and NN, to look up for the most influencing risk features, comparison of each technique and to discover the effect of modification. Chi square Attribute Evaluation has been used to evaluate different risk factors of each patient. The focus of this study was to identify the risk factors has proved to be affective and has achieved a high accuracy. Researchers have used WEKA Software tool for their study.

Arfan *et al.* has worked together in explaining pre-processing of mammograms, which include automated-cropping of mammograms, extracting and removing unnecessary spots. Methods have been presented according to image segmentation method and SVM is selected to run classification. In noise removal, Fuzzy Filter is used to remove the grainy look of breast image. Furthermore, unnecessary background was removed and cropping was implemented. Image enhancement was also performed in this study. Hough transformation has been used to separate Pectoral muscle from Breast. With the help of SVM and MLP classifiers, eight attributes were extracted and it's possible to compare the results. The result shows a high accuracy based on the proposed method.

(Dancea, 2008) used SVM classifier to diagnose patients' with prostate cancer into risk classes. Different AI approach has been applied in Medical Analysis. It includes diagnosing and classification, treatment decisions and conforming risk factors. The researchers therefore applied SVM and genetic algorithms in classifying different types of cancer. Data mining method is used to categorizing patients with prostate cancer into homogenous groups. This will improve clinical decision making. The researchers have once again proved the capabilities of SVM in performing but more balanced dataset is required in order for doctors to make correct judgment of the disease before furthering any treatment.

3. METHODOLOGY

The methodology process may be complex and can be divided into the following steps for a Machine Learning study:

- Data Selection
- Platform Selection
- Data Analysis and Preprocessing
- Data Reduction and Transformation
- Feature Selection and Extraction
- Machine Learning Process
- Evaluation and Knowledge Utilization

3.1. Data Selection

Selection involves creating the target data set, i.e. the data set to about undergo analysis. Modern datasets may be both large and complex. Large datasets which are not particularly complex may generally be subjected in their entirety to the analysis process. Indeed, the larger the amount of available data, the greater the likelihood that an identifiable trend or pattern may be identified and empirically validated.

However, if the dataset is relatively complex, it is often considered impractical to attempt to subject the complete dataset for analysis. It is a common misconception to assume that the complete dataset should be submitted to the data mining software, which in turn will automatically resolve any problems and make sense of any inconsistencies. Subjecting such data to automated analysis may result in the identification of meaningless patterns or trends, which in turn wastes time and effort. Careful thought should therefore be given as to the purpose of the analysis exercise, and a target dataset created which contains data that reflects this purpose.

The data needed for the data mining process may be obtained from many different data sources. This first step obtains the data from various databases and files. This step allows the selection of the datasets needed for the execution of a defined data-mining task. In this phase, data must be evaluated, what is the minimal sub-set of data to be selected, the size of the sample needed.

3.2. Platform Selection: WEKA

Waikato Environment for Knowledge Analysis, called shortly WEKA (Hall, 2009), is a set of state-of-the-art data mining algorithms and tools to in-depth analyses. The author of this environment is University of Waikato in New Zealand. The programming language of WEKA is Java and its distribution is based on GNU General Public License. These complex algorithms may be applied to data set in the aim of detailed analyses and evaluation of data mining examination. There are three main ways of WEKA use. First is analyzing data mining methods' outputs to learn more about the data; next is generation of model for prediction of new instances and finally the last but most important for this master's thesis feature, comparison of data mining methods in order to chose the best one as a predictor e.g. in Medical Decision Support System.

WEKA consists of four user interfaces out of which three are graphical and one command-line. The main interface is called Explorer. It is graphical interface built of menu section and six panels connected to various data mining methods. It enables data preprocessing, classification, clusterization, and mining associations among attributes. Furthermore there is a possibility to select attributes with the attribute evaluator and search method. The last option is visualization plotting the dependencies among attributes.

The next graphical interface, Knowledge Flow is dedicated to selecting components from the tool bar and placing them on the special canvas, connecting them into directed graph than processing and analyzing. Furthermore the data stream data processing can be designed and executed with the usage of this interface.

To compare performance of data mining algorithms it is useful to chose third graphical interface called Experimenter. This module allows one to evaluate how well various data mining methods perform for given datasets. This process is automated and statistics can be saved. This module is a most important part of the experiment. It makes in-depth statistics which are useful in case of medical datasets. After the selection of various methods, their parameters and datasets, it is possible to prepare statistic which are priceless in case of medical diagnosis support.

Experimenter and Explorer are two mainly used interfaces during Master's Thesis Experiments. The comprehensive and deep analysis is possible with the use of WEKA environment. The state-of-art techniques implementations make analyses accurate and precise.

3.3. Data Analysis and Preprocessing

The data to be used is checked in terms of correctness, missing data, noise, duplication, incompleteness data and inconsistency. These problems need to be handled by choosing the appropriate way. From this study, there are some medical datasets needed to handle from missing value and noise, these datasets must be preprocessed to transform into a form that is presentable to the classification techniques.

The detected missing values, noise data and irrelevant data are removed from the collection. Some of the chosen datasets have missing values. Missing data values cause problem during both the training phase and the classification process. Thus missing values must be handled appropriately, otherwise results may be inaccurate.

3.4. Data Reduction and Transformation

Data from different sources will be converted into a common format or appropriate forms for processing purpose. In this study, data may be encoded or transformed into more usable formats.

3.4.1. Data Transformation

Some of attributes in the datasets come in nominal or string format. These shall be converted into numerical values for analysis purposes.

3.4.2. Data Discretization

Data discretization is a necessary procedure when using classification algorithms. Its purpose is to reduce the number of values for a given continuous attribute. Discretization is usually performed by dividing the range of the attribute into intervals. Interval labels are used to replace actual data values.

Equal frequency binning has been used for discretization through WEKA software.

3.5. Feature Selection and Extraction

Before implementing the classification algorithms it is recommended that the incomplete, noisy or inconsistent datasets are pre-processed to make the knowledge discovery process easier and more qualified. The most well known steps are summarization, cleaning, integrations and transformations, data and dimensionality reduction and discretization (Han, 2005). Discretization and dimension reduction are within the scope of this study.

Data discretization techniques can be used to reduce the number of values for a given continuous variable by splitting the range of the variable into intervals. Binning, for example, is a type of discretization technique where variable is splitted into a particular number of bins.

Dimension reduction is another pre-processing technique to obtain a reduced dataset representing the original dataset. The most commonly used dimension reduction technique

is Principal Component Analysis. “PCA searches for k n-dimensional orthogonal vectors that can best be used to represent the data where k <= n. The original data are thus projected onto a smaller space.” (Han, 2005).

3.6. Machine Learning Process

This step analyzes the discretized dataset based on an appropriate data mining task. Examples of data mining tasks are statistical models, classification, predictive modeling, clustering, finding association rules and sequence analysis. Classification technique has been applied in this study.

A small set of rules from training dataset is obtained by learning the objects that has attributes and predetermined target specified in classification. Different classification techniques can be used to extract relevant relationship in the data. Among them are Decision Trees, Statistical and Bayesian approaches, Neural Networks and Support Vector Machines.

The classification techniques which have been examined in this thesis are J48 – Decision Tree, Support Vector Machine, Radial Basis Function, Multilayer Perceptron, k-Nearest Neighbors, Naïve Bayes, Bayes Net and Logistic Regression.

Classification is one of the most important tasks in data mining. This problem involves the need to find the rules that can partition the data into disjoint groups. Often classification involves supervised data mining tools in which the user is heavily involved in the definition of the different groups and the specification of the rules that can be used to determine to which group a data item belongs. Examples of such tools include decision trees and rule-based techniques. Example-based data mining methods such as nearest neighbor classification, regression algorithms and case-based reasoning are also examples of solutions to data classification problems.

The aim of a classification problem is to build a classifier based on some cases with some attributes to describe the objects and one attribute to describe the group of the objects. Then, the classifier is used to predict the group attribute of new cases from the

domain based on the values of other attributes or to give us a better understanding of the available data.

There are eight classification techniques, which have been used in this thesis, to make comparison between them and find the most suitable of them to solve medical problems.

3.6.1. Logistic Regression

Suppose first that there are only two classes. Logistic regression replaces the original target variable.

This cannot be approximated accurately using a linear function.

The resulting values are no longer constrained to the interval from 0 to 1 but can lie anywhere between negative infinity and positive infinity. (a) plots the transformation function, which is often called the logit transformation.

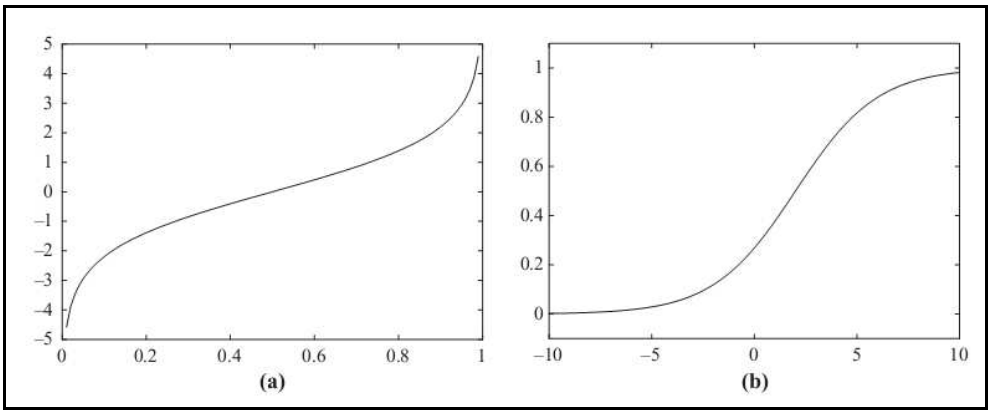


Figure 3.1. Logistic Regression.

The transformed variable is approximated using a linear function just like the ones generated by linear regression. The resulting model is with weights *w*. (b) shows an example of this function in one dimension, with two weights $w_0 = -1.25$ and $w_1 = 0.5$.

Just as in linear regression, weights must be found that fit the training data well. Linear regression measures goodness of fit using the squared error. In logistic regression the log-likelihood of the model is used instead. This is given by where the $x^{(i)}$ are either 0 or 1.

The weights w_i need to be chosen to maximize the log-likelihood. There are several methods for solving this maximization problem. A simple one is to iteratively solve a sequence of weighted least-squares regression problems until the log-likelihood converges to a maximum, which usually happens in a few iterations.

To generalize logistic regression to several classes, one possibility is to proceed in the way described above for multiresponse linear regression by performing logistic regression independently for each class. Unfortunately, the resulting probability estimates will not sum to 1. To obtain proper probabilities it is necessary to couple the individual models for each class. This yields a joint optimization problem, and there are efficient solution methods for this.

3.6.2. Support Vector Machine

A hyperplane separating the two classes might be written as in the two-attribute case, where a_1 and a_2 are the attribute values and there are three weights w_i to be learned

However, the equation defining the maximum-margin hyperplane can be written in another form, in terms of the support vectors. Write the class value y of a training instance as either 1 (for yes, it is in this class) or –1 (for no, it is not). Then the maximum-margin hyperplane can be written as below.

Above, y_i is the class value of training instance a(i), while b and α_i are numeric parameters that have to be determined by the learning algorithm. Note that a(i) and a are vectors. The vector a represents a test instance—just as the vector $[a_1, a_2]$ represented a test instance in the earlier formulation.

The vectors a(i) are the support vectors, those circled in Figure 3.2. are selected members of the training set. The term a(i) • a represents the dot product of the test instance with one of the support vectors.

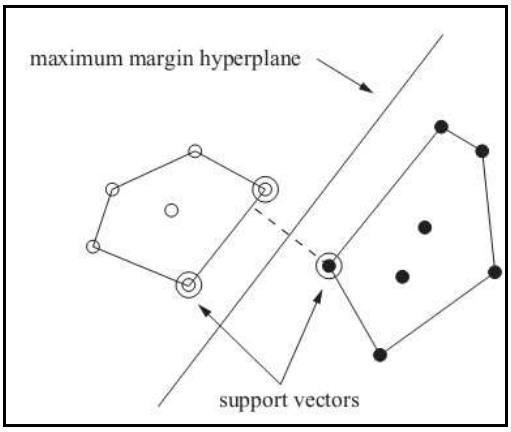


Figure 3.2. A maximum-margin hyperplane.

If you are not familiar with dot product notation, you should still be able to understand the gist of what follows: Just think of a(i) as the whole set of attribute values for the *i*th support vector. Finally, b and α_i are parameters that determine the hyperplane, just as the weights w_0, w_1, and w_2 are parameters that determine the hyperplane in the earlier formulation.

$$a(i) \bullet a = \Sigma_j a(i)_j a_j. \qquad (3.1)$$

It turns out that finding the support vectors for the training instances and determining the parameters b and α_i belongs to a standard class of optimization problems known as constrained quadratic optimization. There are off-the-shelf software packages for solving these problems (see Fletcher, 1987, for a comprehensive and practical account of solution methods). However, the computational complexity can be reduced, and learning accelerated, if special-purpose algorithms for training support vector machines are applied.

3.6.3. Radial Basis Function

RBF has two layers, not counting the input layer, and differs from a multilayer perceptron in the way that the hidden units perform computations. Each hidden unit essentially represents a particular point in input space, and its output, or activation, for a given instance depends on the distance between its point and the instance, which is just another point. Intuitively, the closer these two points, which are the stronger in the activation. This is achieved by using a nonlinear transformation function to convert the distance into a similarity measure. A bell-shaped Gaussian activation function, of which the width may be different for each hidden unit, is commonly used for this purpose. The hidden units are called RBFs because the points in instance space for which a given hidden unit produces the same activation form a hypersphere or hyperellipsoid. (In a multilayer perceptron, this is a hyperplane.)

The output layer of an RBF network is the same as that of a multilayer perceptron: It takes a linear combination of the outputs of the hidden units and—in classification problems—pipes it through the sigmoid function (or something with a similar shape).

The parameters that such a network learns are (a) the centers and widths of the RBFs and (b) the weights used to form the linear combination of the outputs obtained from the hidden layer. A significant advantage over multilayer perceptrons is that the first set of parameters can be determined independently of the second set and still produce accurate classifiers.

A disadvantage of RBF networks is that they give every attribute the same weight because all are treated equally in the distance computation, unless attribute weight parameters are included in the overall optimization process. Thus, they cannot deal effectively with irrelevant attributes, in contrast to multilayer perceptrons. Support vector machines share the same problem. In fact, support vector machines with Gaussian kernels (i.e., “RBF kernels”) are a particular type of RBF network, in which one basis function is centered on every training instance, all basis functions have the same width, and the outputs are combined linearly by computing the maximum-margin hyperplane. This has the effect that only some of the RBFs have a nonzero weight—the ones that represent the support vectors.

3.6.4. Multi Layer Perceptron

Multilayer Perceptron (MLP) is a nonparametric neural network structure and used for both classification and regression. Feedforward MLPs are the most widely used Artificial Neural Network (ANN) models. MLP is composed of three layers: an input layer, hidden layers and an output layer. A two hidden layer MLP is shown in Figure 3. In MLP, using one hidden layer is generally preferred in the case of reducing the complexity. Furthermore, large number of hidden units may cause overfitting, thus hidden layer may contain either predefined number of hidden units or optimal number of hidden units can be determined during learning.

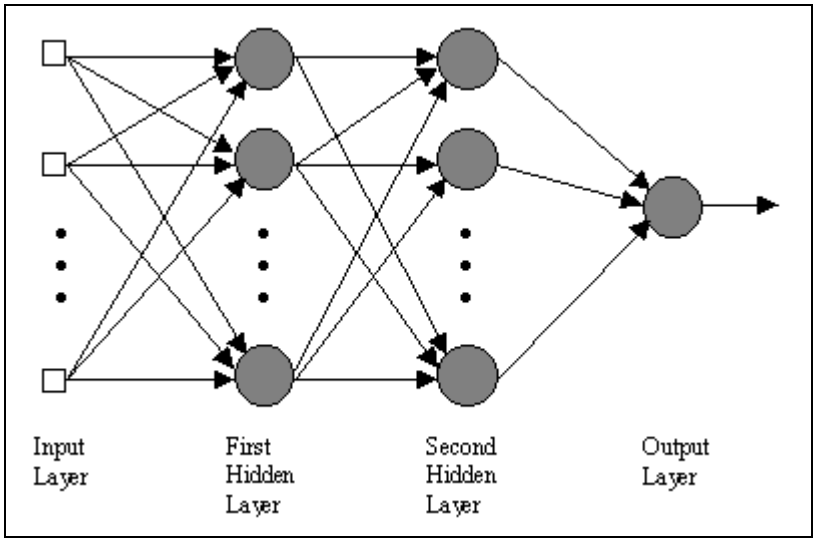


Figure 3.3. Block diagram of a two hidden layer Multiplayer Perceptron (MLP).

MLP learning process starts at the input layer where no calculation is applied. Briefly, hidden units nonlinearly transform the d dimensional input space to h dimensional space. The output units produce the output values as linear combinations of the h dimensional activation values computed by hidden units.

As activation function, sigmoid is used which is given in below equation. Sigmoid produces the output in the [0, 1] range.

$$f(u) = 1/(1 + e^{-u}) \tag{3.2}$$

3.6.5. K-Nearest Neighbors

k-Nearest Neighbor (k-NN) algorithm is a method for classifying objects based on the closest training examples in the feature space. The measure of closeness is in terms of d dimensional input space. There are different measurements such as Euclidean Distance or Mahalanobis Distance. Euclidean distance is a linear distance between two points which is given below.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.3}$$

Mahalanobis distance calculates the distance between two data points by the variation in each component of the points which is given below.

$$d(x, y) = \sqrt{(x - y)\Sigma^{-1}(x - y)} \tag{3.4}$$

After distances between training data and new instance are calculated, k nearest neighbors are determined. Then, the class probabilities are calculated as a proportion of the number of training instances which belong to class *i* to the total number of training instances.

3.6.6. Naïve Bayes

Naive Bayes is a simple probabilistic classifier based on Bayes' theorem, where features are assumed to be independent given the class. The assumption of independence makes it much easier to estimate these probabilities since each attribute can be treated separately. For example, an animal may be considered to be a dog if it is barking and has four legs. Even if these features depend on each other or upon the existence of the other features, a Naive Bayes classifier considers all of these properties to independently contribute to be the probability that this animal is a dog.

Naive Bayes algorithm works as follows: for each decision class it computes the conditional probability that decision class is the correct one, given an object's information vector. The algorithm assumes that the object's attributes are independent. The probabilities involved in producing the final estimate are computed as frequency counts from a "master" decision table.

Given the above description of NB, we can say that the probability of getting the string of feature values like this.

$$P(X_j^1 = a_1, X_j^2 = a_2, ..., X_j^n = a_n \mid \mathrm{C_i}) \tag{3.5}$$

This is just equal to the product of multiplying together all of the individual probabilities which is much easier to compute as well as reducing the curse of dimensionality.

$$P(X_j^1 = a_1 \mid \mathrm{C_i}) \times P(X_j^2 = a_2 \mid \mathrm{C_i}), ..., P(X_j^n = a_n \mid \mathrm{C_i}) = \prod_k P(X_j^k = a_k \mid \mathrm{C_i}) \tag{3.6}$$

NB classifier selects the class C_i for which the following computation is maximum:

$$P(C_i \mid \mathrm{x}) \alpha P(C_i) \prod_k P(X_j^k = a_k \mid \mathrm{C_i}) \tag{3.7}$$

Despite its simplicity, NB is successful in many applications. Its advantage is that it requires a small amount of training data to estimate the parameters necessary for classification.

3.6.7. Bayesian Network

A Bayesian network is a directed acyclic graphical model that encodes probabilistic relationships among variables of interest (Heckerman, 1996).

Bayesian networks allow efficient representation of the joint probability distribution over a set of random variables. The network structure is used to characterize a probability distribution for each node depending on its parents. And posterior probabilities are computed in the form of local conditional distributions.

A Bayesian network is represented by this formula.

$$B = \langle G, \Theta \rangle \tag{3.8}$$

G is a directed acyclic graph. The nodes of the graph correspond to the random variables X_1, X_2, ..., X_n which are the dataset features and edges represent direct dependencies between the associated variables. The graph G encodes the independence assumption where each variable X_i is independent of its nondescendants given its parents $\prod_i$ in G. The second component Θ represents the conditional probability distribution that quantifies the dependency between the nodes.

A Bayesian network defines a unique joint probability distribution over the set of random variables X_i in the network given by this formula.

$$P(X_1,....,X_n) = \prod_{i=1}^{n} P(X_i \mid \prod_{x_i}) \tag{3.9}$$

$\prod_i$ denotes the set of parents of X_i in the network.

In practice, the components of the Bayesian networks are unknown and must be inferred from the data. Learning a Bayesian network from data involves two subtasks, structure learning, which is necessary to identify the topology of the network, and parameter learning, that identifies the statistical parameters for a given network topology.

Most studies concentrate on structure learning which a complex procedure is when there are lots of input features. Learning the parameters in conditional probability tables is recognized as a trivial task based on frequency counts of data points when the observed frequencies are optimal in a sufficiently large database. Here, we review the main approaches for construction of the network structure and estimation of parameters when learning Bayesian networks from data.

Structure learning is a search for encoding appropriate dependencies between the features of a given dataset. It has been argued that Bayesian network structure learners are computationally expensive requiring an exponential number of conditional independence tests (Cheng, 2002). There are two main approaches to learn the network structure from data efficiently reducing the search space: constraint based methods and methods that maximize a selected score.

Simple learning algorithm (SLA) (Cheng, 2002) and three-phase dependency analysis (TPDA) (Cheng, 2002) are examples of constraint based methods that make use of information theory concept in order to reduce the computational complexity of the structure learning procedure. Reiz and Csato also propose a mutual information based approach where direct causal relations encoded by the BN are interpreted as the maximum of conditional mutual information between nodes (Reiz, 2008).

The algorithms based on a scoring function attempt to find a graph that maximizes the selected score, which evaluates how well a given network matches the data. Different learning algorithms can be obtained depending on the definitions of the scoring function

and on the search procedure used. Meloni *et al.* propose a variation of standard search-and-score approach that computes a square matrix containing the mutual information among all pairs of variables (Cheng, 2002). The matrix is binarized to find what relationships must be prevented. This approach prevents the inference of too many connections.

Furthermore, there are well known simple Bayesian network classifiers with highly constrained dependency structures: Naive Bayesian network assuming mutual independence of the feature variables given the class variable and Tree Augmented Network (TAN) representing a tree-like dependency structure over the feature variables (Lucas, 2002).

Parameter learning in Bayesian networks is often based on Frequency Estimates (FE) which determines the conditional probabilities by computing the frequencies of instances from the data. The FE method is efficient since it counts each data point in the training set only once. The parameters estimated using FE method maximizes the likelihood of the model given the data and thus FE is known as a generative learning method (Su, 2008).

The relative frequencies in the CPT are obtained as follows.

$$\hat{P}(X_i = x \mid \prod_{x_i} = \vec{u}) = \frac{count(X_i = x, \prod_{x_i} = \vec{u})}{count(\prod_{x_i} = \vec{u})} \tag{3.10}$$

The classification capability of FE method is argued because of the generative property. Grainer and Zhou proposed a gradient descent based discriminative parameter learning method, ELR that significantly outperforms FE method with a high computational cost (Greiner, 2002).

A Discriminative Frequency Estimate (DFE) is proposed to maximize the generalization accuracy of classification rather than likelihood (Su, 2008). The authors compared the DFE and FE methods based on Naive Bayesian network structure and showed that DFE significantly improve the performance of classification in terms of accuracy. However, it has been widely accepted that accuracy is not an appropriate

performance measure especially for imbalanced datasets. On the other hand, the training time of DFE method is significantly higher than FE method.

3.6.8. C4.5 Decision Tree Algorithm

The decision tree program C4.5 and its successor C5.0 were devised by Ross Quinlan over a 20-year period beginning in the late 1970s. A complete description of C4.5, the early 1990s version, appears as an excellent and readable book, along with the full source code. The more recent version, C5.0, is available commercially. Its decision tree induction seems to be essentially the same as that used by C4.5, and tests show some differences but negligible improvements. However, its rule generation is greatly sped up and clearly uses a different technique, although this has not been described in the open literature.

C4.5 works essentially as the default confidence value is set at 25% and works reasonably well in most cases; possibly it should be altered to a lower value, which causes more drastic pruning, if the actual error rate of pruned trees on test sets is found to be much higher than the estimated error rate. There is one other important parameter whose effect it is to eliminate tests for which almost all of the training examples have the same outcome. Such tests are often of little use. Consequently, tests are not incorporated into the decision tree unless they have at least two outcomes that have at least a minimum number of instances. The default value for this minimum is 2, but it is controllable and should perhaps be increased for tasks that have a lot of noisy data.

Another heuristic in C4.5 is that candidate splits on numeric attributes are only considered if they cut off a certain minimum number of instances: at least 10% of the average number of instances per class at the current node, or 25 instances—whichever value is smaller (but the minimum just mentioned, 2 by default, is also enforced).

C4.5 Release 8, the last noncommercial version of C4.5, includes an MDL-based adjustment to the information gain for splits on numeric attributes. More specifically, if there are S candidate splits on a certain numeric attribute at the node currently considered for splitting, $\log_2(S)/ N$ is subtracted from the information gain, where N is the number of instances at the node. This heuristic is designed to prevent overfitting. The

information gain may be negative after subtraction, and tree growing will stop if there are no attributes with positive information gain—a form of prepruning.

3.7. Evaluation and Knowledge Utilization

Results obtained from data mining step are examined based on Percent Correct, True Positive Rate, False Positive Rate, Precision, Recall, F-Measure and AUC.

The classification technique with the related benchmarking criteria is evaluated such as the most suitable classification technique.

4. DATA SETS USED IN THESIS

The study presents the particular case of analysis of eleven datasets containing data associated to several Healthcare datasets. The datasets that are analyzed in this study are eleven medical datasets that are based on UCI database (Frank, 2010), H1N1 research done by (Cowling, 2010), Obesity research done by Hacettepe University Department of Biostatistics (Biostatistics, 2012), and Real Medical Dataset of Ludmila I. Kuncheva (Kissiov, 2005).

By using 11 datasets in the medical area, the target data will consist of the following information. Here are the brief summary of the datasets:

Table 4.1. Brief Summary about used datasets in this Thesis.

Dataset Name	Number of Classes	Number of Instances	Number of Attributes
Acute Inflammations Data Set	2	120	6
Breast Cancer - Survival from Surgery Data Set	2	306	4
Breast Cancer Wisconsin (Original) Data Set	2	699	11
Dermatology Data Set	2	366	34
Echocardiogram Data Set	2	132	13
H1N1 Data Set	3	348	18
Hepatitis Data Set	2	155	20
Liver Disorders Data Set	2	347	7
Obesity Data Set	2	50	26
Pima Indians Diabetes Data Set	2	768	9
Respiratory Data Set	2	85	18

4.1. Acute Inflammations Data Set (CZERNIAK, 2002)

The main idea of this data set is to prepare the algorithm of the expert system, which will perform the presumptive diagnosis of two diseases of urinary system. It will be the example of diagnosing of the acute inflammations of urinary bladder and acute nephritises. For better understanding of the problem let us consider definitions of both diseases given by medics. Acute inflammation of urinary bladder is characterized by sudden occurrence of pains in the abdomen region and the urination in form of constant urine pushing, micturition pains and sometimes lack of urine keeping. Temperature of the body is rising, however most often not above 38C. The excreted urine is turbid and sometimes bloody. At proper treatment, symptoms decay usually within several days. However, there is inclination to returns. At persons with acute inflammation of urinary bladder, we should expect that the illness will turn into protracted form.

Acute nephritis of renal pelvis origin occurs considerably more often at women than at men. It begins with sudden fever, which reaches, and sometimes exceeds 40C. The fever is accompanied by shivers and one- or both-side lumbar pains, which are sometimes very strong. Symptoms of acute inflammation of urinary bladder appear very often. Quite not infrequently there are nausea and vomiting and spread pains of whole abdomen.

The data was created by a medical expert as a data set to test the expert system, which will perform the presumptive diagnosis of two diseases of urinary system. The basis for rules detection was Rough Sets Theory. Each instance represents a potential patient.

Contains 120 number of instances, 8 number of attributes and no missing values

Table 4.2. Data information about Acute Inflammations Data Set.

No	Physical Attribute	Features – only used ones are listed
1	Temperature	real
2	Occurrence of nausea	{ yes, no }
3	Lumbar pain	{ yes, no }
4	Urine pushing	{ yes, no }
5	Micturition pains	{ yes, no }
6	Burning of urethra, itch	{ yes, no }
7	decision: Inflammation of urinary bladder	{ yes, no }
8	decision: Nephritis of renal pelvis origin	{ yes, no }

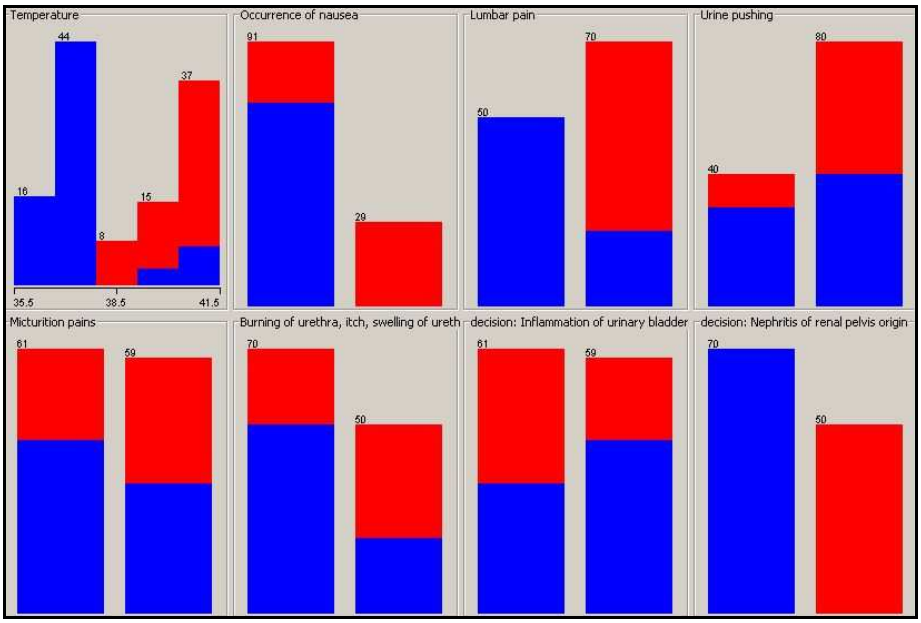


Figure 4.1. Acute Inflammations Attributes' Instances distribution.

4.2. Breast Cancer - Survival from Surgery Data Set

The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

Contains 306 number of instances, 4 number of attributes and no missing values

Table 4.3. Data information about Breast Cancer - Survival from Surgery Data Set.

No	Physical Attribute	Features – only used ones are listed
1	Age	real
2	year-of-operation	real
3	number-of-positive-axillary-nodes-detected	real
4	Survival	{ 1, 2 }

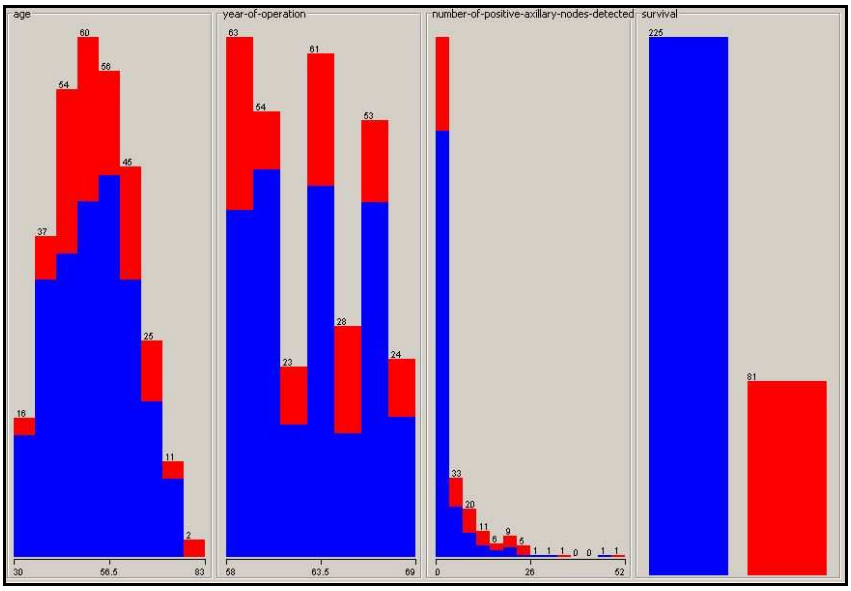


Figure 4.2. Breast Cancer for Attributes' Instances distribution.

4.3. Breast Cancer Wisconsin (Original) Data Set (WOLBERG, 1990)

This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. Samples arrive periodically as Dr. Wolberg reports his clinical cases.

The database therefore reflects this chronological grouping of the data. This grouping information appears immediately below, which have 699 points (as of the donated database on 15 July 1992).

Contains 699 number of instances, 11 number of attributes and it has 16 missing values

Table 4.4. Data information about Breast Cancer Wisconsin (Original) Data Set.

No	Physical Attribute	Features – only used ones are listed
1	Sample code number	Real
2	Clump Thickness	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
3	Uniformity of Cell Size	Real
4	Uniformity of Cell Shape	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}

Table 4.5. Data information about Breast Cancer Wisconsin (Original) Data Set cont.

5	Marginal Adhesion	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
6	Single Epithelial Cell Size	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
7	Bare Nuclei	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
8	Bland Chromatin	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
9	Normal Nucleoli	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
10	Mitoses	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
11	Class (2 for benign, 4 for malignant)	{2, 4}

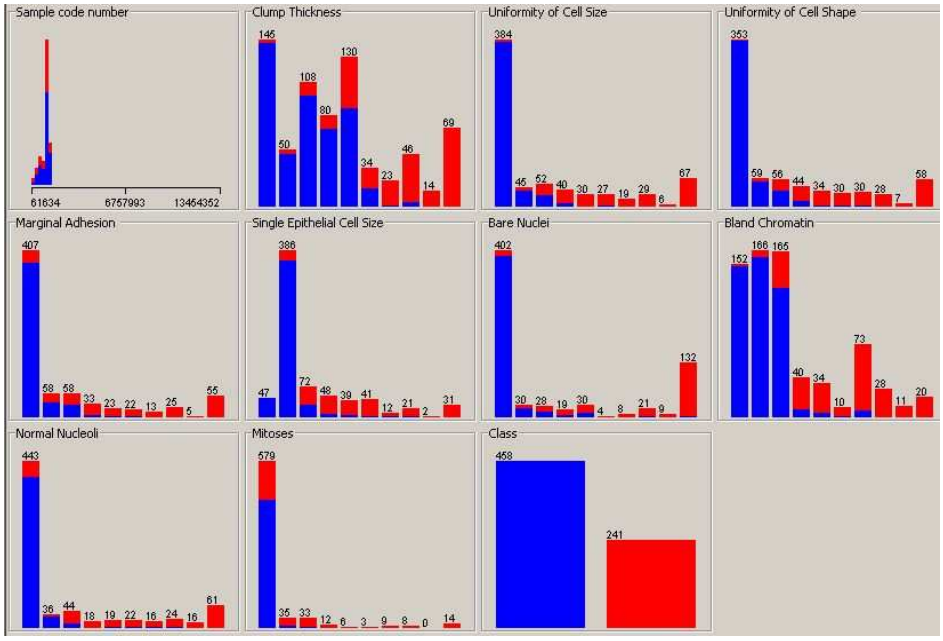


Figure 4.3. Breast Cancer Wisconsin Attributes' Instances distribution.

4.4. Dermatology Data Set (DEMIROZ, 1998)

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus,

pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope.

In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.

The class codes are assigned as follows:

1: psoriasis, 2: seboreic dermatitis, 3: lichen planus, 4: pityriasis rosea, 5: cronic dermatitis, 6: pityriasis rubra pilaris

The names and id numbers of the patients were recently removed from the database.

Contains 366 number of instances, 35 number of attributes and it has 8 missing values

Table 4.6. Data information about Dermatology Data Set.

No	Physical Attribute	Features – only used ones are listed
1	Erythema	{0, 1, 2, 3}
2	Scaling	{0, 1, 2, 3}
3	definite borders	{0, 1, 2, 3}
4	Itching	{0, 1, 2, 3}
5	koebner phenomenon	{0, 1, 2, 3}
6	polygonal papules	{0, 1, 2, 3}
7	follicular papules	{0, 1, 2, 3}
8	oral mucosal involvement	{0, 1, 2, 3}

Table 4.7. Data information about Dermatology Data Set cont.

9	knee and elbow involvement	{0, 1, 2, 3}
10	scalp involvement	{0, 1, 2, 3}
11	family history	{0, 1}
12	melanin incontinence	{0, 1, 2, 3}
13	eosinophils in the infiltrate	{0, 1, 2, 3}
14	PNL infiltrate	{0, 1, 2, 3}
15	fibrosis of the papillary dermis	{0, 1, 2, 3}
16	Exocytosis	{0, 1, 2, 3}
17	Acanthosis	{0, 1, 2, 3}
18	Hyperkeratosis	{0, 1, 2, 3}
19	Parakeratosis	{0, 1, 2, 3}
20	clubbing of the rete ridges	{0, 1, 2, 3}
21	elongation of the rete ridges	{0, 1, 2, 3}
22	thinning of the suprapapillary epidermis	{0, 1, 2, 3}
23	spongiform pustule	{0, 1, 2, 3}
24	munro microabcess	{0, 1, 2, 3}
25	focal hypergranulosis	{0, 1, 2, 3}
26	disappearance of the granular layer	{0, 1, 2, 3}
27	vacuolisation and damage of basal layer	{0, 1, 2, 3}
28	Spongiosis	{0, 1, 2, 3}
29	saw-tooth appearance of retes	{0, 1, 2, 3}
30	follicular horn plug	{0, 1, 2, 3}
31	perifollicular parakeratosis	{0, 1, 2, 3}
32	inflammatory monoluclear infiltrate	{0, 1, 2, 3}
33	band-like infiltrate	{0, 1, 2, 3}
34	Age	Real
35	Class	{1, 2, 3, 4, 5, 6}

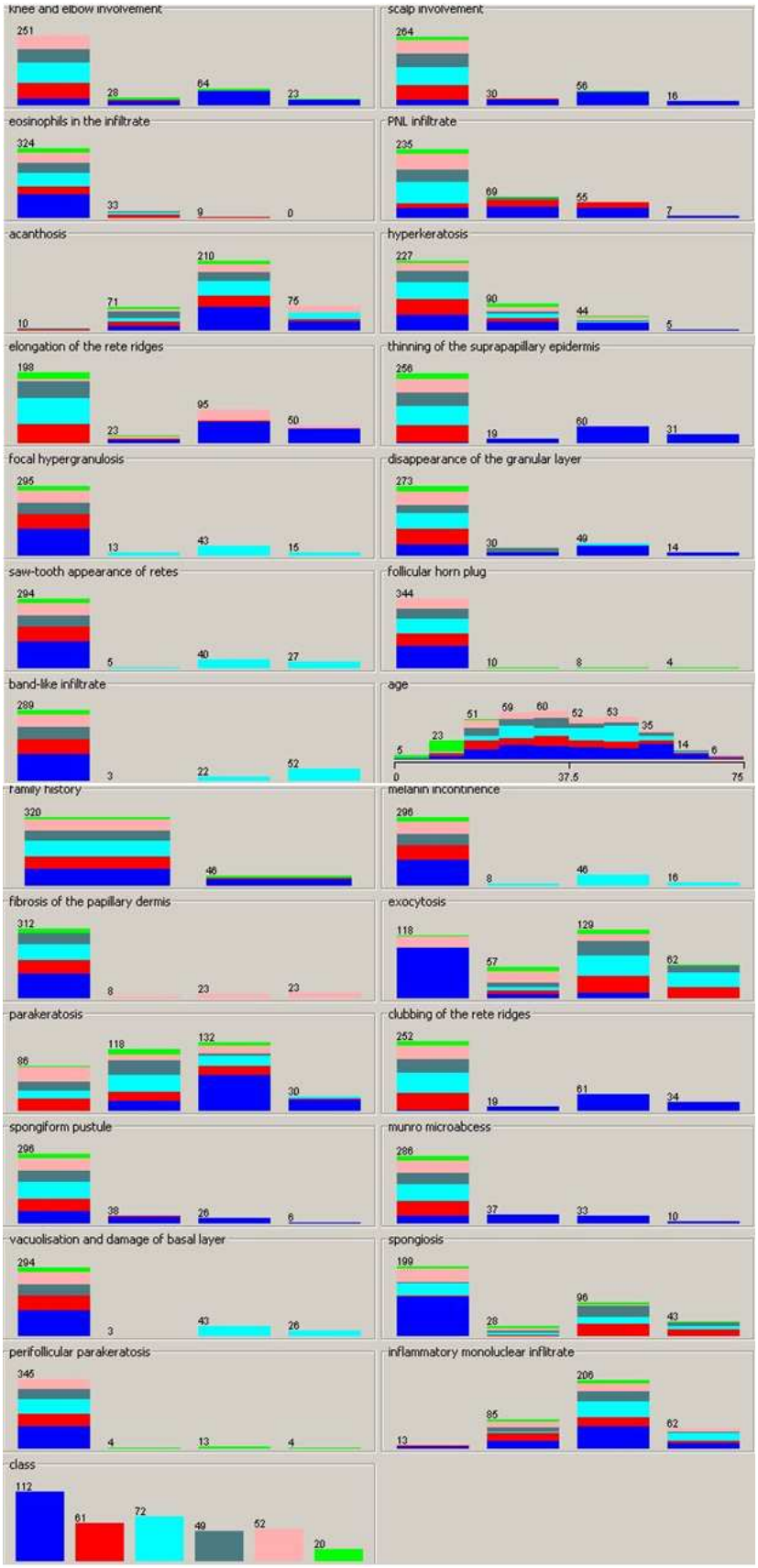


Figure 4.4. Dermatology Data Set Column Chart for Attributes' Instances distribution.

4.5. Echocardiogram Data Set

All the patients suffered heart attacks at some point in the past. Some are still alive and some are not. The survival and still-alive variables, when taken together, indicate whether a patient survived for at least one year following the heart attack. The problem addressed by past researchers was to predict from the other variables whether or not the patient will survive at least one year. The most difficult part of this problem is correctly predicting that the patient will NOT survive.

The class codes are assigned as follows:

0: patient was either dead after 1 year or had been followed for less than 1 year, 1: patient was alive at 1 year.

Contains 132 number of instances, 13 number of attributes and it has 132 missing values

Table 4.8. Data information about Echocardiogram Data Set.

No	Physical Attribute	Features – only used ones are listed
1	survival	Real
2	still-alive	{0, 1}
3	age-at-heart-attack	Real
4	pericardial-effusion	{0, 1}
5	fractional-shortening	Real
6	Epss	Real
7	Lvdd	Real
8	wall-motion-score	Real
9	knee and elbow involvement	Real
10	wall-motion-index	Real
11	Mult	Real
12	Name	Real
13	Group	Real
14	alive-at-1	{0, 1}

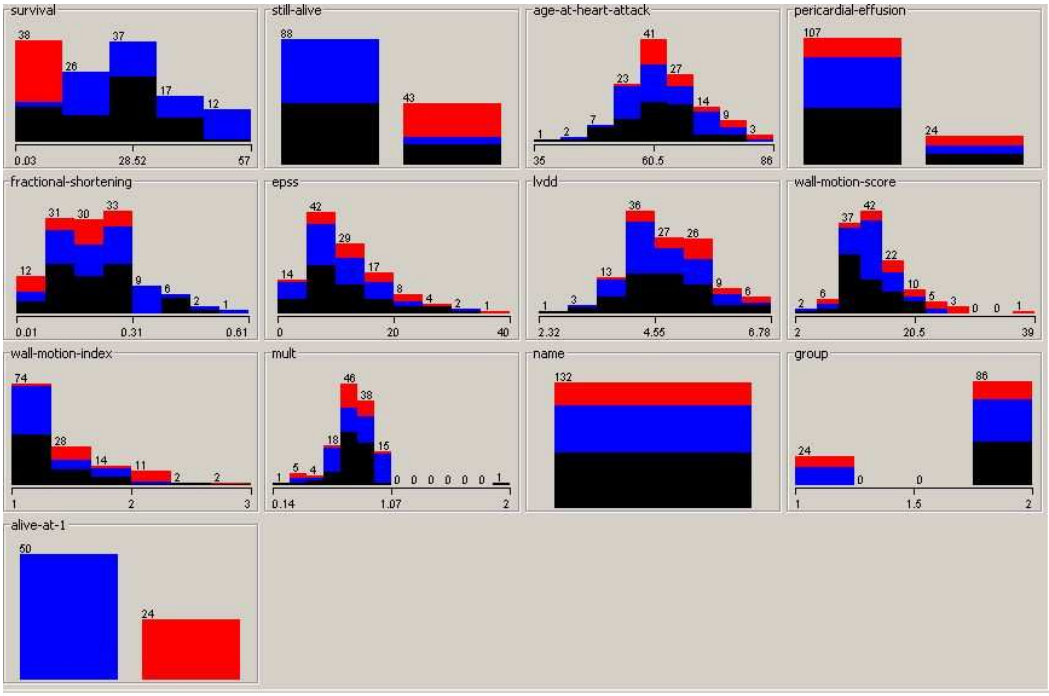


Figure 4.5. Echocardiogram Attributes' Instances distribution.

4.6. H1N1 Data Set (COWLING, 2010)

Following the emergence of pandemic 2009 influenza A (H1N1), Researchers conducted a household transmission study based on a similar protocol to our earlier trial of non-pharmaceutical interventions, to establish the basic household epidemiology of novel H1N1 and compare with seasonal influenza.

In July and August 2009, Researchers recruited subjects presenting to outpatient clinics (in both the private and public sectors across Hong Kong) with influenza-like-illness of <48 hours duration. After influenza was confirmed in an index case by the QuickVue Influenza A+B rapid test the household of the index subject was recruited to our household transmission study. Researchers aimed to conduct an initial home visit within 36 hours of recruitment, and evaluate subsequent infections by self-reported daily symptom diaries and home visits after 3 and 6 days. Nose and throat swabs were collected from index subjects and all household contacts at each home visit and tested by RT-PCR. The primary outcome measure was laboratory-confirmed influenza infection in a household contact by RT-PCR; the secondary outcome was clinically diagnosed influenza by self-

reported symptoms. Researchers collected acute and convalescent sera from a subset of participants and tested for antibody titers to seasonal and pandemic influenza.

Contains 348 number of instances, 18 number of attributes and it has 353 missing values.

Table 4.9. Data information about H1N1 Data Set.

No	Physical Attribute	Features – only used ones are listed
1	hhID	Real
2	Male	{0,1}
3	Age	Real
4	Measure	{0,1}
5	Headache	{0,1}
6	Sthroat	{0,1}
7	Cough	{0,1}
8	Aches	{0,1}
9	Rnose	{0,1}
10	Phlegm	{0,1}
11	Onsettime	{1,2,3,4,5}
12	QVres	{1,2,3}
13	Bodytemp	Real
14	Antiviral	{0,1}
15	Antibiotics	{0,1}
16	pcr.H1	{0,1}
17	pcr.H3	{0,1}
18	pcr.pH1N1	{0,1}

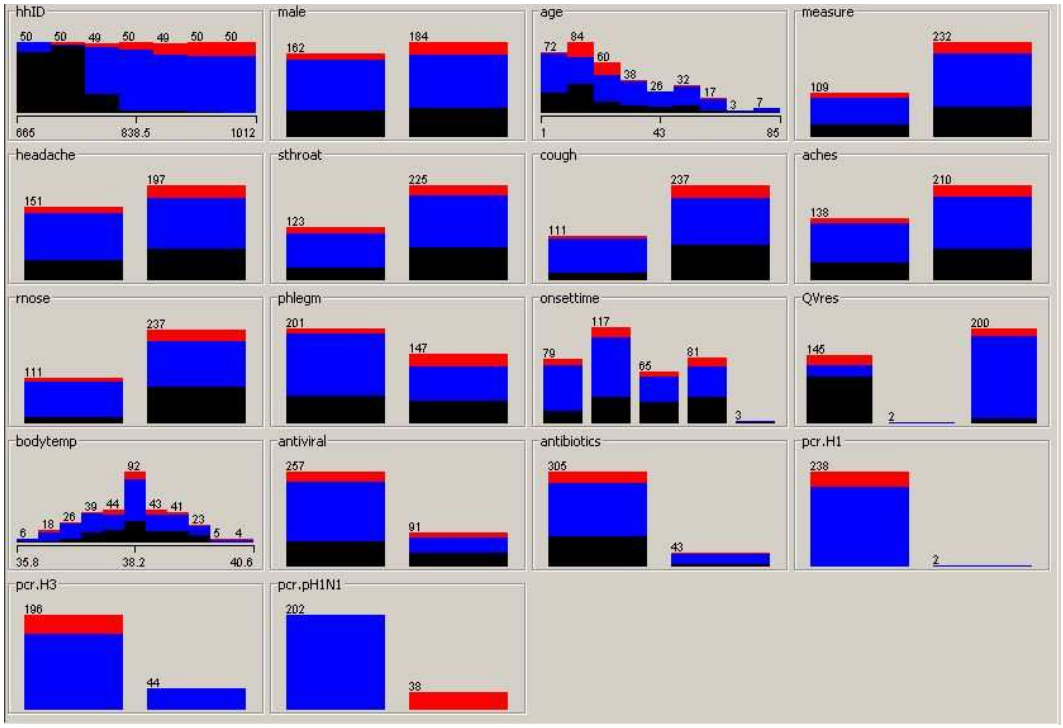


Figure 4.6. H1N1 Data Set Column Chart for Attributes' Instances distribution.

4.7. Hepatitis Data Set

Hepatitis Prognostic Database can predict either a patient is infected with Hepatitis or not. According to the patients' inputs, this data set can gives a prognosis result either the patient can stay live or die.

Contains 155 number of instances, 20 number of attributes and it has 167 missing values

The class codes are assigned as follows. 1: patient was dead, 2: patient was alive

Table 4.10. Data information about Hepatitis Data Set.

No	Physical Attribute	Features – only used ones are listed
1	DIE-LIVE	{2, 1}
2	Age	Real
3	Male-Female	{2, 1}
4	Steroid	{2, 1}
5	ANTIVIRALS	{2, 1}
6	FATIGUE	{2, 1}
7	MALAISE	{2, 1}
8	ANOREXIA	{2, 1}
9	LIVER BIG	{2, 1}
10	LIVER FIRM	{2, 1}
11	SPLEEN PALPABLE	{2, 1}
12	SPIDERS	{2, 1}
13	ASCITES	{2, 1}
14	VARICES	{2, 1}
15	BILIRUBIN	Real
16	ALK PHOSPHATE	Real
17	SGOT	Real
18	ALBUMIN	Real
19	PROTIME	Real
20	HISTOLOGY	{2, 1}

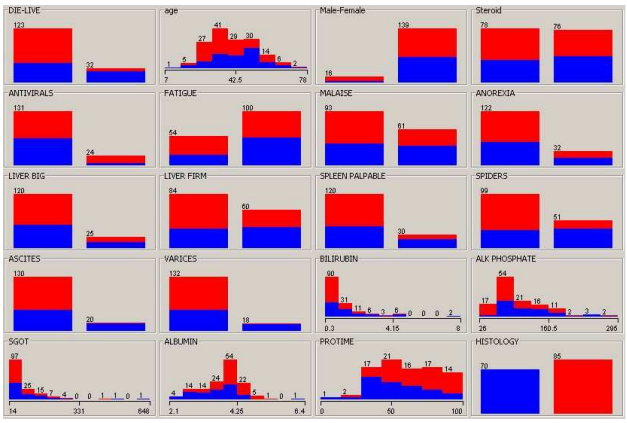


Figure 4.7. Hepatitis Data Set Column Chart for Attributes' Instances distribution.

4.8. Liver Disorders Data Set (FORSYTH, 1990)

The first 5 variables are all blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. Each line in the dataset file constitutes the record of a single male individual.

Contains 347 number of instances, 7 number of attributes and it has no missing values. The class codes are assigned as follows for selector attribute: 1: first set, 2: second set.

Table 4.11. Data information about Liver Disorders Data Set.

No	Physical Attribute	Features – only used ones are listed
1	mean corpuscular volume	real
2	alkaline phosphotase	real
3	alamine aminotransferase	real
4	aspartate aminotransferase	real
5	gamma-glutamyl transpeptidase	real
6	number of half-pint equivalents of alcoholic beverages drunk per day	real
7	Selector	{1, 2}

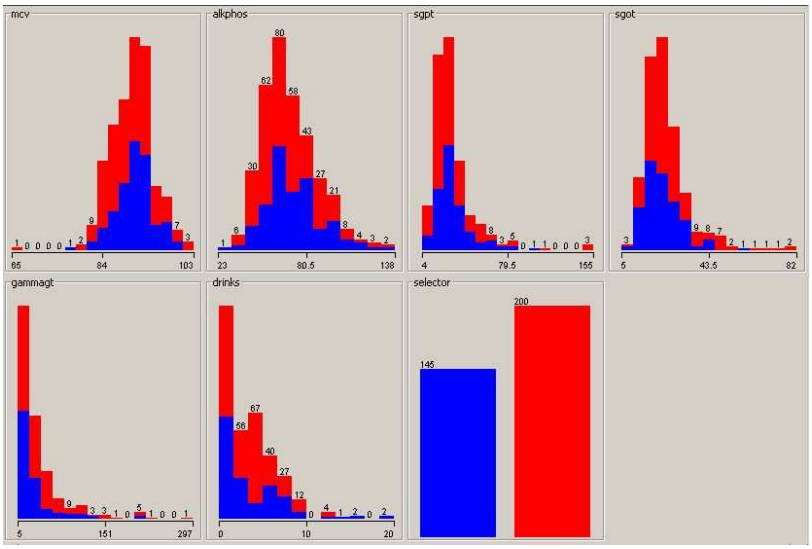


Figure 4.8. Liver Disorders Attributes' Instances distribution.

4.9. Obesity Data Set (BIOSTATISTICS, 2012)

This data was generated by Hacettepe University Bioinformatics Department to illustrate the power of the exercise and nutrition in the human life.

Contains 50 number of instances, 26 number of attributes and it has no missing values. The class codes are assigned for two class distribution types.

Table 4.12. Data information about Obesity Data Set.

No	Physical Attribute	Features – only used ones are listed
1	Sirano	Numeric
2	Cinsiyet	{1, 2}
3	Egzersiztip	{1, 2, 3}
4	Yas	Numeric
5	Boy	Numeric
6	Kilo_EO	Numeric
7	Kilo_ES_1ay	Numeric
8	Kilo_ES_4ay	Numeric
9	BKI_EO	Numeric
10	BKI_ES_1ay	Numeric
11	BKİ_ES_4ay	Numeric
12	bki_gr_EO	{1, 2, 3, 4}
13	bki_gr_ES_1ay	{1, 2, 3, 4}
14	bki_gr_ES_4ay	{1, 2, 3, 4}
15	Yagyuzdesi_EO	Numeric
16	Yagyuzdesi_ES_1ay	Numeric
17	Yagyuzdesi_ES_4ay	Numeric
18	RiskGrubu_EO	{0, 1}
19	RiskGrubu_ES_1ay	{0, 1}
20	Yagkg_EO	Numeric
21	Yagkg_ES_1ay	Numeric
22	Yagkg_ES_4ay	Numeric
23	Protein_EO	Numeric
24	Protein_ES_1ay	Numeric
25	Protein_ES_4ay	Numeric
26	RiskGrubu_ES_4ay	{0, 1}

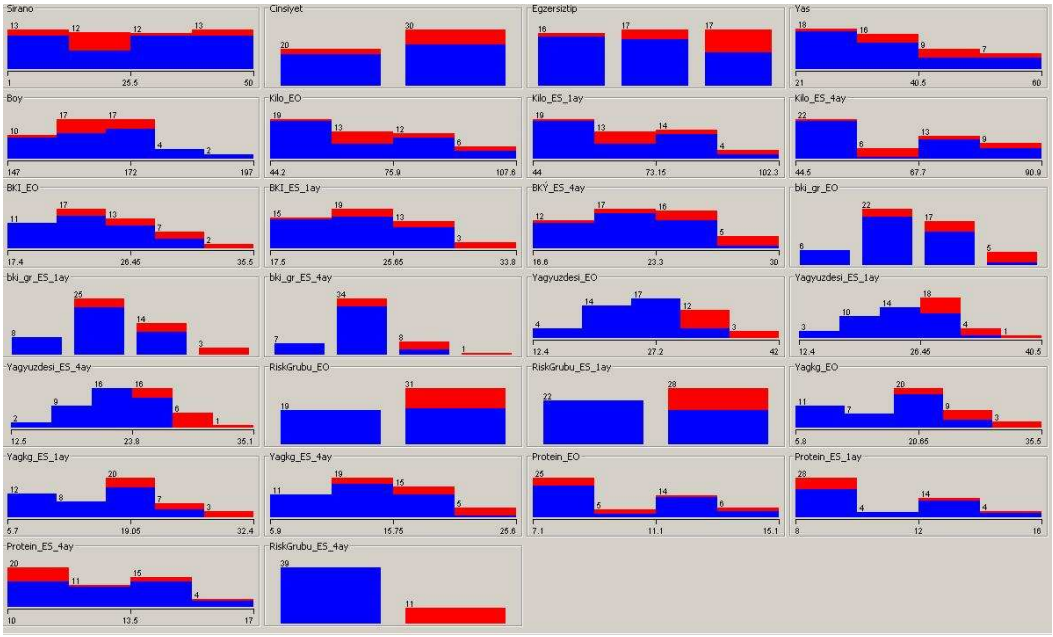


Figure 4.9. Obesity Attributes' Instances distribution.

4.10. Pima Indians Diabetes Data Set

The dataset shows that the diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.

Results: Their ADAP algorithm makes a real-valued prediction between 0 and 1. This was transformed into a binary decision using a cutoff of 0.448. Using 576 training instances, the sensitivity and specificity of their algorithm was 76% on the remaining 192 instances.

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of Perceptron-like devices.

Contains 768 number of instances, 9 number of attributes and it has 5 missing values. The class codes are assigned as follows: 0 for tested negative for diabetes, 1 for tested positive for diabetes.

Table 4.13. Data information about Pima Indians Diabetes Data Set.

No	Physical Attribute	Features – only used ones are listed
1	Number of times pregnant	Real
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Real
3	Diastolic blood pressure	Real
4	Triceps skin fold thickness	Real
5	2-Hour serum insulin	Real
6	Body mass index	Real
7	Diabetes pedigree function	Real
8	Age	Real
9	Class	{0, 1}

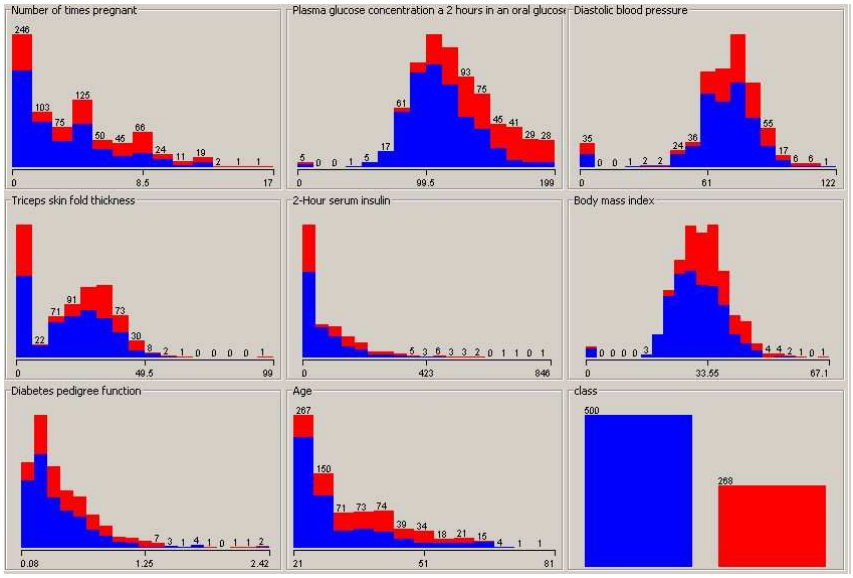


Figure 4.10. Pima Indians Diabetes Attributes' Instances distribution.

4.11. Respiratory Data Set (KISSIOV, 2005)

The data set consists of the clinical records (17 features) for 85 premature newborn children with two types of respiratory distress syndrome (RDS): Hyaline Membrane Disease (HMD) and non-HMD. The two classes need urgent and completely different treatment; therefore an accurate RDS classification is crucial within the first few hours after delivery.

Contains 85 number of instances, 18 number of attributes and it has no missing values

The class codes are assigned as follows: 1: non- Hyaline Membrane Disease, 2: Hyaline Membrane Disease

Table 4.14. Data information about Respiratory Data Set.

No	Physical Attribute	Features – only used ones are listed
1	Pathology during pregnancy	Real
2	Antenatal administration of corticosteroids	real
3	Preterm rupture of the foetal membranes	real
4	Chorionamnionitis	real
5	Tocolysis (suppression of the prematue delivery by i.v. infusion of beta-mimetics)	real
6	Gestation age (weeks)	real
7	Birth weight (grams)	real
8	Morphological maturity (weeks)	real
9	APGAR score at 5 minutes after delivery (scale from 0 to 10)	real
10	pH from umbilical artery (before ressuscitation of the newborn infant)	real
11	pH after ressuscitation (at least after	real

	1 hour)	
12	BE (Basis excess) after resuscitation	real
13	Heart rate at the 1 hour (bpm) [measured one hour after ressuscitation in the delivery room]	real
14	Respiratory insufficiency 6 hours after delivery (Silvermann index, scale from 0 to 10)	real
15	Cyanosis 6 hours after delivery	real
16	Crepitation rales by lung auscultation 6 hours after delivery	real
17	Heart rate 6 hours after delivery (bpm)	real
18	Class 1 - non-HMD, Class 2 – HMD	{1, 2}

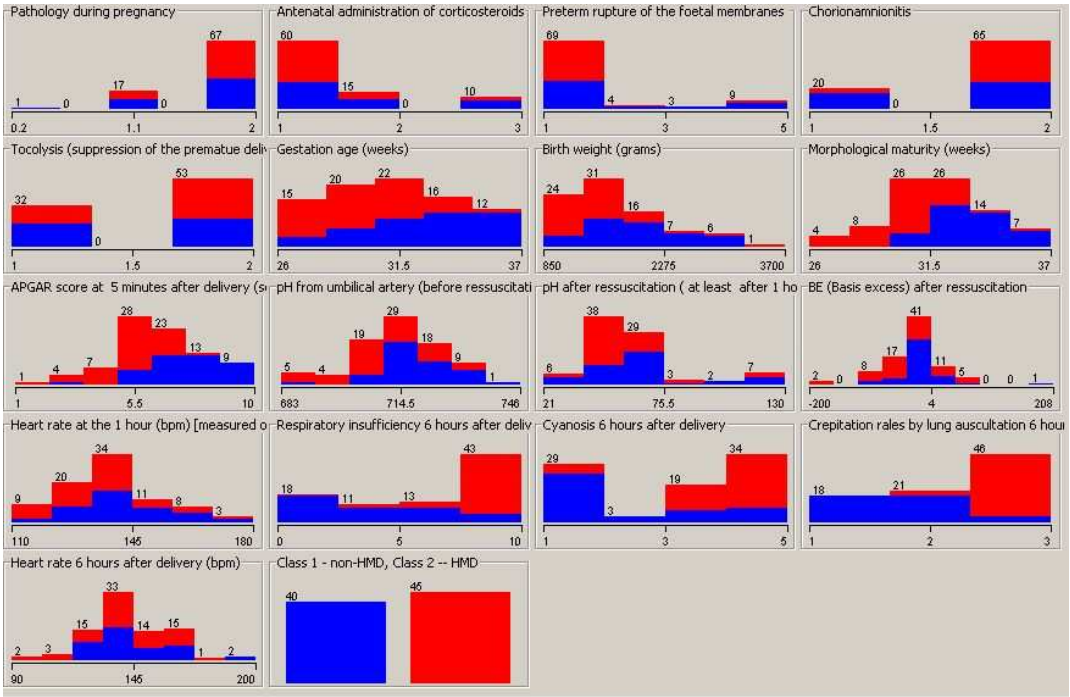


Figure 4.11. Respiratory Data Set Column Chart for Attributes' Instances distribution.

5. METHODS OF EVALUATION OF DATA MINING ALGORITHMS

(İrsoy, 2012) had a study for a Survey and Case Studies in Bioinformatics area. In this paper, they first review the performance measures used in classification, the basics of experiment design and statistical tests. Then they give the results of the survey over 1,500 papers published in the last two years in three bioinformatics journals. Although the basics of experiment design are well understood, such as resampling instead of using a single training set and the use of different performance metrics instead of error, only 21 percent of the papers use any statistical test for comparison. At last, they analyze four different scenarios which they encounter frequently in the bioinformatics literature, discussing the proper statistical methodology as well as showing an example case study for each.

Table 5.1. Number of Papers researched in (İrsoy, 2012).

Journal	**All Papers**	**ML Related**	**Classification**
BMC Bioinformatics 2010	466	167	71
Bioinformatics 2010	334	167	65
IEEE/ACM TCBB 2010	69	34	20
Total (2010)	**869**	**368**	**156**
BMC Bioinformatics 2011	266	85	41
Bioinformatics 2011	272	99	28
IEEE/ACM TCBB 2011	125	54	21
Total (2011)	**663**	**238**	**90**
Grand Total	**1532**	**606**	**246**

Especially in Medical Machine Learning Studies the rate of correct and incorrect diagnoses must be analyzed. It is important to know what part of cases was classified correctly. Medical Diagnosing is an important responsible task and the costs of mistaken

diagnosis may be really high. The cost of a missed diagnosis may be higher than long waiting testing.

In order to assess the significance of a mining method it is necessary to view the data mining as a process of examining data, learning solutions and then evaluating them. A very important aspect in method's performance evaluation is sample data, in case of this research - the medical data records. That is why classification of medical records to appropriate classes (diagnosis) is usually prediction (not certainty). A method of evaluating data mining method relies on dividing the sample data set into two subsets: training data and testing data.

This chapter describes techniques of evaluating performance of data mining methods with the special consideration of medical data and diagnosis. Here we based (İrsoy, 2012) research paper and prepare the experiment accordingly.

To evaluate the accuracy of the data mining method it is essential to start with assessment of the made by this method hypotheses (assumption) accuracy. To achieve this goal the statistical methods are used. The hypothesis accuracy is taken into consideration while learning from limited-size databases. When the hypothesis about medical treatment is prepared one must estimate the effectiveness of different learned solutions. What is more the evaluation is an integral component of data mining methods, for instance the impact of possible pruning steps that let avoid overfitting on the results given by decision tree (Domingos, 1997). Estimating accuracy is trouble-free in case of plentiful data but in real-case situation one have only limited-size database. In such situation there arise bias in estimate that is the difference between mean value and real value of the hypothesis. Moreover there is a variance in the estimation which measure statistical dispersion, indicating how far from the expected value the random variables values typically are. It is a reason why it is impossible to estimate extract accuracy of the method and statistical solutions must be applied.

The accuracy of the hypothesis concerns the correctness of classification of future instances. The accuracy of the test shows how precisely the test set were classified with the method. The results of the classification are being compared with the real (if it is possible) classification made by the physician. The accuracy of the test is ratio of true positive test examples to all test examples. This measure is very

common in science, however it cannot be applied in medical cases because it deform the medical reality perception – it does not take False Negative (FN) case into consideration. That is why in medicine there must be the other measure of training set accuracy applied.

$$Accuracy = \frac{TP}{total} \tag{5.1}$$

The above formula could be incorrect in medical databases. With medical data when only two possible answers are taken into consideration (yes/no), the other measures of accuracy should be applied (Cios, 2002).

The sensitivity is a percentage proportion of True Positive cases to the sum of all positive hypotheses (True Positive and False Negative).

$$Sensitivit\,y = \frac{(TP)}{(TP + FN)} \tag{5.2}$$

The specificity, the next measure, shows the possibility of correct, negative classification of the patient to as an healthy to all negative hypothesis. The specificity is a percentage proportion of True Negative cases to the sum of all negative hypotheses (False Positive and True Negative).

$$Specificit\,y = \frac{(TN)}{(TN + FP)} \tag{5.3}$$

The third measure of accuracy is called predictive accuracy. This measure shows the ratio of correctly classified cases to all cases in the set. The larger predictive accuracy is the better situation.

$$Predictive\ \ Accuracy\ = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{5.4}$$

While estimating the effectiveness and accuracy of data mining technique it is essential to measure the error rate of each method. In the case of binary classification tasks the error rate takes FalsePositives and FalseNegatives components under consideration. The ROC analysis which stands for Receiver Operating Characteristics is applied for these kinds of conditions.

Additionally the AUC (Area Under ROC Curve) is taken under consideration. It measures the model's ability to find the difference between two outcomes. The AUC method is called discrimination. In the perfect case the discrimination is equal 1, however in two outcomes has discrimination rate equal 0.5 because the area under diagonal axis (when no model applied).

The idea of ROC curve originally comes from signal detection theory. It plots the curve which consists of x-axis presenting false positive rate and y-axis which plots the true positive rate. This curve model selects the optimal model on the basis on assumed class distribution. The ROC curves are applicable e.g. in decision tree models or rule sets. The modification of each classification method may leads to better prediction performance and what consequently better medical decision supporting process.

Usually medical diagnosis is a binary classification problem. One has to conclude whether the patient is ill or not. There are four possible results of classification. Different combination of these four error and correct situations are presented in the scientific literature on topic. Here three popular notions are presented. The introduction of these classifiers is explained by the possibility of high accuracy by negative type of data. To avoid such situation recall and precision of the classification are introduced. The F measure is the harmonic mean of precision and recall. In an application like medical diagnosis, more than the true negatives, i.e., the large proportion of healthy individuals, it's important to detect the sick, and it is better to focus on Precision and Recall.

The formal definitions of precision, recall and f-measure are as follow (Ye, 2003):

$$Precision = \frac{(TP)}{(TP + FP)} \tag{5.5}$$

$$Recall = \frac{(TP)}{(TP + FN)} \tag{5.6}$$

$$F - Measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \tag{5.7}$$

6. RESULTS AND DISCUSSION

This section presents the results of the study using WEKA machine learning tools to evaluate the classifiers. The results of varying classifiers and the effects of these classifiers on the datasets are discussed and presented in this chapter.

The used HW platform for the Thesis is as follows:

- OS: Microsoft Windows XP Version 2002 Service Pack 3
- CPU: Intel® Core ™ i5-2520M CPU @ 2.50GHz
- RAM: 3 GB

The used software platform details are as follows:

- WEKA – Version 3.6.6 © 1999-2011(Hall, 2009)
- The University of Waikato
- Hamilton, New Zealand

The analyses are performed as follows:

- Each of the algorithms is calibrated with the use of several parameters available in the WEKA.
- These algorithms are performed on each of the selected datasets.
- The outcomes are measured with the use of the metrics described in Chapter 5.
- The results are presented in Paired T-Tester format.
- The ultimate step encompasses the comparison of the performance of each of machine learning models.

6.1. Algorithms Calibration

This section is dedicated to the analyses of calibration of individual algorithms. The purpose of these analyses is to determine what parameters settings yield the best models. The experiments are conducted with the use of all the eleven datasets mentioned in the Thesis. The calibration aims at finding optimal settings which maximize the performance of each of the algorithms. The tests are done with the use of 66% Train-Test split and 10-fold cross-validations testing configuration.

Making the decision about the split of the available data one must think about the quality of the generated model and the quality of this model's tests. The more data is used in a training phase the better the generated model may be. On the other hand, the more data is taken for testing the more accurate the output model may be. One has to decide about the split of the available data to reach a compromise. The previous researches claim that the most accurate results are obtained for the 66% split of the dataset. The 66% of the data should be taken to build the model and the rest is to test it. The previous researches also say that in order to improve the training phase the n-fold cross-validation should be applied. Their experiments have shown that the best performance is achieved when n equals 10.

Algorithms calibration details are as follows:

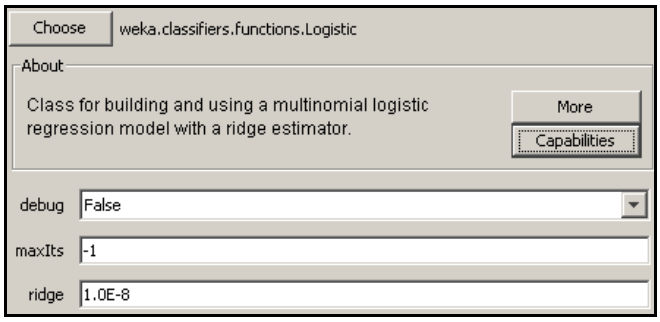


Figure 6.1. Logistic Regression Classifier Calibration.

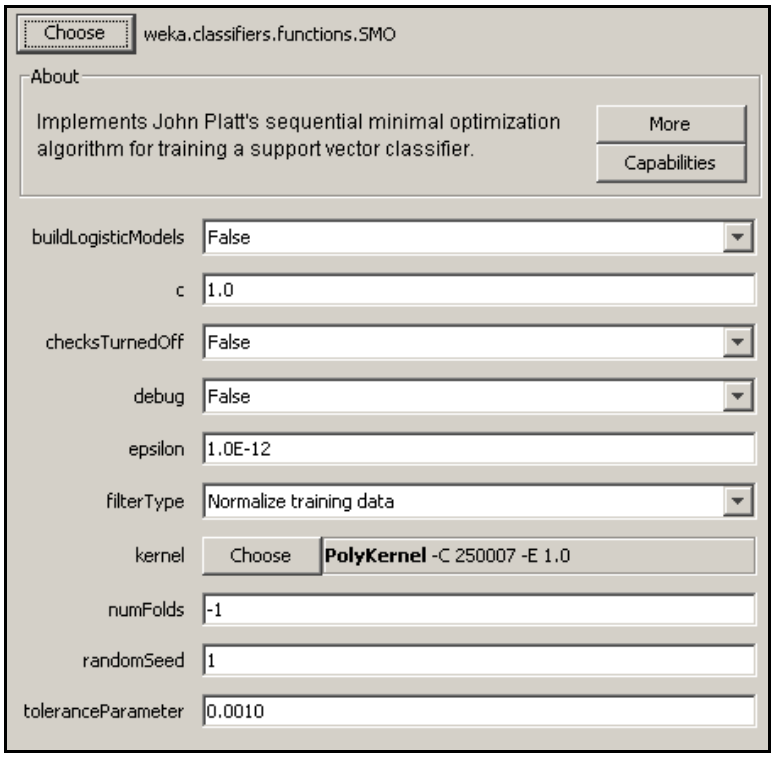


Figure 6.2. Support Vector Machine Classifier Calibration.

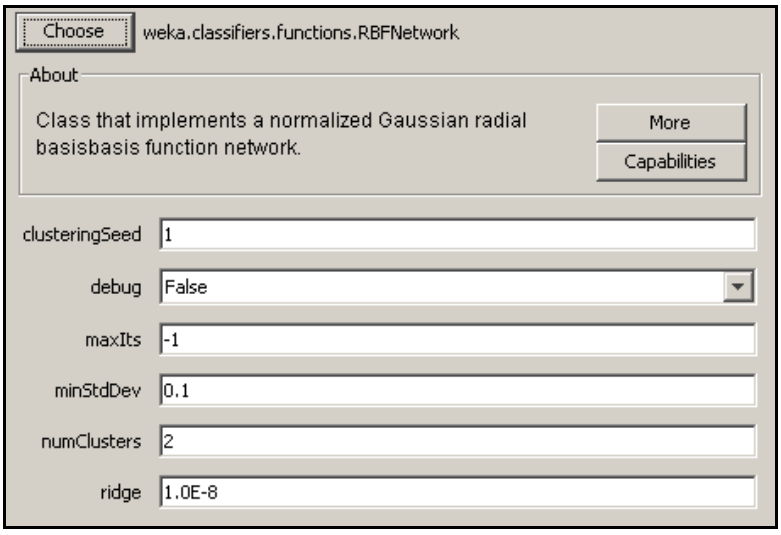


Figure 6.3. Radial Basis Function Classifier Calibration.

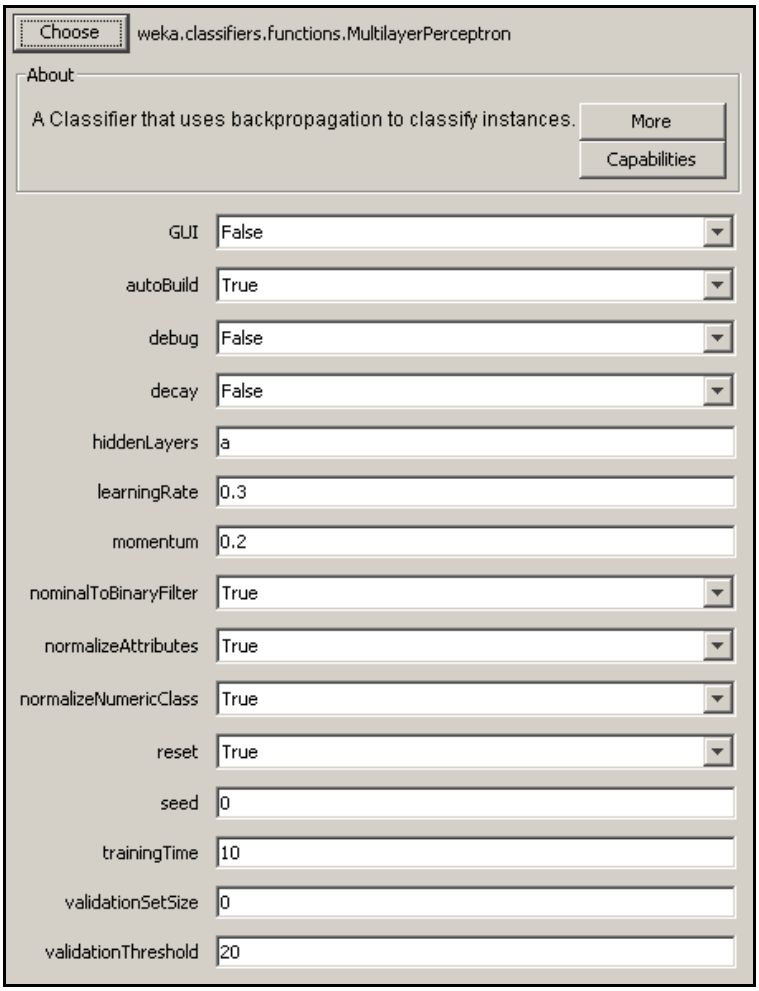


Figure 6.4. Multi Layer Perceptron Classifier Calibration.

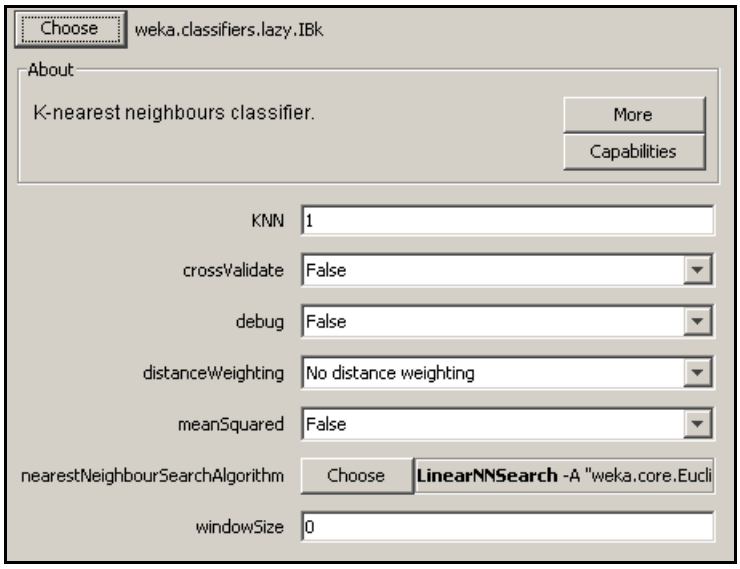


Figure 6.5. k-NN Classifier Calibration.

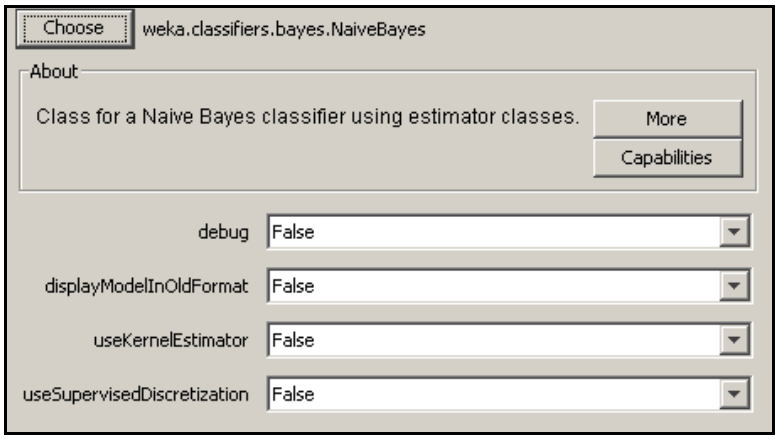


Figure 6.6. Naive Bayes Classifier Calibration.

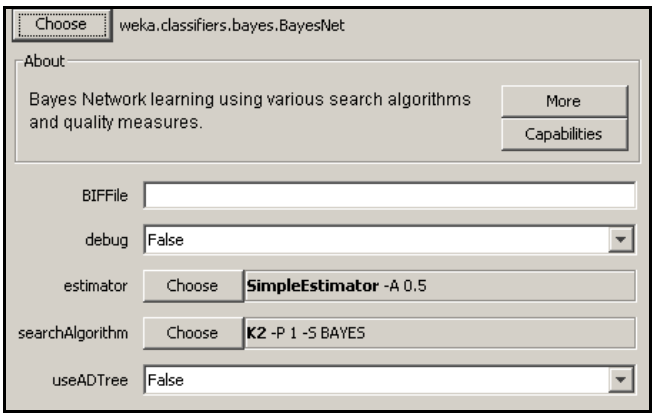


Figure 6.7. Bayesian Network Classifier Calibration.

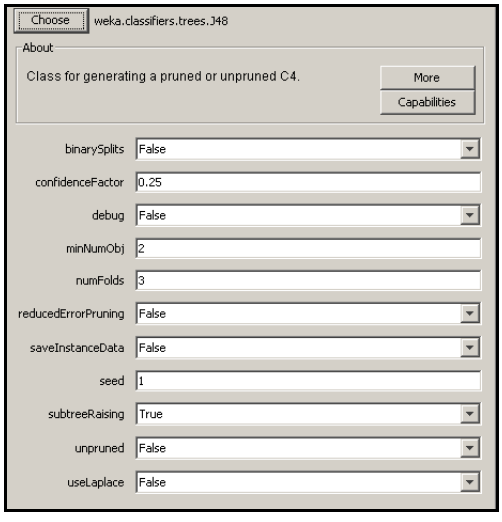


Figure 6.8. J48- Decision Tree Classifier Calibration.

6.2. Evaluation and Comparison of the Data Mining Algorithms

This section is dedicated to analysis of the results obtained during the calibration of the algorithms. Here the comparison of the algorithms in terms of performance. This part of the examinations was conducted in the Experimenter graphical interface of the WEKA environment.

10-fold cross-validation and 2/3 Train Test Split has been chosen. For Feature Extraction, we use PCA after we implemented "Pure Implementations". The results of the comparison of the algorithms are presented in the below figures. The figures show the ranking of the algorithm with Paired T-Tests in case of each Percent Correct, True Positive Rate, Precision, Recall, F-Measure, AUC and Error.

We have prepared special evaluation criteria, which looks at number of Best Averages, Wins Most and Test Base (actually a minus point) in comparisons. There are Four Results, which has been presented:

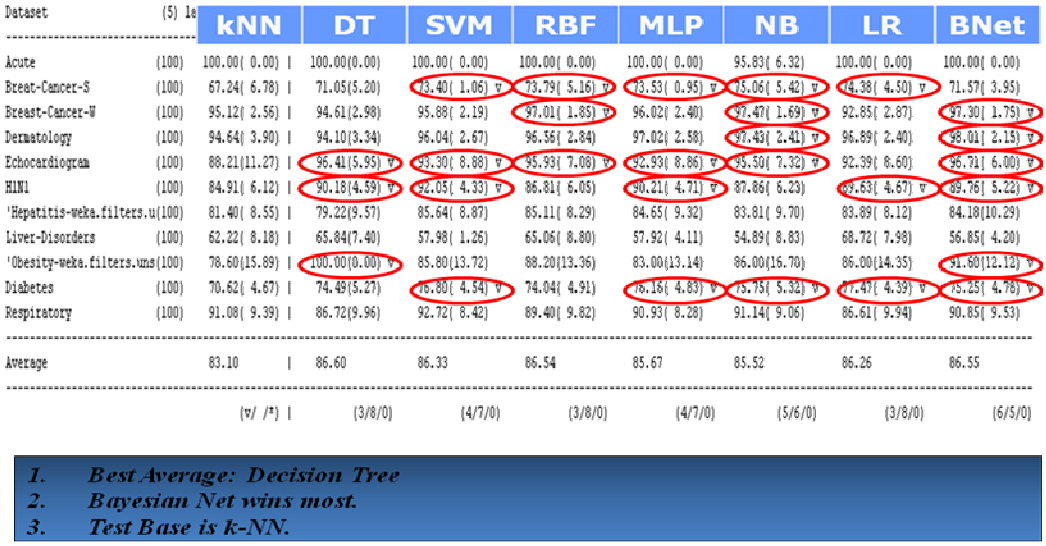
- Pure 10 Fold: Naïve Bayes is the Best Classifier
- Pure Train Test Split: Naïve Bayes is the Best Classifier
- PCA 10 Fold: Naïve Bayes is the Best Classifier
- PCA Train Test Split: Naïve Bayes is the Best Classifier

The unquestionable leader in majority of cases is the Naïve Bayes. Although when we look at Error Rates, we see that Naïve Bayes has worst results in Acute Dataset for Pure 10 Fold and Pure Train Test Splits. On the other side, this disadvantage has been resolved after PCA has been applied in the study. Nevertheless, overall performance was always better in comparison to other algorithms. For most of the databases and metrics the results gained by this algorithm were slightly worse than for the Naïve Bayes in most of the cases.

When it comes to the Bayes Net, it wins the second place in terms of the benchmarking for Pure Implementations. It's obvious that Bayes Net has not succeeded in PCA experiments. Logistic Regression and Support Vector Machine take the second place in terms of benchmarking with the PCA implementations.

Finally, the worst results were yielded by the k-NN classifier. Its results were the worst in most of the cases. The reason for this may the nature of medical data. Its complexity and heterogeneity of values of attributes can hinder data mining.

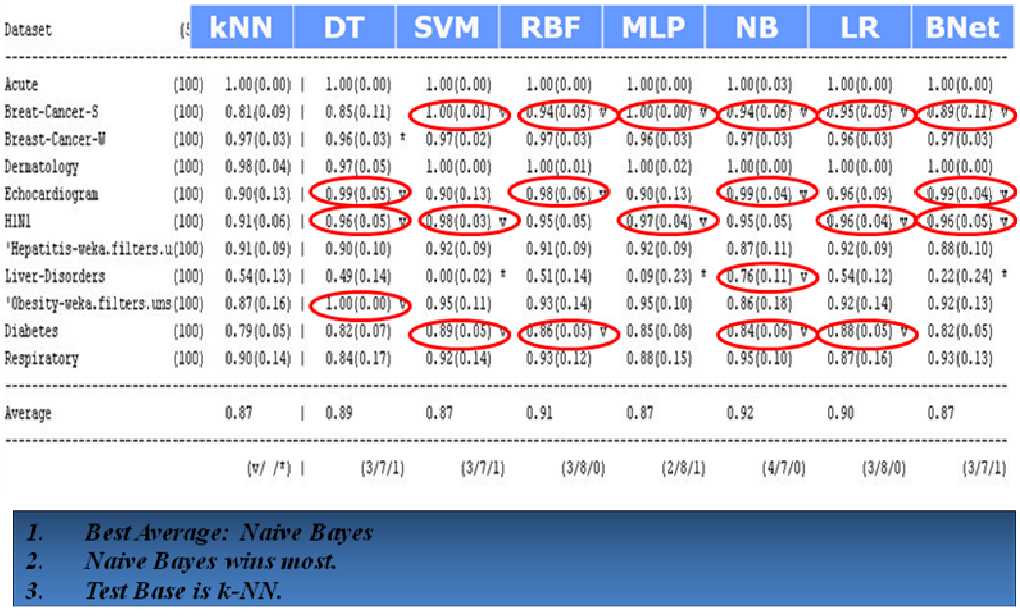
The results of the experiment are demonstrated in following figures:



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)	95.83(6.32)	100.00(0.00)	100.00(0.00)
Breat-Cancer-S	(100)	67.24(6.78)	71.05(5.20)	73.40(1.06) v	73.79(5.16) v	73.53(0.95) v	75.06(5.42) v	74.38(4.50) v	71.57(3.95)
Breast-Cancer-W	(100)	95.12(2.56)	94.61(2.98)	95.88(2.19)	97.01(1.85) v	96.02(2.40)	97.47(1.69) v	92.85(2.87)	97.30(1.75) v
Dermatology	(100)	94.64(3.90)	94.10(3.34)	96.04(2.67)	96.56(2.84)	97.02(2.58)	97.43(2.41) v	96.89(2.40)	98.01(2.13) v
Echocardiogram	(100)	88.21(11.27)	96.41(5.95) v	93.30(8.88) v	95.93(7.08) v	92.93(8.86) v	95.50(7.32) v	92.39(8.60)	96.71(6.00) v
H1N1	(100)	84.91(6.12)	90.18(4.59) v	92.05(4.33) v	86.81(6.05)	90.21(4.71) v	87.86(6.23)	89.63(4.67) v	89.76(5.22) v
'Hepatitis-weka.filters.u	(100)	81.40(8.55)	79.22(9.57)	85.64(8.87)	85.11(8.29)	84.65(9.32)	83.81(9.70)	83.89(8.12)	84.18(10.29)
Liver-Disorders	(100)	62.22(8.18)	65.84(7.40)	57.98(1.26)	65.06(8.80)	57.92(4.11)	54.89(8.83)	68.72(7.98)	56.85(4.20)
'Obesity-weka.filters.uns	(100)	78.60(15.89)	100.00(0.00) v	85.80(13.72)	88.20(13.36)	83.00(13.14)	86.00(16.70)	86.00(14.35)	91.60(12.12) v
Diabetes	(100)	70.62(4.67)	74.49(5.27)	76.88(4.54) v	74.04(4.91)	76.18(4.83) v	75.75(5.32) v	77.47(4.39) v	75.25(4.78) v
Respiratory	(100)	91.08(9.39)	86.72(9.96)	92.72(8.42)	89.40(9.82)	90.93(8.28)	91.14(9.06)	86.61(9.94)	90.85(9.53)
Average		83.10	86.60	86.33	86.54	85.67	85.52	86.26	86.55
		(v/ /*)	(3/8/0)	(4/7/0)	(3/8/0)	(4/7/0)	(5/6/0)	(3/8/0)	(6/5/0)

1. *Best Average: Decision Tree*
2. *Bayesian Net wins most.*
3. *Test Base is k-NN.*

Figure 6.9. Pure 10 Fold – Percent Correct.



Dataset	(5	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.03)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(100)	0.81(0.09)	0.85(0.11)	1.00(0.01) v	0.94(0.05) v	1.00(0.00) v	0.94(0.06) v	0.95(0.05) v	0.89(0.11) v
Breast-Cancer-W	(100)	0.97(0.03)	0.96(0.03) *	0.97(0.02)	0.97(0.03)	0.96(0.03)	0.97(0.03)	0.96(0.03)	0.97(0.03)
Dermatology	(100)	0.98(0.04)	0.97(0.05)	1.00(0.00)	1.00(0.01)	1.00(0.02)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Echocardiogram	(100)	0.90(0.13)	0.99(0.05) v	0.90(0.13)	0.98(0.06) v	0.90(0.13)	0.99(0.04) v	0.96(0.09)	0.99(0.04) v
H1N1	(100)	0.91(0.06)	0.96(0.05) v	0.98(0.03) v	0.95(0.05)	0.97(0.04) v	0.95(0.05)	0.96(0.04) v	0.96(0.05) v
'Hepatitis-weka.filters.u	(100)	0.91(0.09)	0.90(0.10)	0.92(0.09)	0.91(0.09)	0.92(0.09)	0.87(0.11)	0.92(0.09)	0.88(0.10)
Liver-Disorders	(100)	0.54(0.13)	0.49(0.14)	0.00(0.02) *	0.51(0.14)	0.09(0.23) *	0.76(0.11) v	0.54(0.12)	0.22(0.24) *
'Obesity-weka.filters.uns	(100)	0.87(0.16)	1.00(0.00) v	0.95(0.11)	0.93(0.14)	0.95(0.10)	0.86(0.18)	0.92(0.14)	0.92(0.13)
Diabetes	(100)	0.79(0.05)	0.82(0.07)	0.89(0.05) v	0.86(0.05) v	0.85(0.08)	0.84(0.06) v	0.88(0.05) v	0.82(0.05)
Respiratory	(100)	0.90(0.14)	0.84(0.17)	0.92(0.14)	0.93(0.12)	0.88(0.15)	0.95(0.10)	0.87(0.16)	0.93(0.13)
Average		0.87	0.89	0.87	0.91	0.87	0.92	0.90	0.87
		(v/ /*)	(3/7/1)	(3/7/1)	(3/8/0)	(2/8/1)	(4/7/0)	(3/8/0)	(3/7/1)

1. *Best Average: Naive Bayes*
2. *Naive Bayes wins most.*
3. *Test Base is k-NN.*

Figure 6.10. Pure 10 Fold – True Positive Rate.

Dataset	(2)	BNet	DT	SVM	RBF	MLP	kNN	NB	LR
Acute	(100)	1.00(0.00) |	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.94(0.08) *	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(100)	0.73(0.01) |	0.78(0.05) v	0.76(0.04) v	0.74(0.01)	0.78(0.04) v	0.77(0.04) v	0.76(0.03) v	0.73(0.04) v
Breast-Cancer-W	(100)	0.97(0.03) |	0.96(0.03)	0.99(0.02) v	0.98(0.03)	0.95(0.03)	0.99(0.01) v	0.93(0.03) *	0.95(0.01) v
Dermatology	(100)	0.99(0.02) |	0.97(0.06)	1.00(0.01)	0.99(0.02)	0.99(0.02)	1.00(0.00)	0.98(0.04)	1.00(0.00)
Echocardiogram	(100)	1.00(0.00) |	0.97(0.07)	0.96(0.08)	0.99(0.04)	0.93(0.09) *	0.95(0.08) *	0.94(0.09) *	0.97(0.07)
H1N1	(100)	0.93(0.04) |	0.93(0.05)	0.90(0.05) *	0.92(0.04)	0.91(0.04)	0.91(0.05)	0.92(0.04)	0.93(0.04)
'Hepatitis-weka.filters.u	(100)	0.90(0.06) |	0.85(0.06) *	0.91(0.06)	0.90(0.07)	0.87(0.06)	0.92(0.06)	0.89(0.06)	0.92(0.07)
Liver-Disorders	(100)	0.04(0.20) |	0.63(0.13) v	0.60(0.14) v	0.12(0.27)	0.56(0.10) v	0.48(0.07) v	0.66(0.12) v	0.22(0.25)
'Obesity-weka.filters.uns	(100)	0.89(0.12) |	1.00(0.00) v	0.93(0.11)	0.86(0.12)	0.87(0.13)	0.96(0.10)	0.92(0.11)	0.98(0.07) v
Diabetes	(100)	0.78(0.04) |	0.80(0.05)	0.77(0.04)	0.80(0.05) v	0.76(0.04)	0.80(0.04)	0.79(0.04) v	0.81(0.04)
Respiratory	(100)	0.94(0.10) |	0.90(0.13)	0.88(0.13)	0.94(0.10)	0.93(0.12)	0.89(0.12)	0.87(0.13)	0.91(0.13)
Average		0.83 |	0.89	0.88	0.84	0.87	0.87	0.88	0.86
		(v/ /*) |	(3/7/1)	(3/7/1)	(1/10/0)	(2/8/1)	(3/6/2)	(3/6/2)	(3/8/0)

1. *Best Average: DT*
2. *DT, SVM, k-NN and NB win most. DT and SVM are better because of no losses.*
3. *Test Base is SVM*

Figure 6.11. Pure 10 Fold – Precision.

Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	1.00(0.00) |	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.03)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(100)	0.81(0.09) |	0.85(0.11)	1.00(0.01) v	0.94(0.05) v	1.00(0.00) v	0.94(0.06) v	0.95(0.05) v	0.89(0.11) v
Breast-Cancer-W	(100)	0.97(0.03) |	0.96(0.03) *	0.97(0.02)	0.97(0.03)	0.96(0.03)	0.97(0.03)	0.96(0.03)	0.97(0.03)
Dermatology	(100)	0.98(0.04) |	0.97(0.05)	1.00(0.00)	1.00(0.01)	1.00(0.02)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Echocardiogram	(100)	0.90(0.13) |	0.99(0.05) v	0.90(0.13)	0.98(0.06) v	0.90(0.13)	0.99(0.04) v	0.96(0.09)	0.99(0.04) v
H1N1	(100)	0.91(0.06) |	0.96(0.05) v	0.98(0.03) v	0.95(0.05)	0.97(0.04) v	0.95(0.05)	0.96(0.04) v	0.96(0.05) v
'Hepatitis-weka.filters.u	(100)	0.91(0.09) |	0.90(0.10)	0.92(0.09)	0.91(0.09)	0.92(0.09)	0.87(0.11)	0.92(0.09)	0.88(0.10)
Liver-Disorders	(100)	0.54(0.13) |	0.49(0.14)	0.00(0.02) *	0.51(0.14)	0.09(0.23) *	0.76(0.11) v	0.54(0.12)	0.22(0.24) *
'Obesity-weka.filters.uns	(100)	0.87(0.16) |	1.00(0.00) v	0.95(0.11)	0.93(0.14)	0.95(0.10)	0.86(0.18)	0.92(0.14)	0.92(0.13)
Diabetes	(100)	0.79(0.05) |	0.82(0.07)	0.89(0.05) v	0.86(0.05) v	0.85(0.08)	0.84(0.06) v	0.88(0.05) v	0.82(0.05)
Respiratory	(100)	0.90(0.14) |	0.84(0.17)	0.92(0.14)	0.93(0.12)	0.88(0.15)	0.95(0.10)	0.87(0.16)	0.93(0.13)
Average		0.87 |	0.89	0.87	0.91	0.87	0.92	0.90	0.87
		(v/ /*) |	(3/7/1)	(3/7/1)	(3/8/0)	(2/8/1)	(4/7/0)	(3/8/0)	(3/7/1)

1. *Best Average: NB*
2. *NB win most.*
3. *Test Base is k-NN*

Figure 6.12. Pure 10 Fold – Recall.

Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	1.00(0.00) |	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.97(0.05)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(100)	0.78(0.05) |	0.81(0.04)	0.85(0.01) v	0.84(0.03) v	0.85(0.01) v	0.85(0.03) v	0.85(0.03) v	0.82(0.04)
Breast-Cancer-W	(100)	0.96(0.02) |	0.96(0.02)	0.97(0.02)	0.98(0.01) v	0.97(0.02)	0.98(0.01) v	0.95(0.02)	0.98(0.01) v
Dermatology	(100)	0.99(0.02) |	0.97(0.03)	1.00(0.01)	1.00(0.01)	0.99(0.01)	1.00(0.00)	0.99(0.02)	1.00(0.00)
Echocardiogram	(100)	0.91(0.09) |	0.97(0.04) v	0.94(0.08) v	0.97(0.05) v	0.94(0.08) v	0.97(0.05) v	0.94(0.07)	0.98(0.04) v
H1N1	(100)	0.91(0.04) |	0.94(0.03) v	0.95(0.02) v	0.92(0.04)	0.94(0.03) v	0.93(0.04)	0.94(0.03) v	0.94(0.03) v
'Hepatitis-weka.filters.u	(100)	0.88(0.06) |	0.87(0.06)	0.91(0.06)	0.91(0.06)	0.90(0.06)	0.89(0.07)	0.90(0.05)	0.90(0.07)
Liver-Disorders	(100)	0.54(0.10) |	0.54(0.11)	0.01(0.03) *	0.54(0.13)	0.08(0.18) *	0.59(0.07)	0.59(0.11)	0.22(0.24) *
'Obesity-weka.filters.uns	(100)	0.86(0.11) |	1.00(0.00) v	0.91(0.09)	0.92(0.10)	0.90(0.08)	0.90(0.13)	0.91(0.10)	0.94(0.09) v
Diabetes	(100)	0.78(0.04) |	0.81(0.04)	0.83(0.03) v	0.81(0.04)	0.82(0.04) v	0.82(0.04) v	0.84(0.03) v	0.81(0.04) v
Respiratory	(100)	0.91(0.10) |	0.85(0.12)	0.92(0.10)	0.89(0.10)	0.90(0.10)	0.91(0.09)	0.86(0.11)	0.91(0.10)
Average		0.87 |	0.88	0.84	0.89	0.84	0.89	0.89	0.86
		(v/ /*) |	(3/8/0)	(4/6/1)	(3/8/0)	(4/6/1)	(4/7/0)	(3/8/0)	(5/5/1)

1. *Best Average: RBF, NB and LR*
2. *Bayes Net wins most.*
3. *Test Base is k-NN*

Figure 6.13. Pure 10 Fold – F-Measure.

Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	1.00(0.00) |	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.01)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(100)	0.57(0.08) |	0.58(0.08)	0.50(0.00) *	0.67(0.12) v	0.64(0.12)	0.64(0.12)	0.68(0.12) v	0.63(0.09)
Breast-Cancer-W	(100)	0.94(0.03) |	0.97(0.03) v	0.95(0.03)	0.99(0.01) v	0.99(0.01) v	0.99(0.01) v	0.95(0.03)	0.99(0.01) v
Dermatology	(100)	0.99(0.02) |	0.98(0.02)	1.00(0.01)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.01)	1.00(0.00)
Echocardiogram	(100)	0.66(0.09) |	0.71(0.07) v	0.68(0.09) v	0.71(0.12)	0.71(0.12) v	0.69(0.10)	0.70(0.09)	0.69(0.12)
H1N1	(100)	0.66(0.06) |	0.75(0.13) v	0.90(0.05) v	0.79(0.09) v	0.83(0.07) v	0.80(0.08) v	0.93(0.04) v	0.89(0.05) v
'Hepatitis-weka.filters.u	(100)	0.68(0.14) |	0.67(0.18)	0.77(0.14)	0.87(0.12) v	0.85(0.13) v	0.86(0.13) v	0.84(0.15) v	0.88(0.11) v
Liver-Disorders	(100)	0.61(0.08) |	0.65(0.09)	0.50(0.01) *	0.68(0.10)	0.62(0.10)	0.64(0.10)	0.72(0.09) v	0.52(0.05) *
'Obesity-weka.filters.uns	(100)	0.69(0.25) |	1.00(0.00) v	0.75(0.24)	0.88(0.21)	0.91(0.15) v	0.94(0.15) v	0.89(0.22) v	0.97(0.08) v
Diabetes	(100)	0.67(0.05) |	0.75(0.07) v	0.71(0.05) v	0.79(0.06) v	0.83(0.05) v	0.82(0.05) v	0.83(0.05) v	0.81(0.05) v
Respiratory	(100)	0.91(0.09) |	0.91(0.10)	0.93(0.09)	0.91(0.13)	0.98(0.04) v	0.96(0.08)	0.94(0.08)	0.97(0.06)
Average		0.76 |	0.81	0.79	0.84	0.85	0.85	0.86	0.85
		(v/ /*) |	(5/6/0)	(3/6/2)	(5/6/0)	(7/4/0)	(5/6/0)	(6/5/0)	(5/5/1)

1. *Best Average: LR*
2. *MLP wins most.*
3. *Test Base is k-NN*

Figure 6.14. Pure 10 Fold – AUC.

Dataset	(	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	0.01 |	0.00 *	0.00 *	0.00 *	0.08 v	0.13 v	0.00 *	0.02 v
Breat-Cancer-S	(100)	0.57 |	0.44 *	0.52 *	0.43 *	0.44 *	0.44 *	0.42 *	0.43 *
Breast-Cancer-W	(100)	0.21 |	0.21	0.19	0.15 *	0.17	0.14 *	0.26	0.14 *
Dermatology	(100)	0.12 |	0.12	0.31 v	0.09	0.08	0.07 *	0.09	0.06 *
Echocardiogram	(100)	0.27 |	0.11 *	0.17 *	0.10 *	0.22	0.12 *	0.19	0.10 *
H1N1	(100)	0.38 |	0.29 *	0.26 *	0.31 *	0.29 *	0.29 *	0.28 *	0.26 *
'Hepatitis-weka.filters.u	(100)	0.42 |	0.40	0.35	0.32 *	0.33 *	0.35	0.35	0.35
Liver-Disorders	(100)	0.61 |	0.52 *	0.65	0.47 *	0.49 *	0.51 *	0.46 *	0.50 *
'Obesity-weka.filters.uns	(100)	0.39 |	0.00 *	0.29	0.25	0.34	0.27	0.27	0.17 *
Diabetes	(100)	0.54 |	0.44 *	0.48 *	0.42 *	0.40 *	0.42 *	0.40 *	0.42 *
Respiratory	(100)	0.23 |	0.28	0.19	0.27	0.24	0.22	0.32	0.24
Average		0.34 |	0.26	0.31	0.26	0.28	0.27	0.28	0.24
		(v/ /*) |	(0/4/7)	(1/5/5)	(0/3/8)	(1/5/5)	(1/3/7)	(0/6/5)	(1/2/8)

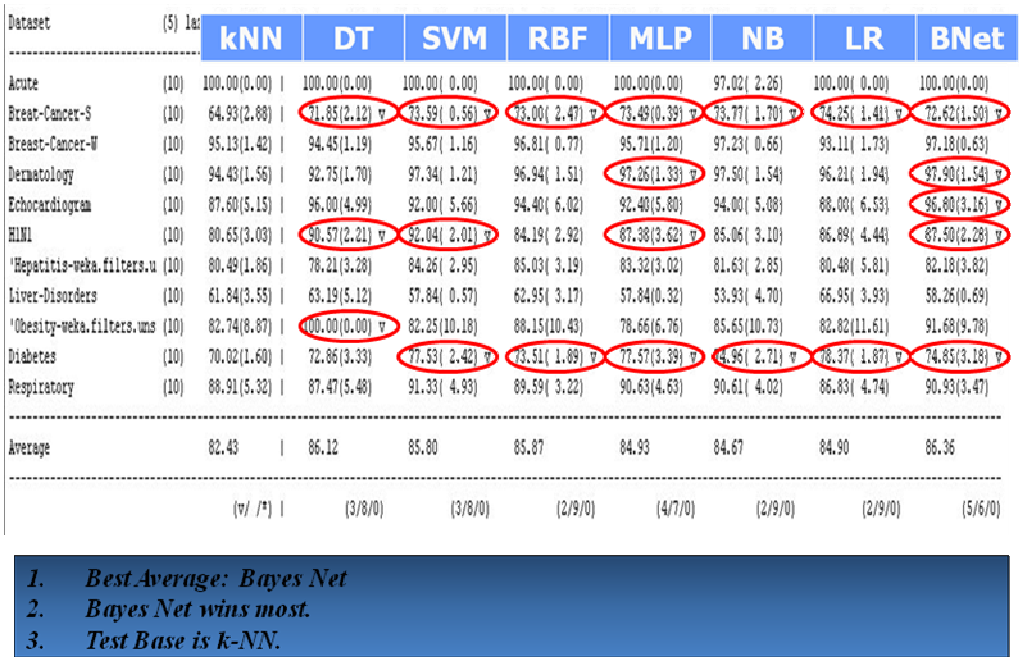
1. *Best Average: Bayes Net*
2. *Bayes Net wins most.*
3. *Test Base is k-NN*

Figure 6.15. Pure 10 Fold – Error Rates.

	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Best Average		2		1		3	2	1
Wins Most		1	1		1	2		3
Test Base	6		3					

Naive Bayes is the best classifier.

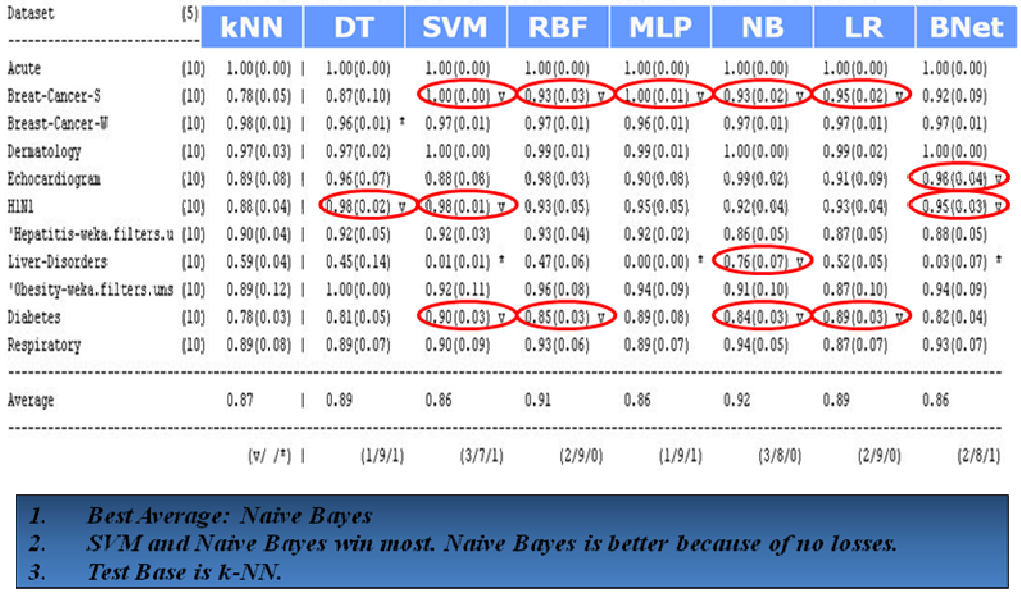
Figure 6.16. Pure 10 Fold – Summary.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)	97.02(2.26)	100.00(0.00)	100.00(0.00)
Breast-Cancer-S	(10)	64.93(2.88)	71.85(2.12) ▽	73.59(0.56) ▽	73.00(2.47) ▽	73.49(0.39) ▽	73.77(1.70) ▽	74.25(1.41) ▽	72.62(1.50) ▽
Breast-Cancer-W	(10)	95.13(1.42)	94.45(1.19)	95.67(1.16)	96.81(0.77)	95.71(1.20)	97.23(0.66)	93.11(1.73)	97.18(0.63)
Dermatology	(10)	94.43(1.56)	92.75(1.70)	97.34(1.21)	96.94(1.51)	97.26(1.33) ▽	97.50(1.54)	96.21(1.94)	97.90(1.54) ▽
Echocardiogram	(10)	87.60(5.15)	96.00(4.99)	92.00(5.66)	94.40(6.02)	92.40(5.80)	94.00(5.08)	88.00(6.53)	96.80(3.16) ▽
H1N1	(10)	80.65(3.03)	90.57(2.21) ▽	92.04(2.01) ▽	84.19(2.92)	87.38(3.62) ▽	85.06(3.10)	86.89(4.44)	87.50(2.28) ▽
'Hepatitis-weka.filters.u	(10)	80.49(1.86)	78.21(3.28)	84.26(2.95)	85.03(3.19)	83.32(3.02)	81.63(2.85)	80.48(5.81)	82.18(3.82)
Liver-Disorders	(10)	61.84(3.55)	63.19(5.12)	57.84(0.57)	62.95(3.17)	57.84(0.32)	53.93(4.70)	66.95(3.93)	58.26(0.69)
'Obesity-weka.filters.uns	(10)	82.74(8.87)	100.00(0.00) ▽	82.25(10.18)	88.15(10.43)	78.66(6.76)	85.65(10.73)	82.82(11.61)	91.68(9.78)
Diabetes	(10)	70.02(1.60)	72.86(3.33)	77.53(2.42) ▽	73.51(1.89) ▽	77.57(3.39) ▽	74.96(2.71) ▽	78.37(1.87) ▽	74.85(3.18) ▽
Respiratory	(10)	88.91(5.32)	87.47(5.48)	91.33(4.93)	89.59(3.22)	90.63(4.63)	90.61(4.02)	85.83(4.74)	90.93(3.47)
Average		82.43	86.12	85.80	85.87	84.93	84.67	84.90	86.36
		(▽/ /*)	(3/8/0)	(3/8/0)	(2/9/0)	(4/7/0)	(2/9/0)	(2/9/0)	(5/6/0)

1. *Best Average: Bayes Net*
2. *Bayes Net wins most.*
3. *Test Base is k-NN.*

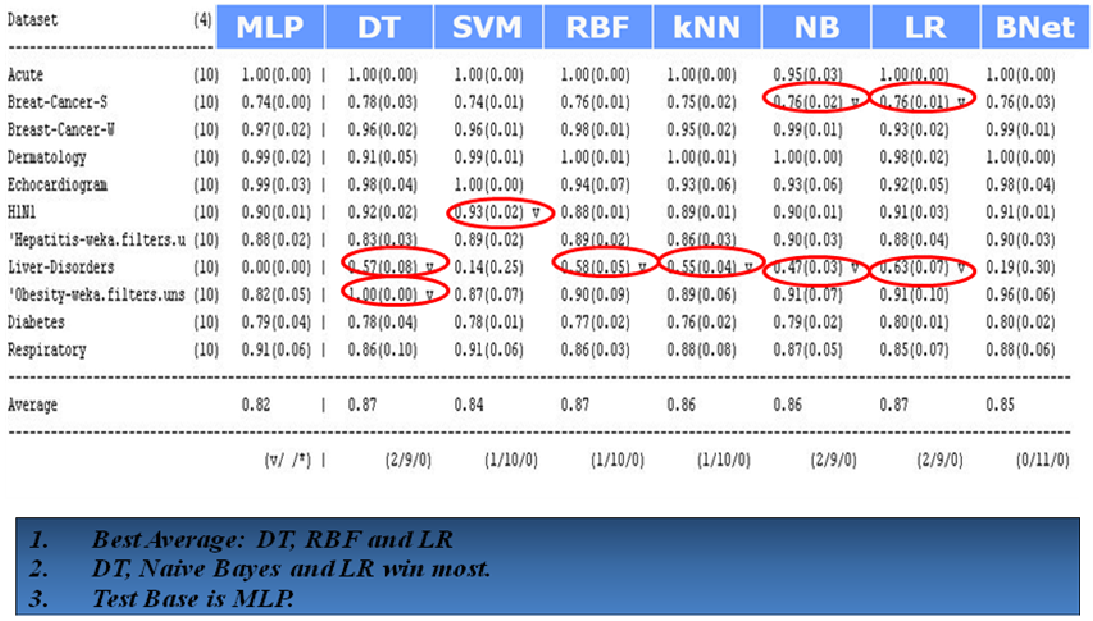
Figure 6.17. Pure Train Test Split – Percent Correct.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Breast-Cancer-S	(10)	0.78(0.05)	0.87(0.10)	1.00(0.00) ▽	0.93(0.03) ▽	1.00(0.01) ▽	0.93(0.02) ▽	0.95(0.02) ▽	0.92(0.09)
Breast-Cancer-W	(10)	0.98(0.01)	0.96(0.01) *	0.97(0.01)	0.97(0.01)	0.96(0.01)	0.97(0.01)	0.97(0.01)	0.97(0.01)
Dermatology	(10)	0.97(0.03)	0.97(0.02)	1.00(0.00)	0.99(0.01)	0.99(0.01)	1.00(0.00)	0.99(0.02)	1.00(0.00)
Echocardiogram	(10)	0.89(0.08)	0.96(0.07)	0.88(0.06)	0.98(0.03)	0.90(0.08)	0.99(0.02)	0.91(0.09)	0.98(0.04) ▽
H1N1	(10)	0.88(0.04)	0.98(0.02) ▽	0.98(0.01) ▽	0.93(0.05)	0.95(0.05)	0.92(0.04)	0.93(0.04)	0.95(0.03) ▽
'Hepatitis-weka.filters.u	(10)	0.90(0.04)	0.92(0.05)	0.92(0.03)	0.93(0.04)	0.92(0.02)	0.86(0.05)	0.87(0.05)	0.88(0.05)
Liver-Disorders	(10)	0.59(0.04)	0.45(0.14)	0.01(0.01) *	0.47(0.06)	0.00(0.00) *	0.76(0.07) ▽	0.52(0.05)	0.03(0.07) *
'Obesity-weka.filters.uns	(10)	0.89(0.12)	1.00(0.00)	0.92(0.11)	0.96(0.08)	0.94(0.09)	0.91(0.10)	0.87(0.10)	0.94(0.09)
Diabetes	(10)	0.78(0.03)	0.81(0.05)	0.90(0.03) ▽	0.85(0.03) ▽	0.89(0.08)	0.84(0.03) ▽	0.89(0.03) ▽	0.82(0.04)
Respiratory	(10)	0.89(0.08)	0.89(0.07)	0.90(0.09)	0.93(0.06)	0.89(0.07)	0.94(0.05)	0.87(0.07)	0.93(0.07)
Average		0.87	0.89	0.86	0.91	0.86	0.92	0.89	0.86
		(▽/ /*)	(1/9/1)	(3/7/1)	(2/9/0)	(1/9/1)	(3/8/0)	(2/9/0)	(2/8/1)

1. *Best Average: Naive Bayes*
2. *SVM and Naive Bayes win most. Naive Bayes is better because of no losses.*
3. *Test Base is k-NN.*

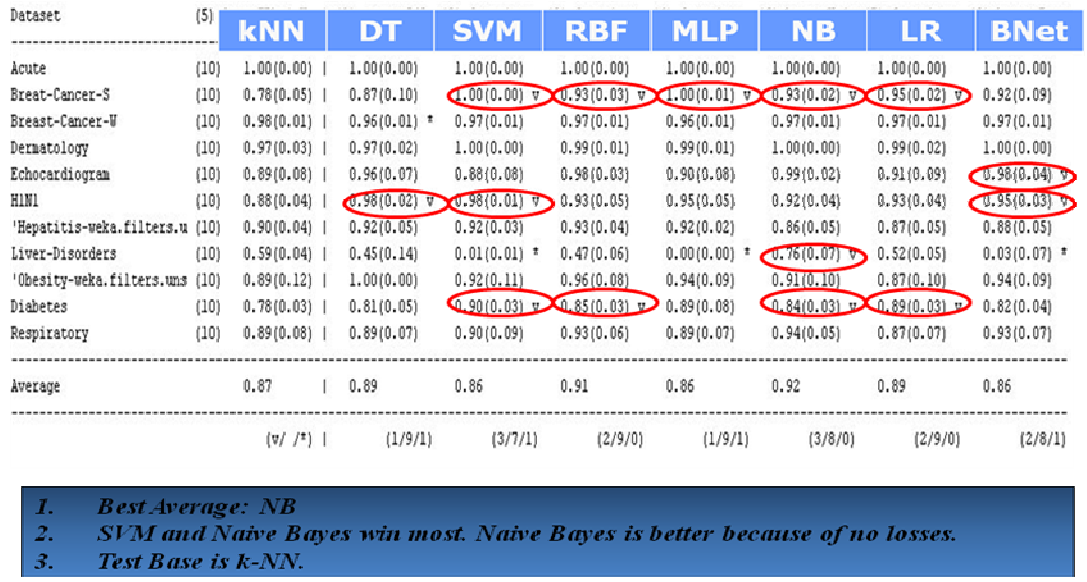
Figure 6.18. Pure Train Test Split – True Positive Rate.



Dataset	(4)	MLP	DT	SVM	RBF	kNN	NB	LR	BNet
Acute	(10)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.95(0.03)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(10)	0.74(0.00)	0.78(0.03)	0.74(0.01)	0.76(0.01)	0.75(0.02)	0.76(0.02) ▽	0.76(0.01) ▽	0.76(0.03)
Breast-Cancer-W	(10)	0.97(0.02)	0.96(0.02)	0.96(0.01)	0.98(0.01)	0.95(0.02)	0.99(0.01)	0.93(0.02)	0.99(0.01)
Dermatology	(10)	0.99(0.02)	0.91(0.05)	0.99(0.01)	1.00(0.01)	1.00(0.01)	1.00(0.00)	0.98(0.02)	1.00(0.00)
Echocardiogram	(10)	0.99(0.03)	0.98(0.04)	1.00(0.00)	0.94(0.07)	0.93(0.06)	0.93(0.06)	0.92(0.05)	0.98(0.04)
H1N1	(10)	0.90(0.01)	0.92(0.02)	0.93(0.02) ▽	0.88(0.01)	0.89(0.01)	0.90(0.01)	0.91(0.03)	0.91(0.01)
'Hepatitis-weka.filters.u	(10)	0.88(0.02)	0.83(0.03)	0.89(0.02)	0.89(0.02)	0.86(0.03)	0.90(0.03)	0.88(0.04)	0.90(0.03)
Liver-Disorders	(10)	0.00(0.00)	0.57(0.08) ▽	0.14(0.25)	0.58(0.05) ▽	0.55(0.04) ▽	0.47(0.03) ▽	0.63(0.07) ▽	0.19(0.30)
'Obesity-weka.filters.uns	(10)	0.82(0.05)	1.00(0.00) ▽	0.87(0.07)	0.90(0.09)	0.89(0.06)	0.91(0.07)	0.91(0.10)	0.96(0.06)
Diabetes	(10)	0.79(0.04)	0.78(0.04)	0.78(0.01)	0.77(0.02)	0.76(0.02)	0.79(0.02)	0.80(0.01)	0.80(0.02)
Respiratory	(10)	0.91(0.06)	0.86(0.10)	0.91(0.06)	0.86(0.03)	0.88(0.08)	0.87(0.05)	0.85(0.07)	0.88(0.06)
Average		0.82	0.87	0.84	0.87	0.86	0.86	0.87	0.85
		(▽/ /*)	(2/9/0)	(1/10/0)	(1/10/0)	(1/10/0)	(2/9/0)	(2/9/0)	(0/11/0)

1. *Best Average: DT, RBF and LR*
2. *DT, Naive Bayes and LR win most.*
3. *Test Base is MLP.*

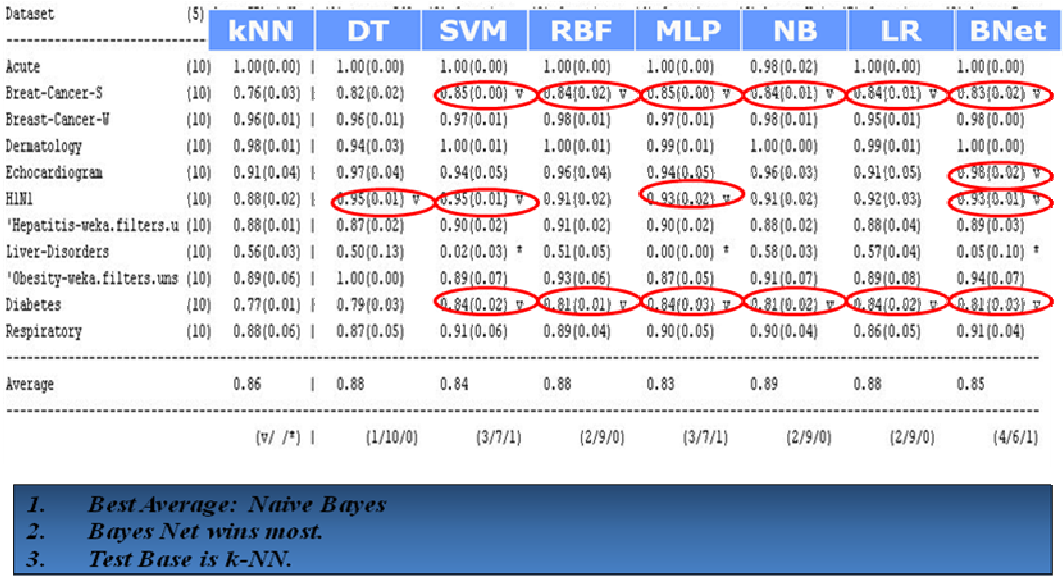
Figure 6.19. Pure Train Test Split – Precision.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(10)	0.78(0.05)	0.87(0.10)	1.00(0.00) ▽	0.93(0.03) ▽	1.00(0.01) ▽	0.93(0.02) ▽	0.95(0.02) ▽	0.92(0.09)
Breast-Cancer-W	(10)	0.98(0.01)	0.96(0.01) *	0.97(0.01)	0.97(0.01)	0.96(0.01)	0.97(0.01)	0.97(0.01)	0.97(0.01)
Dermatology	(10)	0.97(0.03)	0.97(0.02)	1.00(0.00)	0.99(0.01)	0.99(0.01)	1.00(0.00)	0.99(0.02)	1.00(0.00)
Echocardiogram	(10)	0.89(0.08)	0.96(0.07)	0.88(0.08)	0.98(0.03)	0.90(0.08)	0.99(0.02)	0.91(0.09)	0.98(0.04) ▽
H1N1	(10)	0.88(0.04)	0.98(0.02) ▽	0.98(0.01) ▽	0.93(0.05)	0.95(0.05)	0.92(0.04)	0.93(0.04)	0.95(0.03) ▽
'Hepatitis-weka.filters.u	(10)	0.90(0.04)	0.92(0.05)	0.92(0.03)	0.93(0.04)	0.92(0.02)	0.86(0.05)	0.87(0.05)	0.88(0.05)
Liver-Disorders	(10)	0.59(0.04)	0.45(0.14)	0.01(0.01) *	0.47(0.06)	0.00(0.00) *	0.76(0.07) ▽	0.52(0.05)	0.03(0.07) *
'Obesity-weka.filters.uns	(10)	0.89(0.12)	1.00(0.00)	0.92(0.11)	0.96(0.08)	0.94(0.09)	0.91(0.10)	0.87(0.10)	0.94(0.09)
Diabetes	(10)	0.78(0.03)	0.81(0.05)	0.90(0.03) ▽	0.85(0.03) ▽	0.89(0.08)	0.84(0.03) ▽	0.89(0.03) ▽	0.82(0.04)
Respiratory	(10)	0.89(0.08)	0.89(0.07)	0.90(0.09)	0.93(0.06)	0.89(0.07)	0.94(0.05)	0.87(0.07)	0.93(0.07)
Average		0.87	0.89	0.86	0.91	0.86	0.92	0.89	0.86
		(▽/ /*)	(1/9/1)	(3/7/1)	(2/9/0)	(1/9/1)	(3/8/0)	(2/9/0)	(2/8/1)

1. *Best Average: NB*
2. *SVM and Naive Bayes win most. Naive Bayes is better because of no losses.*
3. *Test Base is k-NN.*

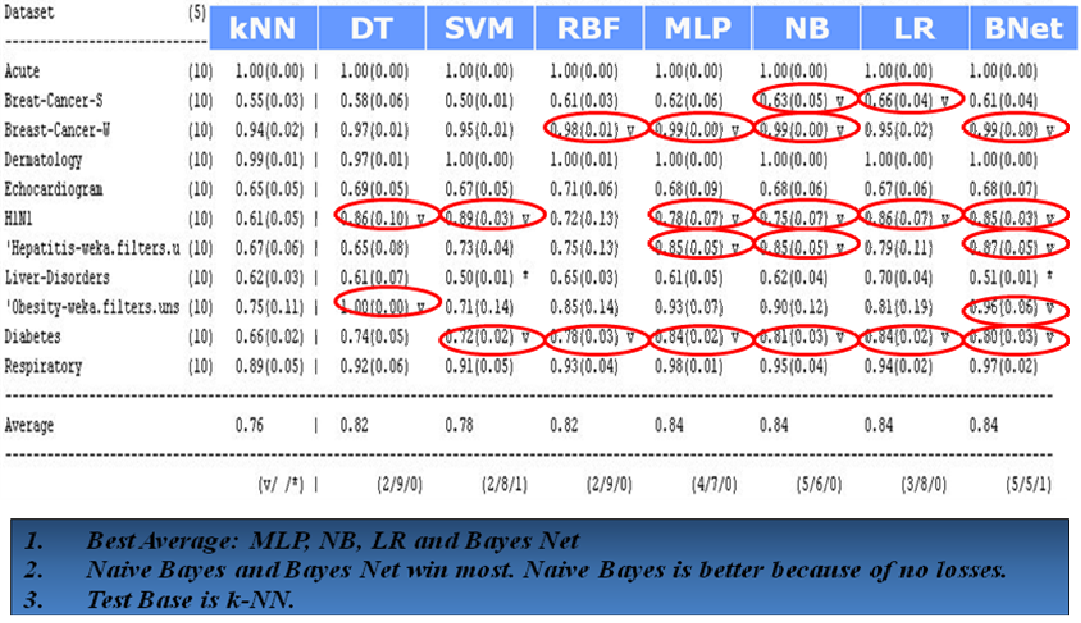
Figure 6.20. Pure Train Test Split – Recall.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	0.98(0.02)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(10)	0.76(0.03)	0.82(0.02)	0.85(0.00) v	0.84(0.02) v	0.85(0.00) v	0.84(0.01) v	0.84(0.01) v	0.83(0.02) v
Breast-Cancer-W	(10)	0.96(0.01)	0.96(0.01)	0.97(0.01)	0.98(0.01)	0.97(0.01)	0.98(0.01)	0.95(0.01)	0.98(0.00)
Dermatology	(10)	0.98(0.01)	0.94(0.03)	1.00(0.01)	1.00(0.01)	0.99(0.01)	1.00(0.00)	0.99(0.01)	1.00(0.00)
Echocardiogram	(10)	0.91(0.04)	0.97(0.04)	0.94(0.05)	0.96(0.04)	0.94(0.05)	0.96(0.03)	0.91(0.05)	0.98(0.02) v
H1N1	(10)	0.88(0.02)	0.95(0.01) v	0.95(0.01) v	0.91(0.02)	0.93(0.02) v	0.91(0.02)	0.92(0.03)	0.93(0.01) v
'Hepatitis-weka.filters.u	(10)	0.88(0.01)	0.87(0.02)	0.90(0.02)	0.91(0.02)	0.90(0.02)	0.88(0.02)	0.88(0.04)	0.89(0.03)
Liver-Disorders	(10)	0.56(0.03)	0.50(0.13)	0.02(0.03) *	0.51(0.05)	0.00(0.00) *	0.58(0.03)	0.57(0.04)	0.05(0.10) *
'Obesity-weka.filters.uns	(10)	0.89(0.06)	1.00(0.00)	0.89(0.07)	0.93(0.06)	0.87(0.05)	0.91(0.07)	0.89(0.08)	0.94(0.07)
Diabetes	(10)	0.77(0.01)	0.79(0.03)	0.84(0.02) v	0.81(0.01) v	0.84(0.03) v	0.81(0.02) v	0.84(0.02) v	0.81(0.03) v
Respiratory	(10)	0.88(0.06)	0.87(0.05)	0.91(0.06)	0.89(0.04)	0.90(0.05)	0.90(0.04)	0.86(0.05)	0.91(0.04)
Average		0.86	0.88	0.84	0.88	0.83	0.89	0.88	0.85
	(v/ /*)		(1/10/0)	(3/7/1)	(2/9/0)	(3/7/1)	(2/9/0)	(2/9/0)	(4/6/1)

1. *Best Average: Naive Bayes*
2. *Bayes Net wins most.*
3. *Test Base is k-NN.*

Figure 6.21. Pure Train Test Split – F-Measure.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(10)	0.55(0.03)	0.58(0.06)	0.50(0.01)	0.61(0.03)	0.62(0.06)	0.63(0.05) v	0.66(0.04) v	0.61(0.04)
Breast-Cancer-W	(10)	0.94(0.02)	0.97(0.01)	0.95(0.01)	0.98(0.01) v	0.99(0.00) v	0.99(0.00) v	0.95(0.02)	0.99(0.00) v
Dermatology	(10)	0.99(0.01)	0.97(0.01)	1.00(0.00)	1.00(0.01)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Echocardiogram	(10)	0.65(0.05)	0.69(0.05)	0.67(0.05)	0.71(0.06)	0.68(0.09)	0.68(0.06)	0.67(0.06)	0.68(0.07)
H1N1	(10)	0.61(0.05)	0.86(0.10) v	0.89(0.03) v	0.72(0.13)	0.78(0.07) v	0.75(0.07) v	0.86(0.07) v	0.85(0.03) v
'Hepatitis-weka.filters.u	(10)	0.67(0.06)	0.65(0.08)	0.73(0.04)	0.75(0.13)	0.85(0.05) v	0.85(0.05) v	0.79(0.11)	0.87(0.05) v
Liver-Disorders	(10)	0.62(0.03)	0.61(0.07)	0.50(0.01) *	0.65(0.03)	0.61(0.05)	0.62(0.04)	0.70(0.04)	0.51(0.01) *
'Obesity-weka.filters.uns	(10)	0.75(0.11)	1.00(0.00) v	0.71(0.14)	0.85(0.14)	0.93(0.07)	0.90(0.12)	0.81(0.19)	0.96(0.06) v
Diabetes	(10)	0.66(0.02)	0.74(0.05)	0.72(0.02) v	0.78(0.03) v	0.84(0.02) v	0.81(0.03) v	0.84(0.02) v	0.80(0.03) v
Respiratory	(10)	0.89(0.05)	0.92(0.06)	0.91(0.05)	0.93(0.04)	0.98(0.01)	0.95(0.04)	0.94(0.02)	0.97(0.02)
Average		0.76	0.82	0.78	0.82	0.84	0.84	0.84	0.84
	(v/ /*)		(2/9/0)	(2/8/1)	(2/9/0)	(4/7/0)	(5/6/0)	(3/8/0)	(5/5/1)

1. *Best Average: MLP, NB, LR and Bayes Net*
2. *Naive Bayes and Bayes Net win most. Naive Bayes is better because of no losses.*
3. *Test Base is k-NN.*

Figure 6.22. Pure Train Test Split – AUC.

Dataset	(	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	0.01	0.00 *	0.00 *	0.00 *	0.08 v	0.13 v	0.00 *	0.02
Breat-Cancer-S	(10)	0.59	0.44 *	0.51 *	0.44 *	0.44 *	0.44 *	0.43 *	0.43 *
Breast-Cancer-W	(10)	0.22	0.21	0.21	0.17	0.19	0.16	0.26	0.16
Dermatology	(10)	0.13	0.15	0.31 v	0.10	0.09 *	0.07 *	0.10	0.06 *
Echocardiogram	(10)	0.34	0.14	0.25	0.19	0.25 *	0.19	0.33	0.13 *
H1N1	(10)	0.44	0.29 *	0.28 *	0.34 *	0.31 *	0.32 *	0.32 *	0.30 *
'Hepatitis-weka.filters.u	(10)	0.44	0.42	0.40	0.35 *	0.35 *	0.39	0.42	0.39
Liver-Disorders	(10)	0.61	0.53 *	0.65	0.49 *	0.49 *	0.52 *	0.46 *	0.49 *
'Obesity-weka.filters.uns	(10)	0.38	0.00 *	0.41	0.30	0.35	0.33	0.36	0.20
Diabetes	(10)	0.55	0.45 *	0.47 *	0.42 *	0.40 *	0.42 *	0.39 *	0.42 *
Respiratory	(10)	0.32	0.31	0.29	0.31	0.26	0.29	0.35	0.27
Average		0.37	0.27	0.34	0.28	0.29	0.30	0.31	0.26
	(v/ /*)		(0/5/6)	(1/6/4)	(0/5/6)	(1/3/7)	(1/5/5)	(0/6/5)	(0/5/6)

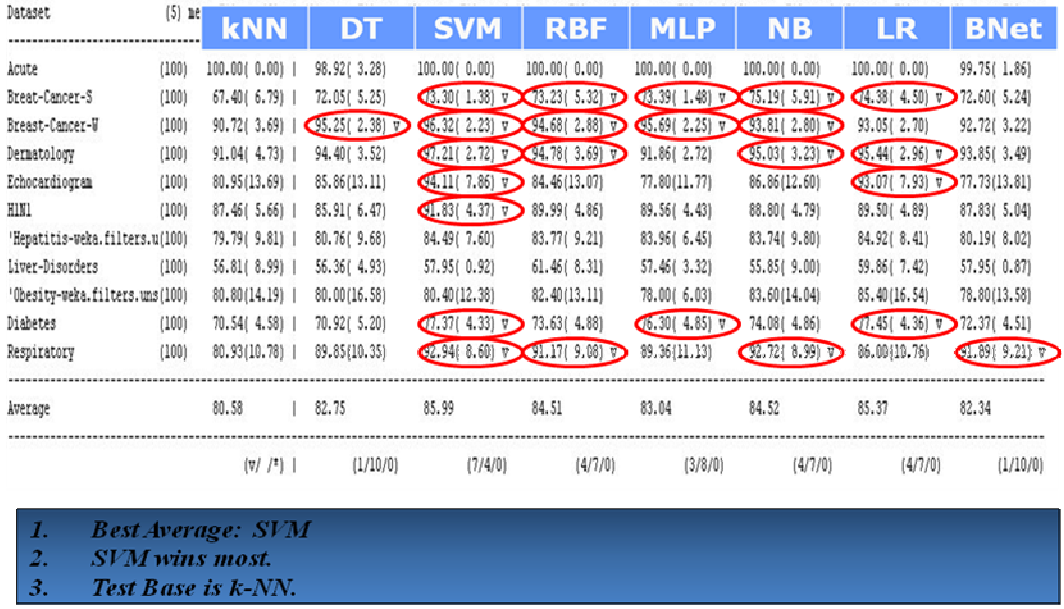
1. *Best Average: Bayes Net*
2. *MLP wins most.*
3. *Test Base is k-NN*

Figure 6.23. Pure Train Test Split – Error Rates.

	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Best Average		1		1	1	4	2	3
Wins Most		1			1	4	1	2
Test Base	6				1			

Naive Bayes is the best classifier.

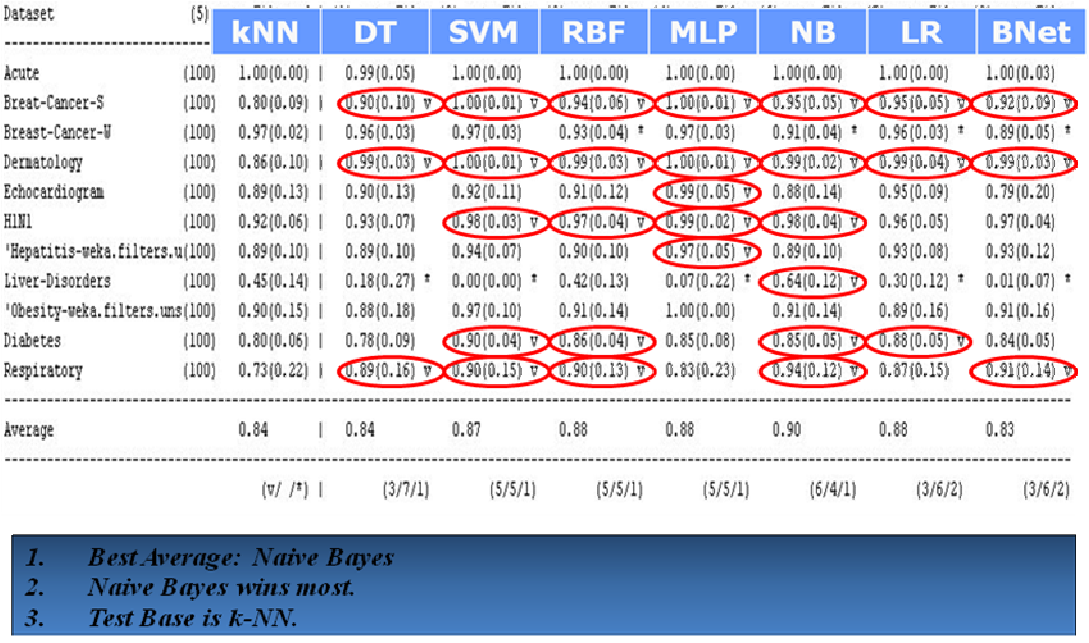
Figure 6.24. Pure Train Test Split – Summary.



Dataset	(5) nt	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	100.00(0.00) |	98.92(3.28)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)	99.75(1.86)
Breat-Cancer-S	(100)	67.40(6.79) |	72.05(5.25)	73.30(1.38) v	73.23(5.32) v	73.39(1.48) v	75.19(5.91) v	74.38(4.50) v	72.60(5.24)
Breast-Cancer-W	(100)	90.72(3.69) |	95.25(2.38) v	96.32(2.23) v	94.68(2.88) v	95.69(2.25) v	93.81(2.80) v	93.05(2.70)	92.72(3.22)
Dermatology	(100)	91.04(4.73) |	94.40(3.52)	97.21(2.72) v	94.78(3.69) v	91.86(2.72)	95.03(3.23) v	95.44(2.96) v	93.85(3.49)
Echocardiogram	(100)	80.95(13.69) |	85.86(13.11)	84.11(7.86) v	84.46(13.07)	77.80(11.77)	86.86(12.60)	83.07(7.93) v	77.73(13.81)
H1N1	(100)	87.46(5.66) |	85.91(6.47)	91.83(4.37) v	89.99(4.86)	89.56(4.43)	88.80(4.79)	89.50(4.89)	87.83(5.04)
'Hepatitis-weka.filters.u	(100)	79.79(9.81) |	80.76(9.68)	84.49(7.60)	83.77(9.21)	83.96(6.45)	83.74(9.80)	84.92(8.41)	80.19(8.02)
Liver-Disorders	(100)	56.81(8.99) |	56.36(4.93)	57.95(0.92)	61.46(8.31)	57.46(3.32)	55.85(9.00)	59.86(7.42)	57.95(0.87)
'Obesity-weka.filters.uns	(100)	80.80(14.19) |	80.00(16.58)	80.40(12.38)	82.40(13.11)	78.00(6.03)	83.60(14.04)	85.40(16.54)	78.80(13.58)
Diabetes	(100)	70.54(4.58) |	70.92(5.20)	77.37(4.33) v	73.63(4.88)	76.30(4.85) v	74.08(4.86)	77.45(4.36) v	72.37(4.51)
Respiratory	(100)	80.93(10.78) |	89.85(10.35)	92.94(8.60) v	91.17(9.08) v	89.36(11.13)	92.72(8.99) v	86.00(10.76)	91.89(9.21) v
Average		80.58 |	82.75	85.99	84.51	83.04	84.52	85.37	82.34
		(v/ /*) |	(1/10/0)	(7/4/0)	(4/7/0)	(3/8/0)	(4/7/0)	(4/7/0)	(1/10/0)

1. *Best Average: SVM*
2. *SVM wins most.*
3. *Test Base is k-NN.*

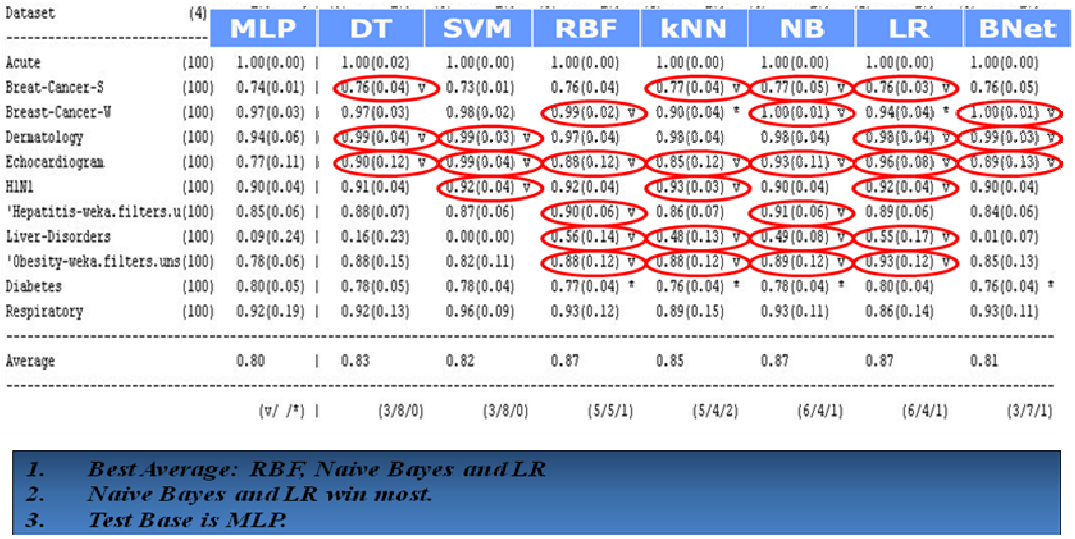
Figure 6.25. PCA 10 Fold – Percent Correct.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	1.00(0.00) |	0.99(0.05)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.03)
Breat-Cancer-S	(100)	0.80(0.09) |	0.90(0.10) v	1.00(0.01) v	0.94(0.06) v	1.00(0.01) v	0.95(0.05) v	0.95(0.05) v	0.92(0.09) v
Breast-Cancer-W	(100)	0.97(0.02) |	0.96(0.03)	0.97(0.03)	0.93(0.04) *	0.97(0.03)	0.91(0.04) *	0.96(0.03) *	0.89(0.05) *
Dermatology	(100)	0.86(0.10) |	0.99(0.03) v	1.00(0.01) v	0.99(0.03) v	1.00(0.01) v	0.99(0.02) v	0.99(0.04) v	0.99(0.03) v
Echocardiogram	(100)	0.89(0.13) |	0.90(0.13)	0.92(0.11)	0.91(0.12)	0.99(0.05) v	0.88(0.14)	0.95(0.09)	0.79(0.20)
H1N1	(100)	0.92(0.06) |	0.93(0.07)	0.98(0.03) v	0.97(0.04) v	0.99(0.02) v	0.98(0.04) v	0.96(0.05)	0.97(0.04)
'Hepatitis-weka.filters.u	(100)	0.89(0.10) |	0.89(0.10)	0.94(0.07)	0.90(0.10)	0.97(0.05) v	0.89(0.10)	0.93(0.08)	0.93(0.12)
Liver-Disorders	(100)	0.45(0.14) |	0.18(0.27) *	0.00(0.00) *	0.42(0.13)	0.07(0.22) *	0.64(0.12) v	0.30(0.12) *	0.01(0.07) *
'Obesity-weka.filters.uns	(100)	0.90(0.15) |	0.88(0.18)	0.97(0.10)	0.91(0.14)	1.00(0.00)	0.91(0.14)	0.89(0.16)	0.91(0.16)
Diabetes	(100)	0.80(0.06) |	0.78(0.09)	0.90(0.04) v	0.86(0.04) v	0.85(0.08)	0.85(0.05) v	0.88(0.05) v	0.84(0.05)
Respiratory	(100)	0.73(0.22) |	0.89(0.16) v	0.90(0.15) v	0.90(0.13) v	0.83(0.23)	0.94(0.12) v	0.87(0.15)	0.91(0.14) v
Average		0.84 |	0.84	0.87	0.88	0.88	0.90	0.88	0.83
		(v/ /*) |	(3/7/1)	(5/5/1)	(5/5/1)	(5/5/1)	(6/4/1)	(3/6/2)	(3/6/2)

1. *Best Average: Naive Bayes*
2. *Naive Bayes wins most.*
3. *Test Base is k-NN.*

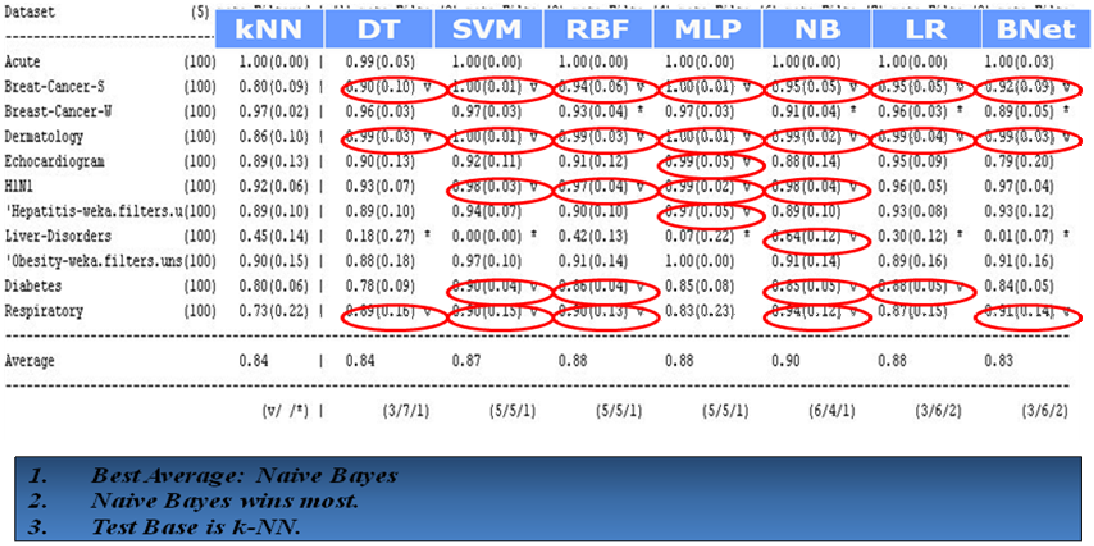
Figure 6.26. PCA 10 Fold – True Positive Rate.



Dataset	(4)	MLP	DT	SVM	RBF	kNN	NB	LR	BNet
Acute	(100)	1.00(0.00)	1.00(0.02)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(100)	0.74(0.01)	0.76(0.04) v	0.73(0.01)	0.76(0.04)	0.77(0.04) v	0.77(0.05) v	0.76(0.03) v	0.76(0.05)
Breast-Cancer-W	(100)	0.97(0.03)	0.97(0.03)	0.98(0.02)	0.99(0.02) v	0.90(0.04) *	1.00(0.01) v	0.94(0.04) *	1.00(0.01) v
Dermatology	(100)	0.94(0.06)	0.99(0.04) v	0.99(0.03) v	0.97(0.04)	0.98(0.04)	0.98(0.04)	0.98(0.04) v	0.99(0.03) v
Echocardiogram	(100)	0.77(0.11)	0.90(0.12) v	0.99(0.04) v	0.88(0.12) v	0.85(0.12) v	0.93(0.11) v	0.96(0.08) v	0.89(0.13) v
H1N1	(100)	0.90(0.04)	0.91(0.04)	0.92(0.04) v	0.92(0.04)	0.93(0.03) v	0.90(0.04)	0.92(0.04) v	0.90(0.04)
'Hepatitis-weka.filters.u	(100)	0.85(0.06)	0.88(0.07)	0.87(0.06)	0.90(0.06) v	0.86(0.07)	0.91(0.06) v	0.89(0.06)	0.84(0.06)
Liver-Disorders	(100)	0.09(0.24)	0.16(0.23)	0.00(0.00)	0.56(0.14) v	0.48(0.13) v	0.49(0.08) v	0.55(0.17) v	0.01(0.07)
'Obesity-weka.filters.uns	(100)	0.78(0.06)	0.88(0.15)	0.82(0.11)	0.88(0.12) v	0.88(0.12) v	0.89(0.12) v	0.93(0.12) v	0.85(0.13)
Diabetes	(100)	0.80(0.05)	0.78(0.05)	0.78(0.04)	0.77(0.04) *	0.76(0.04) *	0.78(0.04) *	0.80(0.04)	0.76(0.04) *
Respiratory	(100)	0.92(0.19)	0.92(0.13)	0.96(0.09)	0.93(0.12)	0.89(0.15)	0.93(0.11)	0.86(0.14)	0.93(0.11)
Average		0.80	0.83	0.82	0.87	0.85	0.87	0.87	0.81
	(v/ /*)		(3/8/0)	(3/8/0)	(5/5/1)	(5/4/2)	(6/4/1)	(6/4/1)	(3/7/1)

1. *Best Average: RBF, Naive Bayes and LR*
2. *Naive Bayes and LR win most.*
3. *Test Base is MLP.*

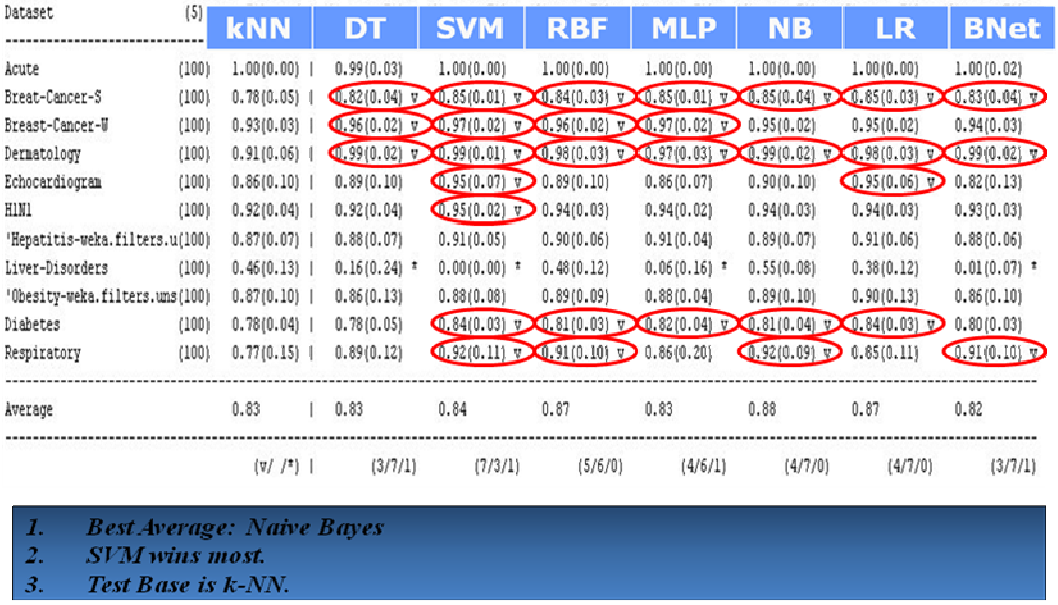
Figure 6.27. PCA 10 Fold – Precision.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	1.00(0.00)	0.99(0.05)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.03)
Breat-Cancer-S	(100)	0.80(0.09)	0.90(0.10) v	1.00(0.01) v	0.94(0.06) v	1.00(0.01) v	0.95(0.05) v	0.95(0.05) v	0.92(0.09) v
Breast-Cancer-W	(100)	0.97(0.02)	0.96(0.03)	0.97(0.03)	0.93(0.04) *	0.97(0.03)	0.91(0.04) *	0.96(0.03) *	0.89(0.05) *
Dermatology	(100)	0.86(0.10)	0.99(0.03) v	1.00(0.01) v	0.99(0.03) v	1.00(0.01) v	0.99(0.02) v	0.99(0.04) v	0.99(0.03) v
Echocardiogram	(100)	0.89(0.13)	0.90(0.13)	0.92(0.11)	0.91(0.12)	0.99(0.05) v	0.88(0.14)	0.95(0.09)	0.79(0.20)
H1N1	(100)	0.92(0.06)	0.93(0.07)	0.98(0.03) v	0.97(0.04) v	0.99(0.02) v	0.98(0.04) v	0.96(0.05)	0.97(0.04)
'Hepatitis-weka.filters.u	(100)	0.89(0.10)	0.89(0.10)	0.94(0.07)	0.90(0.10)	0.97(0.05) v	0.89(0.10)	0.93(0.08)	0.93(0.12)
Liver-Disorders	(100)	0.45(0.14)	0.18(0.27) *	0.00(0.00) *	0.42(0.13)	0.07(0.22) *	0.64(0.12) v	0.30(0.12) *	0.01(0.07) *
'Obesity-weka.filters.uns	(100)	0.90(0.15)	0.88(0.18)	0.97(0.10)	0.91(0.14)	1.00(0.00)	0.91(0.14)	0.89(0.16)	0.91(0.16)
Diabetes	(100)	0.80(0.06)	0.78(0.09)	0.90(0.04) v	0.86(0.04) v	0.85(0.08)	0.85(0.05) v	0.88(0.05) v	0.84(0.05)
Respiratory	(100)	0.73(0.22)	0.89(0.16) v	0.90(0.15) v	0.90(0.13) v	0.83(0.23)	0.94(0.12) v	0.87(0.15)	0.91(0.14) v
Average		0.84	0.84	0.87	0.88	0.88	0.90	0.88	0.83
	(v/ /*)		(3/7/1)	(5/5/1)	(5/5/1)	(5/5/1)	(6/4/1)	(3/6/2)	(3/6/2)

1. *Best Average: Naive Bayes*
2. *Naive Bayes wins most.*
3. *Test Base is k-NN.*

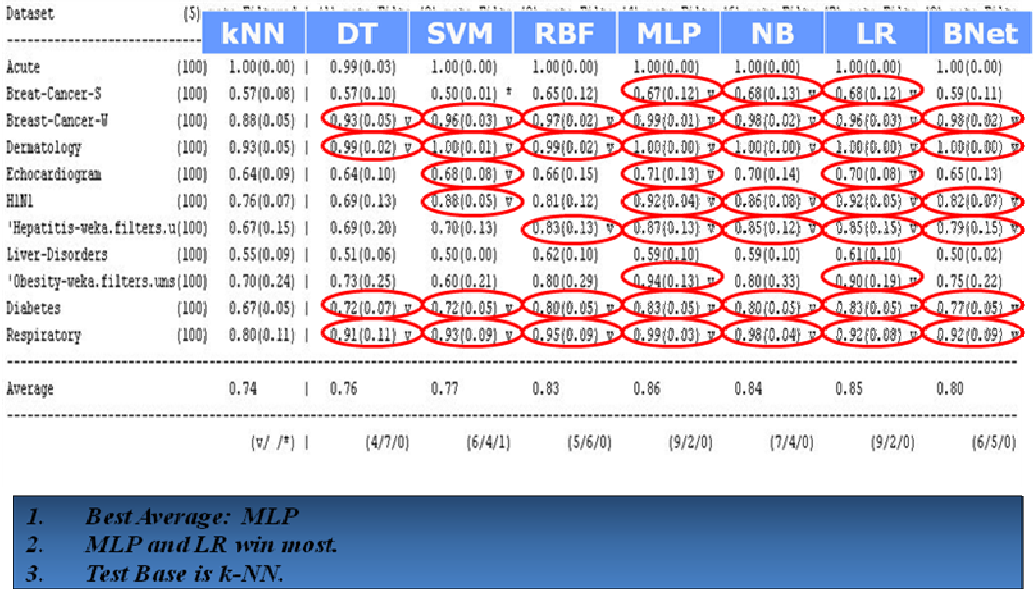
Figure 6.28. PCA 10 Fold – Recall.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	1.00(0.00)	0.99(0.03)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.02)
Breat-Cancer-S	(100)	0.78(0.05)	0.82(0.04) v	0.85(0.01) v	0.84(0.03) v	0.85(0.01) v	0.85(0.04) v	0.85(0.03) v	0.83(0.04) v
Breast-Cancer-W	(100)	0.93(0.03)	0.96(0.02) v	0.97(0.02) v	0.96(0.02) v	0.97(0.02) v	0.95(0.02)	0.95(0.02)	0.94(0.03)
Dermatology	(100)	0.91(0.06)	0.99(0.02) v	0.99(0.01) v	0.98(0.03) v	0.97(0.03) v	0.99(0.02) v	0.98(0.03) v	0.99(0.02) v
Echocardiogram	(100)	0.86(0.10)	0.89(0.10)	0.95(0.07) v	0.89(0.10)	0.86(0.07)	0.90(0.10)	0.95(0.06) v	0.82(0.13)
HlN1	(100)	0.92(0.04)	0.92(0.04)	0.95(0.02) v	0.94(0.03)	0.94(0.02)	0.94(0.03)	0.94(0.03)	0.93(0.03)
'Hepatitis-weka.filters.u	(100)	0.87(0.07)	0.88(0.07)	0.91(0.05)	0.90(0.06)	0.91(0.04)	0.89(0.07)	0.91(0.06)	0.88(0.06)
Liver-Disorders	(100)	0.46(0.13)	0.16(0.24) *	0.00(0.00) *	0.48(0.12)	0.06(0.16) *	0.55(0.08)	0.38(0.12)	0.01(0.07) *
'Obesity-weka.filters.uns	(100)	0.87(0.10)	0.86(0.13)	0.88(0.08)	0.89(0.09)	0.88(0.04)	0.89(0.10)	0.90(0.13)	0.86(0.10)
Diabetes	(100)	0.78(0.04)	0.78(0.05)	0.84(0.03) v	0.81(0.03) v	0.82(0.04) v	0.81(0.04) v	0.84(0.03) v	0.80(0.03)
Respiratory	(100)	0.77(0.15)	0.89(0.12)	0.92(0.11) v	0.91(0.10) v	0.86(0.20)	0.92(0.09) v	0.85(0.11)	0.91(0.10) v
Average		0.83	0.83	0.84	0.87	0.83	0.88	0.87	0.82
		(v/ /*)	(3/7/1)	(7/3/1)	(5/6/0)	(4/6/1)	(4/7/0)	(4/7/0)	(3/7/1)

1. *Best Average: Naive Bayes*
2. *SVM wins most.*
3. *Test Base is k-NN.*

Figure 6.29. PCA 10 Fold – F-Measure.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	1.00(0.00)	0.99(0.03)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(100)	0.57(0.08)	0.57(0.10)	0.50(0.01) *	0.65(0.12)	0.67(0.12) v	0.68(0.13) v	0.68(0.12) v	0.59(0.11)
Breast-Cancer-W	(100)	0.88(0.05)	0.93(0.05) v	0.96(0.03) v	0.97(0.02) v	0.99(0.01) v	0.98(0.02) v	0.96(0.03) v	0.98(0.02) v
Dermatology	(100)	0.93(0.05)	0.99(0.02) v	1.00(0.01) v	0.99(0.02) v	1.00(0.00) v	1.00(0.00) v	1.00(0.00) v	1.00(0.00) v
Echocardiogram	(100)	0.64(0.09)	0.64(0.10)	0.68(0.08) v	0.66(0.15)	0.71(0.13) v	0.70(0.14)	0.70(0.08) v	0.65(0.13)
HlN1	(100)	0.76(0.07)	0.69(0.13)	0.88(0.05) v	0.81(0.12)	0.92(0.04) v	0.86(0.08) v	0.92(0.05) v	0.82(0.07) v
'Hepatitis-weka.filters.u	(100)	0.67(0.15)	0.69(0.20)	0.70(0.13)	0.83(0.13) v	0.87(0.13) v	0.85(0.12) v	0.85(0.15) v	0.79(0.15) v
Liver-Disorders	(100)	0.55(0.09)	0.51(0.06)	0.50(0.00)	0.62(0.10)	0.59(0.10)	0.59(0.10)	0.61(0.10)	0.50(0.02)
'Obesity-weka.filters.uns	(100)	0.70(0.24)	0.73(0.25)	0.60(0.21)	0.80(0.29)	0.94(0.13) v	0.80(0.33)	0.90(0.19) v	0.75(0.22)
Diabetes	(100)	0.67(0.05)	0.72(0.07) v	0.72(0.05) v	0.80(0.05) v	0.83(0.05) v	0.80(0.05) v	0.83(0.05) v	0.77(0.05) v
Respiratory	(100)	0.80(0.11)	0.91(0.11) v	0.93(0.09) v	0.95(0.09) v	0.99(0.03) v	0.98(0.04) v	0.92(0.08) v	0.92(0.09) v
Average		0.74	0.76	0.77	0.83	0.86	0.84	0.85	0.80
		(v/ /*)	(4/7/0)	(6/4/1)	(5/6/0)	(9/2/0)	(7/4/0)	(9/2/0)	(6/5/0)

1. *Best Average: MLP*
2. *MLP and LR win most.*
3. *Test Base is k-NN.*

Figure 6.30. PCA 10 Fold – AUC.

Dataset		kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(100)	0.01	0.04	0.00 *	0.00 *	0.12 v	0.01	0.00 *	0.02
Breat-Cancer-S	(100)	0.57	0.44 *	0.52 *	0.43 *	0.43 *	0.43 *	0.42 *	0.44 *
Breast-Cancer-W	(100)	0.30	0.21 *	0.18 *	0.21 *	0.18 *	0.24 *	0.25	0.26
Dermatology	(100)	0.17	0.12 *	0.31 v	0.12 *	0.15	0.11 *	0.11 *	0.12 *
Echocardiogram	(100)	0.38	0.30	0.15 *	0.32	0.37	0.27 *	0.18 *	0.37
H1N1	(100)	0.34	0.35	0.27	0.28 *	0.29	0.29 *	0.29	0.32
'Hepatitis-weka.filters.u	(100)	0.43	0.40	0.38	0.34 *	0.34 *	0.35 *	0.34 *	0.35 *
Liver-Disorders	(100)	0.65	0.51 *	0.65	0.49 *	0.49 *	0.51 *	0.49 *	0.50 *
'Obesity-weka.filters.uns	(100)	0.37	0.37	0.40	0.36	0.39	0.33	0.27	0.36
Diabetes	(100)	0.54	0.46 *	0.47 *	0.42 *	0.40 *	0.42 *	0.40 *	0.43 *
Respiratory	(100)	0.40	0.24 *	0.18 *	0.22 *	0.37	0.18 *	0.31	0.22 *
Average		0.38	0.31	0.32	0.29	0.32	0.28	0.28	0.31
		(v/ /*)	(0/5/6)	(1/4/6)	(0/2/9)	(1/5/5)	(0/2/9)	(0/4/7)	(0/5/6)

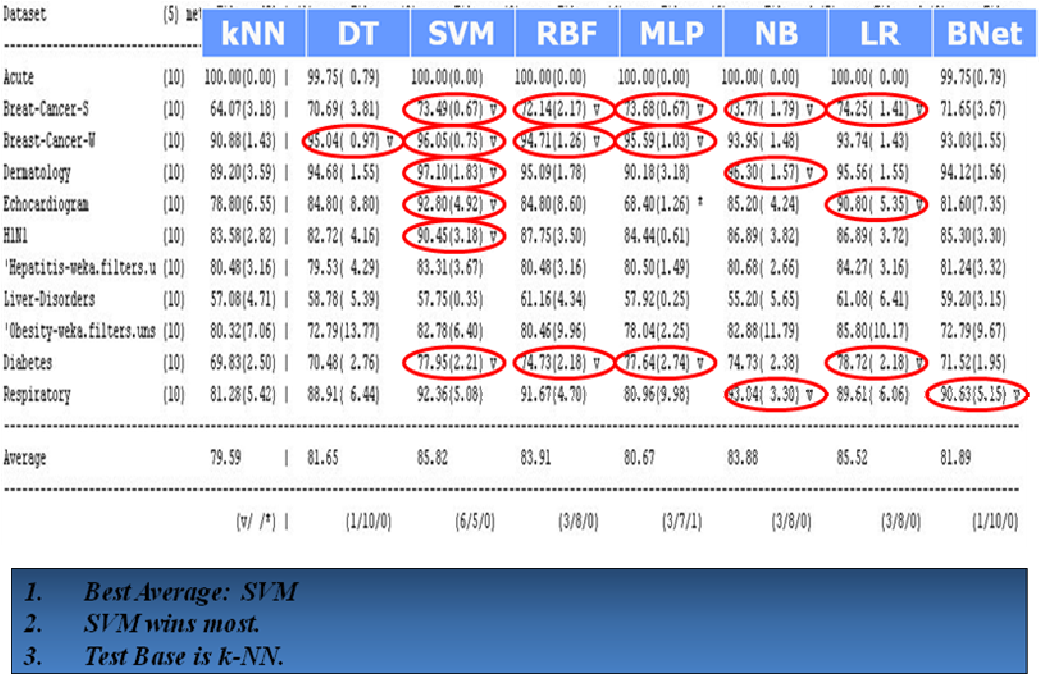
1. *Best Average: Naive Bayes and Logistic Regression*
2. *Naive Bayes and Radial Basis Function wins most.*
3. *Test Base is k-NN*

Figure 6.31. PCA 10 Fold – Error Rates.

	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Best Average			1	1	1	5	2	
Wins Most			2	1	1	4	2	
Test Base	6				1			

Naive Bayes is the best classifier.

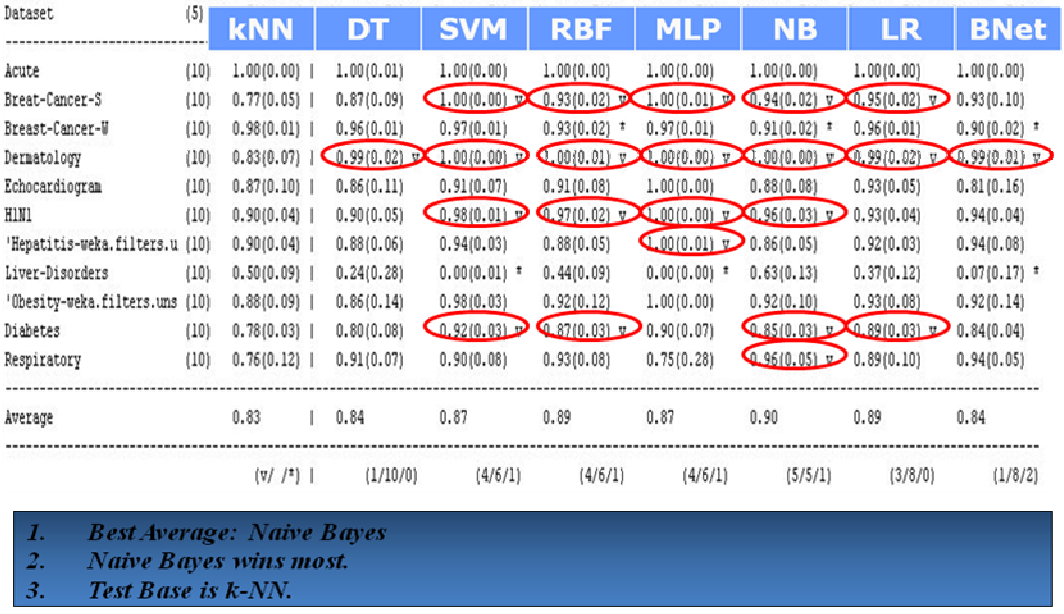
Figure 6.32. PCA 10 Fold – Summary.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	100.00(0.00)	99.75(0.79)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)	99.75(0.79)
Breat-Cancer-S	(10)	64.07(3.18)	70.69(3.81)	73.49(0.67) v	72.14(2.17) v	73.68(0.67) v	73.77(1.79) v	74.25(1.41) v	71.65(3.67)
Breast-Cancer-W	(10)	90.88(1.43)	95.04(0.97) v	96.05(0.75) v	94.71(1.26) v	95.59(1.03) v	93.95(1.48)	93.74(1.43)	93.03(1.55)
Dermatology	(10)	89.20(3.59)	94.68(1.55)	97.10(1.83) v	95.09(1.78)	90.18(3.18)	96.30(1.57) v	95.56(1.55)	94.12(1.56)
Echocardiogram	(10)	78.80(6.55)	84.80(8.80)	92.80(4.92) v	84.80(8.60)	68.40(1.26) *	85.20(4.24)	90.80(5.35) v	81.60(7.35)
H1N1	(10)	83.58(2.82)	82.72(4.16)	90.45(3.18) v	87.75(3.50)	84.44(0.61)	86.89(3.82)	86.89(3.72)	85.30(3.30)
'Hepatitis-weka.filters.u	(10)	80.48(3.16)	79.53(4.29)	83.31(3.67)	80.48(3.16)	80.50(1.49)	80.68(2.66)	84.27(3.16)	81.24(3.32)
Liver-Disorders	(10)	57.08(4.71)	58.78(5.39)	57.75(0.35)	61.16(4.34)	57.92(0.25)	55.20(5.65)	61.08(6.41)	59.20(3.15)
'Obesity-weka.filters.uns	(10)	80.32(7.06)	72.79(13.77)	82.78(6.40)	80.46(9.96)	78.04(2.25)	82.88(11.79)	85.80(10.17)	72.79(9.67)
Diabetes	(10)	69.83(2.50)	70.48(2.76)	77.95(2.21) v	74.73(2.18) v	77.64(2.74) v	74.73(2.38)	78.72(2.18) v	71.52(1.95)
Respiratory	(10)	81.28(5.42)	88.91(6.44)	92.36(5.08)	91.67(4.70)	80.96(9.98)	93.04(3.30) v	89.61(6.06)	90.63(5.15) v
Average		79.59	81.65	85.82	83.91	80.67	83.88	85.52	81.89
		(v/ /*)	(1/10/0)	(6/5/0)	(3/8/0)	(3/7/1)	(3/8/0)	(3/8/0)	(1/10/0)

1. *Best Average: SVM*
2. *SVM wins most.*
3. *Test Base is k-NN.*

Figure 6.33. PCA Train Test Split – Percent Correct.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	1.00(0.00)	1.00(0.01)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(10)	0.77(0.05)	0.87(0.09)	1.00(0.00) v	0.93(0.02) v	1.00(0.01) v	0.94(0.02) v	0.95(0.02) v	0.93(0.10)
Breast-Cancer-W	(10)	0.98(0.01)	0.96(0.01)	0.97(0.01)	0.93(0.02) *	0.97(0.01)	0.91(0.02) *	0.96(0.01)	0.90(0.02) *
Dermatology	(10)	0.83(0.07)	0.99(0.02) v	1.00(0.00) v	1.00(0.01) v	1.00(0.00) v	1.00(0.00) v	0.99(0.02) v	0.99(0.01) v
Echocardiogram	(10)	0.87(0.10)	0.86(0.11)	0.91(0.07)	0.91(0.08)	1.00(0.00)	0.88(0.08)	0.93(0.05)	0.81(0.16)
H1N1	(10)	0.90(0.04)	0.90(0.05)	0.98(0.01) v	0.97(0.02) v	1.00(0.00) v	0.96(0.03) v	0.93(0.04)	0.94(0.04)
'Hepatitis-weka.filters.u	(10)	0.90(0.04)	0.88(0.06)	0.94(0.03)	0.88(0.05)	1.00(0.01) v	0.86(0.05)	0.92(0.03)	0.94(0.08)
Liver-Disorders	(10)	0.50(0.09)	0.24(0.28)	0.00(0.01) *	0.44(0.09)	0.00(0.00) *	0.63(0.13)	0.37(0.12)	0.07(0.17) *
'Obesity-weka.filters.uns	(10)	0.88(0.09)	0.86(0.14)	0.98(0.03)	0.92(0.12)	1.00(0.00)	0.92(0.10)	0.93(0.08)	0.92(0.14)
Diabetes	(10)	0.78(0.03)	0.80(0.08)	0.92(0.03) v	0.87(0.03) v	0.90(0.07)	0.85(0.03) v	0.89(0.03) v	0.84(0.04)
Respiratory	(10)	0.76(0.12)	0.91(0.07)	0.90(0.08)	0.93(0.08)	0.75(0.28)	0.96(0.05) v	0.89(0.10)	0.94(0.05)
Average		0.83	0.84	0.87	0.89	0.87	0.90	0.89	0.84
		(v/ /*)	(1/10/0)	(4/6/1)	(4/6/1)	(4/6/1)	(5/5/1)	(3/8/0)	(1/8/2)

1. *Best Average: Naive Bayes*
2. *Naive Bayes wins most.*
3. *Test Base is k-NN.*

Figure 6.34. PCA Train Test Split – True Positive Rate.

Dataset	(4)	MLP	DT	SVM	RBF	kNN	NB	LR	BNet
Acute	(10)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.01)
Breat-Cancer-S	(10)	0.74(0.01)	0.76(0.03)	0.74(0.01)	0.75(0.01)	0.75(0.02)	0.76(0.02) v	0.76(0.01) v	0.75(0.02)
Breast-Cancer-W	(10)	0.96(0.02)	0.96(0.01)	0.97(0.01)	0.99(0.01)	0.89(0.02) *	1.00(0.00) v	0.95(0.01)	1.00(0.00) v
Dermatology	(10)	0.95(0.03)	0.99(0.01)	0.99(0.01)	0.97(0.02)	0.97(0.02)	0.98(0.02)	0.98(0.01)	0.99(0.01)
Echocardiogram	(10)	0.68(0.01)	0.91(0.07) v	0.98(0.03) v	0.88(0.09) v	0.83(0.07) v	0.92(0.08) v	0.94(0.05) v	0.93(0.09) v
H1N1	(10)	0.84(0.01)	0.89(0.02) v	0.91(0.03) v	0.90(0.03) v	0.91(0.03) v	0.89(0.02) v	0.91(0.03) v	0.89(0.03) v
'Hepatitis-weka.filters.u	(10)	0.80(0.01)	0.87(0.05)	0.86(0.03)	0.88(0.04)	0.86(0.02) v	0.89(0.03) v	0.89(0.03) v	0.85(0.05)
Liver-Disorders	(10)	0.00(0.00)	0.27(0.29)	0.05(0.16)	0.54(0.06) v	0.49(0.05) v	0.48(0.05) v	0.56(0.11) v	0.13(0.27)
'Obesity-weka.filters.uns	(10)	0.78(0.02)	0.81(0.09)	0.83(0.05)	0.85(0.06)	0.88(0.07)	0.87(0.08)	0.89(0.07)	0.77(0.04)
Diabetes	(10)	0.79(0.04)	0.76(0.05)	0.78(0.02)	0.77(0.01)	0.76(0.02)	0.78(0.02)	0.80(0.01)	0.75(0.02)
Respiratory	(10)	0.88(0.13)	0.86(0.07)	0.94(0.05)	0.90(0.07)	0.83(0.06)	0.90(0.05)	0.89(0.05)	0.87(0.07)
Average		0.77	0.83	0.82	0.86	0.83	0.86	0.87	0.81
	(v/ /*)		(2/9/0)	(2/9/0)	(3/8/0)	(4/6/1)	(6/5/0)	(5/6/0)	(3/8/0)

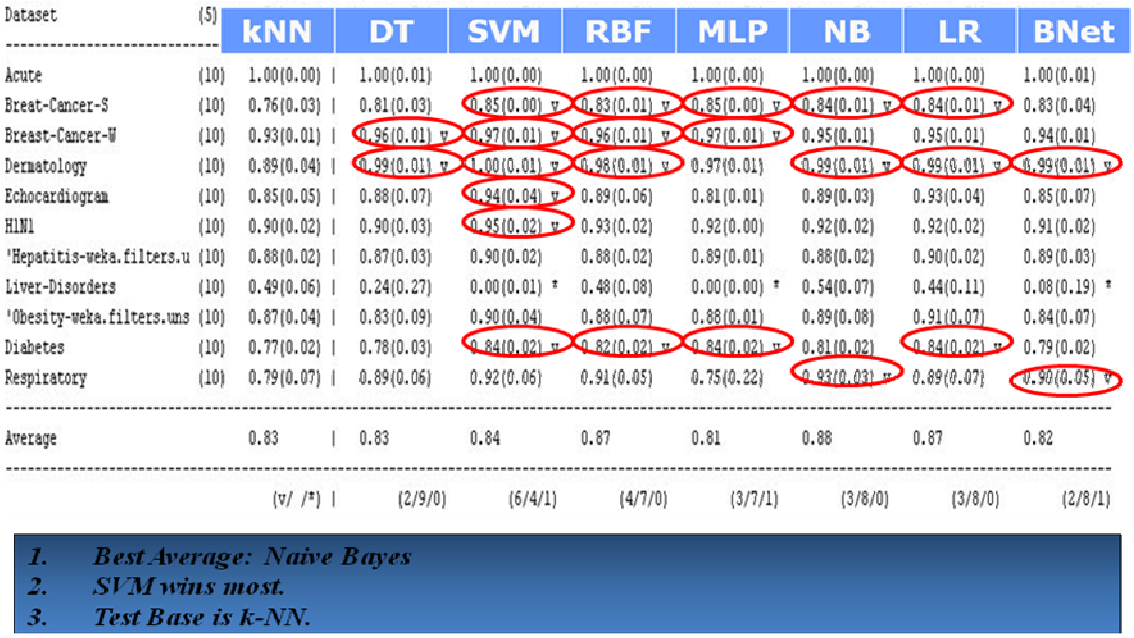
1. *Best Average: Logistic Regression*
2. *Naive Bayes wins most.*
3. *Test Base is MLP.*

Figure 6.35. PCA Train Test Split – Precision.

Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	1.00(0.00)	1.00(0.01)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(10)	0.77(0.05)	0.87(0.09)	1.00(0.00) v	0.93(0.02) v	1.00(0.01) v	0.94(0.02) v	0.95(0.02) v	0.93(0.10)
Breast-Cancer-W	(10)	0.98(0.01)	0.96(0.01)	0.97(0.01)	0.93(0.02) *	0.97(0.01)	0.91(0.02) *	0.96(0.01)	0.90(0.02) *
Dermatology	(10)	0.83(0.07)	0.99(0.02) v	1.00(0.00) v	1.00(0.01) v	1.00(0.00) v	1.00(0.00) v	0.99(0.02) v	0.99(0.01) v
Echocardiogram	(10)	0.87(0.10)	0.86(0.11)	0.91(0.07)	0.91(0.08)	1.00(0.00)	0.88(0.08)	0.93(0.05)	0.81(0.16)
H1N1	(10)	0.90(0.04)	0.90(0.05)	0.98(0.01) v	0.97(0.02) v	1.00(0.00) v	0.96(0.03) v	0.93(0.04)	0.94(0.04)
'Hepatitis-weka.filters.u	(10)	0.90(0.04)	0.88(0.06)	0.94(0.03)	0.88(0.05)	1.00(0.01) v	0.86(0.05)	0.92(0.03)	0.94(0.08)
Liver-Disorders	(10)	0.50(0.09)	0.24(0.28)	0.00(0.01) *	0.44(0.09)	0.00(0.00) *	0.63(0.13)	0.37(0.12)	0.07(0.17) *
'Obesity-weka.filters.uns	(10)	0.88(0.09)	0.86(0.14)	0.98(0.03)	0.92(0.12)	1.00(0.00)	0.92(0.10)	0.93(0.08)	0.92(0.14)
Diabetes	(10)	0.78(0.03)	0.80(0.08)	0.92(0.03) v	0.87(0.03) v	0.90(0.07)	0.85(0.03) v	0.89(0.03) v	0.84(0.04)
Respiratory	(10)	0.76(0.12)	0.91(0.07)	0.90(0.08)	0.93(0.08)	0.75(0.28)	0.96(0.05) v	0.89(0.10)	0.94(0.05)
Average		0.83	0.84	0.87	0.89	0.87	0.90	0.89	0.84
	(v/ /*)		(1/10/0)	(4/6/1)	(4/6/1)	(4/6/1)	(5/5/1)	(3/8/0)	(1/8/2)

1. *Best Average: Naive Bayes*
2. *Naive Bayes wins most.*
3. *Test Base is k-NN.*

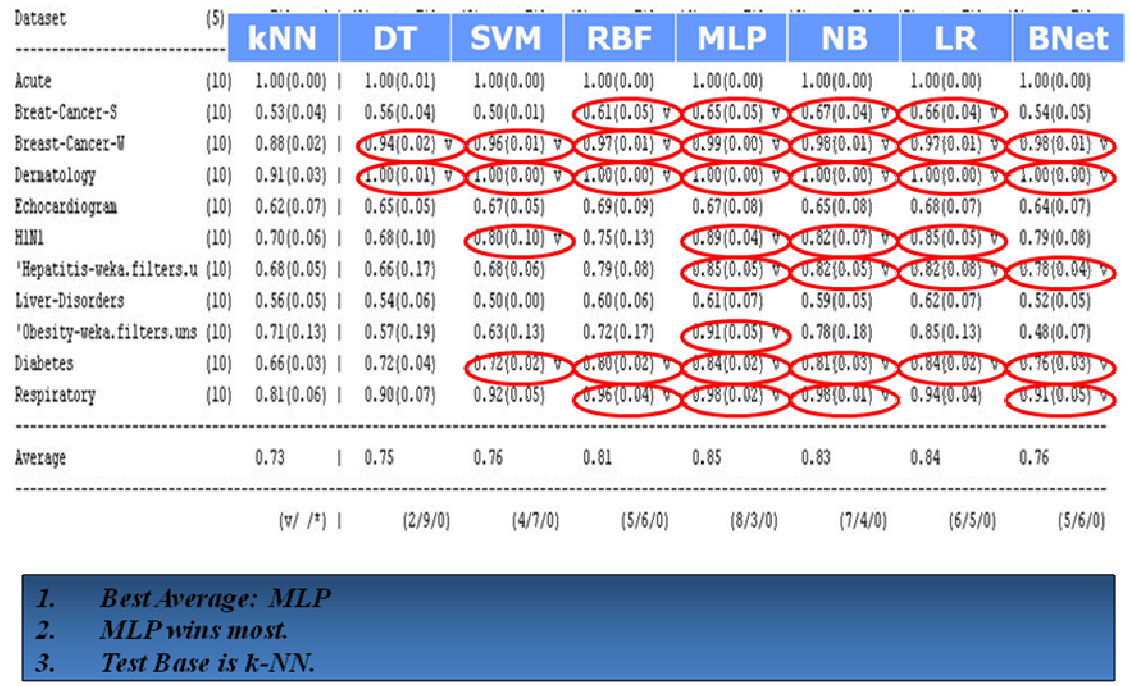
Figure 6.36. PCA Train Test Split – Recall.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	1.00(0.00)	1.00(0.01)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.01)
Breat-Cancer-S	(10)	0.76(0.03)	0.81(0.03)	0.85(0.00) v	0.83(0.01) v	0.85(0.00) v	0.84(0.01) v	0.84(0.01) v	0.83(0.04)
Breast-Cancer-W	(10)	0.93(0.01)	0.96(0.01) v	0.97(0.01) v	0.96(0.01) v	0.97(0.01) v	0.95(0.01)	0.95(0.01)	0.94(0.01)
Dermatology	(10)	0.89(0.04)	0.99(0.01) v	1.00(0.01) v	0.98(0.01) v	0.97(0.01)	0.99(0.01) v	0.99(0.01) v	0.99(0.01) v
Echocardiogram	(10)	0.85(0.05)	0.88(0.07)	0.94(0.04) v	0.89(0.06)	0.81(0.01)	0.89(0.03)	0.93(0.04)	0.85(0.07)
H1N1	(10)	0.90(0.02)	0.90(0.03)	0.95(0.02) v	0.93(0.02)	0.92(0.00)	0.92(0.02)	0.92(0.02)	0.91(0.02)
'Hepatitis-weka.filters.u	(10)	0.88(0.02)	0.87(0.03)	0.90(0.02)	0.88(0.02)	0.89(0.01)	0.88(0.02)	0.90(0.02)	0.89(0.03)
Liver-Disorders	(10)	0.49(0.06)	0.24(0.27)	0.00(0.01) *	0.48(0.08)	0.00(0.00) *	0.54(0.07)	0.44(0.11)	0.08(0.19) *
'Obesity-weka.filters.uns	(10)	0.87(0.04)	0.83(0.09)	0.90(0.04)	0.88(0.07)	0.88(0.01)	0.89(0.08)	0.91(0.07)	0.84(0.07)
Diabetes	(10)	0.77(0.02)	0.78(0.03)	0.84(0.02) v	0.82(0.02) v	0.84(0.02) v	0.81(0.02)	0.84(0.02) v	0.79(0.02)
Respiratory	(10)	0.79(0.07)	0.89(0.06)	0.92(0.06)	0.91(0.05)	0.75(0.22)	0.93(0.03) v	0.89(0.07)	0.90(0.05) v
Average		0.83	0.83	0.84	0.87	0.81	0.88	0.87	0.82
		(v/ /*)	(2/9/0)	(6/4/1)	(4/7/0)	(3/7/1)	(3/8/0)	(3/8/0)	(2/8/1)

1. *Best Average: Naive Bayes*
2. *SVM wins most.*
3. *Test Base is k-NN.*

Figure 6.37. PCA Train Test Split – F-Measure.



Dataset	(5)	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	1.00(0.00)	1.00(0.01)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
Breat-Cancer-S	(10)	0.53(0.04)	0.56(0.04)	0.50(0.01)	0.61(0.05) v	0.65(0.05) v	0.67(0.04) v	0.66(0.04) v	0.54(0.05)
Breast-Cancer-W	(10)	0.88(0.02)	0.94(0.02) v	0.96(0.01) v	0.97(0.01) v	0.99(0.00) v	0.96(0.01) v	0.97(0.01) v	0.98(0.01) v
Dermatology	(10)	0.91(0.03)	1.00(0.01) v	1.00(0.00) v	1.00(0.00) v	1.00(0.00) v	1.00(0.00) v	1.00(0.00) v	1.00(0.00) v
Echocardiogram	(10)	0.62(0.07)	0.65(0.05)	0.67(0.05)	0.69(0.09)	0.67(0.08)	0.65(0.08)	0.68(0.07)	0.64(0.07)
H1N1	(10)	0.70(0.06)	0.68(0.10)	0.80(0.10) v	0.75(0.13)	0.89(0.04) v	0.82(0.07) v	0.85(0.05) v	0.79(0.08)
'Hepatitis-weka.filters.u	(10)	0.68(0.05)	0.66(0.17)	0.68(0.06)	0.79(0.08)	0.85(0.05) v	0.82(0.05) v	0.82(0.08) v	0.78(0.04) v
Liver-Disorders	(10)	0.56(0.05)	0.54(0.06)	0.50(0.00)	0.60(0.06)	0.61(0.07)	0.59(0.05)	0.62(0.07)	0.52(0.05)
'Obesity-weka.filters.uns	(10)	0.71(0.13)	0.57(0.19)	0.63(0.13)	0.72(0.17)	0.91(0.05) v	0.78(0.18)	0.85(0.13)	0.48(0.07)
Diabetes	(10)	0.66(0.03)	0.72(0.04)	0.72(0.02) v	0.80(0.02) v	0.84(0.02) v	0.81(0.03) v	0.84(0.02) v	0.76(0.03) v
Respiratory	(10)	0.81(0.06)	0.90(0.07)	0.92(0.05)	0.96(0.04) v	0.98(0.02) v	0.98(0.01) v	0.94(0.04)	0.91(0.05) v
Average		0.73	0.75	0.76	0.81	0.85	0.83	0.84	0.76
		(v/ /*)	(2/9/0)	(4/7/0)	(5/6/0)	(8/3/0)	(7/4/0)	(6/5/0)	(5/6/0)

1. *Best Average: MLP*
2. *MLP wins most.*
3. *Test Base is k-NN.*

Figure 6.38. PCA Train Test Split – AUC.

Dataset	(	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Acute	(10)	0.01	0.02	0.00 *	0.00 *	0.16 v	0.03	0.00 *	0.02
Breat-Cancer-S	(10)	0.60	0.45 *	0.51 *	0.44 *	0.44 *	0.44 *	0.43 *	0.44 *
Breast-Cancer-W	(10)	0.30	0.22 *	0.20 *	0.21 *	0.18 *	0.24	0.24	0.25
Dermatology	(10)	0.19	0.13	0.31 v	0.12	0.19	0.10 *	0.12	0.12 *
Echocardiogram	(10)	0.45	0.35	0.24 *	0.37	0.42	0.33 *	0.28 *	0.36
HlN1	(10)	0.40	0.40	0.30 *	0.33	0.33 *	0.32 *	0.33 *	0.34 *
'Hepatitis-weka.filters.u	(10)	0.44	0.43	0.41	0.39	0.36 *	0.39	0.38	0.36 *
Liver-Disorders	(10)	0.65	0.51 *	0.65	0.49 *	0.49 *	0.51 *	0.48 *	0.49 *
'Obesity-weka.filters.uns	(10)	0.43	0.49	0.41	0.43	0.40	0.37	0.32	0.46
Diabetes	(10)	0.55	0.46 *	0.47 *	0.42 *	0.40 *	0.42 *	0.39 *	0.44 *
Respiratory	(10)	0.42	0.31	0.25	0.26	0.41	0.23 *	0.30	0.28 *
Average		0.40	0.34	0.34	0.31	0.34	0.31	0.30	0.32
	(v/ /*)		(0/7/4)	(1/4/6)	(0/6/5)	(1/4/6)	(0/4/7)	(0/5/6)	(0/4/7)

1. *Best Average: Logistic Regression*
2. *Naive Bayes and Bayes Net wins most.*
3. *Test Base is k-NN*

Figure 6.39. PCA Train Test Split – Error Rates.

	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Best Average			1		1	3	2	
Wins Most			2		1	4		1
Test Base	6				1			

Naive Bayes is the best classifier.

Figure 6.40. PCA Train Test Split – Summary.

As it's seen that the highest score of the performance for the medical database and machine learning algorithm is achieved for the Naïve Bayes classifier. Following Naive Bayes, following classifiers are Bayes Net, Logistic Regression and Support Vector Machine algorithm.

6.3. Comparison of the Results with the Previous Researches

There are several similar researchers done with the same datasets and we would like to compare this research with the previous researchers within the field.

We see that our research has the same Acute – AUC values with all used classifiers in PCA 10 Fold, Pure Train Test Split, PCA 10 Fold and PCA Train Test Split. Our research has received 1.00 in all classifiers, which is better than One Class SVMs and LOF shown above.

For Breast Cancer Wisconsin, Noto *et al.*'s researches show that the Best Value is 0.95. In our research,

- Pure 10 Fold: DT, RBF, MLP, NB and Bayesian Net
- Pure Train Test Split: DT, RBF, MLP, NB and Bayesian Net
- PCA 10 Fold: SVM, RBF, MLP, NB, LR and Bayesian Net
- PCA Train Test Split: SVM, RBF, MLP, NB, LR and Bayesian Net

Has better values comparing Noto *et al.*'s research.

For Dermatology Data Set, our research again has the received 1.00 in several classifiers as we see Noto *et al.*'s research. SVM, RBF, MLP, NB, LR and Bayesian Net are comparable classifiers, which they have received 1.00 as output.

For Echocardiogram Data Set, Noto *et al.*'s researches show that the Best AUC is 0.73. In our research, we have similar values and even better ones. For comparison manner, the better AUC values show their difference with standard deviations. That's why we don't want to mention these differences in the thesis. As an example, Multilayer Perceptron has 0.71 (0.12±) AUC with Pure 10 Fold.

For Hepatitis Data Set, Noto *et al.*'s researches show that the Best AUC Value is 0.87. In our research,

- Pure 10 Fold: RBF and Bayesian Net
- Pure Train Test Split: Bayesian Net
- PCA 10 Fold: MLP - 0.85 (0.13±)

- PCA Train Test Split: MLP - 0.85 (0.05±)

Has better values comparing Noto *et al.*'s research. Bayesian Net shows that it's significantly better than Noto *et al.* in without PCA and MLP is better with small fractions.

For Pima Indians Diabeters, Noto *et al.*'s researches show that the Best AUC Value is 0.73. In our research,

- Pure 10 Fold: DT, RBF, MLP, NB, LR, Bayes Net
- Pure Train Test Split: RBF, MLP, NB, LR, Bayes Net
- PCA 10 Fold: RBF, MLP, NB, LR, Bayes Net
- PCA Train Test Split: RBF, MLP, NB, LR, Bayes Net

Has better values comparing Noto *et al.*'s research. MLP and LR show that it's significantly better than Noto *et al.* in all experiments.

We see that (Dogan, 2010) had a study for the evaluation of classification algorithms. We see Acute and Breast Cancer are similar datasets that we used in our experiment. Breast Cancer has 97.00% Accuracy with Logistics Regression Algorithm and Train Test Split in (Dogan, 2010) experiment. When we look at our results:

- Pure 10 Fold: k-NN, DT, SVM, RBF, MLP, NB and Bayes Net has better values. Naive Bayes is the best classifier: 97.47% (± 1.69)
- Pure Train Test Split: RBF, NB and Bayes Net have better Accuracy values. Naïve Bayes is the best classifier: 97.23 % (± 0.66)
- PCA 10 Fold: DT, SVM, RBF and MLP have better Accuracy values. SVM is the best classifier: 96.32 % (± 2.23)

It's obvious that Naïve Bayes performs better than Logistic Regression and shows better Accuracy results in our experiment.

We see that (Salim, 2005) had a study for the evaluation of classification algorithms. We see that they use Wisconsin Breast Cancer and Hepatitis Data Sets from UCI. Breast Cancer has 96.94% Accuracy with SVM Algorithm. Hepatitis has 87.23% Accuracy with SVM, MLP and RBF.

When we look at our results for Breast Cancer:

- Pure 10 Fold: RBF, NB and Bayes Net has better values. Naive Bayes is the best classifier: 97.47% (± 1.69)
- Pure Train Test Split: NB and Bayes Net have better Accuracy values. Naïve Bayes is the best classifier: 97.23 % (± 0.66)

It's obvious that Naïve Bayes performs better than SVM and shows better Accuracy results in our experiment.

When we look at our results for Hepatitis, our research is comparable with (Salim, 2005) results. They had 87.23% Accuracy, and our research is comparable with the help of standard deviations within the Classifiers.

In this way, k-NN, DT, SVM, RBF, MLP, NB, LR and Bayes Net have better results.

	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
'Hepatitis-weka.filters.u(100)	81.40(8.55) |	79.22(9.57)	85.64(8.87)	85.11(8.29)	84.65(9.32)	83.81(9.70)	83.89(8.12)	84.18(10.29)

Figure 6.41. Hepatitis Data Set Accuracy Results for Pure 10 Fold Experiment.

7. CONCLUSION

Medical Industry is combining IT techniques in their daily usage and researchers a lot. We showed in our previous chapters that there are 606 papers from more than 1,500 papers in Bioinformatics Domain, which are interested in Machine Learning. From this interest, these knowledge and such medical databases may contain valuable information contained in nontrivial dependencies among symptoms and diagnoses. With the use of medical systems the process of uncovering such relationships in historical data is much easier to conduct. This knowledge can be used in diagnosis of future cases.

The main goal of the research was to identify the most common data mining algorithms, implemented in modern Machine Learning Algorithms, and evaluate their performance on several medical datasets. Eight algorithms were chosen in the Thesis: are J48 – Decision Tree, Support Vector Machine, Radial Basis Function, Multilayer Perceptron, k-Nearest Neighbors, Naïve Bayes, Bayes Net and Logistic Regression. For the evaluation there are eleven data sets were used: Acute Inflammations Data Set, Breast Cancer - Survival from Surgery Data Set, Breast Cancer Wisconsin (Original) Data Set, Dermatology Data Set, Echocardiogram Data Set, H1N1 Data Set, Hepatitis Data Set, Liver Disorders Data Set, Obesity Data Set, Pima Indians Diabetes Data Set and Respiratory Data Set.

Several performance metrics were utilized: Percent Correct, True Positive Rate, False Positive Rate, Precision, Recall, F-Measure, AUC and Error Rates. The underlying reason for such a research was the fact that to look at different datasets, which some of them have not been used in Machine Learning Study, combining different Machine Learning Techniques as much as possible such as (i.e. Cross Validation, Train Test Split and Feature Extraction: PCA), use several classifiers and present a Benchmarking Study.

Below are the answers to the research questions stated at the beginning of the research.

Research Question 1 (RQ1):

Does implementing the same classification algorithms on multiple datasets and with different implementation techniques result in significantly different performance indicators?

Research Answer 1 (RA1):

We have used 11 datasets with 8 classification techniques within this experiment. On top of this 10-Fold Cross Validation and Train Test Split are also applied within the methodology of the experiment. Also, we have applied Feature Extraction with PCA within the experiment. So we can easily find four different implementation techniques with the same classification algorithms on multiple datasets.

If we look at Accuracy for all these tests and make a comparison for the answer of this question, we see:

Acute	(100)	100.00(0.00)	100.00(0.00)	100.00(0.00)	100.00(0.00)
Breat-Cancer-S	(100)	67.24(6.78)	64.93(2.88)	67.40(6.79)	64.07(3.18)
Breast-Cancer-W	(100)	95.12(2.56)	95.13(1.42)	90.72(3.69)	90.88(1.43)
Dermatology	(100)	94.64(3.90)	94.43(1.56)	91.04(4.73)	89.20(3.59)
Echocardiogram	(100)	88.21(11.27)	87.60(5.15)	80.95(13.69)	78.80(6.55)
H1N1	(100)	84.91(6.12)	80.65(3.03)	87.46(5.66)	83.58(2.82)
'Hepatitis-weka.filters.u	(100)	81.40(8.55)	80.49(1.86)	79.79(9.81)	80.48(3.16)
Liver-Disorders	(100)	62.22(8.18)	61.84(3.55)	56.81(8.99)	57.08(4.71)
'Obesity-weka.filters.ums	(100)	78.60(15.89)	82.74(8.87)	80.80(14.19)	80.32(7.06)
Diabetes	(100)	70.62(4.67)	70.02(1.60)	70.54(4.58)	69.83(2.50)
Respiratory	(100)	91.08(9.39)	88.91(5.32)	80.93(10.78)	81.28(5.42)
Average		83.10	82.43	80.58	79.59

Figure 7.1. Four Different Implementation Techniques in k-NN classifier.

To remember Acute Data Set, the decision criteria is perfectly aligned inside the data set and in all our Machine Learning Algorithms, all classifiers are easily found the perfect match because of Data Set.

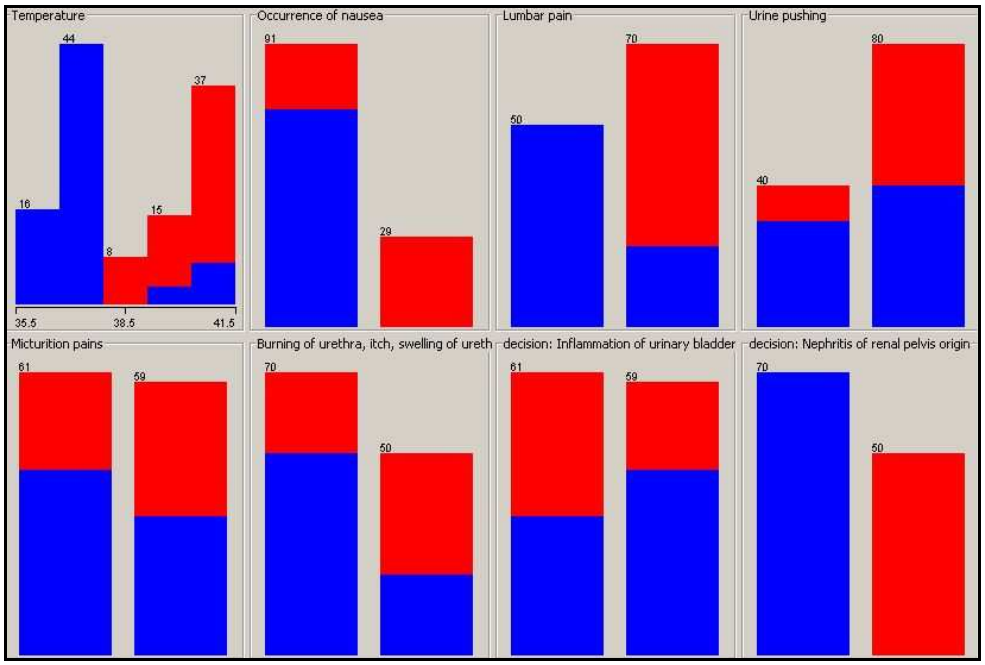


Figure 7.2. Acute Inflammations Attributes' Instances distribution.

Other than Acute Data Set, all remaining 10 data sets has different Accuracy results in k-NN Classification Algorithm. That proves us the same classification algorithms on multiple datasets and with different implementation techniques result in significantly different performance indicators.

Research Question 2 (RQ2):

How is the evaluation of the performance in terms of Percent Correct, True Positive Rate, False Positive Rate, Precision, Recall, F-Measure, AUC and Error Rate for different datasets and algorithms?

Research Answer 2 (RA2):

We have tried to explain all these answers in Chapter 6.2. There are 28 different tables, which shows 2,464 results that we have showed in our thesis. We have tried to generate a different Benchmarking Procedure, where we give ONE point for every classifier, who is Best Average (+ point), Wins Most (+ point), and who's the Test Base (- point), which is the worst performed one in terms of Ranking Results.

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   11
Resultsets: 8
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       1/20/13 3:02 PM


>-<   >   < Resultset
  7  12   5 bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E bayes.net.estimate.SimpleEstimator -- -A 0.5' 746037443258775954
  7  12   5 bayes.NaiveBayes '' 59952312017856976555
  6   9   3 functions.RBFNetwork '-B 2 -S 1 -R 1.0E-8 -M -1 -W 0.1' -3669814959712675720
  3   9   6 functions.Logistic '-R 1.0E-8 -M -1' 39321170325465537 27
  3  13  10 trees.J48 '-C 0.25 -M 2' -217733168393644444
  2   9   7 functions.SMO '-C 1.0 -L 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKernel -C 250007 -E 1.0\"' -6585883636378691736
  0   6   6 functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 10 -V 0 -S 0 -E 20 -H a' -59906078170482107 79
-28   0  28 lazy.IBk '-K 1 -W 0 -A \"weka.core.neighboursearch.LinearNNSearch -A \\\"weka.core.EuclideanDistance -R first-last\\\"\"' -3080186098777067172
```

Figure 7.3. Ranking Result for Accuracy in Pure 10 Fold Experiment.

Here we see that Bayes Net performed best in 12 wins and 5 losses and k-NN performed worst with 28 losses. From this experiment, we define k-NN is the Test Base for Paired T-Test. We are also looking at Error Rates with Root Mean Squared Error and this is also reviewed as Test Base in our Summaries.

If we look at the Results of each different implementation, we see these results:

	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Best Average		2		1		3	2	1
Wins Most		1	1		1	2		3
Test Base	6		3					

Naive Bayes is the best classifier.

Figure 7.4. Pure 10 Fold – Summary.

	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Best Average		1		1	1	4	2	3
Wins Most		1			1	4	1	2
Test Base	6				1			

Naive Bayes is the best classifier.

Figure 7.5. Pure Train Test Split – Summary.

	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Best Average			1	1	1	5	2	
Wins Most			2	1	1	4	2	
Test Base	6				1			

Naive Bayes is the best classifier.

Figure 7.6. PCA 10 Fold – Summary.

	kNN	DT	SVM	RBF	MLP	NB	LR	BNet
Best Average			1		1	3	2	
Wins Most			2		1	4		1
Test Base	6				1			

Naive Bayes is the best classifier.

Figure 7.7. PCA Train Test Split – Summary.

So, we see Naïve Bayes is the best Classifier in the Experiments.

Research Question 3 (RQ3):

What is the best implementation within this experiment?

Research Answer 3 (RA3):

It’s not possible to choose best implementation within the Experiments. The reason is there are 11 different datasets and we are looking seven different results to make a benchmarking study.

If we want to look from the Statistical Tests manner, one of the Future Works of this Thesis is to apply Friedman’s Test and posthoc test with Nemenyi’s Test.

It would be also good to apply Multi2Test to get a full ordering.

Research Question 4 (RQ4):

Do Probabilistic classifiers have any advantages comparing other classifiers used in this thesis?

Research Answer 4 (RA4):

Yes, Naïve Bayes and Bayesian Networks show a clear outperforming result within the experiment. Their Error Rates are also in acceptable ranges of the experiment.

REFERENCES

Al-Aidaroos, K. M., 2012, "Medical Data Classification with Naive Bayes Approach.", *IT Journal - Asian Network for Scientific Information*.

Ali, A., 2010, "Analyzing Potential of SVM Based Classifiers for Intelligent and Less Invasive Breast Cancer Prognosis.", *Computer Engineering and Applications (ICCEA)*.

Alty, S. 2003, "Cardiovascular disease prediction usign support vector machines.", *Micro-Nano Mechatronics and Human Science, IEEE International Symposium*.

Atienza, F., 2000, "Risk Stratification in Heart Failure Using Artificial Neural Networks.", *AMIA, Inc.* , pp. 32-36.

Auer, P., 2005, "A Learning Rule for Very Simple Universal Approximators Consisting of a Single Layer of Perceptrons.", *Neural networks*, *Vol. 21, No. 5*, pp. 786-795.

Badelescu, L., 2007, "The Choice of the Best Attribute Selection Measure in Decision Tree Induction.", *Annuals of Universtiy of Craiova, Math. Comp. Sci. Ser.* , pp. 88-93.

Berson, A., 1999, "Building Data Mining Applications for CRM.", *McGraw Hill*, pp. 4-14.

Biostatistics, H. U., 2012, Obesity Dataset, http://www.biyoistatistik.hacettepe.edu.tr/lisans/beslenme/, accessed at January 2013.

Brause, R., 2001, "Medical Analysis and Diagnosis by Neural Networks.", *In: Medical data analysis. Springer Berlin Heidelberg*, pp. 1-13.

CART, 2010, "Machine Learning in Real World: CART.".

Cheng, J. R., 2002, "Learning Bayesian Networks from Data: An Information-Theory Based Approach.", *Artificial Intelligence, Vol. 137*, pp. 43-90.

Cios, K., 2007, "Data Mining A Knowledge Discovery Approach.", *Springer*.

Cios, K., 2002, "Uniqueness of Medical Data Mining.", *Artificial Intelligence in Medicine Vol. 26*, pp. 1-24.

Cowburn, P. J., 1998, "Risk stratification in chronic heart failure.", *European Heart Journal* , pp. 696–710.

Cowling, B. J., 2010, "Comparative Epidemiology of Pandemic and Seasonal Influenza A in households.", *New England Journal of Medicine* , *Vol. 362, No. 23*, pp. 2175-2184.

Czerniak, J., 2002, "Application of Rough Sets in the Presumptive Diagnosis of Urinary System Diseases.", *Artifical Inteligence and Security in Computing Systems,* pp. 41-51.

Dancea, O, 2008, “Postoperatory Risk Classification of Prostate Cancer Patients using Support Vector Machines.”, *AQTR - IEEE International Conference*.

Dehariya, A., 2011, “An Effective Approach for Medical Diagnosis Preceded by Artificial Neural Network Ensemble.”, *Proceedings of the 5th National Conference, INDIACom-2011*.

Demiroz, G., 1998, “Learning Differential Diagnosis of Eryhemato-Squamous Diseases using Voting Feature Intervals.”, *Aritificial Intelligence in Medicine* .

Dogan, N, 2010, “A Comparative Framework for Evaluating Classification Algorithms.”, *Proceedings of the World Congress on Engineering*.

Domingos, P., 1997, “On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss.”, *Machine learning*, *Vol. 29, No. 2-3,* pp. 103-130.

Dunham, M. H., 2002, “Data Mining: Introduction and Advanced Topics.”, *Pearson Education.*

Florea, F., 2004, “Comparison of Feature-Selection and Classication Techniques for Medical Images Modality Categorization.”, *Rapport Technique*.

Forman, G., 2004, “Learning from Little: Comparison of Classifiers Given Little Training.”, *15th European Conference on Machine Learning and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases* .

Forsyth, R. S., 1990, “BUPA Liver Disorders.”*, BUPA Medical Research Ltd.*

Frank, A., 2010, UCI Machine Learning Repository., http://archive.ics.uci.edu/ml, accessed at January 2013.

Fuster, V. A., 2001, “Hurst’s the heart”, *McGraw Hill Professional.*

Gamble, A., 2001, “The Dummy's Guide to Data Analysis Using SPSS.”, *Scripps College*.

Garcia-Orellana, C.J, 2007, “SVM and Neural Networks Comparison in Mammographic CAD.”, *Engineering in Medicine and Biology Society*, pp. 3204-3207.

German, G. W., 1999, *“*Statistical and AI Techniques in GIS Classification: A Comparison.”, *Proceedings of SIRC99-The 11th Annual Colloquium of the Spatial Information Research Centre*.

Glotsos, D., 2003, “A Hierarchical Decision Tree Classification Scheme for Brain Tumors Aztrocytoma Gradign Using Support Vector Machines.”, *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis* .

Gonçalves, T., 2004, “The Impact of NLP Techniques in the Multilabel Text Classification Problem.”, *Intelligent Information Processing and Web Mining. Springer Berlin Heidelberg*, pp. 424-428.

Greiner, R., 2002, “Structural Extension to Logistic Regression: Discriminative Parameter Learning of Belief Net Classifiers.”, *AAAI/IAAI*.

Hafner, M., 2007, “Comparison of K-NN, SVM and NN in Pit Pattern Classification of Zoom-Endoscopic Colon Images Usign Co-Occurance Histograms.”, *Image and Signal Processing and Analysis*, pp. 516-521.

Hall, M., 2009, “The WEKA Data Mining Software: An Update.”, *ACM SIGKDD Explorations Newsletter, Vol. 11, No.1*, pp. 10-18.

Han, J., 2005, “Data Mining Concepts and Techniques”, *Academic Press, Morgan Kaufmarm Publishers.*

Hand, D., 2001, “Principles of Data Mining.”, *The MIT Press.*

Heckerman, D., 1996, “A Tutorial on Learning with Bayesian Networks.”, *Technical report, Microsoft Research Advanced Technology Division Microsoft Corporation*.

Hong, Z., 1991, “Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane.”, *Pattern Recognition,* pp. 317-324.

İrsoy, O., 2012, “Design and Analysis of Classifier Learning Experiments in Bioinformatics: Survey and Case Studies.”, *IEEE/ACM Transactions on ComputationalBiology and Bioinformatics*, pp. 1663-1675.

Keogh, E. J., 1999, “Learning Augmented Bayesian Classifiers: A Comparison of Distribution-based and Classification-based Approaches.”, *Proceedings of the seventh international workshop on artificial intelligence and statistics*, pp. 225-230.

Kissiov, V., 2005, “Respiratory Database.”, *Central Lab on Biomedical Engineering, Bulgarian Academy of Sciences; Neonates Clinics, University Hospital "Maichin Dom", Sofia, Bulgaria.*

Kononenko, I., 1993, “Inductive and Bayesian Learning in Medical Diagnosis.”, *Applied Artificial Intelligence an International Journal, Vol. 7, No. 4*, pp. 317-337.

Koutroumbas, K., 2001, “Comparison of Computational Learning Methods on a Diagnostic Cytological Application.", *Eunite 2001*, pp. 500-508.

Lanzarini, L., 2000, “Pattern Recognition in Medical Images using Neural Networks.”, *IEEE Transaction on Image and Signal Processing Analysis*.

Larsen, O., 2002, “Constructing X-of-N Attributes with a Genetic Algorithm.”, *GECCO Late Breaking Papers*, pp. 316-322.

Lucas, P., 2002, “Restricted Bayesian Network Structure Learning.”, *Springer-Verlag* , pp. 217-232.

Maimon, O., 2005, “The Data Mining and Knowledge Discovery Handbook.”*, Springer*.

Noto, K., 2010, “Anomaly Detection Using an Ensemble of Feature Models.”, *IEEE International Conference on Data Mining*.

Pal, M., 2003, “An Assessment of the Effectiveness of Decision Tree Methods for Land Cover Classification.”, *Remote Sensing of Environment*.

Pomi, A., 2006, "BMC Medical Infromatics and Decision Making. Context Sensitive Auto Associative Memories as Expert System in Medical Diagnosis.", *BioMed Central*.

Ramaswamy, S. R., 2000, "Efficient Algorithms for Mining Outliers from Large Data Sets.", *ACM,* pp. 427 – 438.

Reiz, B., 2008, "Tree-Like Bayesian Network Classifers for Surgery Survival Chance Prediction.", *International Journal of Computers, Communications and Control,* pp. 470-474.

Richards, D., 1998, "Taking Up the Situated Cognition Challenge with Ripple-Down Rules.", *International Journal of Human-Computer Studies* , pp. 895–926.

Salim, N. B., 2005," Neural Networks Classification Performance for Medical Dataset.", *PhD Thesis, Universiti Utara Malaysia.*

Seker, M. O., 2001, "Prognostic Comparison of Statistical, Neural and Fuzzy Methods of Analysis Breast Cancer Image Cytometric Data.", *Engineering in Medicine and Biology Society, Proceedings of the 23rd Annual International Conference of IEEE*.

Sewak, M., 2007, "SVM Approach to Breast Cancer Classification.", *Second International Multisymposium on Computer and Computational Sciences*.

Shih, Y. S., 2005, "QUEST Classification Tree.".

SPSS., 2010, "CHAID and Exhaustive CHAID Algorithms.".

Su, J. H., 2008, "Discriminative Parameter Learning for Bayesian Networks.", *25th International Conference on Machine Learning*.

Tan, A.C., 2003, "An Empirical Comparison of Supervised Machine Learning Techniques in Bioinformatics.", *Proceedings of the First Asia Pacific Bioinformatics Conference*.

Tian, Z. G., 2010, "Health Condition Prediction of Gears Using a Recurrent Neural Network Approach Reliability.", *IEEE Transactions*.

Todorova, L. A., 2004, "Weaning from Long-Term Mechanical Ventilation: A Nonpulmonary Weaning Index.", *Journal of Clinical Monitoring and Computing* , pp. 275-281.

Ultsch, A., 1995, "Integration of Neural Networks and Knowledge-Based Systems in Medicine.", *AIME '95 Proceedings of the 5th Conference on Artificial Intelligence in Medicine in Europe: Artificial Intelligence Medicine*.

Witten, I. H., 2011, "Data Mining : Practical Machine Learning Tools and Techniques.", *Elsevier*.

Wolberg, O. L., 1990, "Cancer Diagnosis via Linear Programming.", *SIAM News, Vol. 23, No. 5* , pp. 1-18.

Ye, N., 2003, "The Handbook of Data Mining.", *Lawrence Earlbaum Associates.*

Yuan, X., 2003, "A Comparison Study of Decision Tree and SVM to Classify Gene Sequence.", *Electrical Engineering and Computer Science Department, Tulane University*.

Zadeh, L. A., 1969, "Biological Application of the Theory of Fuzzy Sets and System.", *The Proceedings of an International Symposium on Biocybernetics of the Central Nervous System*, pp. 199-206.

Zenko, B, 2001, "A Comparison of Stacking with MDTs to Bagging, Boosting, and Other Stacking Methods.", *ICDM* .

Zhang, W., 2009, "A Comparative Study of Ensemble Learning Approaches in the Classification of Breast Cancer Metastatis.", *International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*.

Zhang, Z., 2003, "SEP: Score for Expression Profile-A Novel Method for Predicting Clinical Outcome in Breast Cancer.", *Engineering in Medicine and Biology Society. Proceedings of the 25th Annual International Conference of the IEEE*.

Zhou, Z., 2002, "Ensembling Neural Networks: Many Could Be Better Than All.", *Artificial Intelligence,* pp. 239-263.

Zhou, Z., 2002, "Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles.", *Artificial Intelligence in Medicine, Vol. 24, No. 1,* pp. 25-36.