CHAPTER ONE: INTRODUCTION

1.1Overview

This chapter demonstrates an overview for the research section 1.2 presents background , section 1.3 presents problem statement, section 1.4 presents the significance of the research, section 1.5 presents hypothesis ,section 1.6 presents research objectives ,section 1.7 presents the scope , and tools and section 1.8 presents thesis structure research.

1.2 Background

The amount of data collected by the bank has grown rapidly in recent years I addition to the information contained in this data can be very important . The creation of knowledge base and its utilization for the benefit of the bank is a strategy tool to compete and the wide availability of huge amounts of data and the need for transforming such data into knowledge encourage Information technology (IT) industry to use data mining. The banking industry around the world has undergone a great change in the way business is conducted. It has started realizing the need of the techniques like data mining which can help them to compete in the market. Leading banks are using Data Mining (DM) tools for customer segmentation and profitability, credit scoring and approval, predicting payment default, marketing, detecting fraudulent transactions, etc (Moin and Ahmed, 2012).

Data mining or Knowledge Discovery in data is process of identifying non-trivial, valid, novel, potentially useful and ultimately understandable patterns in data .Data mining is more than collection and managing data; it also includes analysis and prediction. People are often do mistakes while analyzing or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems. Machine learning can often be successfully applied to these problems, improving the efficiency of systems and the designs of machines. There are several applications for machine learning (ML), the most significant of which is data mining. numerous ML applications involve tasks that can be set up as supervised (Kumar and Verma, August 2012).

1.3 Problem statement

Nowadays, there are many risks related to bank loans, for the bank and for those who get the loans. The analysis of risk in bank loans need understanding what is the meaning of risk. in addition, the number of transactions in banking sector is rapidly growing and huge data volumes are available which represent the customers behavior and the risks around loan are increased(Hamid and Ahmed, March 2016). Therefore, they have to be more careful toward loan granting decisions and avoid manual analysis of borrowers' applications by loan officers because it depends on loan officers' characters and experience accordingly ,subjective, inaccurate, inconsistent, and informal decision might be made.

1.4 Research significance

This research aims to avoid the real risks in banks due to non-payment, which consider a burden on the bank, also offers models to classify the borrowers according to their payment risk, so that the bank can focus only on the reliable borrowers to increase the profitability.

1.5 Research Questions

We can formulate some questions, such as:

1. Are the models that resulting from data mining techniques will enable the administrator to reduce the bank's losses?

2. Does the use of data mining techniques will increase the performance of the bank?

1.6 Research objectives

The main aim of this research is to enhance the process of loan granting in banks by using of preprocessing techniques to achieve a suitable dataset to improve the performance of classification models. This is achieved by fulfilling the following objectives:

1. To apply (J48) classifier (using original dataset and preprocessed dataset).

2. To Compare the performance before using preprocessing techniques (using original dataset) and after using preprocessing techniques (using preprocessed dataset).

3. To apply (Naive bayes), (J48), (IBK), (MLP) and (SMO) using preprocessed dataset.

1.7 Research scope

This research focuses on applying DM techniques for supporting banks loans by extracting knowledge from a Taiwanese customers data, which obtained from the UCI machine learning repository website (repository, 2005)

1.8 Thesis Structure

This research contains five chapters. Chapter one consists of introduction, research problem and objective, chapter two presents the literature review and related works. Chapter three describes the methodology and techniques, chapter four presents the description of data set, data preprocessing and the Experiments and results of the research, and the last chapter contains the conclusion and future work.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This chapter presents the literature review and the work done by other researchers related to this work. Section 2.2 presents the knowledge discovery, section 2.3 presents data mining definition, section 2.4 presents the data mining models; section 2.5 presents data mining applications in banking and Section 2.6 presents previous studies and related works.

2.2 The Knowledge Discovery

Knowledge discovery from databases(KDD) is defined as the process of identifying valid, novel, potentially useful and ultimately understandable patterns of data .One of the crucial steps in Knowledge discovery is Data Mining and often they are used as synonyms(Pulakkazhy and Balan, 2013).

Data mining is the core part of the knowledge discovery process. This process consists of seven steps which are : data selection, data cleaning, data transformation, pattern searching (data mining), finding presentation, finding interpretation and evaluation. The data mining and KDD often used interchangeably because Data mining is the key part of KDD process. The term Knowledge Discovery in Databases or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is interest of researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large data bases. It does this by using data mining methods(algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformations of that database(Priyadharsini and Thanamani, March 2014).

The knowledge discovery is an iterative and interactive process. It consist of seven steps as in figure 2.1, these steps are:

1.Data cleaning: It is also known as data cleansing. Noise data and irrelevant data are removed in this phase.

2.Data integration: In this stage, multiple data sources, often heterogeneous, are combined in a common source.

3.Data selection: The data relevant to the analysis is decided on and retrieved from the data collection.

4.Data transformation: It is also known as data consolidation. In this phase the selected data is transformed into forms appropriate for the mining procedure.

5.Data mining: It is the crucial step in which clever techniques are applied to extract potentially useful patterns.

6.Pattern evaluation: In this step, interesting patterns representing knowledge are identified based on given measures.

7.Knowledge representation: It is the final phase in which the discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results (Hemlata Sahu, 2013).

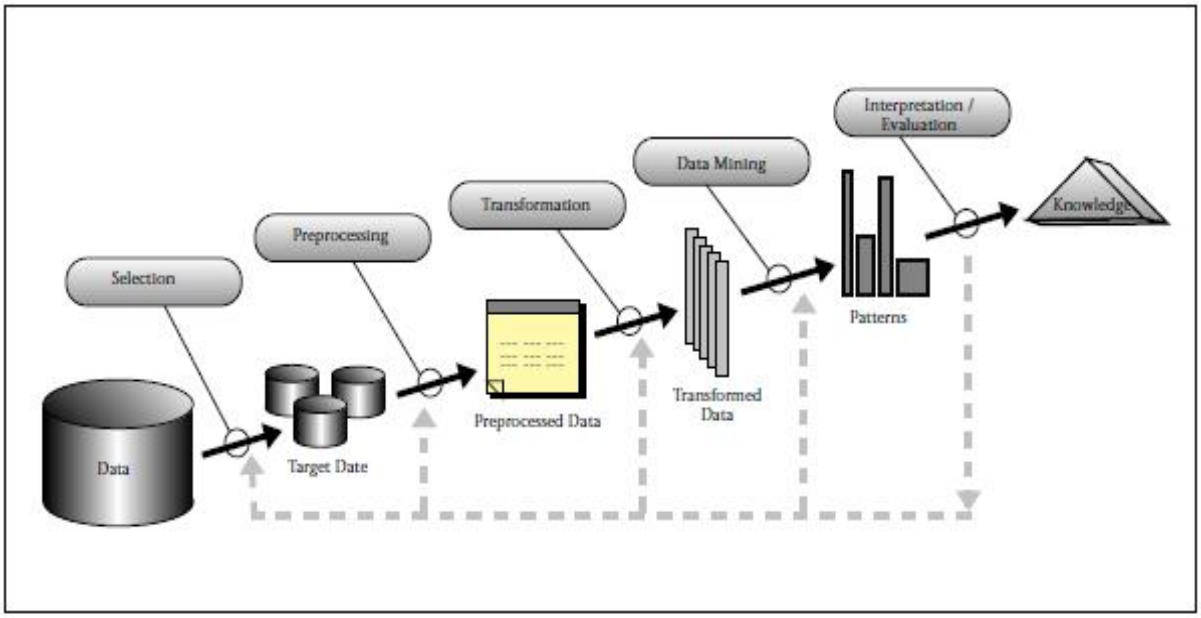


Figure 2. 1 Steps of the KDD Process

2.3Data Mining

Data mining is a process to extract knowledge from existing data. It is used as a tool in banking and finance in general to discover useful information from the operational and historical data to enable better decision-making. It is an interdisciplinary field, gathering of statistics, database technology, Information science, machine learning and visualization(Pulakkazhy and Balan, 2013).

Data mining models are categorized in two types predictive and descriptive, The predictive model makes prediction about unknown or missing data values by using the known

values, such as classification, regression, prediction, time series analysis The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined, such as association rule, summarization, sequence discovery etc(Khan et al., May-Jun 2014), Figure 2.2 shows data mining model and tasks .

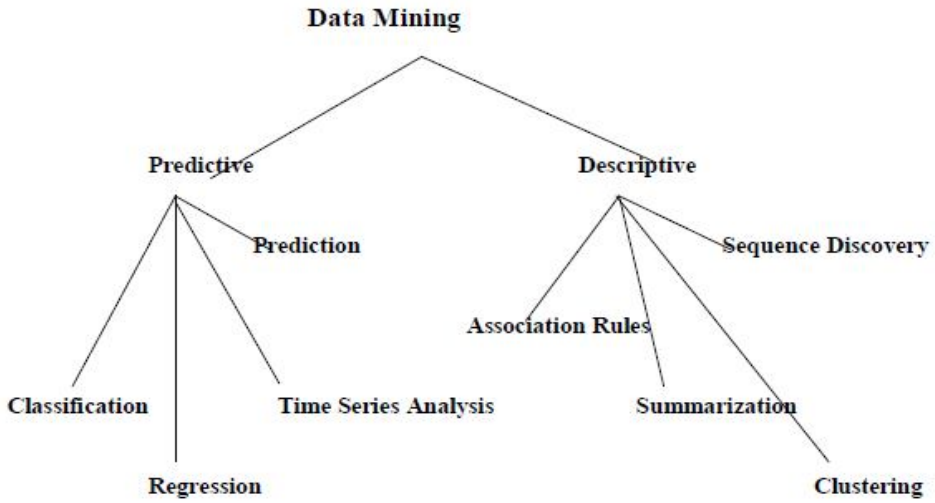


Figure 2. 2: Data mining model and tasks

2.3.1 Classification

Classification is the most common data mining technique, which uses a set of pre-classified examples to develop a model that can classify the records set in general. The classification is used primarily to classify each item in a set of data into one of predefined set of classes or groups. The classification method uses mathematical techniques such as decision trees, linear programming, neural network, and statistics. In classification, we make the software that can learn how to classify the data items into groups. Fraud detection and credit risk applications are particularly suitable for this type of analysis. This approach frequently uses decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In learning, the training data is analyzed through a classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If accuracy is acceptable, rules can apply to new data sets (Moin and Ahmed, 2012).Examples of classification models :Classification by decision tree induction, Bayesian classification ,Neural Networks ,Support Vector Machines (SVM) . and classification based on association (Dr. Maruf Pasha, March 2017)**.**

2.3.2 Prediction

The prediction as the name implied is one of a data mining techniques that discover relationship between independent variables and relationship between dependent and independent

variables. Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. Examples of regression methods: linear regression ,multivariate linear regression ,nonlinear regression and multivariate nonlinear regression (Moin and Ahmed, 2012) .

2.4 Data Mining Applications in Banking

Banking is an area where vast amounts of data are collected .The amount of data collected by banks has grown rapidly in recent years. This data can be generated from bank account transactions, loan applications, loan repayments, credit card repayments, etc. It is assumed that valuable information on the financial profile of customers is hidden within these massive operational databases and this information can be used to improve the performance of the bank this explosive growth has lead to the need for new data analysis techniques and tools in order to find the hidden information in this data. Existing statistical data analysis techniques find it difficult to manage with the large volumes of data now available. (Moin and Ahmed, 2012)**.**

Banking information systems contains huge volumes of data both operational and historical. Data mining can help processes making critical decision in the banks. Banks who apply data mining techniques in their decision making have hugely benefit than others who don't. Banks use data mining in various application areas like marketing, fraud detection, risk management, money laundering detection and investment banking. The patterns detected help the bank to forecast future events that can help in its decision-making processes. More and more banks are investing in data mining technologies to be more competitive(Pulakkazhy and Balan, 2013).

Data might be one of the most valuable resources of any bank but only if it knows how to expose valuable knowledge hidden in raw data. Data mining allows extracting knowledge from the historical data, and predicting outcomes of future situations. It helps optimize business decisions, increase the value of each customer and communication, and improve customer satisfaction (Moin and Ahmed, 2012).

2.4 .1 Risk Management in Banks

Data mining is widely used for risk management in the banking industry. Executive managers for banks need to know whether the customers who are dealing with are reliable or not. Offering new customers credit cards, extending existing customers lines of credit, and approving loans can be risky decisions for banks if they do not know anything about their customers, So banks provide loan to its customers by verifying the various details relating to the loan such as amount of loan, lending rate, repayment period, type of property mortgaged, demography, income and credit history of the borrower. Customers with bank for longer periods, with high

income groups are likely to get loans very easily. Even though, banks are cautious while providing loan, there are chances for loan defaults by customers (Moin and Ahmed, 2012).

There are three main categories of risks in banks(ElHassan, July 2014) :

1. Market Risk: It is defined as the possibility of losing the Bank due to changes in market variables.
2. Operational Risk: defined as the risk of loss arising from inadequate or failed internal processes, people and systems or from external events. In order to minimize this, internal control and internal audit systems are used.

3. Credit Risk: defined as the risk of loss due to the lack of commitment by the creditor to the repayment according to the agreed terms.

2.5 Previous studies and related works

This section presents briefly some of these techniques which are used in loans risk management and their finding:-

2.5.1Developing Prediction Model for banking Loans Risk in Banks Using Data Mining techniques:

This study is an evaluation of the applicability of a new integrated model on a sample data taken from Indian Banks contain of 500 customers data from different banks such as SBI, Andhra, ICICI and Syndicate banks. After they normalized the data they developed a new model by combining the advantages of all these five models (Logistic Regression, Multilayer Perceptron Model (MLP), Radial Basis Neural Network, Support Vector Machine (SVS) and Decision tree (C4.5)) and compared the effectiveness of these techniques for credit approval process. Their findings indicated that SVM, decision tree and logistic regression is the best methodology for classifying the loans. Also, they found that in case of missing data multilayer perceptron model and logistic regression is also good that is why they developed a new integrated model which takes advantages of all the five models(Sudhakar and Reddy, July-August 2014).

This study is attempt to compare the predictive accuracy of customer's default payments using different data mining techniques (Linear Discriminant Analysis (LDA), Naïve Bayes, J48, Logistic Regression, MLP, and IBK) .Their results showed that Multilayer Perceptron performed best with the data-set of default of credit card clients . The utmost accuracy presented by MLP is 81.7% comparing to other algorithm's predictive accuracy. Precision and recall is 79.9% and 81.7% respectively. It's learning by example capability makes it superior to others. It has the highest coefficient of determination. The algorithm is fault tolerant through redundant information coding ability, whereas in other algorithms due to the low power of fault tolerance, partial destruction of network leads other algorithm to low efficient performance(Dr. Maruf Pasha, March 2017).

The researchers of this study used the default of credit card clients dataset in the UCI machine learning repository. The credit card customers were classified if they would do payment or not (yes=1 no=0) for next month by using 23 information about them. Totally 30000 data in the dataset's (66%) was used for training and the rest of them (33%) was used for testing. .Multilayer Perceptron (MLP) and k Nearest Neighbors (kNN) used as machine learning algorithms. With (kNN) estimation success rates for various numbers of neighborhoods value was calculated one by one. The highest success rate was achieved as (80.66%) when the number of neighbor is 10. With MLP neural network model the estimation success rates was calculated when there are different number of neurons in the hidden layer of MLP. The best estimation success rate was achieved as 81.049% when there was only one neuron in the hidden layer(Murat KOKLU, September 2016).

The aim of this study is to built a new model for classifying loan risk in banking sector by using data mining to predict the status of loans . they used three algorithms (j48, bayes Net and naïve Bayes) to build a predictive models that can be used to predict and classify the applications of loans that introduced by the customers to good or bad loan by investigate customer behaviors and previous pay back credit. They found that the best algorithm for loan classification is j48 because it has high accuracy and low mean absolute error(Hamid and Ahmed, March 2016).

In this study a suitable and high performance credit scoring models (CSMs) was developed to assess credit risk of personal loans for the Sudanese commercial banks using data mining techniques, two Sudanese credit datasets were constructed. These datasets were provided by Agricultural Bank of Sudan and Al Salam Commercial Bank. In addition to these two datasets, a German credit dataset was also employed in this research as a benchmarking dataset. Three data mining classification techniques were employed in this research: Artificial Neural Network (ANN), Support vector Machine (SVM) and Decision Tree (DT). Genetic Algorithm (GA) is also applied as a feature selection technique. Two validation methods (split validation with two ratios (70:30 and 60:40) and 10-cross validation) were used to validate the proposed credit scoring models .The researcher found that combining identified single techniques with GA as a feature selection technique will develop the proposed CSMs more than applying these classification techniques to the original datasets, or applying these single techniques to the reduced datasets(ElHassan, July 2014).

2.6 Summary of the related works

Table 2. 1 Summary of the related works

Title of the study	The researchers	The used techniques	The results	The open issue
Credit Evaluation Model of Loan Proposals for Banks Using	Sudhakar M, Dr. C. V. Krishna Reddy	Logistic Regression, Multilayer Perceptron Model, Radial	This research found that SVM, decision tree and	In the case of missing data multilayer perceptron

Data Mining Techniques		Basis Neural Network, Support Vector Machine and Decision tree (C4.5)	logistic regression is the best methodology for classifying the loan applications	model and logistic regression is better
Performance Comparison of Data Mining Algorithms for the Predictive Accuracy of Credit Card Defaulters	Dr. Maruf Pasha, Meherwar Fatima, Abdul Manan Dogar and Furrakh Shahzad	Six data mining techniques (FLDA, Naïve Bayes, J48, Logistic Regression, MLP, and IBK)	The results of this research indicate that the neural network performs best to predict the default of credit card clients and shows the highest accuracy	using large number of features in classification
Estimation of Credit Card Customers Payment Status by Using kNN and MLP	Murat KOKLU, Kadir SABANCI	Multilayer Perceptron (MLP) and k Nearest Neighbors (kNN)	The best estimation success rate was achieved as 81.049% when there was only one neuron in the hidden layer	using large number of features in classification
Developing Prediction Model of Loan Risk in Banks USING Data Mining	Aboobyda Jafar Hamid and Tarig Mohammed Ahmed	j48, bayesNet and naiveBayes	J48 was selected as best algorithm based on accuracy	The model gives low accuracy
Credit Scoring Using Data Mining Classification: Application on Sudanese Banks	Eiman Mohammed El Hassan	Artificial Neural Network (ANN), Support Vector Machine(SVM) and Decision Tree	For all datasets, combining GA as a wrapper-feature selection technique	The problem of imbalanced datasets is not discussed in this research

		(DT). Genetic Algorithm (GA) is also applied as a feature selection technique	with ANN, SVM and DT classification techniques is more beneficial than applying these techniques individually	

CHAPTER THREE: METHODOLOGY

3.1Introduction

This chapter presents data set description which used in our experiments; section 3.3 presents data preprocessing techniques and section3. 4 present the classifiers used in this study.

3.2 Data Set Description

The dataset for this research was chosen from the University of California, Irvine machine learning repository. it consists of 30,000 records of Taiwanese credit card customers and with a total of 24 attributes relating to payment defaults, customer demographics, credit data, history of payments, and bill statements from April to September 2005 from UCI machine learning repository: default of credit card clients Data Set, There are no missing values in this data-set.

Table 3. 2 An overview of the dataset.

ID	ID of each client
LIMIT-BAL	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
SEX	Gender (1 = male; 2 = female)
EDUCATION	(1 = graduate school; 2 = university; 3 = high school; 4 = others).
Age	Age in year
MARRAGE	Marital status (1 = married; 2 = single; 3 = others).
PAY1___PAY6	History of past payment. The past monthly payment records (from April to September, 2005) were tracked as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .; X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
BILL_AMT1__ BILL_AMT6	Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005
PAY AMT1__ PAY AMT6	Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .; X23 = amount paid in April, 2005.
DEFAULT	Default payment (1=yes, 0=no).

3.3 Data Preprocessing

Preprocessing is one of the important steps required in data mining. Feature selection (FS) is a process to identify which features are most advantageous but some features may be redundant, while others may be irrelevant and noisy. When the data set consists of meaningless

data that is incomplete (missing), noisy (outliers) and inconsistent data, preprocessing of the dataset is required. In order to improve the accuracy of the models and gain useful results any irrelevant and correlated data were removed from the dataset then implemented data discretization ,data cleaning ,and applied target class balancing as well to our data to achieve a suitable dataset for our Algorithms. Preprocessing step includes(Srivastava., February 2014.) :

1. Data Cleaning: Dealing with missing values by ignoring those particular records, filling that value with some specific value and handling noisy data using binning methods, clustering, combined human & machine inspection and regression. Inconsistency may be handled manually. The data sets contain values that are not specified by the UCI site; the values of Education attribute were determined from one to four but it also contained values that were greater than four , (331) instances besides (14) instances contained the value zero. Actually, we do not know the actual data so we ignored the records. The advantage of this method is that, the model would be based on actual data and not guessed data. We repeated this procedure to the (marital status) attribute, as it contained the same error as our education column, it contained the value zero to (54) instances, after the data cleaning process our dataset became (29601) instances.
2. Features selection: Feature selection is a technique to reduce the dimensionality. The main use of this method is to extract small subsets of relevant features from the original dataset based on evaluation criterion.
3. Data Transformation: Data Transformation is to transform the data in given format to required format for data mining. Normalization, smoothing, aggregation and generalization are few methods to perform transformation. The dataset values in all the records were in numerical form; for example, categorical data such as (sex) were encoded as “1” and “2” to represent male and female respectively. this was an issue as having numerical values in our data would diminish its effectiveness for our users hence, we need to transform specific columns so that they would be more suitable for analyze the result. We changed these attributes are (sex, education, and marital status) into their string representations.
4. Data Reduction: Data analysis on huge amount of data takes a very long time. It can be performed using data cube aggregation, dimension reduction, data compression, data reduction, discretization and concept hierarchy generation .We converted a continuous variable into a discrete one by applying the discretization to our dataset because a group of researchers were concluded that discretization improves the performance of the naive Bayesian algorithm(Jonathan L. Lustgarten, 2008.),So the Age attribute was split into equal sized intervals of ten years also (Limit_ bal) attribute, was placed into ranges labeled as either Low, Medium or High instead of the ranges ((0-100,000 ,100,001-500,000) , and over (500,001)) , in New Taiwan dollars, to the labels respectively. Our last preprocessing step was data reduction , we reduced the size of the dataset in order to achieve two equal class representation ,the default and no default classes, the dataset was reduced from 30,000 to 13,210 records, with a 50%class representation to be 6,605 records each. In addition to that we removed any irrelevant and correlated attributes from our dataset in order to increase the performance, as a result the attributes becomes only 6 instead of 24 attributes, see Figure (3.3).

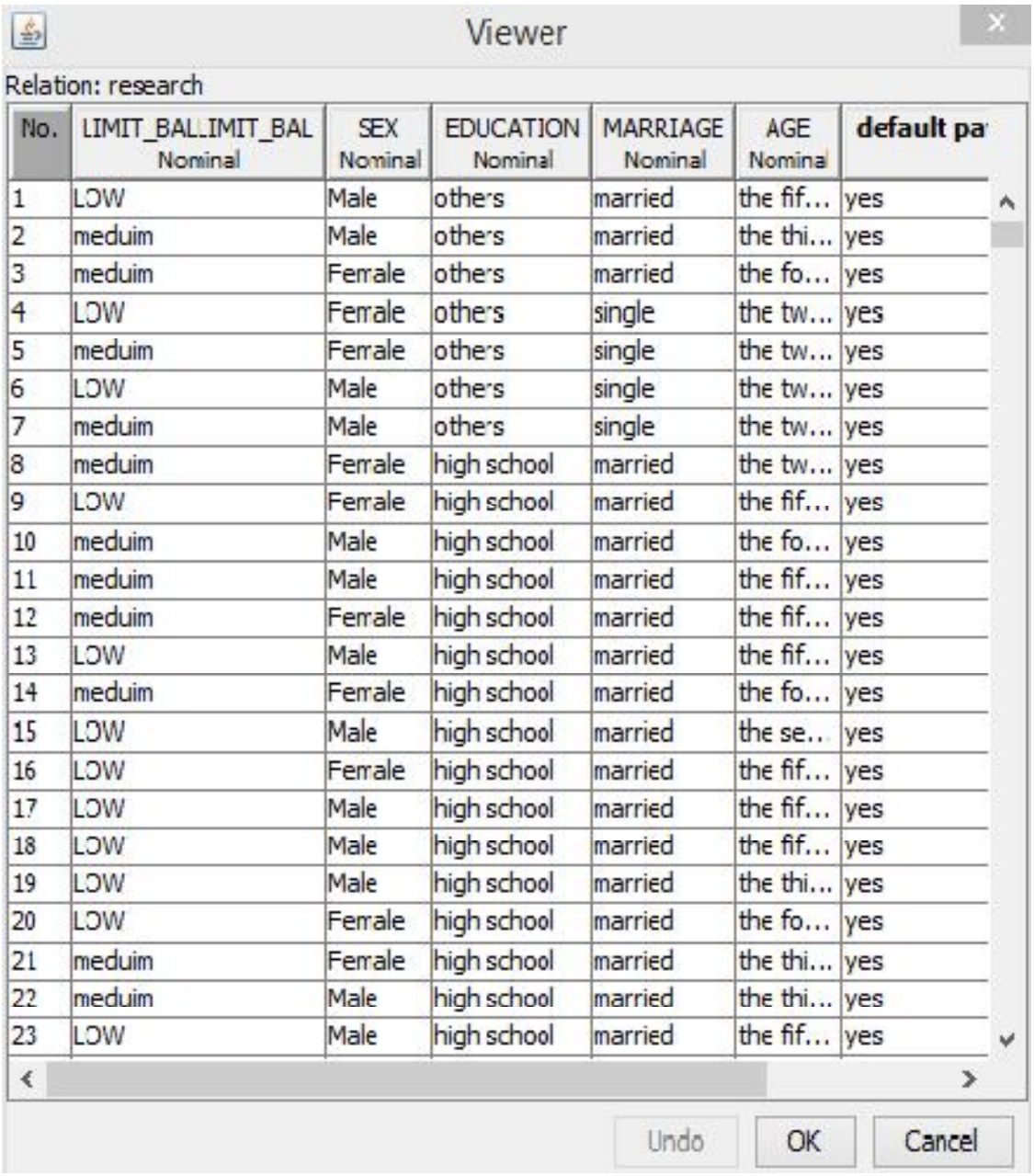


Figure 3.4 The preprocessed dataset

3.4 Classification Using Weka

Weka is a collection of algorithms that help in solving some real world problems. Algorithms can be applied either directly or to a dataset called from own java code. Data processing, classification, clustering, visualization regression and feature selection these techniques are supported by Weka. In Weka data is considered as an instances and features as attributes. In this main user interface is the explorer but essential functionality can be attained by

component based knowledge flow interface and command line whenever simulation is done than the result is divided into several sub items for easy analysis and evolution. One part in correctly or correctly classified instances partitioned into percentage value and numeric value and subsequently kappa statistics mean absolute error and root mean squared error will in numeric value. In data mining, an important problem is large data set for classification. For a database with a set of classes and number of records such that each record belongs to one of the given classes, the classification problem is to decide the class to which a given record. Here, it is concerned with a type of classification called supervised classification. In supervised classification, a training data set of records and for each of this set, the respective class to which it belongs is also known. As they represent rule (Meenakshi, January 2014.). In our experiments five classifiers are used which demonstrate in the following sections.

3.4.1 Naïve Bayes

A Naive Bayes classification is a simple probability class based on applying Bayes' theorem with strong independence assumptions. Naïve Bayes can handle a random number of independent variables, whether continuous or categorical... The algorithm makes predictions using the Bayes theory, which includes evidence or prior knowledge in its predictions. Given the set of variables X = {x1, x2… XD}, the back probability of the Cj among event can be constructed with a set of possible results C = {c1, c2... cd}. Simply, X is the predictor and C is a set of class levels in the dependent variable. Using Bayes rule:

$$p(C_j \mid x_1, x_2, \ldots, x_d) \propto p(x_1, x_2, \ldots, x_d \mid C_j) p(C_j)$$

Where p (Cj | x1, x2… XD) is the posterior probability of class membership.

In simple terms, a Naive Bayes classifier assumes that the presence or absence of a particular feature of a class is unrelated to the presence or absence of any other feature, given the class variable. Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for Naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods. In spite of their naive design and apparently over-simplified assumptions, Naive Bayes classifiers have worked quite well in many complex real-world situations. Analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of Naive Bayes classifiers. An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix(Meenakshi, January 2014.).

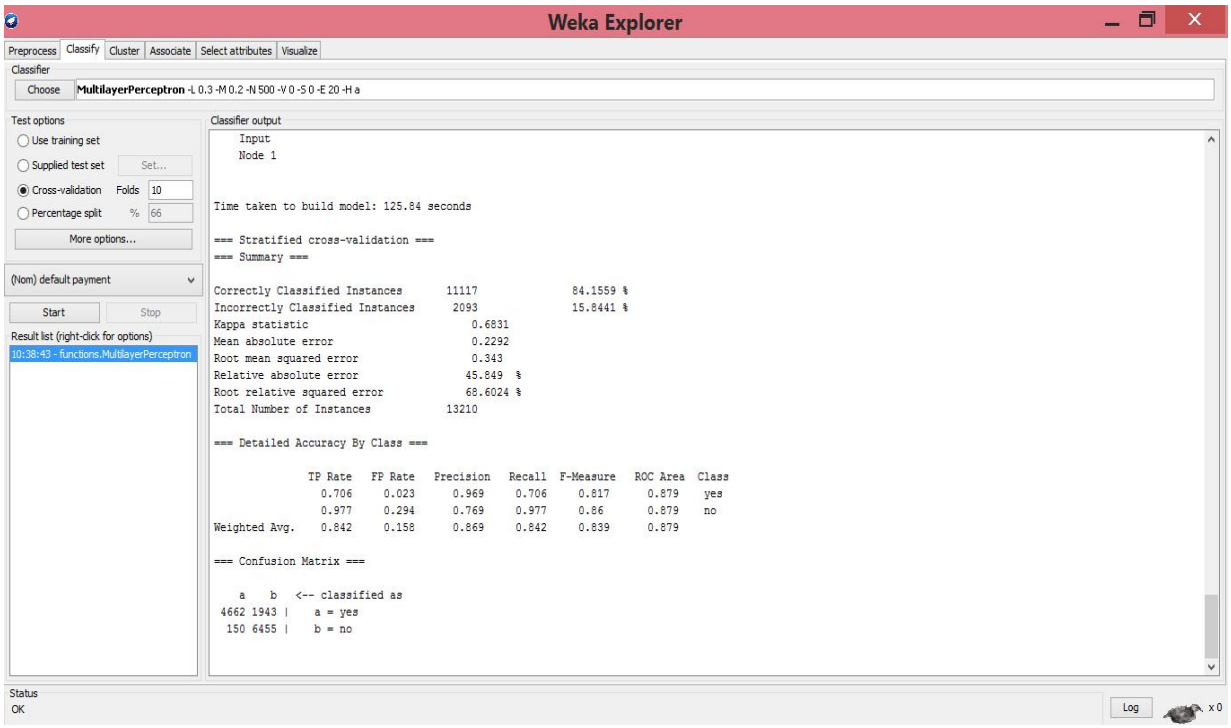


Figure 3.2(Naïve Bayes) classifier with preprocessed dataset

3.4.2. IBK

This algorithm uses the similar distance metric. In object editor, the amount of nearest neighbor can be defined clearly. It can be automatically determined by using leave-one-out cross-validation emphasize to an upper bound given by the particular value. Dissimilar search algorithms are employed to find the nearest neighbors at high speed. Linear search is used as a default search, but there are further options containing “KD-trees,” “ball trees” and “cover trees”, as a parameter, distance function may be employed. The behind thing is similar as “IBL.” That is a Euclidean distance; further choices are “Chebyshev,” “Manhattan,” and “Minkowski distances.” Predictions/Guesses can be weighted, from one or more neighbor, following their distance from the test examples. Distance is converted into weights by applying two dissimilar formulas. Training examples that are held by the classifier can be limited by “window size” option. When novel examples are included, previous examples are removed to retain the number of training examples at this size(Dr. Maruf Pasha, March 2017)

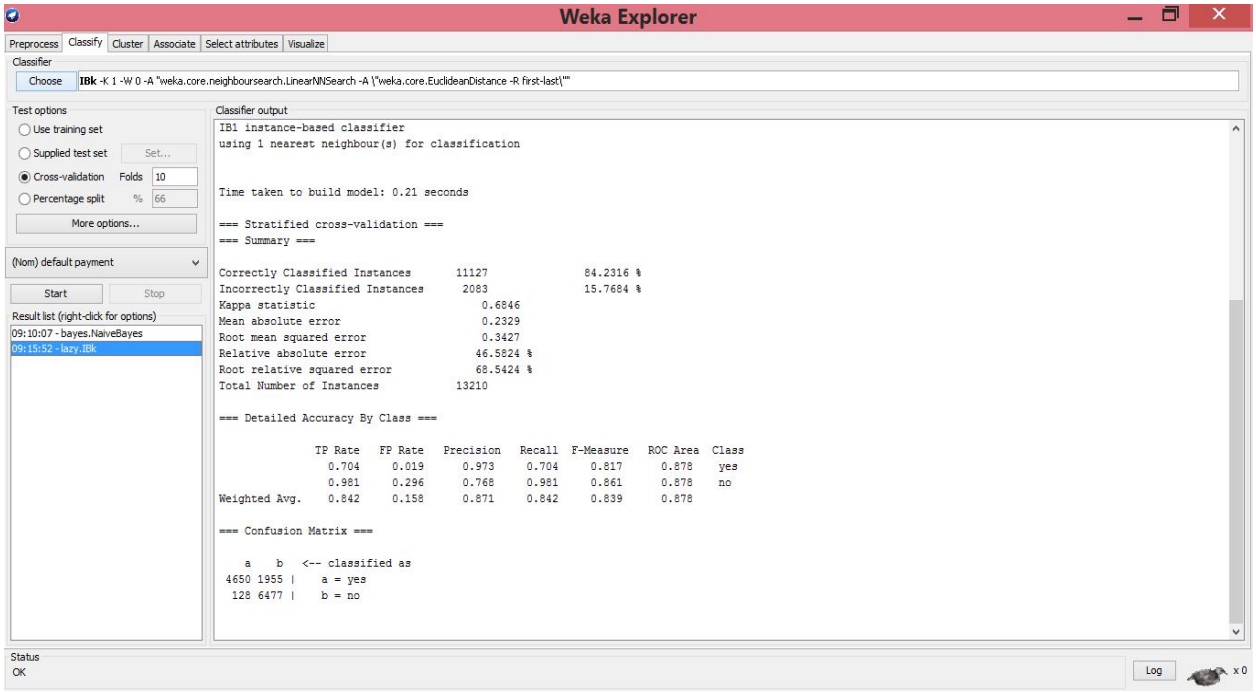


Figure 3.3(IBK) classifier with preprocessed dataset

3.4.3 J48

It is also known as free classifier who accepts nominal classes only. In this prior knowledge should be there while classifying instances. It is used in the construction of decision tree from a set of labeled training data using the information entropy. Attributes which we use helps in building decision tree by splitting it into subset and normalization information gained can be calculated. Splitting process comes to an end when all instances in a subset belong to the same class. Leaf node is also is also present or being created to choose that class a possibility also can be there that none of the feature provides information gain.J48 creates decision nodes up higher in the tree using expected value of the class.J48 can use both discrete and continuous attributes, attributes with differencing lost and training data with missing attribute values (Meenakshi, January 2014.)

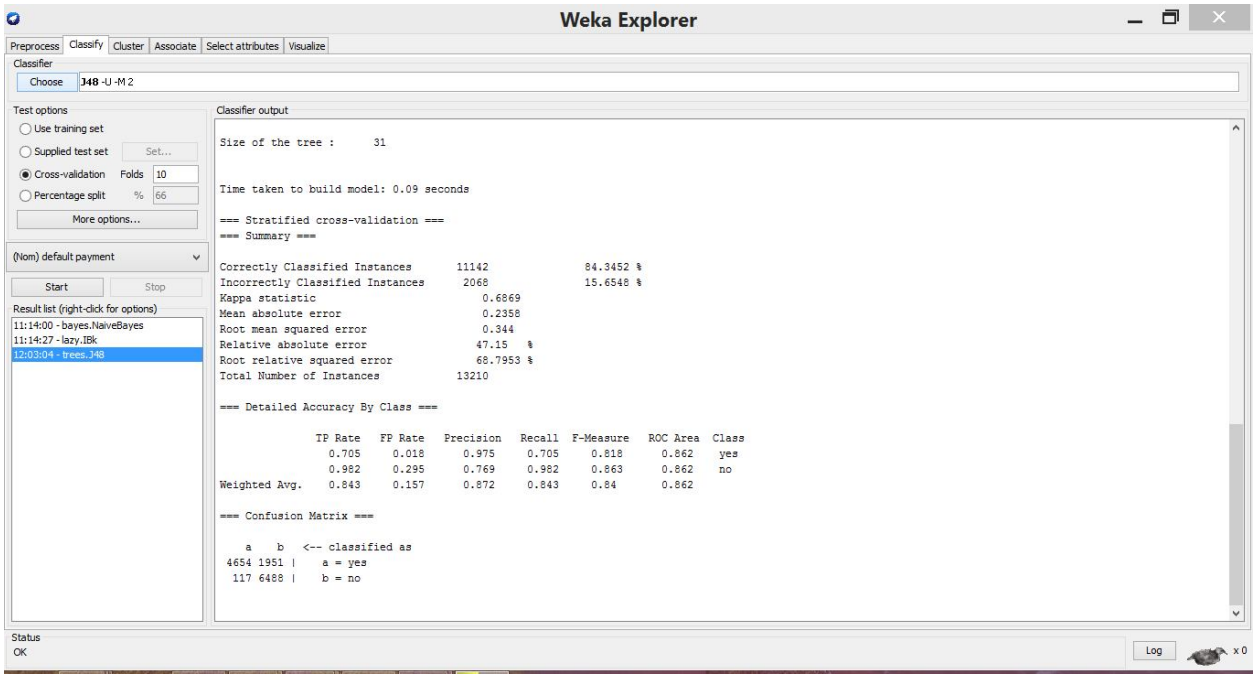


Figure 3.4(J48) classifier with preprocessed dataset

3.4.4 Multilayer Perceptron (MLP)

Multilayer perceptron is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLP consists of multiple layers of nodes, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function.MLP utilizes a supervised learning technique called back propagation If a multilayer perceptron has a linear activation function in all neurons, that is, a simple on-off mechanism to determine whether or not a neuron fires, then it is easily proved with linear algebra that any number of layers can be reduced to the standard two-layer input-output model Each neuron uses a nonlinear activation function which was developed to model the frequency of action potentials.(Sudhakar and Reddy, July-August 2014).

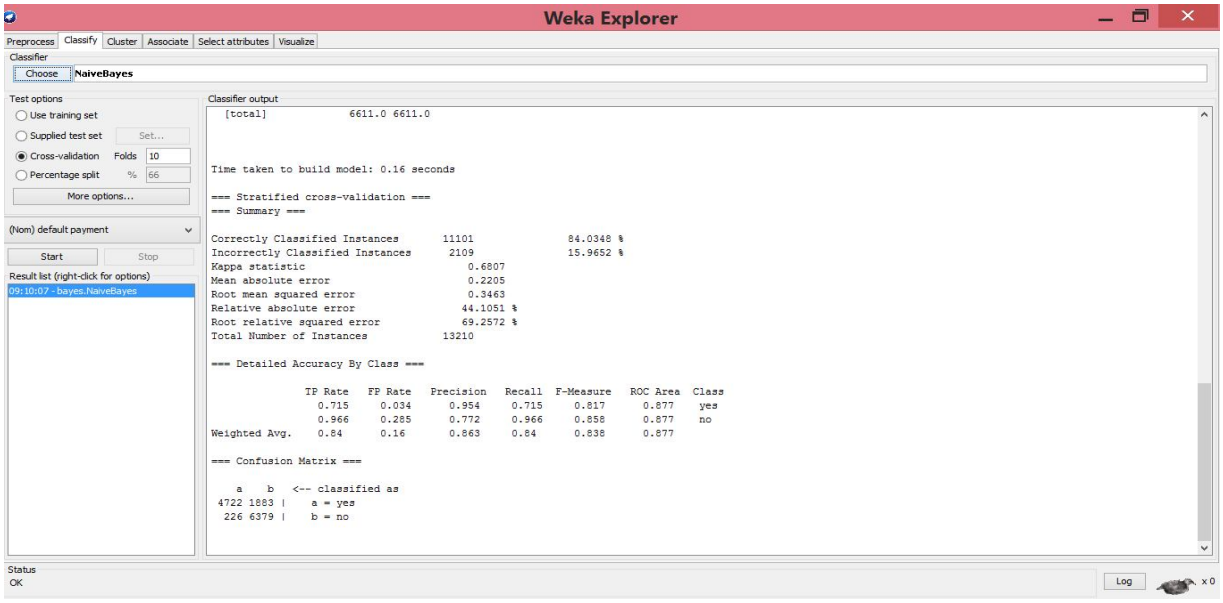


Figure 3.5(MLP) classifier with preprocessed dataset

3.4.5 Support Vector Machine (SVM)

A support vector machine is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. Support vector machine was first proposed by Vapnik (1998). Its main idea is to minimize upper bound of the generalization error and it maps the input vector into high dimensional feature space through some nonlinear mapping. In this space, the optimal separating hyper plane, which separates the two classes of data with maximal margins, is constructed by solving constrained quadratic optimization problem whose solution has an expansion in terms of a subset of training patterns that lie closest to the boundary. Sequential minimal optimization (SMO) is an algorithm for efficiently solving the optimization problem which arises during the training of support vector machines. It was invented by John Platt in 1998 at Microsoft Research. SMO is widely used for training support vector machines and is implemented by the popular libsvm tool. The publication of the SMO algorithm in 1998 has generated a lot of excitement in the SVM community, as previously available methods for SVM training were much more complex (Sudhakar and Reddy, July-August 2014).

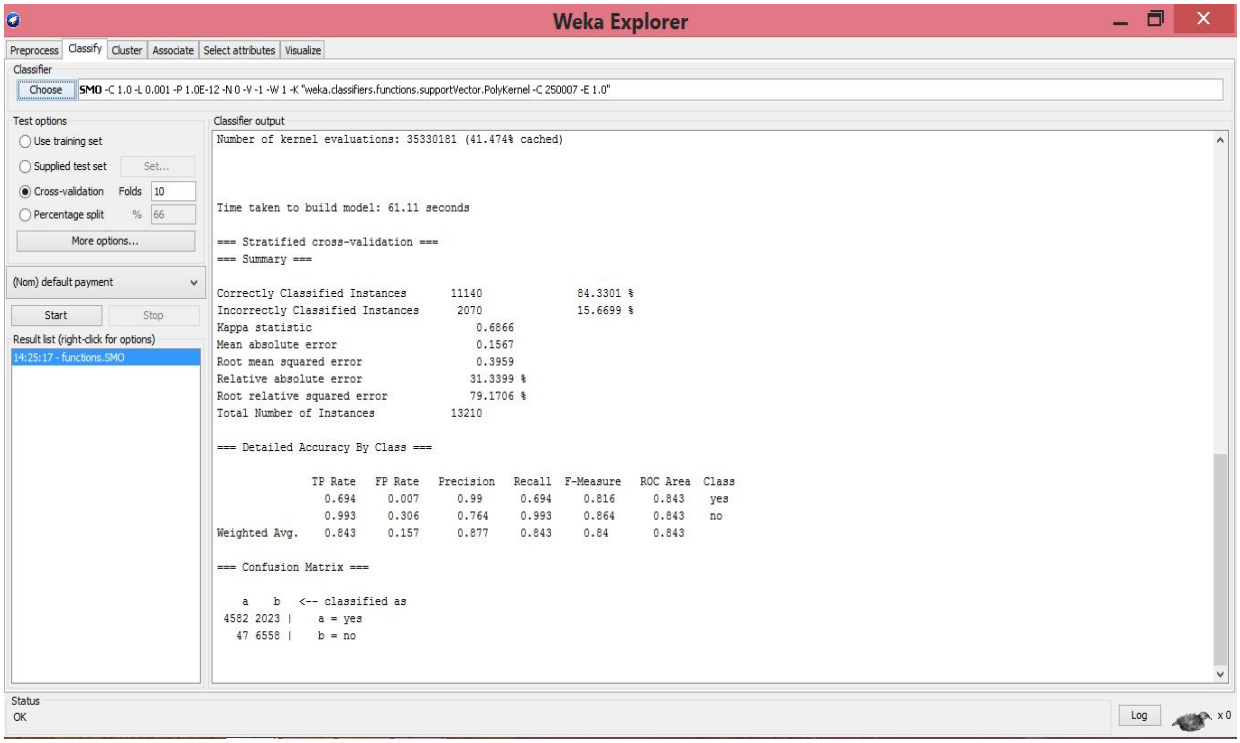


Figure 3.6(SMO)classifier with preprocessed dataset

3.5 Evaluation of classifiers

There are two test modes to evaluate the performance of selected tool, the k-fold cross-validation (k-fold cv) mode and percentage split (holdout method) mode. The k-fold cv refers to a widely used experimental testing procedure where the database is randomly divided in to k disjoint blocks of objects, then the data mining algorithm is trained using k-1 blocks and the remaining block is used to test the performance of the algorithm, this process is repeated k times. At the end, the recorded measures are averaged. It is common to choose k=10 or any other size depending mainly on the size of the original dataset. In percentage split, the database is randomly split in to two disjoint datasets. The first set, which the data mining system tries to extract knowledge from called training set. The extracted knowledge may be tested against the second set which is called test set; it is common to randomly split a data set under the mining task in to 2 parts. It is common to have 66% of the objects of the original database as a training set and the rest of objects as a test set. Once the tests is carried out using the selected datasets, then using the available classification and test modes ,results are collected and an overall comparison is conducted (Nikhil N. Salvithal, October 2013). The k-fold cross-validation (k-fold cv) mode was used in this study.

The Experiments

the first experiments were conducted in two steps:Step1: applying classification model (48 classifier)by using original dataset .

Step2: using preprocessing techniques (data cleaning, feature selection, data transformation data discretization, data target balancing) on the dataset to improve the performance of classification model ,then compare the result that obtained from step1 and step2.

In the second experiment Five classification models were applied by using the preprocessed dataset (1320 records and six features) to evaluate the performance of various classifiers on 10 fold cross validation test mode at WEKA 36.9 .in terms of accuracy, precision and recall.

CHAPTER FOUR: RESULT,ANALYSIS AND DISCUSSION

4.1 Introduction

This chapter presents the results obtained by implementing five classification algorithms on data set of 13210 records for Taiwanese credit card customers with five features related to payment defaults. Finally, the all experimental results for this research were discussed

4.2 The first Experiments

We found in the first classifier with the full data set the accuracy was 77% but the classification rule assumes that no customer belongs to positive class(TP rate 0%). and the model is unable to recognize the positive class(FP rate 0%) as it obvious in the confusion matrix for the model Table (4.1) . Table(4.2) presents the values of ROC area measurement, which is one of the most important values output by weka, the classifier have ROC area values (0.5,0.5), while the values of precision, which is proportion of instances that are truly of a class divided by the total instances classified as that class, are (0.779,0) and the values of recall, which is proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate), are (0,1). It is very clear that this model cannot be relied upon although the percentage of correctly classified instances obtained by the model, It has some disadvantages as a performance estimate (not sensitive to class distribution), see figure (4.1).

Table 4. 1 Confusion Matrix J48 Classifier with Raw data set

a	b	classified as
0	6636	a = yes
0	23364	b = no

Table 4. 2 detailed Accuracy by Class J48 Classifier with Raw data set

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
yes	0	0	0	0	0	0.5
no	1	1	0.779	1	0.876	0.5

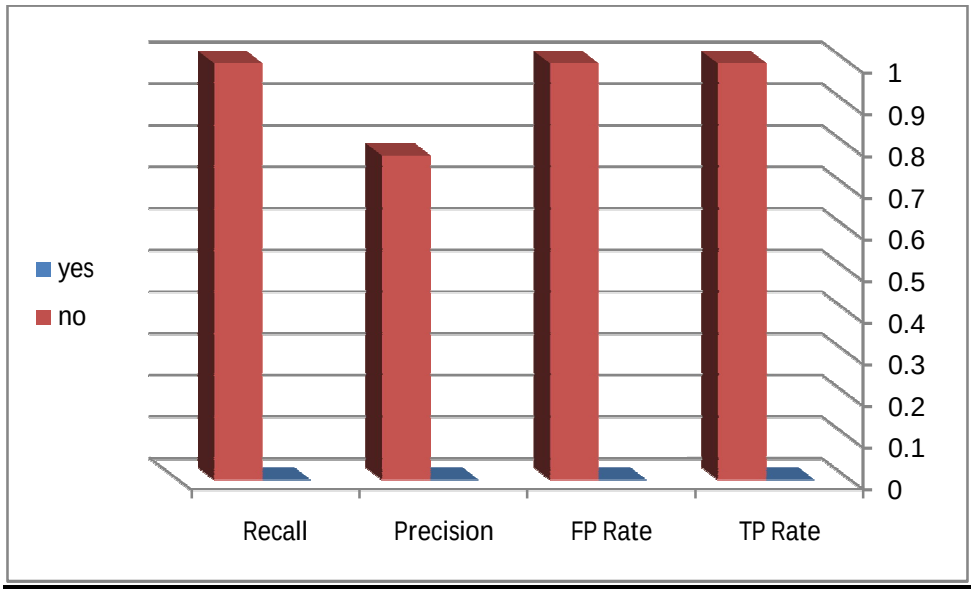


Figure 4. 1detailed accuracy by class using J48 classifier with full data set

Then we built the classifier with preprocessed data set ,see appendix, and found that it have ROC area values (0.858, 0.858) while the values of precision are (0.986, 0.763) and the values of classifier's recall with preprocessed data set are (0.705, 0.99) as it is shown in Table (4.3) , Table (4.3) and figure (4.2). Clearly the classifier with preprocessed data set performed well and given better results than using it in data without any processing, so the preprocessed data set will be used in the rest of the experiments.

Table 4. 3 Confusion Matrix J48 Classifier with preprocessed data set

a	b	classified as
4654	1951	a = yes
117	6488	b = no

Table 4. 4 detailed accuracy by class J48 classifier with preprocessed data set

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
yes	0.705	0.018	0.986	0.705	0.818	0.862
no	0.982	0.295	0.763	0.99	0.863	0.862

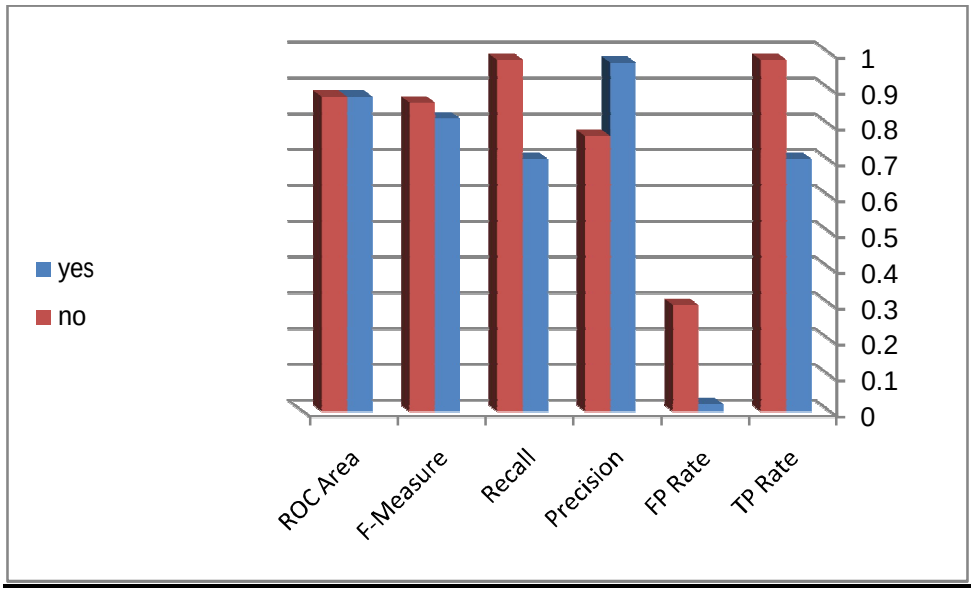


Figure 4. 2 detailed accuracy by class J48 classifier with preprocessed data set

we conducted comparisons between the two classifier In terms of accuracy, precision , Incorrect classification and Recall .It is clear that classifier with preprocessed dataset has got big improvements in terms of performance measures as it is shown in Table (4.5) and Fig(4.3)

Table 4. 5: the evaluation of J48 classifier on different datasets with Cross-validation mode.

j48 Classifier	Correct classification (Accuracy %)	Incorrect classification%	Precision%	Recall%
Full data set	77.88	22.12	60.7	77.9
processed data set	84.36	15.65	87.2	86.2

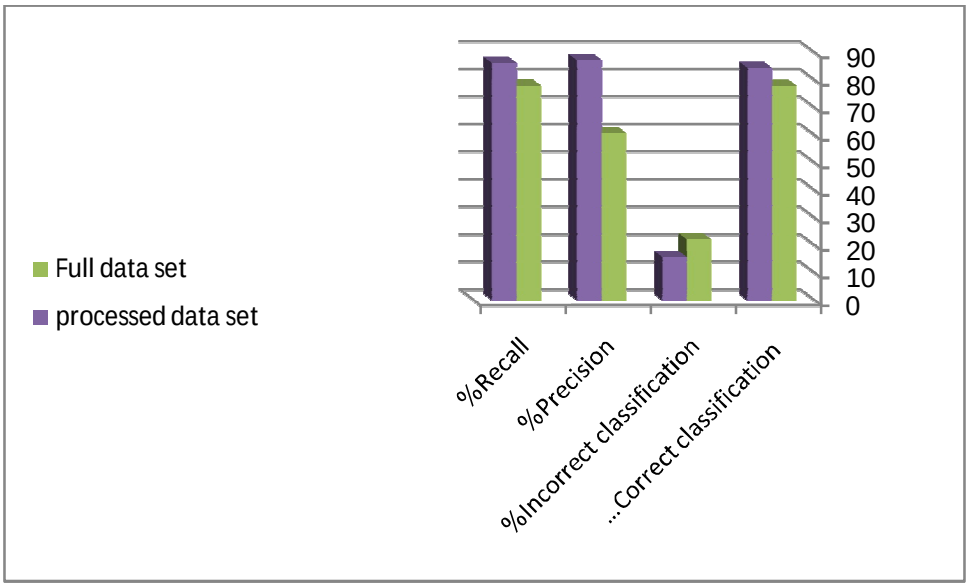


Figure 4. 3comparing between two data sets

From the above results we observed that applying of the preprocess techniques on the data set improved the performance of the classifier .Also observed that "balance" of the data set should be taken into account when interpreting results because unbalanced data sets in which a large amount of instances belong to a certain class may lead to high accuracy rates (majority of the class) even though the classifier may not necessarily be particularly good. So we looked at some of the other measures.

4.3 The second Experiments

Five classification algorithms are applied on the data-set through Weka software (Naive bayes , J48, IBK, MLP , SMO) with (10-fold cross-validation) test mode results were evaluated in term of correctly classified instances, incorrectly classified instances, time taken to build the model precision and recall.

Naive bayes model:

Table 4. 6 Confusion Matrix Naive bayes

a	b	classified as
4722	1883	a = yes
226	6379	b = no

Table 4. 7 Detailed Accuracy by Class Naive bayes

class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
yes	0.715	0.034	0.954	0.715	0.817	0.877
no	0.966	0.285	0.772	0.966	0.858	0.877

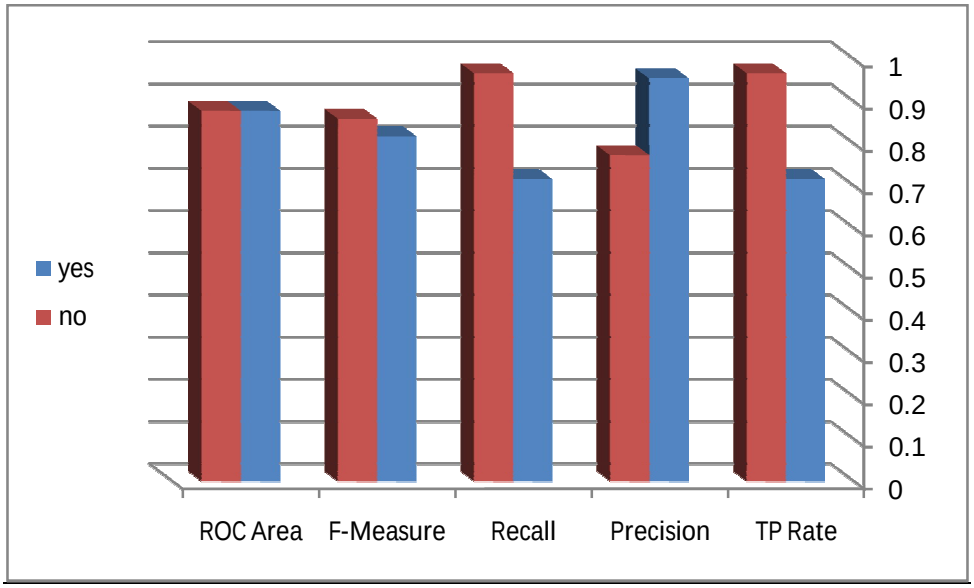


Figure 4. 4 Detailed Accuracy by Class Naive bayes

The accuracy given by naïve bayes model was (84.03%), the recall was (84%) and the precision was (86.3%).

1. J48 model:

Table 4. 8 Confusion Matrix J48

a	b	classified as
4654	1951	a = yes
117	6488	b = no

Table 4. 9 Detailed Accuracy by Class using J48

class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
yes	0.705	0.018	0.975	0.705	0.818	0.862
no	0.982	0.295	0.769	0.982	0.863	0.862

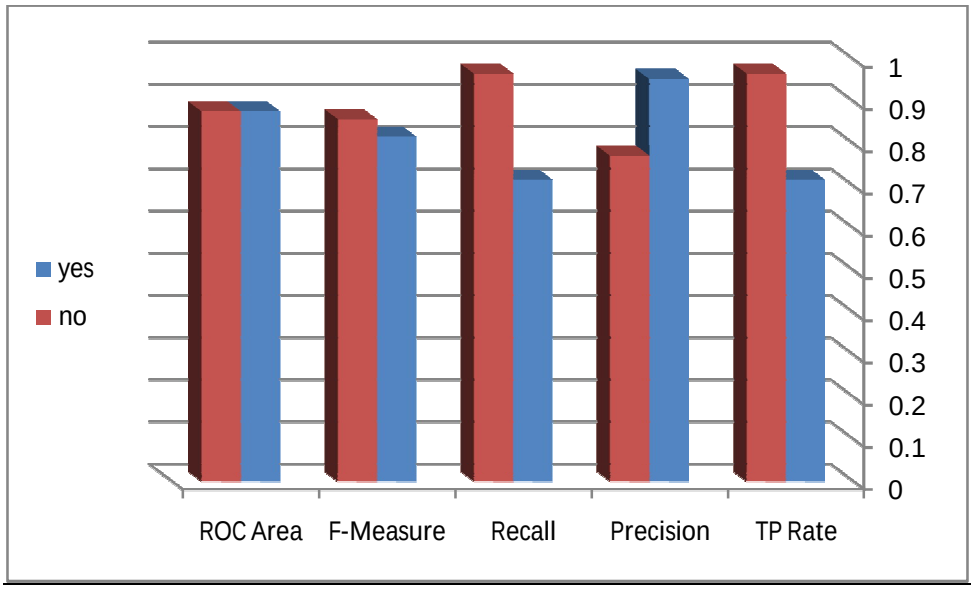


Figure 4. 5 Detailed Accuracy by Class using J48

The accuracy given by J48 model was (84.35%), the recall was (84.3%) and the precision was (87.2%).

2. IBK model:

Table 4. 10 Confusion Matrix IBK

a	b	classified as
4650	1955	a = yes
128	6477	b = no

Table 4. 11 Detailed Accuracy by Class using IBK

class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
yes	0.704	0.019	0.973	0.704	0.817	0.878
no	0.981	0.296	0.768	0.981	0.861	0.878

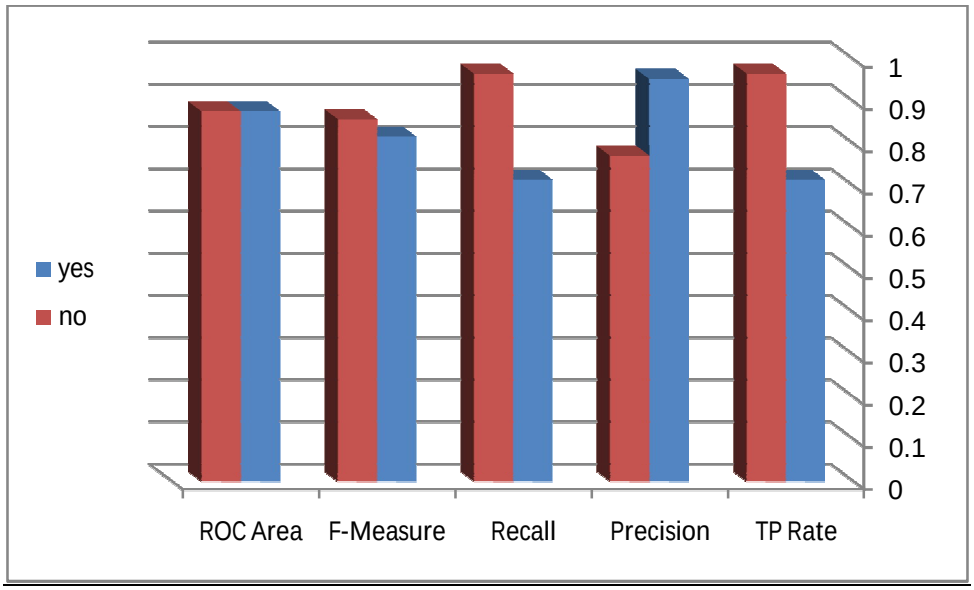


Figure 4. 6 Detailed Accuracy by Class using IBK

The accuracy given by IBK model was (84.23%), the recall was (84.2%) and the precision was (87.1%).

3. IBK model:

Table 4. 12 Confusion Matrix SMO

a	b	classified as
4582	2023	a = yes
47	6558	b = no

Table 4. 13 Detailed Accuracy by Class using SMO

class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
yes	0.694	0.007	0.99	0.694	0.816	0.843
no	0.993	0.306	0.764	0.993	0.864	0.843

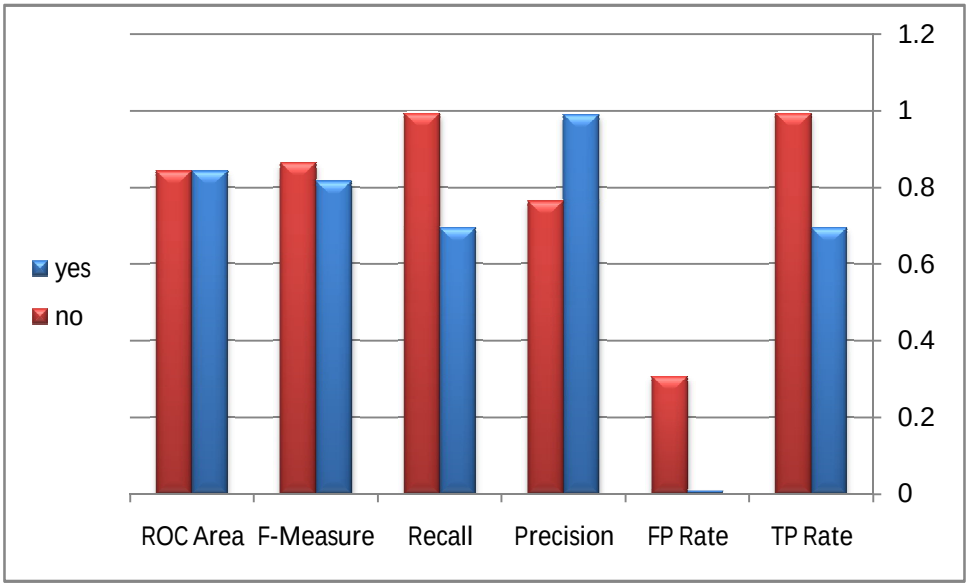


Figure 4. 7 Detailed Accuracy by Class using SMO

The accuracy given by SMO model was (84.33%) , the recall was (84.3%) and the precision was (87.7%) .

5. MLP model

Table 4. 14 Confusion Matrix MLP

a	b	classified as
4662	1943	a = yes
150	6455	b = no

Table 4. 15 Detailed Accuracy by Class using MLP

class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
yes	0.706	0.023	0.969	0.706	0.817	0.879
no	0.977	0.294	0.769	0.977	0.86	0.879

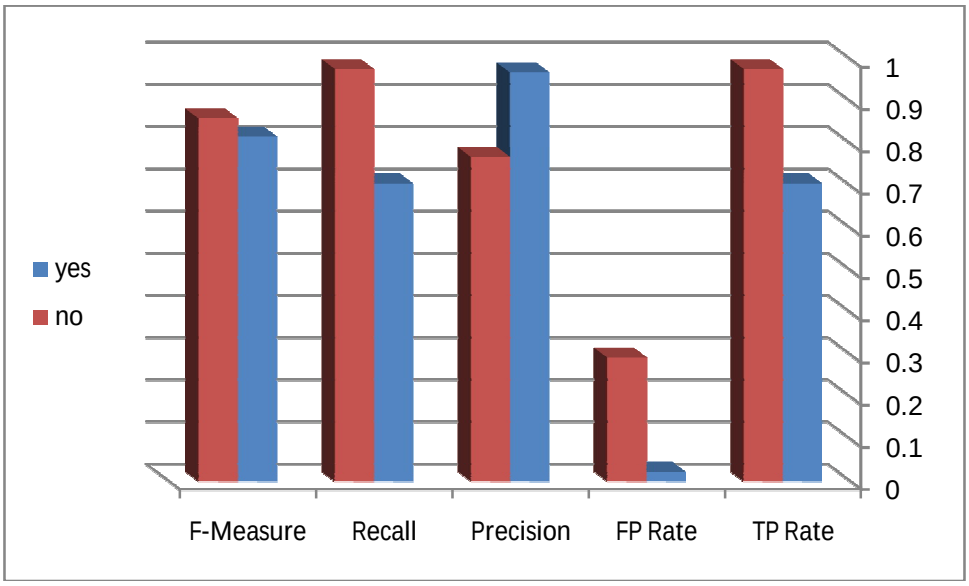


Figure 4. 8 Detailed Accuracy by Class using MLP

The accuracy given by MLP model was (84.16%) , the recall was (84.2%)and the precision was (84.3%) .

Table 4. 16 Evaluation of classifiers on credit card dataset with Cross validation

Classifier	Correct classification %	Incorrect classification%	Precision%	Recall%
Tree (J48)	84.35	15.65	87.2	84.3
Bayes(naïve bayes)	84.03	15.97	86.3	84
Lazy(IBK)	84.23	15.77	87.1	84.2
Function(SMO)	84.33	15.67	87.7	84.3
Function(MLP)	84.16	15.84	84.3	84.2

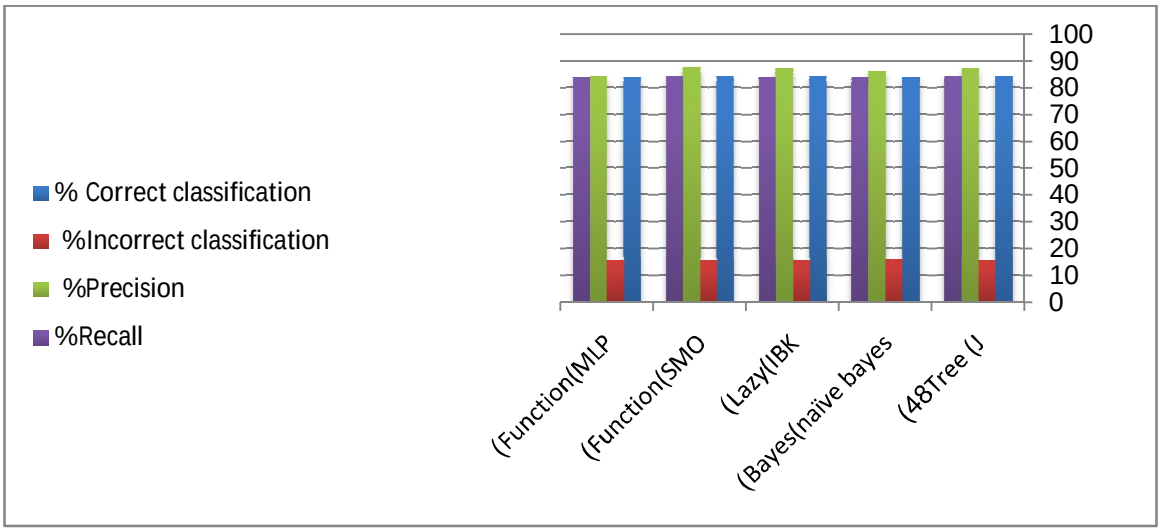


Figure 4. 9 comparing between the classifiers

The experiments show that results of all algorithms are nearly same. The highest accuracy for this dataset (84.35%) was achieved by J48 model .SMO ranked secondly, it yielded (84.33%), then (84.23%), (84.16%) and (84.03%) for IBK, MLP and Naïve bayes respectively .The highest recall for this dataset (84.3%) was achieved by SMO and J48 models. IBK and MLP ranked second, they yielded (84.2%) for recall, then (84) for Naïve bayes .The highest precision for this dataset (87.7%) was achieved by SMO model. J48 ranked secondly, it yielded (87.2) , then (87.1), (86.3) and (84.3) for IBK, Naïve bayes and MLP respectively.

CHAPTER FIVE: CONCLUSION AND RECOMMENDATION

5.1 Introduction

This chapter explains the research conclusion in section 5.2 and recommendation in section 5.3.

5.2 Conclusion

The main purpose of this research is to increase the performance of the bank by building models that can be used to predict trusted and non-trusted borrowers in order to increase the performance of the bank by reducing the costs of non-payment borrowers and decrease the high number of bad loans using data mining techniques.

The dataset of this research was obtained from the UCI machine learning repository website. Some preprocessing techniques (data cleaning, feature selection, data transformation data discretization, data target balancing) were applied to achieve a suitable dataset for our Algorithms , then five data mining classification techniques were used in this research: statistical method (Naive bayes), Decision tree based algorithm (J48), K-nearest neighbor (IBK), Multilayer Perceptron (MLP) and Sequential minimal optimization (SMO).The Weka software from Waikato university with (10-cross validation) was used to model and validate the proposed models, The results carried out in this stage showed that the performance of the five classification algorithms are nearly same . Out of these five classification algorithms, J48classifier had the highest accuracy (84.35%).

5.3 RECOMMENDATION

Expand the size and type of data required from customers in order to add new variables to the data mining process such as the monthly expenditures value, monthly salary value, occupation, number of children, and number of spouses and so on.

Focus on training and developing decision-makers on data mining techniques to make them able to use and apply the data mining software easily.

References:

DR. MARUF PASHA, M. F., ABDUL MANAN DOGAR,FURRAKH SHAHZAD March 2017. Performance Comparison of Data Mining Algorithms for tor tor tor the he he Predictive Accuracy of Credit Card Defaulters. *International Journal of Computer Science and Network Security,* VOL.17 No.3.

DR. SUDHIR B. JAGTAP, D. K. B. G. 2013. Census Data Mining and Data Analysis using WEKA. *(ICETSTM – 2013) International Conference in "Emerging Trends in Science, Technology and Management-2013, Singapore*.

ELHASSAN, E. M. July 2014. Credit Scoring Using Data Mining Classification: Application on Sudanese Banks.

HAMID, A. J. & AHMED, T. M. March 2016. DEVELOPING PREDICTION MODEL OF LOAN RISK IN BANKS USING DATA MINING. *Machine Learning and Applications: An International Journal,* Vol.3, No.1.

HEMLATA SAHU, S. S., SEEMA GONDHALAKAR 2013. A Brief Overview on Data Mining Survey. *International Journal of Computer Technology and Electronics Engineering,* Volume 1.

IAN H. WITTEN, E. F. 2000. Practical Machine Learning Tools and Techniques with Java Implementations. *Morgan Kaufmann Publishers.***,** 2 6 5.

JONATHAN L. LUSTGARTEN, V. G., HIMANSHU GROVER, SHYAM VISWESWARAN. 2008. Improving Classification Performance with Discretization on Biomedical Datasets. *AMIA 2008 Symposium Proceedings Page - 448.*

KHAN, D. Z., KUMAR, A. & KUMAR3, S. May-Jun 2014. A Survey of Data Mining Concepts with Applications and its Future Scope. *International Journal of Computer Science Trends and Technology,* Volume 2.

KUMAR, R. & VERMA, D. R. August 2012. Classification Algorithms for Data Mining: A Survey. *International Journal of Innovations in Engineering and Technology,* Vol. 1.

MEENAKSHI, G. January 2014. Survey on Classification Methods using WEKA. *International Journal of Computer Applications (0975 – 8887).* Volume 86 – No 18.

MOIN, K. I. & AHMED, D. Q. B. 2012. Use of Data Mining in Banking. *International Journal of Engineering Research and Applications,* Vol. 2**,** pp.738-742.

MURAT KOKLU, K. S. September 2016. Estimation of Credit Card Customers Payment Status by Using kNN and MLP. *International Journal of Intelligent Systems and Applications in Engineering*.

NIKHIL N. SALVITHAL, R. B. K. October 2013. "Evaluating Performance of Data Mining Classification Algorithm in Weka". *International Journal of Application or Innovation in Engineering & Management (IJAIEM),* 2.

PRIYADHARSINI & THANAMANI, D. A. S. March 2014. an-overview-of-knowledge-discovery databaseand-data-mining-techniques. *International Journal of Innovative Research in Computer and Communication Engineering,* Vol.2.

PULAKKAZHY, S. & BALAN, R. V. S. 2013. DATA MINING IN BANKING AND ITS APPLICATIONS-A REVIEW. *Journal of Computer Science 9 (10): 1252-1259*.

REPOSITORY, M. L. 2005. default of credit card clients Data Set.

SRIVASTAVA., S. February 2014. Weka: A Tool for Data preprocessing, Classification, Ensemble,Clustering and Association Rule Mining. *International Journal of Computer Applications (0975 – 8887).* Volume 88 – No.10.

SUDHAKAR & REDDY, D. C. V. K. July-August 2014. CREDIT EVALUATION MODEL OF LOAN PROPOSALS FOR BANKS USING DATA MINING TECHNIQUES. *International Journal of Latest Research in Science and Technology,* Volume 3**,** Page No.126-131.

Appendix A.1

Weka

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University Of Waikato, New Zealand. Weka is free software available under the GNU General Public License. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality.

Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code.

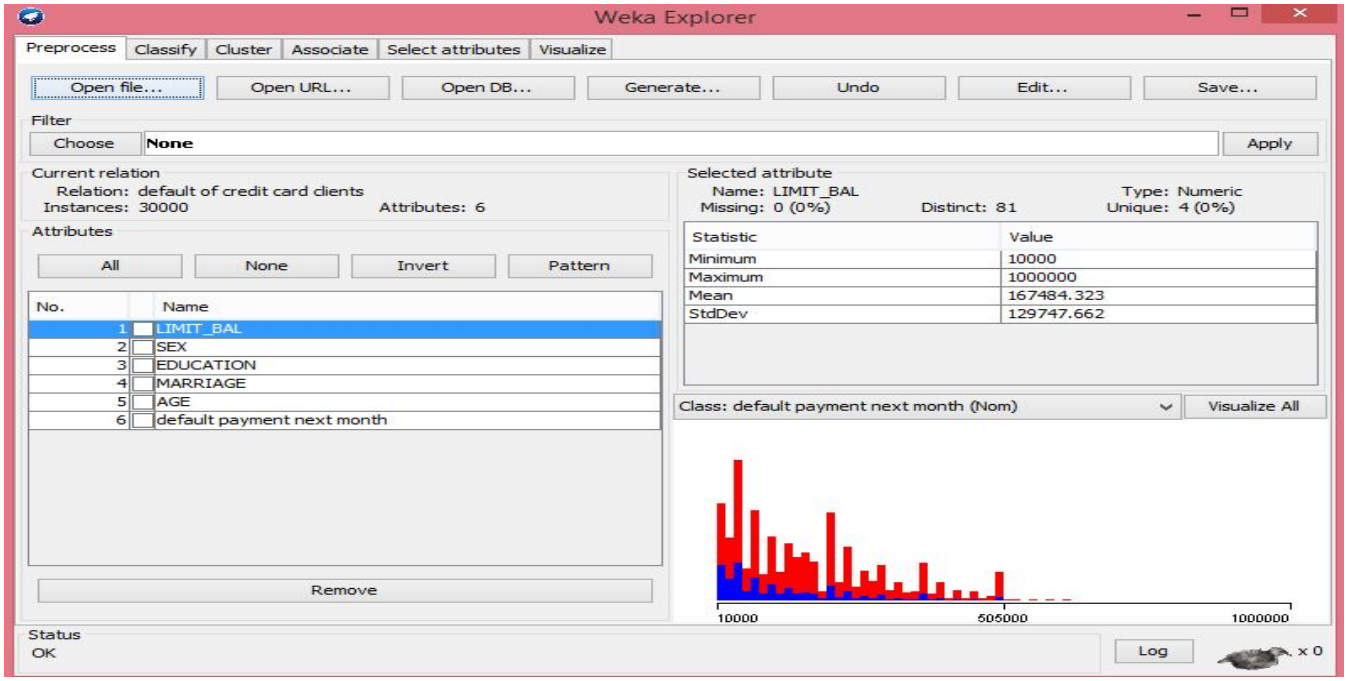
The original non-Java version of Weka was a TCL/TK front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Make file-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research.

Advantages of Weka

1. Free availability under the GNU General Public License.
2. Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
3. A comprehensive collection of data preprocessing and modeling techniques.
4. Ease of use due to its graphical user interfaces.

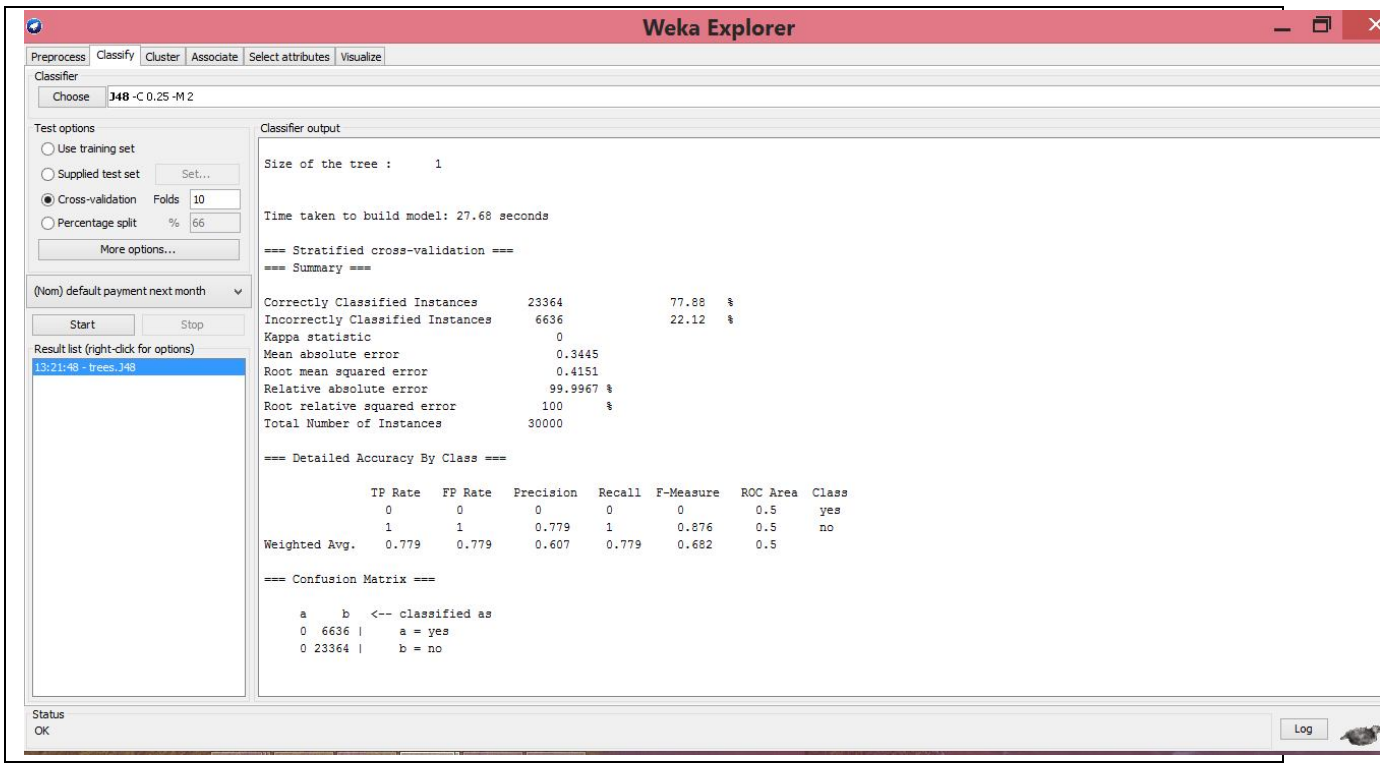
Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All techniques of Weka are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling (Dr. Sudhir B. Jagtap, 2013., Ian H. Witten, 2000).

Appendix A.2



Weka version 3.6.9 main page

Visualization for data set attributes distribution



J48 classifier with full dataset