T.C.

ISTANBUL ALTINBAS UNIVERSITY

GRADUATE SCHOOL OF SCIENCES ENGINEERING

HYBRID DATA MINING APPROACH TO PREDICT THE SUCCESS OF BANK TELEMARKETING

Anas Nabeel Falih AL-Shawi

Master of Information Technology

Thesis

Dr. Sefer Kurnaz

HYBRID DATA MINING APPROACH TO PREDICT THE SUCCESS OF BANK TELEMARKETING

by

Anas Nabeel Falih Al-Shawi

Information Technology

Submitted to the Graduate School of Science and Engineering

in partial fulfillment of the requirements for the degree of

Select Degree

ALTINBAŞ UNIVERSITY

2019

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of .....

Asst. Prof. Dr. Sefer KURNAZ

Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

Asst. Prof. Dr. Osman N. UCAN	School of Engineering and Natural Scince, Altinbaş University	_____
Asst. Prof. Dr. Sefer KURNAZ	School of Engineering and Natural Scince, Altinbaş University	_____
Asst. Prof. Name SURNAME	School of Engineering and Natural Scince, Altinbaş University	_____

I certify that this thesis satisfies all the requirements as a thesis for the degree of .....

Asst. Prof. Dr. OĞUZ ATA

Head of Department

Approval Date of Graduate School of Science and Engineering: ____/____/____

Asst. Prof. Dr.OĞUZ BAYAT

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Anas Nabeel Falih AL-Shawi

DEDICATION

First, I would like to thank Allah Almighty for the power of mind , health, strength , guidance , knowledge and skills to complete this study. This thesis is wholeheartedly dedicated to my beloved grandfather, who have been my source of inspiration, he tells me that" every success in your life will be best gift for me". To my grandmother, who have been supporting me with the kind and pure love. To my parents , there is no words to describe what you mean to me , there is nothing that I can repay for what you have done to me. I will continue to do my best to achieve your expectations. I dedicated this to cherished people who have meant and continue to mean so much to me . And lastly, to the family, relatives and friends who have been encouraging me during this study .

ABSTRACT

HYBRID DATA MINING APPROACH TO PREDICT THE SUCCESS OF BANK TELEMARKETING

Anas Nabeel Falih AL-Shawi

M.Sc., Information Technologies, Altınbaş University

Supervisor: Asst. Prof. Dr. Sefer KURNAZ

Date: March 2019

Pages: 43

Telemarketing is a kind of straightforward marketing in which salesman requests the consumer either face to face or telephone request and influence him to purchase the product. Telemarketing achieves most prevalence in the 20th century and still increasing it. Now, the phone has been broadly accepted. Business promotion and marketing is frequently based on an exhaustive understanding of actual information about the market and the real client demands for the productive bank manner. We recommend a data mining (DM) method to foretell the achievement of telemarketing requests for contracting long-term bank deposits. A local Portuguese bank was labeled, with data gathered from 2011 to 2016, thus involving the effects of the current economic crisis. We examined a comprehensive set of 11 features associated with bank consumer, goods and social-economic characteristics. We also discuss four DM forms with the hybrid model: Naïve Bayes (NB), Decision Trees (DTs), Perceptron Neural Network (NN) and Support Vector Machine (SVM). The four types were tested and compared with proposed hybrid classification methods (Perceptron Neural Network + Decision Tree) on an evaluation set, and we are splitting data into training and testing sets using cross-validation method. The proposed hybrid classification technique presented the best results (Precision 99% and ROC = 97%).

Keywords: *Data mining, Decision Tree, k-means, Support Vector Machine, bank telemarketing and neural network.*

Özet:

BANKADA TELEPAZARLAMANIN BAŞARISINI ÖNGÖRMEDE HİBRİT VERİ MADENCİLİĞİ YAKLAŞIMI

Anas Nabeel Falih AL-Shawi

Yüksek Lisans, Bilgi Teknolojileri, Altınbaş Üniversitesi

Danışman: Yrd. Prof. Dr. Sefer KURNAZ

Tarih: Mart 2019

Sayfalar: 43

Telepazarlama, satıcının müşteriyi yüzyüze veya telefon üzerinden ürünü alması için etkilediği basit bir çeşit pazarlamadır. 20.yüzyıldaki en fazla prevelansı elde etmekte ve hala arttırmaktadır.

Şimdilerde telefon yaygın olarak kabul görmektedir. Bu etkin bir değerdir ve müşterileri güncel tutar. Bankacılık alanında pazarlama, ürün veya hizmet alışverişinde kullanılan en önemli destektir. Şirketlerin tanıtımı ve pazarlaması, genellikle verimli bir banka anlayışı için esas piyasa bilgileri ve gerçek müşteri taleplerinin kapsamlı bir şekilde anlaşılmasına dayanmaktadır. Uzun vadeli banka mevduatları için telefonla pazarlama taleplerinin yerine getirilmesini öngörmede bir veri madenciliği (VM) yöntemi önerilmektedir. Yerel bir Portekiz bankası nitekim mevcut ekonomik krizin etkilerini içererek 2011'den 2016'ya kadar toplanan veriyle etiketlenmiştir. Banka müşterisi, ürün ve sosyo-ekonomik nitelikleri ile ilgili 11 özelliğin kapsamlı bir seti incelenmiştir. Ayrıca, dört VM formu hibrit modelle de tartışılmaktadır: Naïve Bayes (NB), Karar Ağaçları (KA), Algılayıcı Sinir Ağları (SA) and Support Vector Machine (SVM).

Dört tür, bir değerlendirme setinde önerilen hibrit sınıflandırma yöntemleriyle (Algılayıcı Sinir Ağı+ Karar Ağacı) test edilip karşılaştırılmıştır ve veri çapraz doğrulama yöntemi kullanarak eğitim ve test olmak üzere ikiye ayrılmıştır. Önerilen hibrit sınıflandırma tekniği en iyi sonuçları sunmuştur. (Doğruluk %99 ve ROC=%97).

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

1. INTRODUCTION

Marketing is a procedure of detecting the destination consumers to buy or make a deal with a product via fitting systems. It presently promotes the process to purchase the goods or service and even assists in planning the necessary for the product and convince customers to purchase it. The overall purpose is to enhance the selling of goods and services for the industry, marketing, and commercial institutions. It also accommodates to preserve the status of the business [1].

Telemarketing is a kind of straightforward marketing. Telemarketing achieves most prevalence in the 20th century and still increasing it. Now, the phone has been broadly accepted. It is valued efficient and holds the consumers up to date [1].

1.1 GENERAL INTRODUCTION

To be able to satisfy almost all customers there is the need for a competent system that would be used as a business tool to pinpoint, select, acquire and grow the loyal and most profitable customers. Such systems when available would help corporate organizations to analyze huge amounts of data and would be more helpful even when applied on transactional data. To expand customer acquisition and retention, organizations can also hit into the territory of unstructured data such as personal daily life, behaviors on social media, customer suggestions or feedbacks and support requests where conclusions are virtually impossible for the human experts. The marketing departments of organizations play significant role in the introduction of products and services to the outside world. The marketing department is an equally powerful section of any organization and makes significant use of organizational resources. So, the fact is if there can be a drastic decrease in the expenditure in the marketing department then the annual profits of the entire company can also increase. Also, banks and other financial institutions are faced with the problem of frauds and bad loans where loans granted by the banks are not paid back as agreed.

With this, it is quite obvious that marketing decisions on how to pinpoint customers who have high tendency of purchasing a loan so as to limit the number of contacts per person is very crucial for the bank in the district. In order to stay on top of the competition, this study is undertaken to use data mining to enhance the process of marketing and also to analyze loan assessment to prevent bad loans. Figure 1.1 is a graph depicting the number of approved loans between October, 2012 and September, 2015.

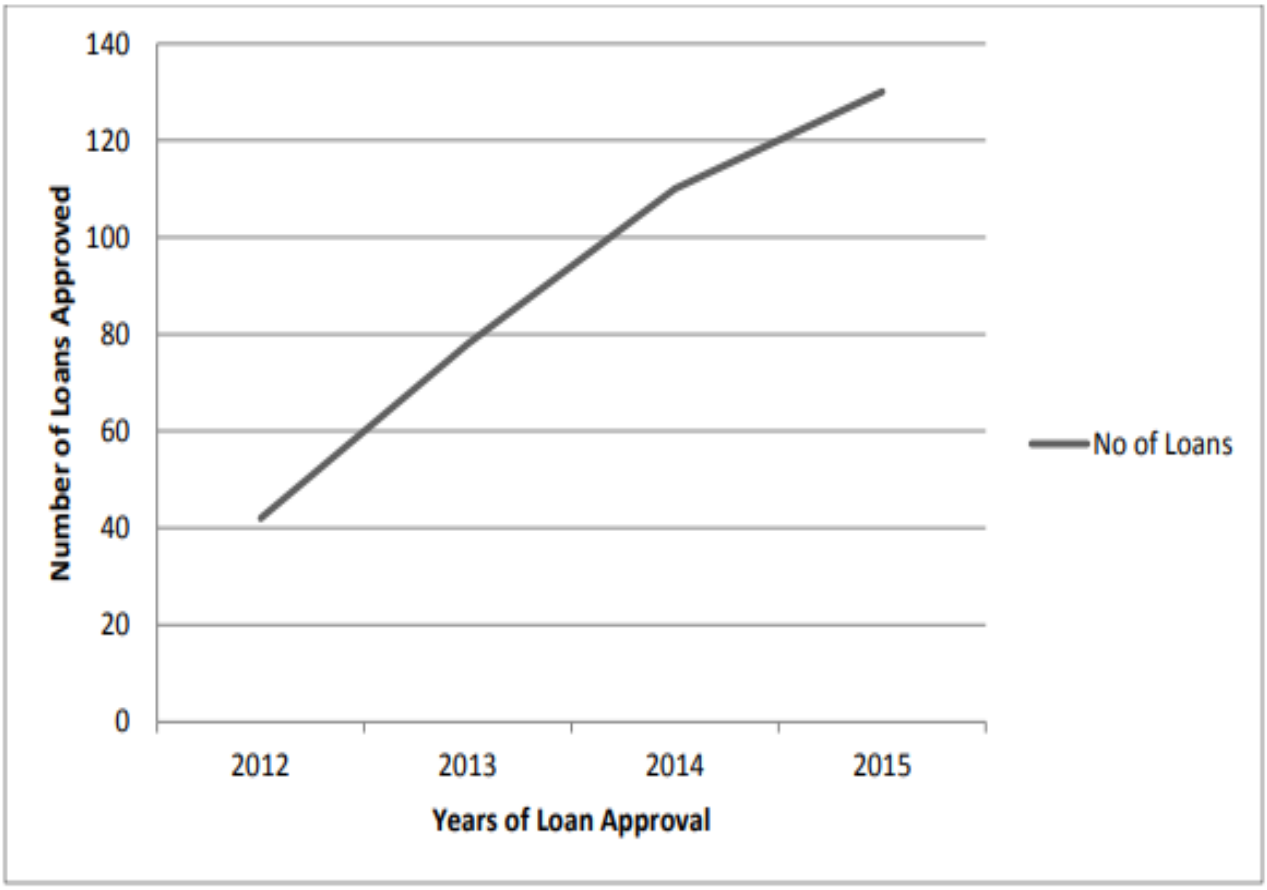


Figure 1.1 : The Number of Loans Approved (Oct 2012 - Sept 2015)

1.2 WHY DATA MINING

data is being and will always be produced regularly by organizations and DM has the power to make sense of these huge data. A significant benefit of Data mining is the discovering of hidden pattern and the identification of relationships among variables and also assisting the companies in seeing each of their customers individually. DM has the capacity to operate on data that are extracted from several sources which include but not limited to Web, databases and data warehouses. The concept of data warehousing has made it possible for the integration of data from different sources and data warehouses are programmed and configured to support data mining [1].

The tools and techniques available in DM could be used by the business to make helpful decisions. By using these data mining techniques, organizations and institutions can mine information that does not exist about their customers and products and by this can simply define the ideals of customers and forecast their future behaviors and requirements. Data mining uses several techniques to achieve its purpose. These techniques are grouped into Supervised and Unsupervised learning. Each of these tools consists of a number of algorithms and has its unique benefits whiles others cut across. In literature, prediction and classification techniques of data mining is found to be very helpful in any organization. Prediction has been proven to accurately predict the success or otherwise of a product by estimating the number and identifying the characteristics of customers who may probably be interested in that product. In a typical retail shop, transactions of every customer are stored in a database. Such database can be mined and prediction can be made on a similar product in terms of cross selling and up selling [2].

1.3 PROBLEM STATEMENT

The fact is, more businesses fail and are eventually shut down because of unsatisfactory customer services and more. Examples include Solyndra in America [3]. Business authorities spend time and huge sums of money to import products that are of low significance to their current customers. Many more financial institutions (credit unions, savings and loans, microfinance and banks) collapse because of high risk in granting loans to non-creditworthy customers and focusing attention on non-profitable customers.

1.4 DECISION SUPPORT SYSTEM

To obtain the best choices in organizational processes are sometimes established numerous difficulty where the quality of choice concerns. Decision Support Systems (DSS) are organized as a collection of automated events and terms that maintains the system or administration into their decision-making activities. The idea of DSS arises from stability which occupies between the data produced by the workstation and the decision of a human [2].

DSS applies statistical and mathematical processes to defeat the losses in data or knowledge and assists the decision producers to choose the best decision. Data mining (DM) represents a necessary function to maintain the DSS which are based on the data collected from the data mining forms: controls, guides, and connection. Data mining is the method of choosing, learning, and forming a large quantity of data and interpret undiscovered patterns. The purpose of data mining in DSS is to recommend a mechanism which is regularly obtainable for the market users to examine the data mining patterns [9].

A technology utilized inside the DSS is Machine learning (ML) that links data and learning on it to correctly foretelling the decisions. The basic principle of ML is to assemble the methods that can get input data and then forecast the results or outputs by applying the statistical interpretation within enough interval. ML allows the DSS to gain new experience which supports it to make the right decisions [2].

Machine Learning can be essentially distinguished into two categories, i.e. supervised training and unsupervised training. In supervised training, the producing of the algorithm is previously known, and we utilize the information data to foretell the result. Examples of supervised training are regression and classification. In opposite, unsupervised training we have input information whereas no same result variables are chosen. The form of unsupervised training is clustering.

Feature selection is the method of choosing the subset of relevant variables from the total features or patterns. It recognizes the most influential properties which accommodate to foretell the output. By using this method, we can decrease the dimensionality of properties, limit the scheme from overfitting and decrease the training time. In this process, the parsimonious form can be accomplished with a smallest amount of parameter with good performance in minimum time [3].

1.5 THESIS CONTRIBUTIONS

Inside our thesis we have focused on data mining classification methods these can forecast a certain consequence based on a specified input. We have utilized four classifiers and create comparative study to analyze a bank telemarketing dataset that recorded previously to take a decision. The main contribution of this study applies hybrid techniques (Neural network and Decision tree) in a proposed framework which had a highest accuracy to classify client's records.

The full model which is used in this study consists of 17 variables. Feature selection approach has been used to select the best subsets of variables and then different type of classification algorithms have been utilized to check their accuracy and performance. Several trials have been constructed to compare the accuracy of the implemented classifiers on a different size full training dataset with 11 attributes. Results showed that hybrid classification in proposed framework outperforms other classifiers with an accuracy rate of 97%, which provided a more effective and comprehensive classification mechanism than other classification techniques.

1.6 ORGANIZATION OF THESIS

The document is organized as follows:

Chapter 1 – introduction of decision support systems and data mining, i.e., gives an overview of the work, depict its importance, the contributions of this thesis, and the objectives behind the work.

Chapter 2 – The literature review which state the previous researches in the field of classification.

Chapter 3 – Comparative study introduced methods of datamining techniques: we describe the datasets, the different model development stages and the computational experiments performed.

Chapter 4 - discusses the research methodology. in this chapter we define the framework development phases, which includes data-preprocessing, the algorithms and the evaluation technique.

Chapter 5 – The results and discussions are obtainable with different model development stages and experimental discussion in this chapter.

Chapter 6 – this chapter represent the final conclusions and future work.

.

2. Related Works

Generally, there many data mining techniques that have been adopted in healthcare such as classification, clustering, association, and regression as shown in figure 2.1 A brief description about each one of them is provided next.

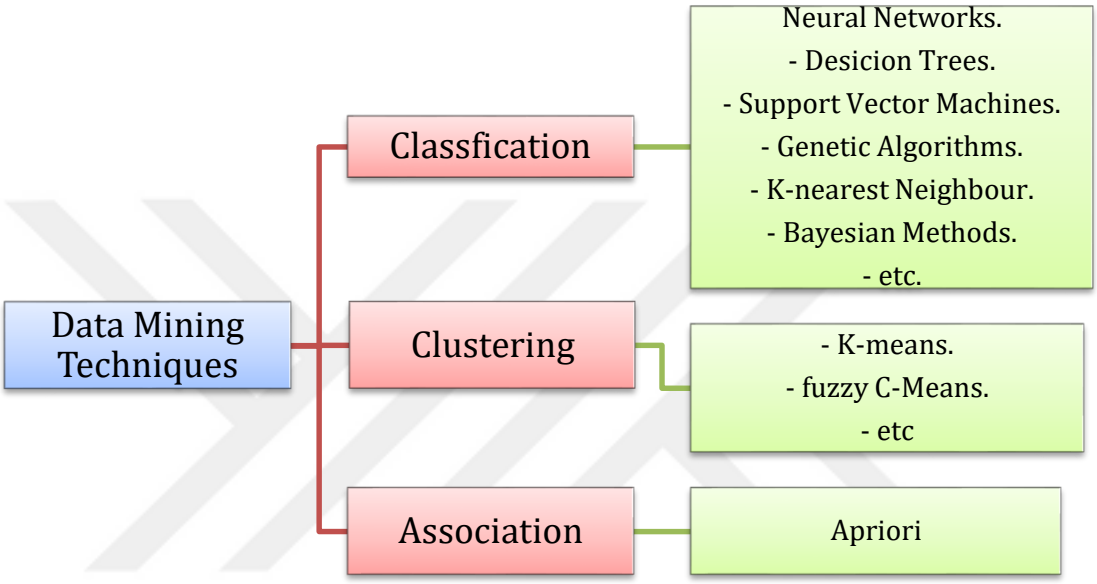


Figure 2.1: The various data mining methods applied in healthcare control.

2.1 CLASSIFICATION TECHNIQUES

Classification divides data samples into target classes. The classification technique predicts the target class for each data points. For example, customer can be classified as “yes, Take the loan” or “No, will not take the loan” using data classification approach. It is a supervised learning approach having known class categories. Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, “yes” or “no” [4].

Classification comprises of two footsteps. First step is model construction, which is used to analyze the training dataset of a database. The second step is model usage where the constructed model is used for classification. The accuracy of the classification is estimated according to the percentage of test samples or test dataset that are correctly classified [5]. Actually, there are a large group of techniques which have been used in bank management in order to perform the classification process which include: neural networks, decision trees, support vector machines, genetic algorithms, K-nearest neighbour, Bayesian methods, etc.

2.2 CLUSTERING

Clustering is a main task in data mining and a general technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bio-informatics [6]. Clustering is an unsupervised learning method that is different from classification. Clustering is unlike to classification since it has no predefined classes. In clustering large database are separated into the form of small different subgroups or clusters. Clustering partitioned the data points based on the similarity measure [4]. Clustering algorithms discovers collections of the data such that objects in the same cluster are more similar to each other than other groups. Many clustering algorithms used in healthcare management such as K-means, fuzzy C-Means, etc.

2.3 ASSOCIATION

Association Rule Mining appeared much later than machine learning and is subject to greater influence from the research area of databases. Although, association rule mining was first introduced as a market basket analysis tool, it has since become one of the most valuable tools for performing unsupervised exploratory data analysis over a wide range of research and commercial areas, including biology and bioinformatics [2]. Generally, Association is one of the most vital approaches of data mining that is used to find out the frequent patterns, interesting relationships among a set of data items in the data repository [4]. [5]. Aperiori algorithm is one of the association algorithms that is widely adopted in healthcare management.

2.4 REGRESSION

Regression is used to find out functions that explain the correlation among different variables. A mathematical model is constructed using training dataset. In statistical modeling two kinds of variables are used where one is called dependent variable and another one is called independent variable and usually represented using ‘Y’ and ‘X’. Regression is widely used in medical field for predicting the diseases or survivability of a patient [4]. Many regression algorithms used in healthcare management such as Logistic Regression Algorithm, Support Vector Regression, etc.

2.5 RELATED WORK BASED ON DATA MINING

This part describes the earlier study work which has been previously made in classification using ML methods. The data which is applied in this investigation work is the data of consumers of a Portuguese banking organization. In manuscript [5], this research proposed to obtain the form that can enhance the achievement rate of telemarketing for the bank. The statistical and analytical procedures of data mining which have been applied in their study are SVM, DT, and Naive Bayes. The achievement of these procedures was examined through the Receiver Operator Characteristics (ROC) curve. Between all these analytical procedures, DT comes up with the various useful effects.

In study [6], the investigation aimed to foretell the achievement of bank telemarketing. The data set which they applied in their study was included of 150 properties and is a comprehensive dataset of the term from 2008 to 2013. They examine four data mining procedures, i.e. Logistic Regression (LR), SVM, and ANN. The ANN achieved the highest result.

In the study [7], authors analyzed the fault diagnosis system for reciprocating compressors. Data was taken from oil corporation (5 years of operational data) and utilizes the SVM to examine it. They come up with the outputs that SVM correctly foretells the 80% correct classification of the possible mistakes in the compressor.

In a study [8], authors researched to foretell the growing of bankruptcy by using the SVM and RF techniques. The input was received from the Salomon Center database regarding North American from the period 1985 to 2013 with notes of more than 10,000. After implementing SVM and RF procedures, they analyze the outputs using discriminant analysis and logistic regression.

To find the risk factors about failure of banks Le & Viviani [11] conducted a research. In their study, a sample of 3000 US banks was analyzed by using 2 traditional statistical methods i.e. discriminant analysis and logistics regression. Then they compare these methods with the machine learning methods i.e. SVM, ANN and k-nearest neighbors.

3. Comparative Study

Data set which is utilized for this research has been taken from University of California, Irvine (UCI) machine learning repository website(http://archive.ics.uci.edu/ml) which is openly available for the public for research purpose. This data set works on the real information associated with the direct marketing campaigns of one of the most famous Portuguese retail banks, from May 2011 to November 2018, with a total of 45211 phone calls (and code notes during the call). The bank marketing was based on phone calls. In many cases, you can contact the same customer several times for the need to know more accurate information about the client and see if it will be taking a loan "Yes" or will not participate "No". A local Portuguese bank was labeled, with data gathered from 2011 to 2016, thus involving the effects of the current economic crisis. We examined a comprehensive set of 11 features associated with bank consumer, goods and social-economic characteristics.

3.1 DATASET DESCRIPTION

Table 3.1: Dataset Statistical Information

Customers Records	1000 Records
Number of Attributes	11 Attributes
Marital Status	220 Single 637 Married 143 Divorced
Education Status	155 Primary 570 Secondary 185 Tertiary 89 Unknown
Range of ages	22 – 61 years

In Table 3.1 we show all the information in the dataset , we use 1000 records 220 from them are single 673 are married and 143 are divorced , 11 Attributes ,155 had primary education 570 secondary education 185 College and 89 unknown , their age between 22 and 61 .

Table 3.2: Attributes Information of Dataset

age	numeric, age of client
job	categorical, type of job (admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services)
marital	categorical, marital status (married, divorced, single. Here "divorced" states the both divorced or widowed)
education	categorical (unknown, secondary, primary and tertiary)
default	binary, customer credit is in default (yes, no)
month	categorical, last contact month of the year
housing	binary, status of housing loan (yes, no)
loan	binary, client's personal loan (yes, no)
contact	categorical, contact communication type (unknown, telephone, cellular)
duration	numeric, last contact duration (in seconds)
pdays	numeric, number of days that passed by after the client was last contacted from a previous campaign

In Table 3.2 all the Attributes that we used in the dataset and we explain it in detail , its contain (age , job , marital , education , default , month , housing, loan , contact , duration , pdays) .

3.2 DATA MINING TECHNIQUES

Data mining is the unified name for all tools that can be used when searching for relationships and trends in large amounts of data, mainly used on data showing no such trends when judged by the human eye. The purpose of data mining is to be able to extract information from and make sense of large amounts of information.

The statistically significant relationships between data points that are extracted with data mining, often referenced as a classifier and will be from now on in this report, can be applied to new data and the chance of each outcome can be extracted.

The classifier is built using a set of training data. Training data is information with an already known outcome for the intended research. After the classifier has been defined, it is most importance that it is applied to data that was not a part of the training data. The evaluation could otherwise result in incorrectly high accuracy for the classifier. An example of training data could be a database containing patients and their medical records. The patients all have the same information recorded and not just sporadic information. The patients also have a yes/no diagnosis for a specific disease.

3.2.1 DECISION TREE ALGORITHM

Decision Trees is a technique very commonly used within data mining. The idea is to create a set of rules which can predict a specific question variable based on a set of input data. A Decision Tree consists of vertices and edges. The edges symbolize a road or a decision leading to the next vertices, possibly a pendant vertices (a pendant vertices is the vertices from which there are no further edges to travel), which could represent the next question or statement.

3.2.2 J48 ALGORITHM

J48 is an open source Java implementation of the C4.5 algorithm, used in the data mining software WEKA. The C4.5 algorithm is an extension to the ID3 algorithm and is used to initialize a Decision Trees that can be used for classification. This is the initialization algorithm used for all Decision Trees tests in this report.

3.2.3 NAIVE BAYESIAN ALGORITHM

Bayesian classification is used in data mining in a similar fashion to Decision Trees and is able to predict the probability of class membership. Bayesian classification is based on Bayes Theorem and is most commonly used in machine learning. There are several different versions of Bayesian classification where Naive-Bayes is the most common. It has been, and still is a very popular method to use when implementing spam filters and other types of text categorization.

3.2.4 SVM ALGORITHM

Support vector machines has been effectively utilized in numerous applications over the recent years. The great speculation capacity of SVM is that, it separate the two classes for which these are generally suitable. Among all the strategies of classification, the SVM has been broadly known. Usually, the performance time of a SVM model for classification takes longer and relates directly to the quantity of support vectors which can be sometimes challenging for some ongoing applications. For the purpose of classifying the data point, a SVM calculates the dot product for the given test point with each of support vector either in the feature space or in the space of input after the conversion through a function i.e. kernel.

3.3 WEKA

Waikato Environment for Knowledge Analysis (WEKA) is an open-source software written in Java developed by the University of Waikato in New Zealand. It is a machine learning software that has many pre-written algorithms, making it possible to use these algorithms without having to implement any algorithm code from scratch. WEKA has support Neural Network, Random Forest, Decision Trees (J48) and Naive Bayesian classification and is the software used to implement these algorithms in this report.

3.3.1 NAÏVE BAYES RESULT FROM WEKA PROGRAM

```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         946
Incorrectly Classified Instances        54
Kappa statistic                          0.4535
Mean absolute error                      0.0263
Root mean squared error                  0.1496
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.986    0.474    0.991      0.986    0.988      0.457    0.959     0.999     0
               0.526    0.014    0.417      0.526    0.465      0.457    0.959     0.270     1
Weighted Avg.  0.946    0.465    0.946      0.946    0.946      0.457    0.759     0.985
```

Figure 3.1 : Naive Bayes Result using WEKA program

İm Figure 3.6 we can seethe result of Naive Bayes with the accuracy of 75% and 829 of correct classified recods but it still to low since there is higher score using other algorithm .

Table 3.3:- Naïve Bayes Result from WEKA program

Result	Values
Correctly Classified Instances	946
In Correct Classified Instances	54
TP Rate	0.94
FP Rate	0.4
Precision	0.94
Recall	0.94
ROC Area	0.75
Time	0.03 Second

In the table 3.3, we discuss the Naïve Bayes algorithm's results. We applied cross validation for splitting records. We noticed that corrected classified records 946 and incorrect classified records 54 from total 1000 records. Also, we noticed that the accuracy ratio is very low with 75% but it's very fast. It is applied in 0.03 second.

3.3.3 DECISION TREE RESULT FROM WEKA PROGRAM

```
Time taken to build model: 0.4  seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         959
Incorrectly Classified Instances       41
Kappa statistic                          0.1523
Mean absolute error                      0.032
Root mean squared error                  0.133
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
                 0.996    0.895    0.983      0.996    0.989      0.179    0.887
                 0.105    0.004    0.333      0.105    0.160      0.179    0.887
Weighted Avg.    0.959    0.333    0.96       0.959    0.96       0.179    0.79
```

Figure 3.2 : Decision Tree Result using WEKA program

İm Figure 3.7 we can seethe result of Decision Tree with the accuracy of 79% and 959 of correct classified recods but it still to low since there is higher score using other algorithm .

Table 3.4:- Decision Tree Result from WEKA program

Result	Values
Correctly Classified Instances	959
In Correct Classified Instances	41
TP Rate	0.959
FP Rate	0.33
Precision	0.96
Recall	0.95
ROC Area	0.79
Time	0.4 second

In the table 3.4, we discuss decision tree algorithm's results. We applied cross validation for splitting records. We noticed that corrected classified records 959 and incorrect classified records 41 from total 1000 records. Also, we noticed that the accuracy ratio is low with 79% but it’s fast. It is applied in 0.4 second.

3.3.4 MULTILAYER PRECEPTRON (NEURAL NETWORK) RESULT

```
Time taken to build model: 12 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         979
Incorrectly Classified Instances        21
Kappa statistic                          0.3125
Mean absolute error                      0.0236
Root mean squared error                  0.1374
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.993    0.737    0.986      0.993    0.989      0.321    0.837     0.996     0
               0.263    0.007    0.417      0.263    0.323      0.321    0.837     0.299     1
Weighted Avg.  0.979    0.723    0.975      0.979    0.977      0.321    0.837     0.982
```

Figure 3.3 : Neural Network Result using WEKA program

İm Figure 3.8 we can seethe result of Neural Network Result algorithm with the accuracy of 83% and 987 of correct classified recods , it still to low since there is higher score using other algorithm but it still higer than Decision Tree and Naïve Bayes algorithm.

Table 3.5:- Multi-layer perceptron Result from WEKA program

Result	Values
Correctly Classified Instances	979
In Correct Classified Instances	21
TP Rate	0.979
FP Rate	0.72
Precision	0.975
Recall	0.979
ROC Area	0.83
Time	12 second

In the table 3.5, we discuss Support Vector Machine algorithm's results. We applied cross validation for splitting records. We noticed that corrected classified records 978 and incorrect classified records 21 from total 1000 records. Also, we noticed that the accuracy ratio is medium with 83% but it’s very slow. It is applied in 12 second.

3.3.5 RANDOM FOREST RESULT

```
Time taken to build model: 0.39 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         979
Incorrectly Classified Instances        21
Kappa statistic                          0.1523
Mean absolute error                      0.0275
Root mean squared error                  0.1221
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 0.996    0.895    0.983      0.996    0.989      0.179
                 0.105    0.004    0.333      0.105    0.160      0.179
Weighted Avg.    0.979    0.878    0.971      0.979    0.974      0.179
```

Figure 3.4 : Random Forest Result using WEKA program

İm Figure 3.9 we can see the result of Random Forest Result algorithm with the accuracy of 82% and 979 of correct classified recods .

Table 3.6: Random Forest Result from WEKA program

Result	Values
Correctly Classified Instances	979
In Correct Classified Instances	21
TP Rate	0.979
FP Rate	0. 878
Precision	0.971
Recall	0.979
ROC Area	0.82
Time	0.39 second

In the table 3.6, we discuss the Random Forest algorithm's results. We applied 10-fold cross validation for splitting records. We noticed that corrected classified records 979 and incorrect classified records 21 from total 1000 records. Also, we noticed that the accuracy ratio is medium with 82%. It is applied in 0.39 second.

3.3.6 SVM RESULT

Table 3.7: SVM Result from WEKA program

Result	Values
Correctly Classified Instances	967
In Correct Classified Instances	33
TP Rate	0.96
FP Rate	0.554
Precision	0.97
Recall	0.95
ROC Area	0.83
Time	0.55 second

In the table 3.7, we discuss the SVM algorithm's results. We applied 10-fold cross validation for splitting records. We noticed that corrected classified records 967 and incorrect classified records 33 from total 1000 records. Also, we noticed that the accuracy ratio is medium with 83%. It is applied in 0.55 second.

3.4 SUMMARY

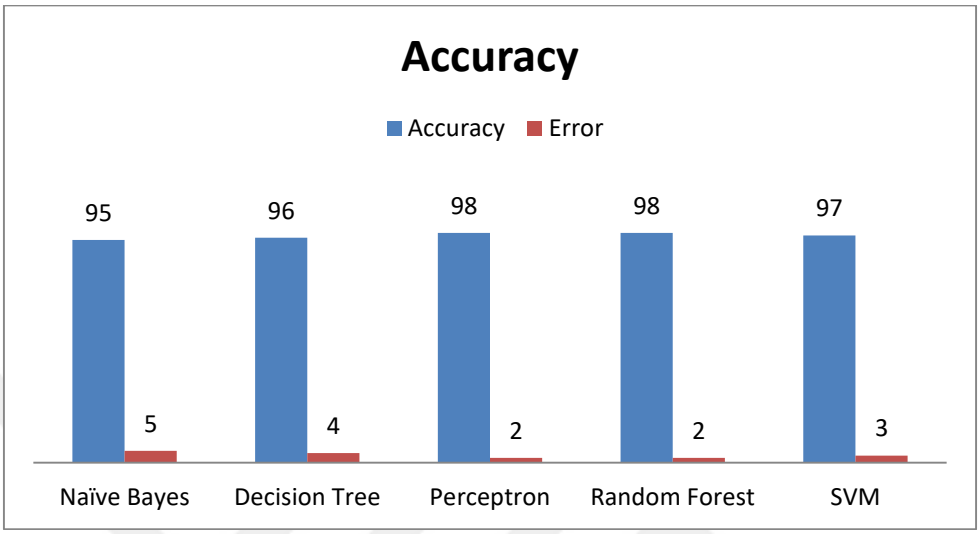


Figure 3.5 : Accuracy and Error Diagram of WEKA

In the Figure 3.10 we show the final result of the four algorithm that we use , like we see the error in the (naïve bayes) is al little bet high in comparison to the other algorithms ,
But at the final the error is too low for all algorithm .

4. Proposed Framework

In this chapter, the block design of the recommended mechanism is presented in Figure 4.1. We moved within two stages toward developing the recommended mechanism: data preprocessing and data classification processing. A separate subsection is dedicated to each stage.

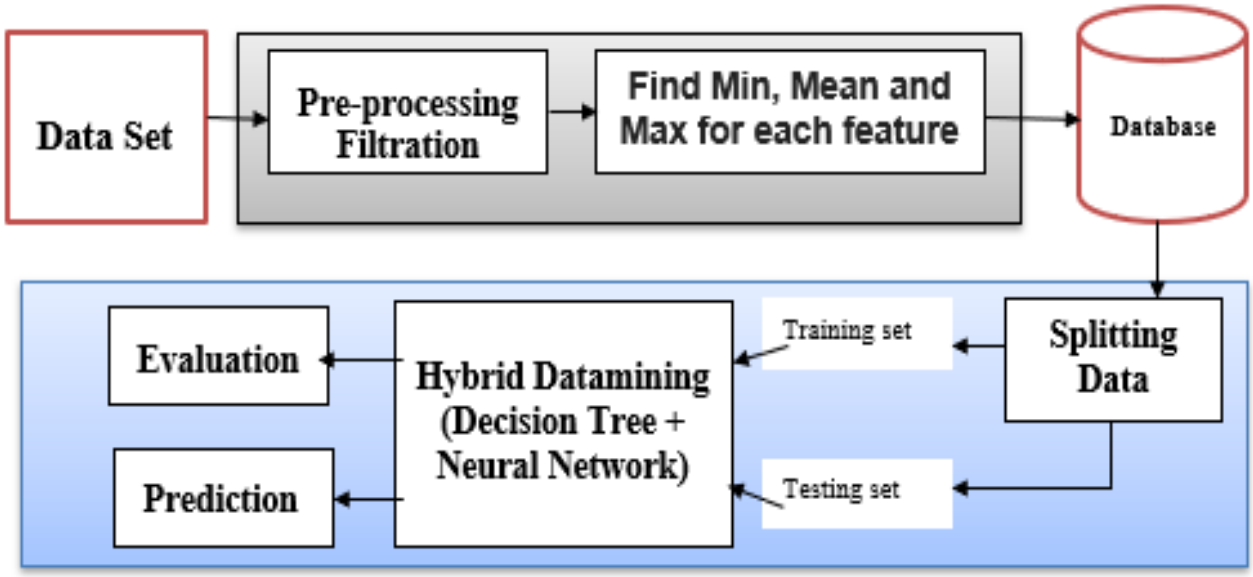


Figure 4.1:- Proposed Framework

Preprocessing (Steps)

1. Filter data (extract each feature) and remove incomplete records
2. Find minimum, average and maximum for each feature
3. Insert each record in database

Data processing (Steps)

1. Divide data into training set and test set
2. Apply different datamining techniques using cross validation
3. Apply hybrid of two best techniques according to accuracy

4.1 FRAMEWORK DESCRIPTION

We have designed a simple proposed framework for evaluating the hybrid algorithms as depicted in Fig. 4.1.

4.1.1 PRE-PROCESSING STAGE

A. Input Filtration
- Parsing Different Attributes from CSV File of data set
- Convert Attributes into suitable format
- Extract useful pattern and features

B. Statistical Processing
- Compute Min, Max, Mean and Standard Deviation

4.1.2 HYBRID CLASSIFICATION PROCESSING STAGE

A. Sampling or Splitting

Various classifiers elective involves distributing the training data in decreased equal subsections from data using the cross-validation procedure. Cross-validation, seldom called rotation evaluation, or out-of-sample measurement is any of several related form validation procedures for evaluating how the outcomes of the mathematical interpretation will conclude to an independent data set [10].

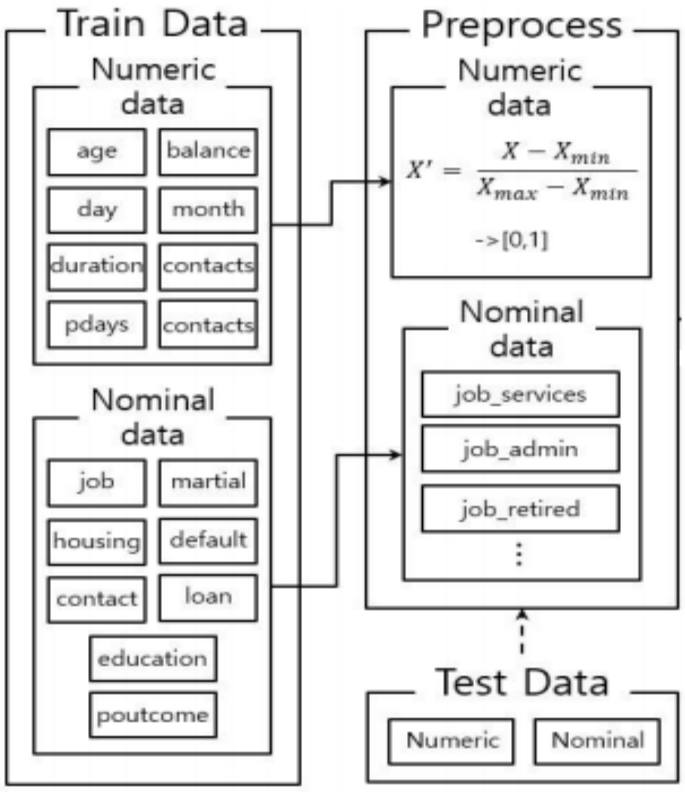


Figure 4.2 : Splitting Diagram Dataset Into Trainning and Testing Data

B. Perceptron Neural Network

Perceptron is an algorithm for supervised learning of binary classifiers. A binary classifier is a function which can decide whether an input, represented by a vector of numbers, belongs to some specific class. It is a type of linear classifier, i.e. a classification algorithm that makes its predictions based on a linear predictor function combining a set of weights with the feature vector [3]. Artificial Neural Network resembles the human brain in learning over the data storage and training. It is made and prepared over a particular input information training pattern. Through-out the procedure of learning, the results of NN is then matched to the target value and in this way, algorithm has been accomplished to reduce the error as minimum between the two values.

C. Decision Tree

In the decision tree method Figure 4.5, we need to pick the excruciating feature that reduces the value from entropy and exploiting the Information Gain. To recognize an excruciating feature from the Decision Tree, should compute the Information Gain to every feature also then choose a feature that exploits an Information Gain.

$$E = \sum_{i=1}^{k} -P_i \, log_2 \, P_i \qquad 1)$$

Where

- k is that value from classes of this objective feature
- Pi is a value of incidences from class i separated via the whole value from occurrences

```
duration <= 559: 0 (914.0)
duration > 559
|   loan = 0
|   |   balance <= 2248: 0 (74.0/10.0)
|   |   balance > 2248: 1 (4.0)
|   loan = 1
|   |   marital = single: 1 (2.0)
|   |   marital = married
|   |   |   balance <= 294: 0 (3.0)
|   |   |   balance > 294: 1 (2.0)
|   |   marital = divorced: 1 (1.0)

Number of Leaves  :     7

Size of the tree :      12
```

Figure 4.3 : Sample Screen Shots of Decision Tree Strucutre

4.2 PREPROCESSING IMPLEMENTATION USING PHP CODE

```php
function processingFile($array){
    $file = fopen("files/file.arff","w+");
    $firstLines="@relation bank_records";
        fwrite($file,$firstLines."\n");
        $firstLines="@attribute 'age' real";
        fwrite($file,$firstLines."\n");
        $firstLines="@attribute 'marital' {single,married,divorced}";
        fwrite($file,$firstLines."\n");
        $firstLines="@attribute 'education' {unknown,primary,secondary,tertiary}";
        fwrite($file,$firstLines."\n");
        $firstLines="@attribute 'default' {0,1}";
        fwrite($file,$firstLines."\n");
        $firstLines="@attribute 'balance' real";
        fwrite($file,$firstLines."\n");
        $firstLines="@attribute 'housing' {0,1}";
        fwrite($file,$firstLines."\n");
        $firstLines="@attribute 'loan' {0,1}";
        fwrite($file,$firstLines."\n");
        $firstLines="@attribute 'job' {management,technician,entrepreneur,blue-colla
        fwrite($file,$firstLines."\n");
        $firstLines="@attribute 'duration' real";
        fwrite($file,$firstLines."\n");
        $firstLines="@attribute 'y' {0,1}";
        fwrite($file,$firstLines."\n");
        $firstLines="@data";
        fwrite($file,$firstLines."\n");

        for($i=0;$i<count($array);$i+=10){
            $firstLines = $array[$i].",".$array[$i+1].",".$array[$i+2].",".$array[$i
            fwrite($file, $firstLines."\n");
```

Figure 4.4 : Screen shot of PHP Preprocessing code

We use the php program to make a small program that weuse it to see the result and the work more easer and we can use any dataset in the future , in figure 4.6 we show the code we use in preprocessing implementation php program .

5. Experimental Result

In the previous chapter, the proposed hybrid ensemble classification algorithm (Neural Network+ Decision Tree) was introduced. In this chapter, we need comparison between the proposed hybrid ensemble classification algorithm and existing algorithms is required to prove its efficiency. Finally, the testing of proposed framework is required to prove its reliability, usability and efficiency.

5.1 DEVICE AND TOOL CAPABILITIES

Table 5.1: Tools and device used to preform proposed framework

Metric	Values
CPU	Intel core i7
RAM	4G
Operating system	Windows 10
Programming Language	PHP v4
Server Platform	Apache server

We needed a high device for our work to get a great result , in table 5.1 we show the requirement for the device that we use in our work .

5.2 UPLOADING SCREEN SHOTS

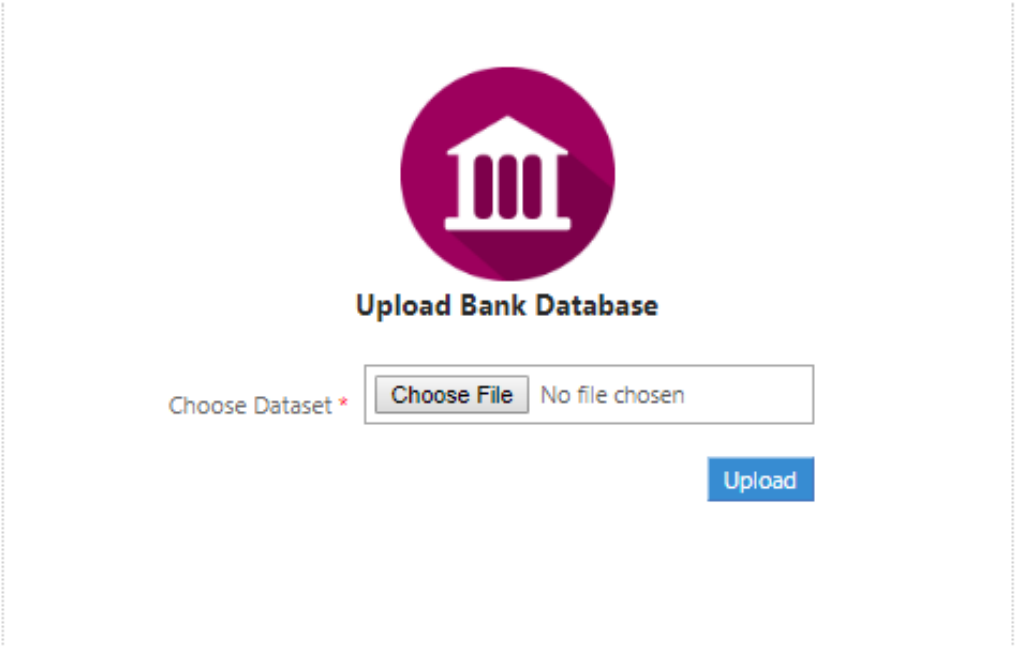


Figure 5.1: Screen shot of Uploading Process in Our System

Like we said befor we made a php program to make the work more easier , so in figure 5.1 is show the first page in the program .

İn choose file we choose the dataset from the net or the computer and then we make an upload .

+ Options

	id	age	job	marital	education	defaultt	balance	housing	loan	contact	dayy	monthh	duration	campaign	pday
Edit Copy Delete	25191	58	management	married	tertiary	0	2143	1	0	unknown	5	may	261	1	-
Edit Copy Delete	25192	44	technician	single	secondary	0	29	1	0	unknown	5	may	151	1	-
Edit Copy Delete	25193	33	entrepreneur	married	secondary	0	2	1	1	unknown	5	may	76	1	-
Edit Copy Delete	25194	47	blue-collar	married	unknown	0	1506	1	0	unknown	5	may	92	1	-
Edit Copy Delete	25195	33	unknown	single	unknown	0	1	0	0	unknown	5	may	198	1	-
Edit Copy Delete	25196	35	management	married	tertiary	0	231	1	0	unknown	5	may	139	1	-
Edit Copy Delete	25197	28	management	single	tertiary	0	447	1	1	unknown	5	may	217	1	-
Edit Copy Delete	25198	42	entrepreneur	divorced	tertiary	1	2	1	0	unknown	5	may	380	1	-
Edit Copy Delete	25199	58	retired	married	primary	0	121	1	0	unknown	5	may	50	1	-
Edit Copy Delete	25200	43	technician	single	secondary	0	593	1	0	unknown	5	may	55	1	-
Edit Copy Delete	25201	41	admin.	divorced	secondary	0	270	1	0	unknown	5	may	222	1	-
Edit Copy Delete	25202	29	admin.	single	secondary	0	390	1	0	unknown	5	may	137	1	-
Edit Copy Delete	25203	53	technician	married	secondary	0	6	1	0	unknown	5	may	517	1	-
Edit Copy Delete	25204	58	technician	married	unknown	0	71	1	0	unknown	5	may	71	1	-
Edit Copy Delete	25205	57	services	married	secondary	0	162	1	0	unknown	5	may	174	1	-
Edit Copy Delete	25206	51	retired	married	primary	0	229	1	0	unknown	5	may	353	1	-
Edit Copy Delete	25207	45	admin.	single	unknown	0	13	1	0	unknown	5	may	98	1	-
Edit Copy Delete	25208	57	blue-collar	married	primary	0	52	1	0	unknown	5	may	38	1	-
Edit Copy Delete	25209	60	retired	married	primary	0	60	1	0	unknown	5	may	219	1	-
Edit Copy Delete	25210	33	services	married	secondary	0	0	1	0	unknown	5	may	54	1	-

Figure 5.2 : Screen shot of Database System

After we chosse the dataset we see it like the figure 5.2 and it shown as the data base with all the artibute .

5.3 PRE-PROCESSING STAGE SCREEN SHOTS

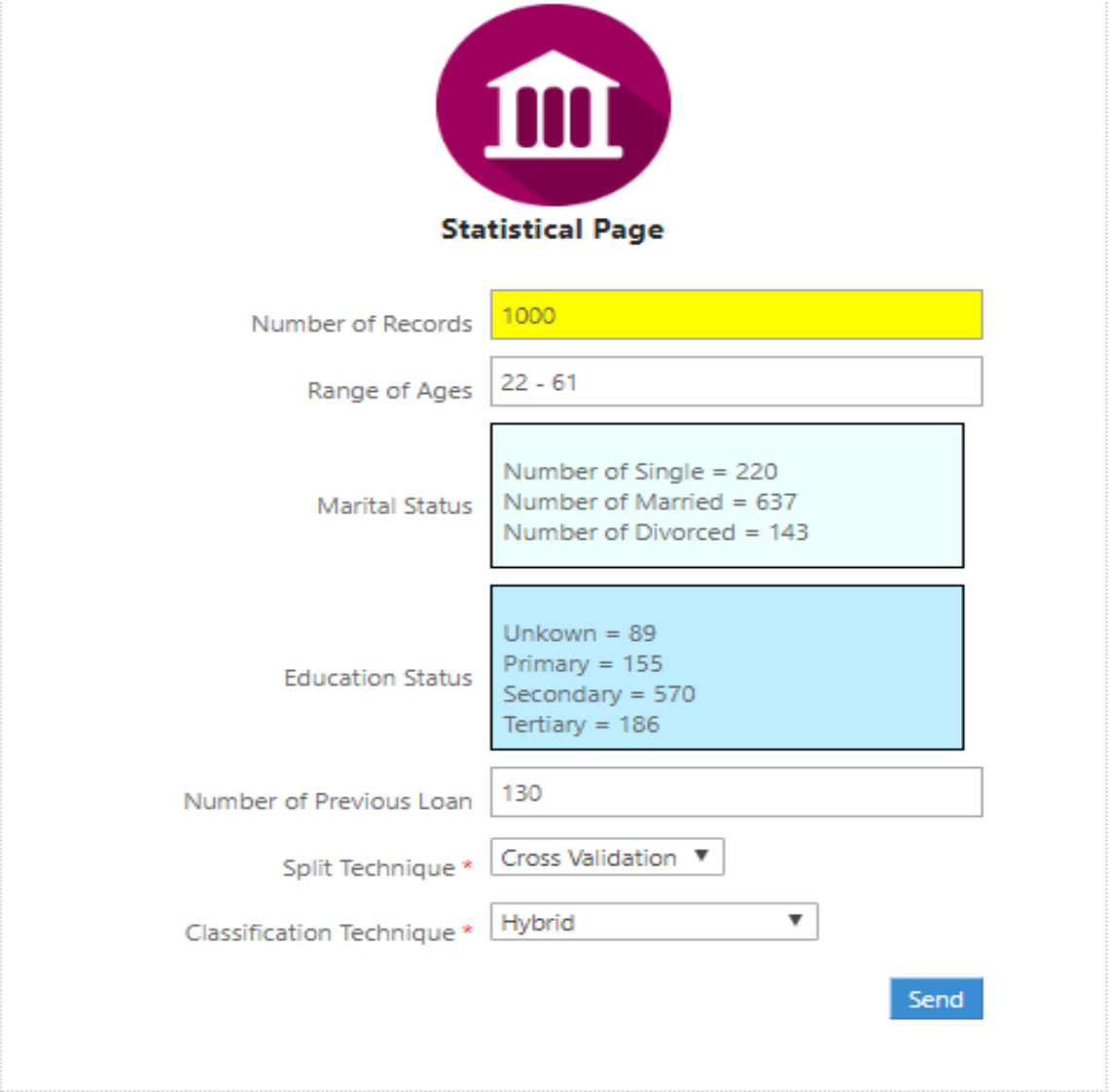


Figure 5.3: Screen shot of Preprocessing Process in Our System

When we complete choosing the data set and all the algorithem we see the ruselt just like shown in figure5.3 , first row is show the number of the record that we enter , the scond row rage of age from dataset ,and the other one is about the marital state for each person is he marrge or single or divorsed , the the education state .

After that we choose the algorthim that we want like its shown .

5.4 DECISION TREE TECHNIQUES SCREEN SHOTS

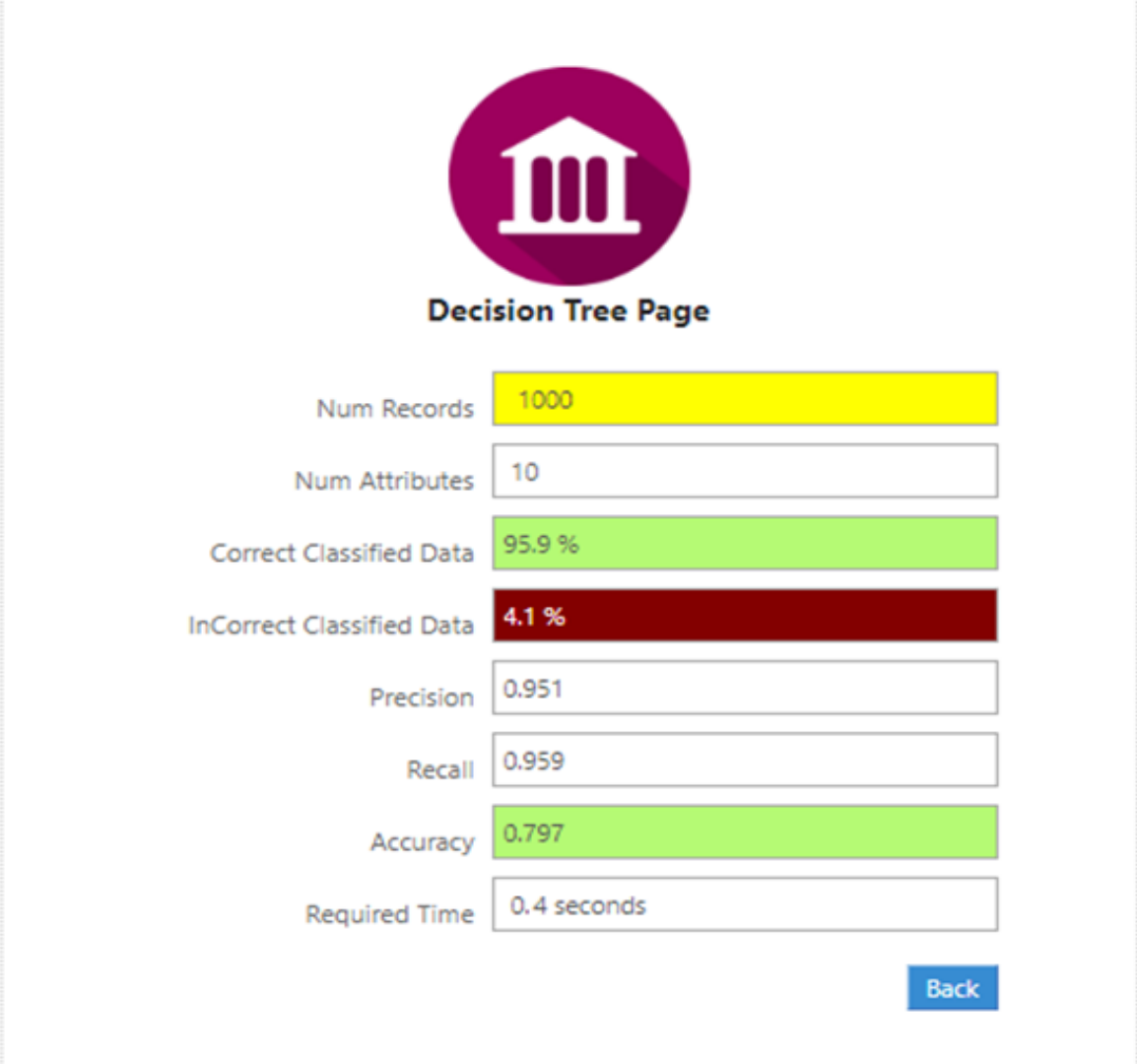


Figure 5.4 : Screen shot of Decision Tree Process in Our System

After we choose the algorithm the result will appear , in figure 5.4 we choose the Decision tree algorithm and it shown the result , the correct classified data 95.9% , incorrect classified data 4.1% , the accuracy is 0.787 in 0.4 seconds .

5.5 ARTIFICIAL NEURAL NETWORK TECHNIQUES SCREEN SHOTS

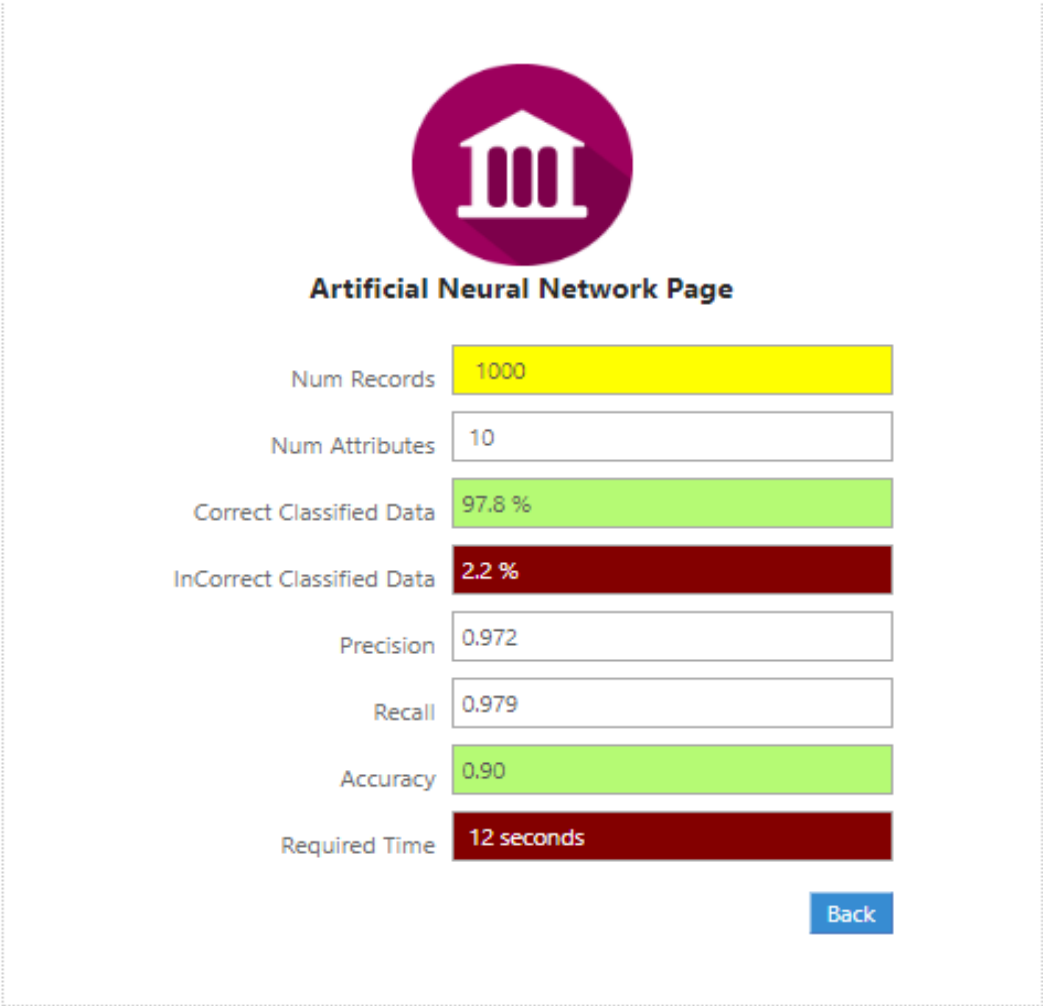


Figure 5.5 : Screen shot of Neural Perceptron Network Process in Our System

In figure 5.5 we choose the Neural perceptron network algorithm and it shown the result , the correct classified data 97.8% , incorrect classified data 2.2% , the accuracy is 0.90 in 12 seconds.

5.6 HYBRID TECHNIQUES SCREEN SHOTS

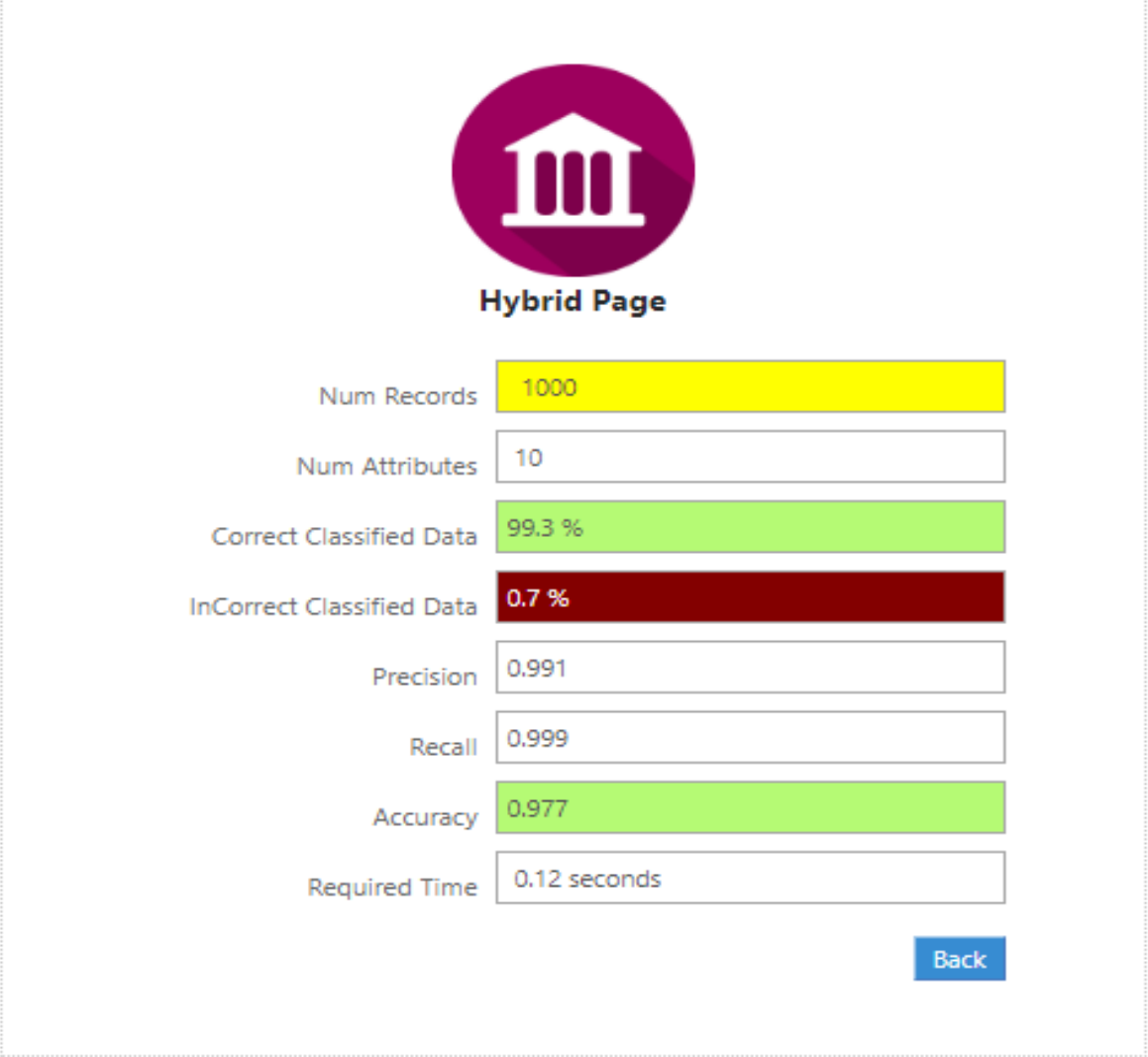
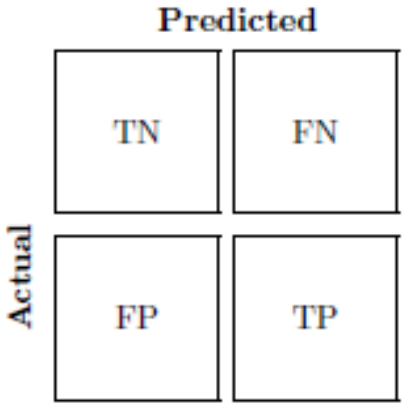


Figure 5.6 : Screen shot of Hybrid Process in Our System

In figure 5.6 we choose the Hybrid process between the Neural Perceptron Network algorithm and Decision Tree ,and the result is correct classified data 99.3% , incorrect classified data 0.7% , the accuracy is 0.977 in 0.12 seconds .

5.7 EXPERIMENTAL RESULT AND DISCUSSION

Table 5.2 : Confusion Matrix



Cross validation is a way to assess the performance of a prognostic model. In general the k-fold cross-validation is used, which is a method of partitioning the original data into k subsamples of equal size. At that point, single sub sample is examined as the validation of data for the purpose of testing the model while the rest of the samples are then utilized as training data. This procedure is then repetitive k times and every k sub sample is then utilized precisely one as per validation data. An average result is generated from the ten runs. The performance of the used algorithms has been evaluated using in terms of correctly classified instances, incorrectly classified instances, TP rate, FP rate, precision, recall, ROC area, CPU Time, accuracy, error. We used different measures to evaluate the methods adopted on the heart diseases dataset as shown below:

- $\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \times 100\%$
- $\text{Specificity} = \frac{\text{TN}}{\text{FP+TN}} \times 100\%$
- $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+ FN}} \times 100\%$

Table 5.3 : Compartive Reuslt between Proposed hybrid algorithm and other algorithms

Classifier	**Sensitivity**	**Specificity**	**Accuracy**	**Time in seconds**
Decision Tree	95%	96%	79%	0.4
Naïve Bayes	94%	95%	75%	0.03
SVM	97%	95%	83%	0.55
Perceptron	97%	98%	90%	12
Hybrid Perceptron + Decision Tree	99%	99%	97%	0.12

We summarize our previous result in clear table 5.3 which present 4 principal argument Sensitivity Specificity Accuracy and time in second .

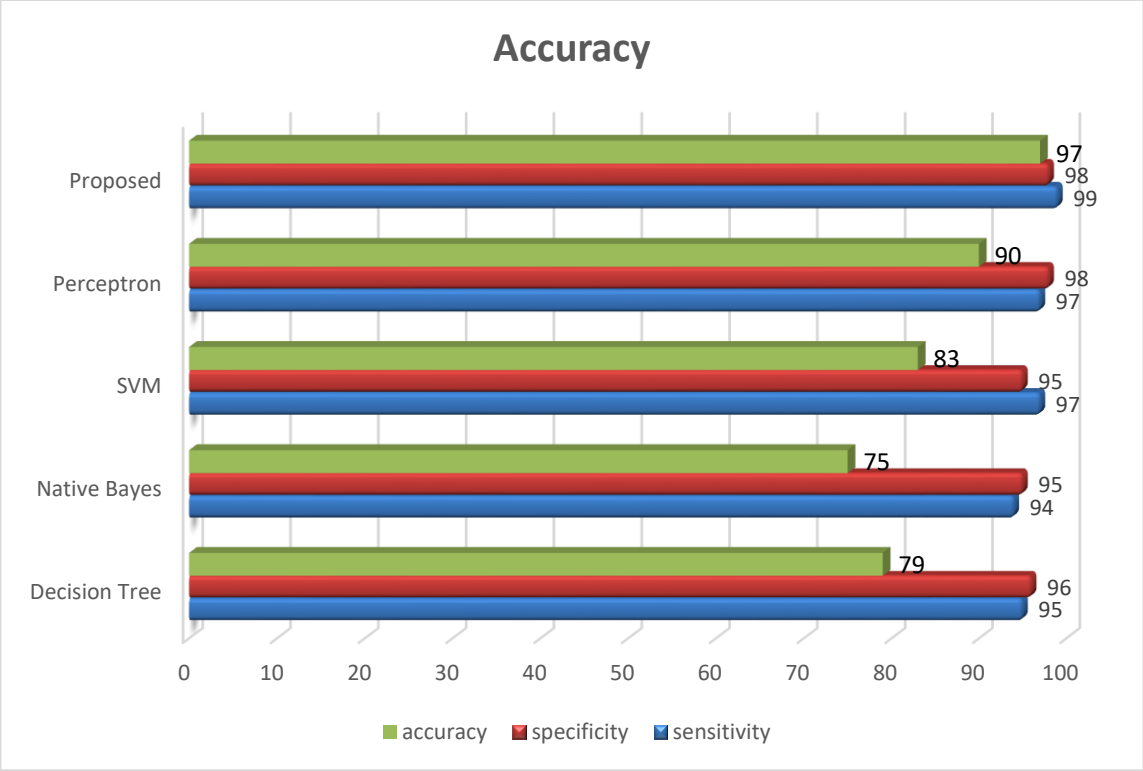


Figure 5.7 : Performance Diagram between proposed hybrid classification method and other methods

In figure5.8 we show all accuracy between all the algorithm , like it appear the Hybrid algorithm has the highest accuracy form all the other algorithm 97% , the perceptron algorithm has 90% , svm has 83% , decision tree 79% , and the lowest accuracy is the native bayes is 75 % .

6. Conclusion

Within the banking enterprise, optimizing targeting for telemarketing is a vital problem, under increasing stress to improve earnings and decrease losses. Inappropriate, Portuguese banks were constrained to develop capital elements (e.g., by catching extra long-term deposits). In this research, we recommend a particular and smart DSS that applies a data mining (DM) procedure for the determination of bank telemarketing consumers. We examined a current and sizeable Portuguese bank dataset, with a whole of 1000 records. We picked a standardized set of 11 related features. Also, four DM procedures were analyzed: naïve Bayes (NB), decision trees (DTs), perceptron neural networks (NNs) and support vector machines (SVMs). These models were compared with proposed hybrid classification methods (Decision Tree + Neural) using four metrics, Precision, recall, ROC and time. For both parameters and phases, the best results were obtained by the hybrid techniques, which resulted in a ROC of 97%.

REFERENCES

1. Sadaf Hossein Javaheri, Mohammad Mehdi Sepehri, Babak Teimourpour, Response modeling in direct marketing: a data mining based approach for target selection, Data Mining Applications with R, Elsevier,2014, pp. 153–178.

2. Rupnik, R. & Kukar, M. (2007), `Decision support system to support decision processes with data mining', Journal of information and organizational sciences 31(1), 217-232.

3. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. (2010), `Variable selection using random forests', Pattern Recognition Letters 31(14), 2225-2236.

4. Izetta, J., Verdes, P. F. & Granitto, P. M. (2017), `Improved multiclass feature selection via list combination', Expert Systems with Applications 88, 205-216.

5. Moro, S., Cortez, P. & Rita, P. (2014), `A data-driven approach to predict the success of bank telemarketing', Decision Support Systems 62, 22-31.

6. Moro, S., Laureano, R. & Cortez, P. (2011), Using data mining for bank direct marketing: An application of the crisp-dm methodology, in `Proceedings of European Simulation and Modelling Conference-ESM'2011', Eurosis, pp. 117-121.

7. Qi, G., Zhu, Z., Erqinhu, K., Chen, Y., Chai, Y. & Sun, J. (2018), `Fault-diagnosis for reciprocating compressors using big data and machine learning', Simulation Modelling Practice and Theory 80, 104-127.

8. Barboza, F., Kimura, H. & Altman, E. (2017), `Machine learning models and bankruptcy prediction', Expert Systems with Applications 83, 405-417.

9. Le, H. H. & Viviani, J.-L. (2017), `Predicting bank failure: An improvement by implementing machine learning approach on classical financial ratios', Research in International Business and Finance.

10. Rohani, A., Taki, M. & Abdollahpour, M. (2018), `A novel soft computing model (gaussian process regression with k-fold cross validation) for daily and monthly solar radiation forecasting (part: I)', Renewable Energy 115, 411-422.

11. Le, H. H. & Viviani, J.-L. (2017), `Predicting bank failure: An improvement by implementing machine learning approach on classical financial ratios', Research in International Business and Finance .

12. Moro, S., Cortez, P. & Rita, P. (2014), `A data-driven approach to predict the success of bank telemarketing', Decision Support Systems 62, 22{31.