Republic of Iraq

Ministry of Higher Education and
Scientific Research
University of Babylon
College of Information Technology
Software Department

A Predictive Modeling Approach to Improve the Banking Operations

A Thesis

Submitted to the Council of the College of Information Technology for Postgraduate Studies of the University of Babylon in Partial Fulfillment of the Requirements for the Degree of Master in Information Technology – Software

By

Taif Ali Talib Jawad

Supervised By

Prof. Dr. Saad Talib Hasson

2023 A. D. 1445 A. H.

بسم الله الرحمن الرحيم

ٱقۡرَأۡ بِٱسۡمِ رَبِّكَ ٱلَّذِى خَلَقَ ١

خَلَقَ ٱلۡإِنسَٰنَ مِنۡ عَلَقٍ

٢ ٱقۡرَأۡ وَرَبُّكَ ٱلۡأَكۡرَمُ ٣

ٱلَّذِى عَلَّمَ بِٱلۡقَلَمِ ٤

عَلَّمَ ٱلۡإِنسَٰنَ مَا لَمۡ يَعۡلَمۡ ٥

صَدَقَ اللهُ العَظِيمْ

سُورَةُ العَلَقِ

الآيات (1-5)

Supervisor Certification

I certify that this thesis entitled

A Predictive Modeling Approach to Improve the Banking Operations

written by

Taif Ali Talib

was prepared under my supervision at the department of Software / College of Information Technology / the University of Babylon as partial fulfillment of the requirements of the degree of Master in Information Technology - Software.

Signature:

Name: Prof. Dr Saad Talib Hasson

Date: / / 2023

Head of the Department Certification

In view of the available recommendations, I forward the thesis entitled "A Predictive Modeling Approach to Improve the Banking Operations" for debate by the examining committee.

Signature:

Assis. Prof. Dr. Ahmed Saleem Abbass

Head of Software Department

Date: / / 2023

Acknowledgements

First of all, I thank God who inspired me with patience and strength to complete this study.

It is not easy except what God makes easy.

I would like to express my sincere thanks and appreciation to Supervisor

Prof. Dr. Saad Talib Hasson

To direct and follow up on it and provide important advice and suggestions for improving this study

My sincere thanks and appreciation to the Dean of the University of Babylon and to the academic and administrative faculty.

I would like to thank the Head of the Department of Software and her academic and staff colleagues.

In the end, I thank everyone who wished me success.

Finally, I apologize to those whose names were not mentioned. .But I am grateful to all of them for their help

Dedication

Praise be to Allah always and forever. Praise be to Allah who we think is good and he honors us with something better than him. Thank Allah for my success in every step of my life and passing my studies with success and excellence...Thank Allah for achieving one of the goals of my father, who was and still is the light of my path and my eyes, may Allah have mercy on you A piece of my heard. My thanks to the first supporter and the true supporter of the human being. If I gave my soul to her, I would not reward a little for a little

My thanks to my dear husband, if it were not for him, I would not have reached this stage. My thanks to my beloved sisters and everyone who stood and supported. My thanks for the gift of studying, my beloved friends.

Abstract

A significant tool for banks and other financial institutions is predictive modeling. In order to determine the realistic possibilities for future outcomes, a method of evaluating the banking data and predicting specific probabilities must be followed. Predictive modeling is a method for using the present data to create an appropriate model to forecast the results of future data.

Examining how successfully machine learning algorithms predict whether a new customer will have a term deposit or not is the goal of this study. It can be used to figure out the best strategy to spot banking company consumers that frequently leave. The primary goal of this thesis is to create a model with precise forecasting capabilities to improve bank operations. This goal can be accomplished with the least amount of error by choosing the most crucial features, along with (Collect a better understanding of the customer's needs based on analyzing different banking datasets, perform feature ranking to indicate the features effectiveness, and indicate the features correlation to show the reduction possibilities).

The data preparation stage and the prediction stage are the two key stages of the proposed system. The suggested model's prediction accuracy is increased by doing data pretreatment utilizing the Data Cleaning (Missing values), Data Transformation (Nominal to Binary, Nominal to Numeric), Normalization (Standardization, Smote), and Data Reduction procedures. The ranking correlation coefficient approach is one of the data reduction techniques. In order to assure the significance of these features and the accuracy that can be attained, the suggested system uses the ranking correlation coefficient approach, which identifies the most advantageous features at each step before incorporating them into the model.

Moreover. This thesis employs statistical techniques, machine learning algorithms, and regression analysis to use and evaluate two datasets (a bank-additional-full and Banking Dataset - Marketing Targets datasets). Predictive modeling techniques include Naive Base, Decision Tree, KNN, and Logistic Regression. Accuracy, precision, recall, F-measure, and error were used as the basis for the evaluation. The outcomes demonstrated that the performance of the suggested system is efficient, DT

(0.9336), KNN (0.9285), Logistic (0.9102), and Naive Bays (0.8606) had the highest prediction accuracy.

List of Contents

List of Figures

List of Table

List of Algorithms

List of Abbreviations

Abbreviation	Meaning
R	Correlation
DS	Dataset
DT	Decision tree
ERR	Error rate
FN	False Negative
FP	False Positive
KNN	K-nearest neighbour
LG	Logistic Regression
NB	Naïve Bayes
ROC	Receiver Operating Characteristics
Z	Standardization
SMOTE	Synthetic Minority Over-sampling Technique
TN	True Negative
TP	True Positive

Chapter One

General Introduction

Chapter One
General Introduction

1.1 Introduction

Predictive Analytics can be identified as a stream of analytics based on historical data to predict future behavior, action, and trends. It involves modeling, data mining, statistics, and machine learning algorithms to create certain expectations [Kuhn, M., & Johnson, K. (2013)]. Predictive Analytics can assist in estimating different required real time issues or report them at the accurate time to develop the suitable correct outcomes. The availability of the information through the media and interne, increase the significant desire of how to utilize the information in making trustworthy decisions. The main concern of predictive modeling is to create precise predictions, and to understand the model and to interpret how it operates [Kumar, et, al, (2018)]. An essential measure to develop the banking operations and also reestablish customers' satisfaction, effectiveness and performance analysis in a bank has got an increase attention [Barga, et, al, (2015].

Predictive Modeling is the essential key to solve all the commercial problems but it can be utilized appropriately by aligning the technology with the commercial objectives. It used to prevent frauds in banking services. Decision Support Systems are based on different statistical and computational methods to assist management in their strategic and tactical decisions [Ali, et, al, (2021)]. To make good predictions, investigators must completely understand the use of the possible techniques when dealing with banking data.

Banking operations involve a wide range of activities, such as customer acquisition, credit risk assessment, fraud detection, and customer retention. In this thesis predict whether a new customer will have a term deposit or not, With the increasing availability of data and advancements in predictive modeling techniques, banks can leverage data-driven approaches to enhance their operations and decision-making processes [Gavurova, et, al, (2017)].

Predictive modeling refers to the use of statistical and machine learning techniques to analyze historical data and make predictions about future outcomes. By employing predictive modeling in banking operations, banks can gain valuable insights into customer behavior, assess risks more accurately, and optimize their operations for better efficiency. Predictive modeling can help banks identify potential customers who are likely to be interested in their products or services [Demraoui, et, al, (2022)]. Analyzing customer data may including transaction history and preferences, predictive modeling can enable banks to offer personalized recommendations and services, such process can improve customer satisfaction, loyalty, and retention [Ashraf, et, al, (2019)].

1.2 Problem statement

Predictive modeling represents one of the vital issues for any financial institution planning and evaluation. It is very challenging to predict a customer if he/she can be a depositor by analyzing related information. Some recent studies demonstrated that economic depression and the continuous decline of the economy negatively impact business organizations and banking sectors. Due to such economic depression, banks cannot attract a customer’s attention. Thus, marketing is preferred to be a handy tool for the banking sector to draw customers’ attention for a term deposit. for this, machine learning algorithms were used which is a suitable approach to predict whether a new customer will have a term deposit or not. The contribution can be used as a depositor prediction system to provide additional support for bank deposit prediction. The problem is to indicate the main marketing campaign factor that can increase the customer’s decision to subscribe to a term deposit? The performance of a machine learning algorithms which is an approach to predict whether a new customer will have a term deposit or not. Identifying the main factor that can increase the customer’s subscription to a term deposit is the main concern.

1.3 Aim of Thesis

This Thesis aims to explore the application of predictive modeling in the banking industry and its potential to improve various aspects of banking operations. The main aim is to propose a suitable predictive model to improve the bank operations. This aim can be achieved by the following objectives:

- Collect a better understanding of the customer’s needs based on analyzing different banking datasets.

- Perform feature ranking to indicate the features effectiveness.
- Indicate the features correlation to show the reduction possibilities.
- Perform predictive approaches to show their capabilities and possibilities.
- Evaluate the performance of the used techniques

1.4 Related works

- Peter Appiahene, et. al, at, 2020, presented a combined Data Envelopment Analysis (DEA) with three machine learning approaches in evaluating bank efficiency and performance using 444 Ghanaian bank branches, Decision Making Units (DMUs). The results were compared with the corresponding efficiency ratl, ings obtained from the DEA. Finally, the prediction accuracies of the three machine learning algorithm models were compared. The results suggested that the decision tree (DT) provided the best predictive model [Appiahene, P.et, al. 2020].
- Meilin Widyastuti, et. al, at, 2019 classified the quality of customer service at Bank BTN Pematangsiantar Branch. Data was obtained from the results of the customer questionnaire. From the results of the study, there were 5 rules for classification in determining the quality of customer service with 3 rules with satisfaction status and 2 rules with dis-satisfied status. The algorithm can be used in the case of determining customer service quality of the Bank. They improved the quality of service to customer satisfaction [Widyastuti, et, al, 2019].
- Jing-Ping Li, et. al, 2020, employed the machine learning techniques in order to predict the credit ratings for the banks in GCC. The quarterly dataset of the macro and bank specific variables was used for a period that spanned between the years 2010 to 2018, with an out of sample prediction, for three years. Their findings suggested that arbitrary forests demonstrate the highest precision, based on the F1 score, specificity, and the accuracy scores. Other findings also revealed that the Artificial Neural Networks are ranked second for the overall predictions that have been made. However, for the speculative and default grades. They suggested that the Classification and Regression Trees (CART) are significantly

relevant, and although their precision is less than the random forests, the difference is not significant. They also proposed that, for the stressed banks, both random forests and the CART should be employed, for a better and more informed risk assessment [Li, J. P., Mirza et, al, 2020].

- Martin Jullum et, al, at 2020, Proposed a machine learning model for prioritizing which financial transactions should be manually investigated for potential money laundering. Their model was applied to a large data set from Norway's largest bank, DNB. A supervised machine learning model was trained by using three types of historic data: "normal" legal transactions; those flagged as suspicious by the bank's internal alert system; and potential money laundering cases reported to the authorities. Their model was used to predict the probability that a new transaction should be reported, using information such as background information about the sender/receiver, their earlier behaviour and their transaction history. Their developed method outperforms the bank's current approach in terms of a fair measure of performance [Jullum et, al, 2020].
- Premkumar Borugadda et, al, at 2021, investigated the demand for the adoption of telemarketing practices for promoting long-term bank deposits to potential bank customers. Explored the demand for long-term bank deposits by employing various machine learning algorithms like Random Forest (RF). The dataset related to direct marketing campaigns (phone calls) of a Portuguese banking institution was considered for analysis. The results of the study also provided insightful information to banks for making telemarketing policy decisions in the success of bank deposits to their existing and prospective bank customers [Borugadda et, al, 2021].
- Bluwstein et, al, at 2021, developed early warning models for financial crisis prediction by applying machine learning techniques to macro financial data for 17 countries over 1870–2016. Most nonlinear machine learning models outperform logistic regression in out-of-sample predictions and forecasting. They identified economic drivers of the developed machine learning models using a ~~novel~~ proposed framework based on Shapley values, uncovering nonlinear relationships between the predictors and crisis risk. Throughout, the most important predictors are credit growth and the slope of the yield curve, both domestically and globally. A flat or inverted yield curve is of most concern when nominal interest rates are low and credit growth is high [Bluwstein et, al, 2021].

- David Mhlanga et, al, at 2021, discovered that artificial intelligence and machine learning have a strong impact on credit risk assessments using alternative data sources such as public data to deal with the problems of information asymmetry, adverse selection, and moral hazard. This allows lenders to do serious credit risk analysis, to assess the behaviour of the customer, and subsequently to verify the ability of the clients to repay the loans, permitting less privileged people to access credit. Therefore, this study recommended that financial institutions such as banks and credit lending institutions invest more in artificial intelligence and machine learning to ensure that financially excluded households can obtain credit [Mhlanga et, al, 2021].
- Moscato et, al, at 2020, proposed a benchmarking study of some of the most used credit risk scoring models to predict if a loan will be repaid in a P2P platform. They dealt with a class imbalance problem and leverage several classifiers among the most used in the literature, which are based on different sampling techniques. A real social lending platform (Lending Club) data-set, composed by 877,956 samples, has been used to perform the experimental analysis considering different evaluation metrics (i.e., AUC, Sensitivity, Specificity), also comparing the obtained outcomes with respect to the state-of-the-art approaches. Finally, the three best approaches have also been evaluated in terms of their explainability by means of different eXplainable Artificial Intelligence (XAI) tools [Moscato et, al, 2020].
- Uthayakumar, et, al, at 2020, presented a cluster-based classification model, comprises of two stages: improved K-means clustering and a fitness-scaling chaotic genetic ant colony algorithm (FSCGACA) based classification model. In the first stage, an improved K-means algorithm was devised to eliminate the wrongly clustered data. Then, a rule-based model was selected to design to fit the given dataset. At the end, FSCGACA was employed for seeking the optimal parameters of the rule-based model. The proposed algorithm was employed to a collection of three benchmark dataset which include qualitative bankruptcy dataset. A detailed statistical analysis of the dataset was also given. The results analysis ensured that the presented FCP model was superior to other classification model based on the different measures and also found to be more appropriate for diverse dataset [Uthayakumar, et, al, 2020].

- Deepak Kikan, et, al, at 2019, analyzed the critical factors that are causing Banking and Financial Services sector adoption to be slower than the other industries and suggest drivers to improve the adoption thus support it in fighting financial frauds through Predictive Analytics and safeguarding itself and its customers' interest. This paper was also discussed various areas of Predictive Analytics being used by Banking and Financial Services companies to detect financial frauds. It helped in setting the base for Banking and Financial Services institutions to understand their potential and as well as help them adopt Predictive Analytics [Kikan, et, al, 2019].
- Broby et, al, at 2022, presented a method-based focused on the predictive analytics domain. The study comprehensively covered classification, regression, clustering, association and time series models. It expands existing explanatory statistical modelling into the realm of computational modelling. The methods explored enable the prediction of the future through the analysis of financial time series and cross-sectional data that is collected, stored and processed in Information Systems. The output of such models allows financial managers and risk oversight professionals to achieve better outcomes. This review brings the various predictive analytic methods in finance together under one domain [Broby et, al, 2022].
- Journal, I. J. C. S. M. C. 2019, propose a data mining approach to predict the success of telemarketing. We are applying the algorithms for the first time on the dataset. The dataset obtained from UCI, which contain the most common machine learning datasets. The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The number of the instance is 45212 with 15 input variables and the output variable. Classification is a data mining technique used to predict group membership for a data instance. we present the comparison of different classification techniques in open-source data mining software which consists of a One-R algorithm methods and Naïve-Bayes algorithm The experiment results show is a bout classification sensitivity, specificity, accuracy. The results on bank marketing data discovered that the One-R algorithm is better in classifying the data comparing with the Naïve-Bayes algorithm; where the error rate is lower [Journal, I. J. C. S. M. C., 2019].

- Oni, J. O. 2020, the resampling technique was used to deal with the imbalance dataset and three classifiers were applied; Logistic regression, Support vector machine, and K-nearest neighbor were used to achieve the set objective. Comparative analysis was performed using correlation heatmap to identify the main factors that can increase customer subscriptions to a term deposit. The outcome shows that ‘Duration’ is the main factor that can increase customer subscriptions in the bank. two experiments were performed in this study. Of all the algorithms used in this work, KNN has the best performance with the accuracy of 91.8% in the first experiment and 91.7% in the second experiment as compared to the Support vector machine and Logistic regression [Oni, J. O. 2020].
- Patwary et, al, at 2021, performance of the three mostly used classification algorithms named Support Vector Machine (SVM), Neural Network (NN), and Naive Bayes (NB) are analyzed. Then the ability of ensemble methods to improve the efficiency of basic classification algorithms is investigated and experimentally demonstrated. Experimental results exhibit that the performance metrics of Neural Network (Bagging) is higher than other ensemble methods. Its accuracy, sensitivity, and specificity are 96.62%, 97.14%, and 99.08%, respectively. Although all input attributes are considered in the classification method, in the end, a descriptive analysis has shown that some input attributes have more importance for this classification. Overall, it is shown that ensemble methods outperformed the traditional algorithms in this domain. We believe our contribution can be used as a depositor prediction system to provide additional support for bank deposit prediction [Patwary et, al, 2021].
- Kinga Włodarczyk & Kingsley Success Ikani, 2020 analyzed data from the UCI Machine Learning Repository called Bank Marketing Data Set. The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (’yes’) or not (’no’) subscribed. There were available four datasets, we chose bank-full.csv data set, which contains all examples for older version this data set [Kinga Włodarczyk & Kingsley Success Ikani, 2020].
- Choi, Y., & Choi, J. 2022, using a machine learning technique on the dataset of direct marketing campaigns of a Portuguese banking

institution which is obtained from the UC Irvine Machine Learning Repository. Based on these results, first, among all variables, age, balance, loan, day, duration, campaign, pdays, and poutcome influence the success of bank telemarketing, while job, martial, education, default, housing, contact, month, and previous have no significance. Second, for the full model, the accuracy rate is 0.784, which implies that the error rate is 0.216. Among the patients who predicted not to have the success of bank telemarketing, the accuracy that would not have the success of bank telemarketing was 75.63%, and the accuracy that had the success of bank telemarketing was 82.61% among the patients predicted to have the success of bank telemarketing [Choi, Y., & Choi, J. 2022].

Table (1.1) presents a summary of the stated related works.

Table (1.1) related works summary

Ref. No, Year	The used algorithm	Goal	Problem	The used Dataset	Results
2022	Decision tree	This paper intends to understand the antecedents in the success of bank telemarketing prediction	combined impact of the variables on the success of bank telemarketing modeling	Banking Dataset - Marketing Targets	decision tree (DT) accuracy 0.784
2021	NN, SVM, NB	study the performance of ensemble learning algorithms which is a novel approach to predict whether a new customer will have a term deposit or not.	Then the ability of ensemble methods to improve the efficiency of basic classification algorithms is investigated and experimentally demonstrated	Banking Dataset - Marketing Targets	NN Accuracy 94.8684 SVM Accuracy 89.8031 NB Accuracy 88.3294
2021	Random Forest (RF), Support Vector	investigate the demand for the adoption of	long-term bank deposits	UCI direct marketing campaigns (phone	The logistic regression model gave better results with 92.72 accuracies and

	Machine (SVM), Gaussian Naive Bayes (GNB), Decision Tree (DT), and Logistic Regression (LR)	telemarketing practices for promoting long-term bank deposits to potential bank customers.		calls) of a Portuguese banking institution	93.62 AUROC Score among five models compared to other models. So, choose a logistic regression model for deployment in real-time applications.
2021	logistic regression, SVM, Neural network	early warning models for financial crisis prediction by applying machine learning techniques	the prediction of crises as a classification	macroeconomic dataset covering 17 countries between 1870 and 2016	Accuracy of 2004 to 2016 Forecasting period of logistic is 0.867, SVM is 0.867, Neural is 0.872
2021	Machine learning and AI frameworks	In banking and finance, credit risk is among the important topics	the creditworthiness of the previously excluded individuals allowing them to also access credit.	UCI repository	financial institutions such as banks and credit lending institutions invest more in artificial intelligence and machine learning to ensure that financially excluded households can obtain credit.
2021	Logistic regression Random forest Multi-layer perceptron	A benchmark of machine learning approaches for credit score prediction	Credit risk assessment	Lending clubs dataset	Classification results (Under-Sampling approach) of Logistic regression 0.65 Random forest is 0.64 Multi-layer perceptron is 0.73
2020	KNN, linear model, logistic model	in order to access if the product (bank term deposit) would be (’yes’) or not (’no’) subscribed.	The efficiency of banks and the safety of depositors in the banking industry	Banking Dataset - Marketing Targets	KNN is Accuracy 0.88 Linear Accuracy 0.89 Logistic Accuracy 0.89

2020	Logistic Regression, KNN, SVM	The primary objective of the video content campaigns is to ensure that the customers understand the products offered by the bank	The endpoint of the campaign be it in any form is to meet the targeted needs of the customers thereby satisfying the customers	A bank-additional-full	Logistic Regression Accuracy 0.848 SVM Accuracy 0.856 KNN Accuracy 0.917
2020	Classification and Regression Trees (CART)	predict the credit ratings for the banks in GCC	unprecedented innovations and improvements for the financial sector	dataset of the macro and bank 2010 to 2018	Classification and Regression Trees (CART) is accuracy 0.86
2020	XGBoost	develop, describe and validate a machine learning model for prioritising which financial transactions	large data sets and increase training time	Norway's largest bank, DNB	Proportion of positive predictions (PPP) is 0.8 and true positive rate (TPR) is 0.95
2020	decision tree (DT), Random Forest, Neural network	Predicting Bank Operation	The efficiency of banks and the safety of depositors in the banking industry	bank branches in Ghana (DMUs)	decision tree (DT) accuracy is 100% of 30 samples. , Random forest accuracy is 98.5, Neural network accuracy is 86.6%.
2019	NB, One-R	Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribe	The efficiency of banks and the safety of depositors in the banking industry	Banking Dataset - Marketing Targets	NB Accuracy 88.4541 One-R Algorithm 89.3875
2019	C.45 algorithm (decision tree DT)	classify the quality of customer service at Bank BTN	Quality of Custumer Service in Bank BTN	Small dataset from Bank BTN Pematangsiantar Branch	C.45 accuracy is 77.78%

		Pematangsiantar Branch.	Pematangsiantar Branch		
2019	Classification bank operation framework	Predictive Analytics Adoption by Banking and Financial Services	credit scoring risk analysis, customer retention, operational decisions, investment decisions, operations optimization, and fraud prevention.	Banking services from HDFC Bank, ICICI Bank, State Bank of India (or other state banks), Punjab National Bank, etc.	Channel used for banking and financial services used by the customers : Answer Use services online / over internet Count 69, Percent is 31.65% Use services at a Branch Count 54, Percent is 24.77%.
2018	K-means clustering and a fitness-scaling chaotic genetic ant colony algorithm (FSCGACA)	develop an efficient prediction model for better classification performance and adaptable to diverse dataset.	Financial crisis prediction	qualitative bankruptcy dataset, Weislaw dataset and Polish dataset.	the proposed model attains the maximum accuracy of 97.93.

1.5 Thesis Outline

The thesis is organized into five chapters

- **In addition to chapter one, Chapter Two:** includes a comprehensive description of the main banking concepts, and preprocessing techniques that included both the handling of missing values and normalization process, in addition, this chapter included the methods used to features selection to reduce data dimensions, the machine learning algorithms, the proposed prediction model, and the evaluation methods used.
- **Chapter Three:** This chapter describes the proposed system. In this chapter, the preprocessing stages are explained, and the following stages select important features in the dataset (features selection) and prediction.

- **Chapter Four:** The fourth chapter explains and discusses the implementation the proposed system on the dataset, and illustrates of the experimental results obtained after applying the proposed model.
- **Chapter Five:** The fifth chapter represents the most important conclusions this study reached based on the results of the thesis. In addition to that, this chapter highlights possible future works.

Chapter Two

Theoretical Background

Chapter Two
Theoretical Background

2.1 Introduction

This chapter displays some definition of the predictive modeling approaches for banking operations. It also states the background for the applications, some banking operations and models.

2.2 Predictive Modeling

Predictive Modeling is a machine learning technique that would work best to predict the future outcomes. It represents a statistical method to analyze the patterns of the data to estimate future outcomes or events. Predictive Modeling can be considered as an essential feature of predictive analytics [Niemann, et, al, 2008]. Selecting the greatest predictive analytics tools in creating informed decisions needs choosing which predictive data modeling methods are perfect for banking.

Predictive modeling techniques are utilizing the available data to form a model that can be used in predicting consequences for new generated data. Applying such procedures enables banking managers in optimizing their decision-making and in making new understandings for more actual and effective commercial actions [Kumar, V., & Garg, M. L. 2018].

Predictive modeling comprises generating a model that yields the probability of the resulted outcomes based on the input parameters as a current state value. Predicting customer behavior is one of the most important contexts in banking operations. Availability of the past customer activity data can be used to build a suitable predictive model to capture the greatest influence features on next future client activity [Ashraf, et, al, 2019].

2.2.1 Predictive Modeling Advantages

Predictive models having various advantages in banking operations [Ian H. Witten, et, al, 2017]:

- Offer a technique to gain a reasonable advantage.
- Get an enriched understanding of the customer demands.
- Evaluate and moderate the financial threats.

- Improve the current products to increase its revenue.
- Reduce the time required to predict the consequences.
- Expect most of the outside components that may have an influence on the service efficiency.

Most of these predictive models are working rapidly and may perform their operations in real time. This is the reason behind using it in calculating the danger of an online mortgage or credit card application in banks and retailers. Banks are utilizing these prediction models to appraisal borrowers' credit scores to confirm reliability, previous defaults and background. It can benefit in predicting the probabilities of misrepresentation, fraud, and dangers involved with a specific customer [Pang-Ning Tan, et, al, 2021].

2.2.2 Predictive Modeling Steps

Predictive modeling process concerning many steps. These steps are [Ian H. Witten, et, al, 2017]:

1. Recognizing the bank process aims.
2. Describe the Modeling process goals.
3. Collect the suitable Data.
4. Arrange and analyze the collected Data.
5. Examine and Convert Variables and perform Sampling.
6. Select suitable Models.
7. Confirm Models (Testing process), Optimize.
8. Implementation, Monitoring and Performance measures.

2.2.3 Predictive analytics

Predictive analytics means Predictive modeling. Predictive modeling is preferred in academic sites, while "predictive analytics" is the favored word for banking applications. It can support financial organizations for recognizing unexpected prospects, optimizing their commercial methods, knowing consumer behavior, and helps in stopping problem before its happened [Martens, et, al, 2016].

"Predictive analytics" is a traditional case of business intelligence (BI) technologies that exposes associations and shapes within data that can be used to forecast performance and actions. Figure (2.1) represents a description for the process of predictive analytics. Traditional BI represents a descriptive model to achieve data analytics process taking place by data collection, organizing report, execution a data analysis step and observing commercial activity. Predictive analytics is performed by looking forward with the previous actions to expect the upcoming [Ian H. Witten, et, al, 2017]. Predictive analytics approaches can reply what is expecting to occur in future?

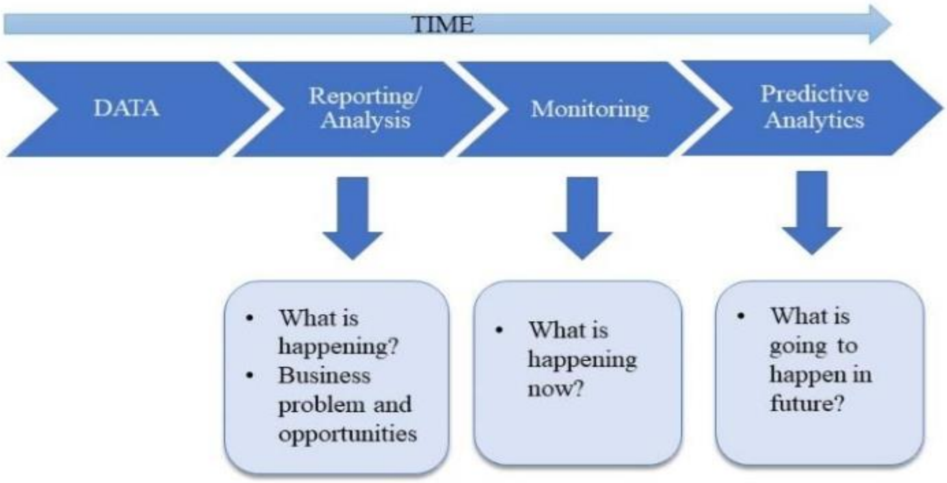


Figure (2.1) Predictive analytic method [Ian H. Witten, et, al, 2017]

Figure (2.2) represents a situation for predictive analytics in data science. Scientists recognize commercial objectives based on consideration of data from numerous bases. Predictors make data preparation desirable and constructing predictive analytics. Further state is to assess the model until receiving the suitable model to be formed [Kuhn, M., & Johnson, K. 2013].

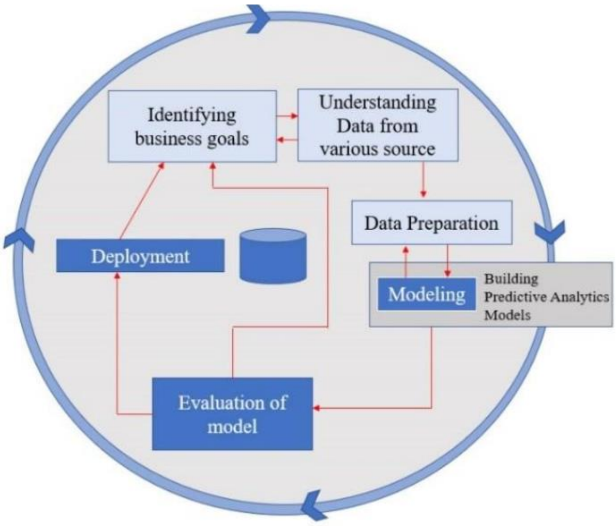

Identifying business goals
Understanding Data from various source
Data Preparation
Modeling
Building Predictive Analytics Models
Deployment
Evaluation of model

Figure (2.2) Predictive analytics in data science [Kuhn, M., & Johnson, K. 2013]

Models can be proposed to determine relations between various performance issues. Figure (2.3). Presents these Predictive Modeling possibilities. Following are some algorithms concerning the predictive modeling process [Pang-Ning Tan, et, al, 2021]:

1) *Classification*: algorithms of Classification is one task of the data mining to predict the value variable (categorical) by constructing a model.
2) *Regression*: another task of the data mining is a Regression. It can be used in predicting the value of a numerical variable by constructing a model based on numerical or categorical variables. Regression models is suitable for expecting the outcome which may be continuous.
3) *Clustering:* A similar data can be collected in a set called cluster. Clustering is the method of isolating a dataset into sets.
4) *Association rules:* A machine learning approach to find certain relations between variables is known as association Rules. It is selected to identify robust rules exposed in databases using some special methods.

An expression X1 → Y1 is an association rule, where X1 and Y1 are sets of objects. The intuitive sense of such a rule is that relations of the database which comprise X1 tend to enclose Y1 [Pang-Ning Tan, et, al, 2021].

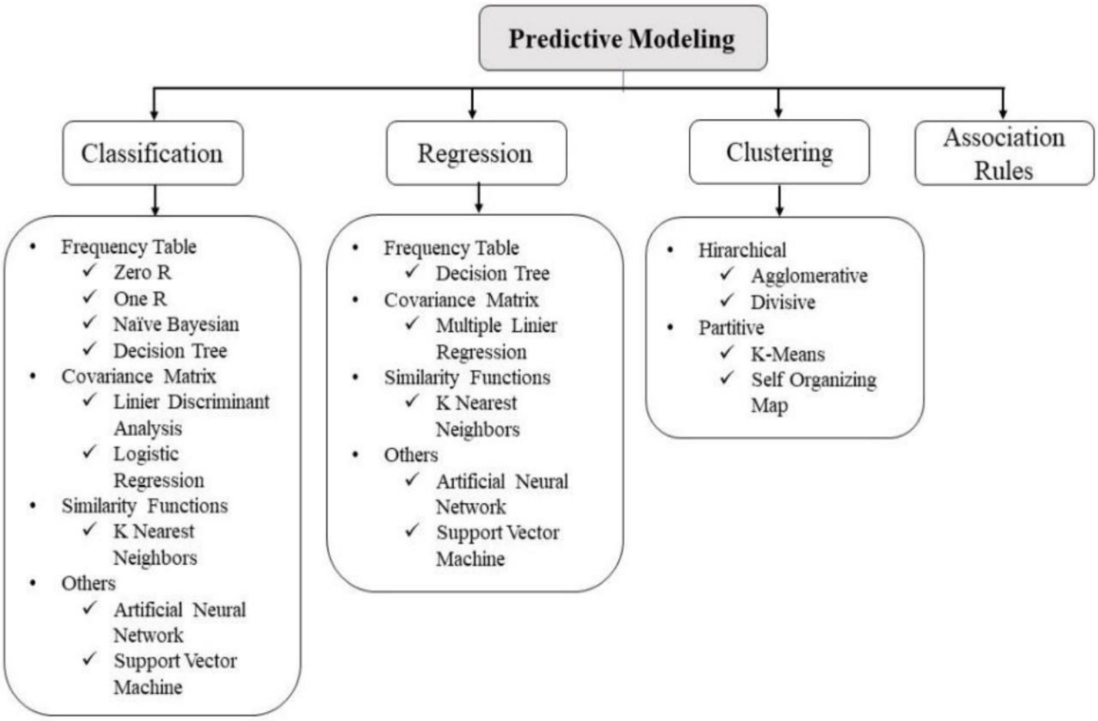


Figure (2.3) Predictive Modeling possibilities [Pang-Ning Tan, et, al, 2021]

2.2.4 Predictive Modeling Limitations

The most indicated limitations in the predictive modeling process are [Pang-Ning Tan, et, al, 2021]:

1. **Data Labeling Errors:** To overcome the data labeling errors, generative adversarial networks or reinforcement learning can be used.
2. **Train the machine learning tool by a shortage of massive datasets: there is a** possibility to repair a small dataset to train the model rather than a massive dataset.
3. **The inability to explain the processes:** computers are not like humans, cannot "learn", or "think". Computations are also being complex that humans cannot follow logically. Valid model is essential.
4. **Learning Generalizability:** Computers have trouble carrying what they've educated forward. They have concern in using what they've learned to a new set of situations.
5. **Data and algorithms Bias: outcomes** can be skewed and lead to human's mistreatment. Biases may tend to self-perpetuate.

2.2.5 Predictive Model Selection

Predictive modeling is the method of getting known outcomes and creating a model that can expect values for new events. It uses past data to forecast future actions. There are many dissimilar forms of predictive modeling methods containing linear regression (ordinary least squares), Analysis of Variance (ANOVA), logistic or ridge regression, neural networks, time series, decision trees, and so on [Kuhn, M., & Johnson, K. 2013]. Choosing the accurate predictive modeling method at the beginning of a task can save time. Selecting the improper modeling technique can end in imprecise predictions and residual plots that knowledge non-constant adjustment. After defining the variables, several types of models can be formed. The greatest common ones for predicting consumer performance are: decision trees and logistic regression. Decision trees utilizes graph or typical decisions to define the conditional probability of a consequence [Ian H. Witten, et, al, 2017].

Logistic regression model can be created to predict the probability of occurrence of an event. Commonly used metrics are Cumulative Gains Chart, Lift Chart, and Receiver Operating Characteristics (ROC) curve. All of these provide metrics by trading off desirable outcomes. These metrics can be obtained by implementing the model on the training data set [Pang-Ning Tan, et, al, 2021].

2.2.6 Predictive model Creation

To create or improve a predictive model, experts or analysts create standard predictive algorithms with statistical models, train them using subsets of the data, and implement them against the entire data set. To create a suitable predictive model [Kuhn, M., & Johnson, K. 2013]:

- Data Collection process
- Data preprocessing (cleaning)
- Perform an Exploratory Data Analysis (EDA)
- Predictive Model Development
- Model Evaluation
- Predictive Model Implementation
- Model Tracking (Check its performance)

2.2.7 Business requirements.

The essential general steps to perform any business requirements are [Gavurova, et, al, 2017]:

- Identify and explore data relevant to the analysis.
- Clean the data and remove any unwanted or redundant data.
- Perform EDA on clean data and build a suitable predictive model using statistical data modeling techniques.
- Validate your model's accuracy and deploy it once the validation is successful.
- Monitor your model regularly to optimize its performance.
  Figure (2.4) Collects more of these steps.

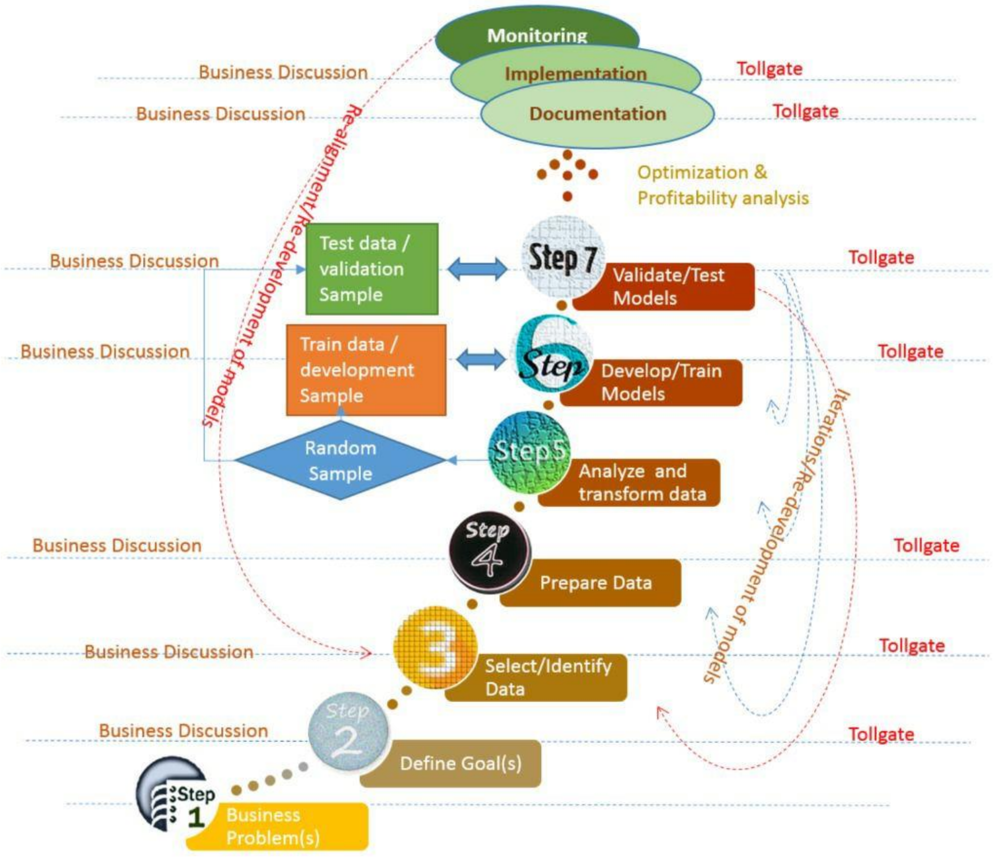


Figure (2.4) Business requirements steps [Kuhn, M., & Johnson, K. 2013]

2.3 Data preprocessing

Data preprocessing methods are important to prepare the dataset. In general, all of these methods fall into two categories: selecting data objects and attributes for the analysis or creating/changing the attributes [P. Tang, M. Steinbach, and V. Kumar.,2006]. These methods include several strategies for handling dataset issues such as noise, missing values, and inconsistent data.

In all cases, the goals of preprocessing are [A. Jović, et, al, 2015]:

- ✓ Reducing the dataset size to achieve a more efficient analysis concerning time, cost, and quality.
- ✓ Adapting the dataset to best suit the selected analysis method.

Today's real-world databases are highly vulnerable to missing and cluttered data because of their huge volume and the origin of this data is likely to be from multiple heterogeneous sources. Low quality data will lead to lower quality mining results. Data preprocessing helps improve data quality and thus improves the efficiency and ease of the mining process. Data preprocessing is an important step in the data mining process and data collection methods are often controlled loosely, resulting in out-of-band values, or missing values. Analyzing data that has not been carefully examined to address these problems can lead to illogical and misleading results [M. Toloo, et, al, 2008]:

There are many data pre-processing techniques, including data cleaning which can be applied to remove inconsistencies in the data and remove noise. Data integration is where data from multiple sources are combined into a cohesive data store. Data reduction can reduce data size through aggregation and deletion of redundant features. Normalization can be applied where data is scaled to fall within a smaller range (0.0 and 1.0). Thus, the accuracy and efficiency of mining can be improved, and the missing values can be estimated [M. Toloo, et, al, 2008]:

After applying the pre-processing of the data and obtaining the appropriate results, the final obtained data set can be considered as a reliable source and can be used in any algorithm that is applied to extract the features [M. Toloo, et, al, 2008]: Table (2.1) shows the different stages of data pre-processing.

Table (2.1) Methods for Pre-processing the Data [M. Toloo, et, al, 2008]:

Data Preprocessing		
Data Cleaning	Data Transformation	Data Reduction
1-Missing Data a- Ignore the tuple b- Fill the missing value 2- Noisy Data a- Regression b- Clustering	1- Normalization 2- Attribute selection 3- Discretization 4- Concept hierarchy Generation	1-Data Cube Aggregation 2- Attribute subset selection 3- Numerosity Reduction 4-Dimensionality Reduction

2.3.1 Handling Missing Values

Usually, some objects missing one or more attribute values. These missing values can adversely affect the performance of prediction models. In [P. Tang, et, al, 2006], several strategies for handling missing values are discussed such as:

- Ignoring missing values during analysis.
- Eliminating data objects.
- Estimating missing values.

If the attribute contains small and widely scattered missing values, then the estimation method can be used. The missing values can be evaluated by using the residual values. If the predictor is categorical, the most occurring predictor value can be taken [Martens, et, al, 2016]. On the other side, if the predictor is continuous, the average predictor value of the closest neighbors is used [S. Wang and W. Shi 2012].

2.3.2 Data Normalization

The unit of measure used may affect the analysis of the data and in many cases this data is in a format that is not suitable for use by data mining techniques. So, the data is normalized by converting the raw data values into another form with properties that fit the model used. Normalization aims to ensure that all features are in the same unit of measurement [L. Al Shalabi and Z. Shaaban 2006], i.e., between [0 ,1] or between [1, -1]. Therefore, it is used to avoid the difference between the influence of small values and large values that dominate the results. In general, normalization methods are applied to the data to reduce the error and increase the accuracy of the model used. Many methods are found for data normalization such as min-max, and z-score normalization [P. Tang, M. Steinbach, and V. Kumar.,2006]. standardization, also known as z-score normalization, is an algorithm used to standardize or normalize a dataset by transforming it into a standard normal distribution. The z-score normalization algorithm calculates the z-score for each data point, representing the number of standard deviations away from the mean. The formula to calculate the z-score for a data point, given a dataset with mean μ and standard deviation σ, is as follows [Giordana, et, al, 2017]:

$$z = \frac{(x - \mu)}{\sigma} \qquad \text{............ (2.1)}$$

Where:

x is the data point

μ is the mean of the dataset

σ is the standard deviation of the dataset

The resulting z-score indicates how many standard deviations a particular data point is away from the mean. If the z-score is positive, the data point is above the mean, while a negative z-score indicates it is below the mean [Alcalde, et, al, 2022].

2.3.3 Handling Imbalanced Data

In banking operations, it is common to have imbalanced datasets, where the occurrence of certain events or classes is significantly higher or lower than others. Techniques such as oversampling (e.g., SMOTE) or under sampling (e.g., Random Under Sampler) can be used to address class imbalance issues and improve the predictive performance. SMOTE (Synthetic Minority Over-sampling Technique) is an algorithm used for imbalanced learning in machine learning. It is specifically designed to address the problem of imbalanced class distribution, where one class has significantly fewer instances compared to the other(s).

In imbalanced datasets, traditional machine learning algorithms may struggle to effectively learn the minority class, resulting in biased predictions and poor performance. SMOTE helps to alleviate this issue by generating synthetic examples of the minority class, effectively oversampling it and balancing the dataset [Ian H. Witten, et, al, 2017].

Here are the steps to implement the SMOTE algorithm:

- Identify the minority class: Determine which class in your dataset is the minority class that you want to oversample.
- Calculate the imbalance ratio: Compute the ratio of the number of majority class samples to the number of minority class samples. This will help determine the amount of oversampling required.
- Determine the number of synthetic samples to generate: Decide on the number of synthetic samples you want to generate to balance the dataset. This can be based on a fixed ratio or a specific number of samples.
- Identify the k nearest neighbors: For each minority class instance, find its k nearest neighbors from the minority class instances. Typically, Euclidean distance is used as the distance metric, but other metrics can also be used.
- Generate synthetic samples: For each minority class instance, select k nearest neighbors randomly, and create synthetic samples by interpolating between the minority instance and its neighbors. This is done by randomly selecting a value between 0 and 1 and multiplying it by the difference between the feature values of the instance and its neighbor. Add this difference to the minority instance to create a synthetic sample.
- Repeat the previous step until the desired number of synthetic samples is generated.
- Combine the original and synthetic samples: Combine the original minority class instances with the synthetic samples to create a balanced dataset.

Here's a simplified example to illustrate the SMOTE algorithm:

Let's say we have a binary classification problem with two classes: Class A (minority) and Class B (majority). We have 100 instances in Class A and 1000 instances in Class B, resulting in an imbalance ratio of 10:1.

To balance the dataset, we decide to generate synthetic samples for Class A until it matches the size of Class B (1000 instances).

- ➢ Identify the minority class: Class A is the minority class.
- ➢ Calculate the imbalance ratio: 10:1 (Class B instances: Class A instances).
- ➢ Determine the number of synthetic samples to generate: We need to generate 900 synthetic samples for Class A to match the size of Class B.

- Identify the k nearest neighbors: Let's set k = 5. For each instance in Class A, find its 5 nearest neighbors from Class A.
- Generate synthetic samples: For each instance in Class A, randomly select 5 nearest neighbors and create synthetic samples by interpolating between the instance and its neighbors.
- Repeat step 5 until 900 synthetic samples are generated.
- Combine the original and synthetic samples: Combine the original 100 instances of Class A with the 900 synthetic samples to create a balanced dataset with 1000 instances of each class.
- This is a simplified explanation of the SMOTE algorithm. There are variations and improvements to SMOTE, such as Borderline-SMOTE and SMOTE-ENN, which address specific challenges in imbalanced datasets.

2.3.4 Handling Categorical Variables

Banking data often contains categorical variables such as account types, transaction types, or customer segments. These variables need to be encoded into numerical values for modeling purposes. Common techniques include one-hot encoding, label encoding, or target encoding, depending on the nature and cardinality of the categorical variables [Pang-Ning Tan, et, al, 2021]. In this thesis will be use his method Nominal to (Numeric or Binary).

- Nominal to Binary**:** The proposed system used this method to convert nominal attribute values into binary state. In this methodology, the used system is based on converting nominal (string values) in attribute big dataset into binary data as (0, 1) as it is optimal to use in the proposed machine learning algorithm to increase accuracy of prediction [Pang-Ning Tan, et, al, 2021].

- Nominal to Numeric: It used to convert nominal values in dataset into numeric values [Pang-Ning Tan, et, al, 2021].

2.4 Data Reduction.

Data reduction in banking refers to the process of reducing the size and complexity of large datasets while preserving essential information. This is often done to improve data processing efficiency, reduce storage requirements, and enhance data analysis capabilities. Therefore, two processes will be used of reducing features in banking operations for this thesis.

2.4.1 Ranking Correlation

Ranking correlation algorithms are statistical techniques used to measure the similarity or agreement between two rankings or orders of items. These algorithms assess how well the relative ordering of items in one ranking match or correlates with the relative ordering in another ranking. Ranking correlation coefficients provide a measure of the strength and direction of this agreement [M. Toloo, et, al, 2008]. The correlation algorithm, also known as correlation analysis, is a statistical technique used to measure the strength and direction of the relationship between two or more variables. It assesses how changes in one variable are associated with changes in another variable. The correlation coefficient is the main output of this algorithm, which quantifies the degree of association between variables. The most commonly used correlation coefficient is the Pearson correlation coefficient (r), which measures the linear relationship between two continuous variables. It ranges from -1 to +1, where [Pang-Ning Tan, et, al, 2021]:

- A correlation coefficient of +1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- A correlation coefficient of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A correlation coefficient of 0 indicates no linear relationship between the variables.

The correlation coefficient is calculated using the following formula:

$$\boldsymbol{r} = \frac{(\boldsymbol{\Sigma}\ ((\boldsymbol{x} - \bar{\boldsymbol{x}}) * (\boldsymbol{y} - \bar{y})))}{(\boldsymbol{n} * \boldsymbol{\sigma x} * \boldsymbol{\sigma y})} \text{ .... } \mathbf{(2.2)}$$

Where:

- x and y are the individual data points of the two variables being analyzed.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, respectively.
- σx and σy are the standard deviations of the x and y variables, respectively.
- n is the number of data points.

Once the correlation coefficient is calculated, it can be interpreted to determine the strength and direction of the relationship between the variables. Other correlation coefficients, such as Spearman's rank correlation coefficient and Kendall's rank correlation coefficient, can be used for variables that are not normally distributed or are measured on an ordinal scale. Correlation analysis is commonly used in various fields, including finance, social sciences, marketing, and data analysis, to understand the relationships between variables, identify patterns, and make predictions. It is important to note that correlation does not imply causation, and additional analysis and domain knowledge are often required to establish causal relationships between variables [Pang-Ning Tan, et, al, 2021].

2.4.2 Feature selection

Feature selection is an important step in predictive modeling that involves choosing the most relevant and informative features from a dataset to build an accurate and efficient model. In the context of improving banking operations through predictive modeling, feature selection can help identify the key variables that have the most significant impact on the desired outcomes or performance measures. Here is an approach for feature selection in the context of banking operations. Select feature selection methods: There are various feature selection techniques you can use, depending on the nature of your data and the objective. Some common methods include [Pang-Ning Tan, 2021]:

- Univariate feature selection: This method involves evaluating each feature independently based on statistical measures such as ranking correlation, chi-squared test, ANOVA, or mutual information. Features with the highest scores are selected. In this thesis l will use his method ranking correlation.
- Recursive feature elimination: This method uses an iterative process to recursively eliminate less important features based on the model's performance. It starts with all features and progressively removes the least significant ones until a desired number of features remains.
- Tree-based feature importance: Tree-based algorithms (e.g., Decision Tree) can provide feature importance scores based on the number of times a feature is used to split the data across multiple trees. Features with higher importance are considered more relevant.
- Domain knowledge: Consider expert knowledge in the banking domain to identify features that are likely to have a significant impact on the desired outcome.

2.5 Predictive modeling Algorithms

Banks play a critical role in the financial sector, providing various services to individuals and businesses. As technology continues to advance, banks are increasingly leveraging predictive modeling approaches to enhance their operations and improve overall efficiency. Predictive modeling involves using historical data and statistical techniques to make predictions or forecast future outcomes [Pang-Ning Tan, 2021]. By applying predictive modeling techniques to banking operations, banks can gain valuable insights, optimize processes, and deliver better services to their customers. To perform a predictive modeling approach, four prediction models are utilized in this thesis (Naïve bayes, Decision tree, KNN and Logistic).

2.5.1 Naïve Bayes

Naive Bayes [S. Jiang, 2014] is a direct and powerful classifier that uses the Bayes theorem. It predicts the probability that a given record or data point belongs to a particular class. The class with the highest probability is considered to be the most likely class. This algorithm assumes that all features are independent and unrelated. The Naive Bayes model is simple and easy to build and particularly useful for large data sets. This model is known to outperform even highly sophisticated classification methods**.**

Naive Bayes (NB) [A. Jović, 2015] is one of the most popular classifiers that is used in the banking sector for classification. Naive Bayes is a classifier that performs probabilistic prediction and predicts probabilities of class membership. Naive Bayes classifier, also known as a statistical classifier, is based on Bayes Theorem with independent assumptions between the predictors [Junqué de Fortuny, 2013].
The Bayes Theorem is:

$$\boldsymbol{p(h\backslash x) = \frac{p(x\backslash h) * p(h)}{p(x)}} \quad \textbf{........... (2.3)}$$

Were,

P(x) = Prior probability of x.

P(h) = Prior probability of h.

P (h\ x) = Posterior probability of h condition on x.

P(x\h) = Posterior probability of x condition on h.

Naïve Bayes used for classification tasks, including banking applications. Here are the steps involved in the Naïve Bayes algorithm:

Step 1: Data Preparation

- Collect a dataset that includes labeled examples of banking transactions or customer data.
- Preprocess the data by cleaning, normalizing, and transforming it into a suitable format for the algorithm.
- Split the dataset into a training set and a test set.

Step 2: Calculate Class Priors

- Calculate the prior probability of each class in the training set.
- The prior probability of a class is the proportion of instances in the training set that belong to that class.

Step 3: Calculate Likelihoods

- For each feature and each class, calculate the likelihood of observing a particular value of the feature given the class.
- Depending on the type of feature, different techniques can be used. For categorical features, you can calculate the probabilities directly. For continuous features, you can assume a probability distribution (e.g., Gaussian) and estimate the parameters (mean and variance) from the training data.

Step 4: Make Predictions

- Given a new instance (e.g., a new banking transaction or customer data), calculate the posterior probability of each class using Bayes' theorem.
- The posterior probability of a class given an instance is proportional to the product of the prior probability of the class and the likelihood of the instance given the class.
- The class with the highest posterior probability is predicted as the output for the new instance.

Step 5: Evaluate Model Performance

- Use the test set to evaluate the performance of the Naïve Bayes model.
- Calculate metrics such as accuracy, precision, recall, and F1 score to assess the model's effectiveness in classifying banking instances.

Finally, you would evaluate the model's performance using the test set, comparing the predicted labels with the true labels. This evaluation would help you understand how well the Naïve Bayes algorithm is performing for the specific banking application.

Remember that Naïve Bayes is a simple and fast algorithm but assumes independence between features, which may not always hold in real-world scenarios. Nonetheless, it can still be effective for many classification tasks, including banking applications.

2.5.2 Decision Tree

DT is commonly classification technique, which depends on creation a structure as tree with each branch representing an association between the values of feature and a class label [Ian H. Witten, 2017]. The results are very interpretable because DT generates rules, which easy to understand, but the outcomes have been represented in categorical data. Hence, DT is less efficient for prediction when the features are numerical [Kuhn, M., & Johnson, K. 2013].

The most famous typical amongst DTs is the C4.5 tree [Pang-Ning Tan, 2021]. The mechanism of C4.5 tree is done by recursively partitioning the training dataset according to tests on the possible feature values in separating the classes. The most important question is how can select the most important features first? Entropy or IG is used to select the most important features, which is automatically considered feature of the lowest entropy or of the highest IG.

Decision tree is a common and popular data mining technique that uses the tree hierarchy for data classification and rule inductions. The internal nodes of the tree represent the attributes' tests, the branches hold the resulted tests' values, and the leaf nodes represent the class labels for the decision attributes. For new object classification, a path for the attribute values of that object is examined according to the decision tree nodes and branches, starting from the root node till reach to the leaf node that holds the class label, such class label is considered as the class prediction for the new object [Sravani, K., & Mahaveerakannan, R. 2023].

Despite the decision tree classification is widely used, and many decision algorithms are developed in order to obtain the minimal set of attributes that are needed to build a more certain decision tree, the problem of finding the best tree representation for specific information space is still NP-hard problem. The tree building is mainly depending on the divide-and-conquer algorithm [P. Tang, 2006] and all current decision tree building algorithms are heuristic; the heuristic is to select and split the attribute with maximum gain ratio based on the associated information. Decision trees are widely used in banking for various applications. Here are the general steps to create a decision tree algorithm for banking:

Step 1: Gather and preprocess data

- Collect a dataset containing relevant banking information, such as customer demographics, credit history, income, loan amount, repayment history, etc.
- Preprocess the data by handling missing values, encoding categorical variables, and normalizing numerical variables.

Step 2: Define the target variable

- Determine the target variable you want to predict using the decision tree algorithm. In the case of credit scoring, the target variable could be a binary variable indicating whether a customer is likely to default on a loan or not.

Step 3: Split the dataset

- Split the dataset into a training set and a test set. The training set will be used to build the decision tree, while the test set will be used to evaluate its performance.

Step 4: Build the decision tree

- Use the training set to build the decision tree. The algorithm will automatically select the best features and split points based on certain criteria (e.g., Gini impurity or information gain).

Step 5: Evaluate the decision tree

- Evaluate the performance of the decision tree using the test set. Common evaluation metrics include accuracy, precision, recall, and F1-score.

Step 6: Fine-tune the decision tree

- If the decision tree's performance is not satisfactory, you can fine-tune it by adjusting hyperparameters (e.g., maximum depth, minimum sample split) or using techniques like pruning to prevent overfitting.

Step 7: Predict and deploy

- Once you are satisfied with the decision tree's performance, you can deploy it to make predictions on new, unseen data.

2.5.3 KNN

KNN can be used for statistical problems both in classification and regression. Nevertheless, it's used more broadly in business classification issues. This is also applied for solving problems of classification as well as regression. KNN is simpler to use, has quick execution time as the data quality determines the accuracy of the model, and the k value (nearest neighbor) must be adequate. The KNN learning algorithm is a simple classification algorithm that works on the basis of the smallest distance from the query instance to the training sample [P. Tang, 2006] to determine the simple majority of KNN as a prediction of the query. KNN is used due to its predictive power and low calculation time, and it usually produces highly competitive results.

This method is introduced in the early 1950s. It is applied in the most common application named pattern recognition. There are many machine learning methods in DM, nevertheless lazy learning algorithm is the simplest one because it does not need for any model in training. The model is built in classification or prediction is required [Pang-Ning Tan, 2021]. Therefore, KNN is one of lazy learning type, which predicts classes of entity based on the K nearest training instances in the feature space. The K-NN classifiers is based learning work by calculating the similarity between the training objects and a specific test object. The training objects consist of N features. Each object defines a point in n-dimensions space. To classify an unidentified object, a KNN classifier observes for the K training objects that are nearest to that object in the n-dimensions space. This K training objects represent the KNNs of the unidentified example [Kuhn, M., & Johnson, K. 2013].

The classification of KNN entity can be done by collect the majority votes for its neighbors. How to determine the correct value of K in this method is considered a problem**.** Figure (2.5) explains the nearest-neighbor classification. Here are the steps involved in implementing the KNN algorithm for a banking example:

Step 1: Collect and preprocess the data Gather a dataset that includes information about bank customers, such as age, income, credit score, loan history, etc. Preprocess the data by cleaning it, handling missing values, and normalizing or standardizing the features if necessary.

Step 2: Choose the value of K Decide on the number of nearest neighbors (K) to consider when making predictions. The selection of K depends on the dataset and problem at hand. A higher value of K provides a smoother decision boundary but may lead to misclassification of data from different classes. Conversely, a lower value of K may be more prone to overfitting.

Step 3: Define a distance metric Select an appropriate distance metric to measure the similarity between data points. Common distance metrics used in KNN include Euclidean distance, Manhattan distance, and cosine similarity. The choice of distance metric depends on the data and the problem domain.

Step 4: Calculate the distances for each data point in the dataset, calculate its distance to all other data points based on the chosen distance metric. This step helps identify the K nearest neighbors for each data point.

Step 5: Find the K nearest neighbors Sort the calculated distances in ascending order and select the K nearest neighbors based on the sorted distances.

Step 6: Make predictions for classification tasks, determine the class of the new data point by majority voting among its K nearest neighbors. The class with the highest frequency among the neighbors is assigned to the new data point. For regression tasks, the predicted value is usually the mean or median value of the K nearest neighbors.

Step 7: Evaluate the model Split the dataset into training and testing sets. Use the training set to train the KNN model and the testing set to evaluate its performance. Common evaluation metrics for classification tasks include accuracy, precision, recall, and F1 score

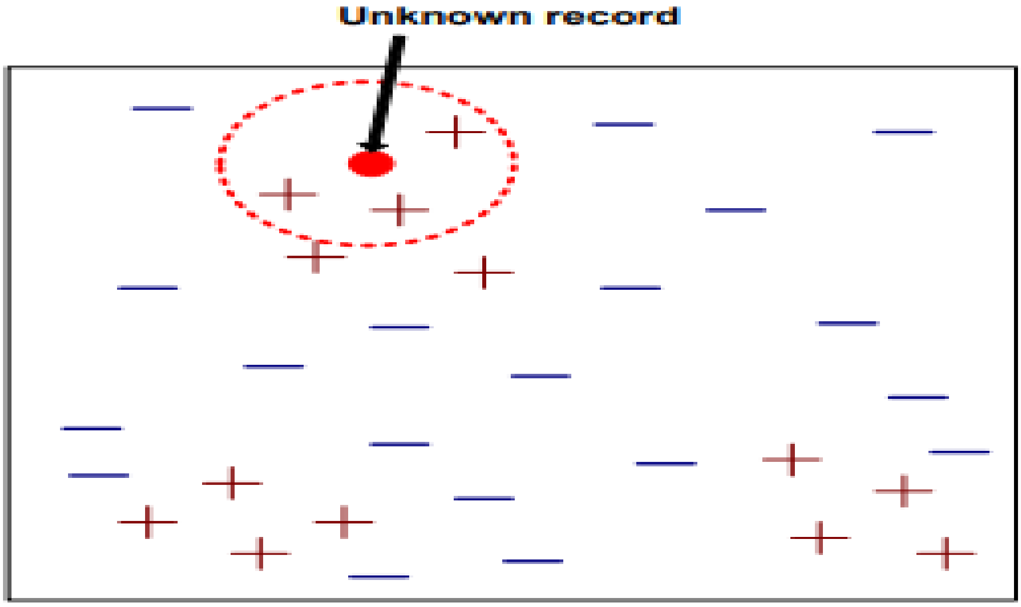


Figure (2.5) nearest-neighbor classification

2.5.4 Logistic Regression

This algorithm is frequently used for analytical function. Empirically, this method increases the probability of logging to allow the data required to perform well (Kuhn, M., & Johnson, K. (2013). It also indicates that it functions much like the linear regression only because it deals with variables that are categorical like Yes or No, 1 and 0, and others.

Logistic Regression is a supervised learning algorithm used for binary classification tasks, where the goal is to predict whether an input belongs to one of two classes. It's called "logistic" because it uses the logistic function (also known as the sigmoid function) to model the relationship between the input features and the binary output. Logistic regression is a popular machine learning algorithm used for binary classification tasks, which can be applied to various domains, including banking [Pang-Ning Tan, 2021].

Here are the steps to perform logistic regression in banking:

Step 1: Data Collection Collect relevant data from the banking domain. This can include features such as customer demographics (age, gender, income), credit history, transaction patterns, loan default status, etc. The data should consist of labeled examples, where each example is associated with a binary outcome (e.g., whether a customer defaulted on a loan or not).

Step 2: Data Preprocessing Clean and preprocess the data to ensure it is in a suitable format for logistic regression. This may involve handling missing values, encoding categorical variables, scaling numerical features, and splitting the data into training and testing sets.

Step 3: Feature Selection Analyze the collected data and select relevant features that are likely to have a significant impact on the binary outcome. This step involves exploratory data analysis (EDA) techniques, such as correlation analysis, feature importance ranking, and domain expertise.

Step 4: Model Training Train a logistic regression model using the training data. In logistic regression, the goal is to learn a set of coefficients that define a linear boundary between the two classes. The model estimates the probability of an instance belonging to the positive class (e.g., loan default) based on the given features.

Step 5: Model Evaluation Evaluate the performance of the trained model using the testing data. Common evaluation metrics for logistic regression include accuracy, precision, recall, and F1-score. These metrics provide insights into the model's ability to correctly classify positive and negative instances.

Step 6: Model Optimization If the model's performance is not satisfactory, you can optimize it by fine-tuning hyperparameters, applying feature engineering techniques, or trying different algorithms. This step aims to improve the model's predictive accuracy and generalization capabilities.

2.6 Evaluation Metric

The performance metrics are used to evaluate the generalization power of the trained model and its quality when it is examined with unseen data. In the classification models, different metrics can be used to evaluate the efficiency of a particular classification algorithm. This encompasses accuracy, F1-measure, precision, and recall.

One of the most common metrics to evaluate the generalization power of models is the accuracy [Pang-Ning Tan, 2021]. Through accuracy, the trained model is evaluated based on the total instances that are correctly predicted by the trained model when it is tested with the unseen data. The recall, precision, and F1-measure metrics are used to deal with imbalanced class problems for optimizing the accuracy performance [Pang-Ning Tan, 2021]. The calculation of these measures is based on computing the confusion matrix. This matrix summarizes the number of instances wrongly or properly predicted by a classification model (see Table (2.2)) [P. Tang, 2006]:

Table (2.2) Two-Dimensional Confusion Matrix

Predicte *Actual*	***Positive***	***Negative***
Positive	TP	FN
Negative	FP	TN

1. True Positive (TP): The positive examples which are properly classified.
2. False Negative (FN): The positive examples which are wrongly classified.
3. False Positive (FP): The negative examples which are wrongly classified.
4. True Negative (TN): The negative examples which are properly classified.

The measures accuracy, recall, precision, and F1- measure are discussed here:

1. Accuracy is the number of correct predictions divided by the total number of predictions. The accuracy can be computed based on Equation 2.2 [P. Tang, 2006].

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{TP+TN+FP + FN}} \quad \dots\dots\dots \ (2.4)$$

2. Precision is the number of TP divided by the number of TP and FP. The precision can be computed based on Equation 2.3 [P. Tang, 2006].

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \dots\dots\dots\dots\dots\dots\dots.. \ (2.5)$$

3. A recall is the number of TP divided by the number of TP and the number of FN. This metric can be computed based on Equation 2.4 [P. Tang, 2006].

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad \dots\dots\dots\dots\dots\dots\dots\dots. \ (2.6)$$

4. F1-measure is the 2*((precision*recall)/ (precision + recall)). An equation of this metric can be computed based on Equation 2.5 [P. Tang, 2006].

$$\text{F1} - \text{score} = \frac{(2*\text{TP})}{(2*\text{TP+FN+FP})} \quad \dots\dots\dots. (2.7)$$

5. Error rate It can be defined as the number of all wrong predictions divided by the entire number of dataset predictions. An equation of this metric can be computed based on Equation 2.6 [P. Tang, 2006].

$$\text{ERR} = \frac{FN\text{+FP}}{\text{TP+ FN+FP+TN}} \quad \dots\dots\dots\dots\dots\dots \ (2.8)$$

Besides in some cases, it is calculated as follow:

Error Rate = Incorrect Predictions / Total Predictions (2.9)

Error Rate = 1 – Accuracy (2.10)

Chapter Three
Proposed Model

Chapter Three
Proposed Model

3.1 Introduction

Predictive modeling is the method of getting known outcomes and creating a model that can expect values for new events. It uses past data to forecast future actions. There are many dissimilar forms of predictive modeling methods in this chapter following methods are used decision trees, KNN, Logistic regression, Naïve Bayes. Choosing the accurate predictive modeling method at the beginning of a task can save time. Selecting the improper modeling technique can end in imprecise predictions and residual plots that knowledge non-constant adjustment. After defining the variables, several types of models can be formed. The greatest common ones for predicting consumer performance are: decision trees and logistic regression. Decision trees utilizes graph or typical decisions to define the conditional probability of a consequence. Logistic regression model can be created to predict the probability of occurrence of an event. Commonly used metrics are Cumulative Gains Chart, Lift Chart, and Receiver Operating Characteristics (ROC) curve. All of these provide metrics by trading off desirable outcomes. These metrics can be obtained by implementing the model on the training data set.

3.2 Proposed Model Architecture

This research is started by downloading an open-source data set. After verifying the data set, next step is preprocessing and data discretization in the form of Data Cleaning (Missing values), Data Transformation (Nominal to Binary, Nominal to Numeric), Normalization (Standardization, smote), Data Reduction, Binning and Select Attributes. And applied ranking feature, Correlation and Predictive model. After applying all these techniques on downloaded dataset, the main technique feature selection is applied later on, following algorithms are applied on the data i.e., Decision Tree, Logistic Regression, KNN, and Naïve Bayes. After applying algorithms and techniques we compare results and discuss about conclusion. Figure (3.1) summaries the research methodology steps implemented in this thesis.

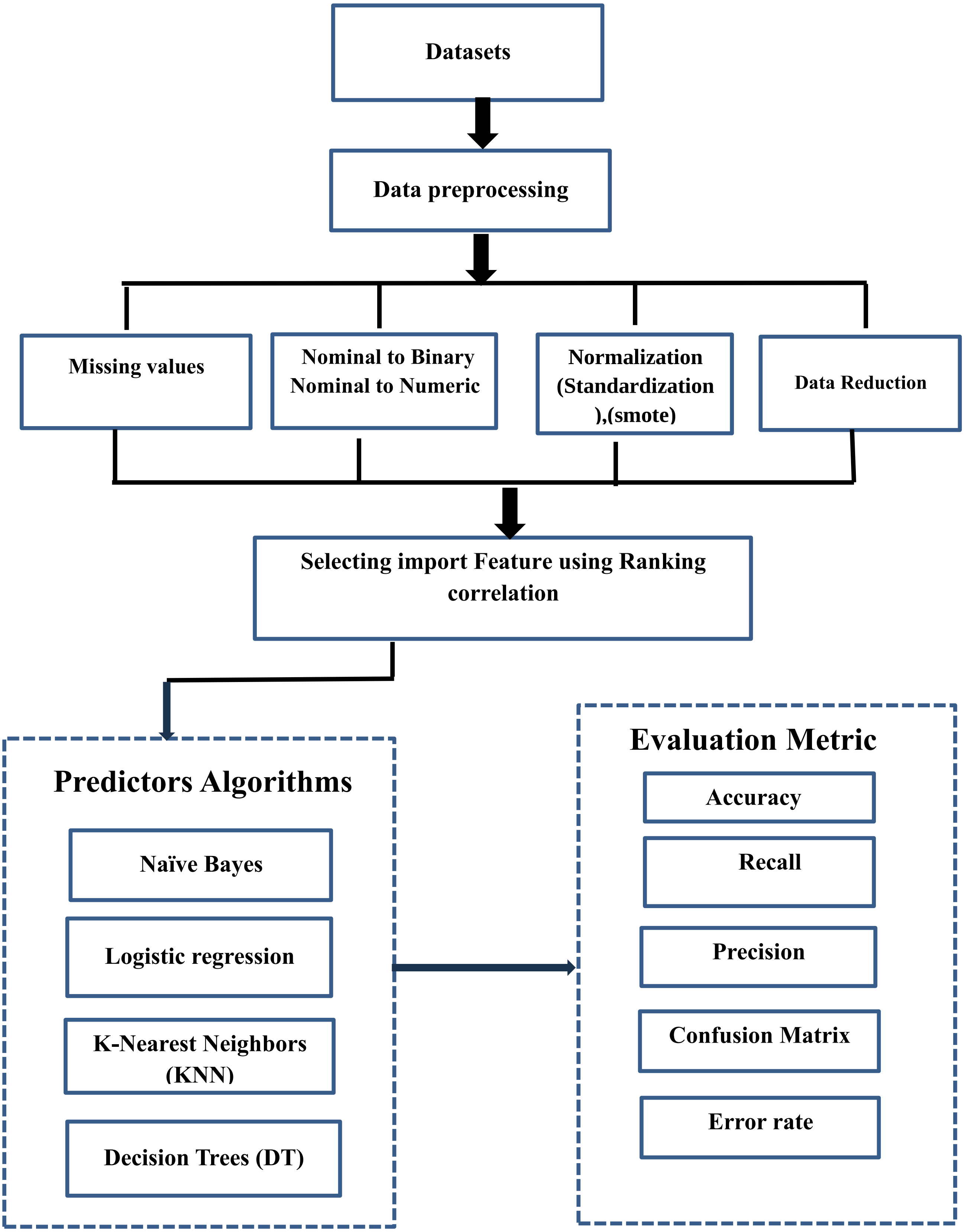


Figure (3.1) Proposed Model Architecture

3.2.1 Select dataset

Two different data sets are selected and downloaded. The 1st dataset is, a bank-additional-full downloaded from "https://www.kaggle.com/datasets/sahistapatel96/bankadditionalfullcsv ". The data is related to direct marketing campaign direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014] .The 2nd dataset is a Banking Dataset - Marketing Targets downloaded from "https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets". This paper used a bank marketing dataset gathered from a Portuguese retail bank. This dataset was created by P. Rita, P. Cortez, and S. Moro from actual bank data. The dataset contains information regarding a Portuguese banking institution's direct marketing campaign, and it contains 45,211 phone contacts. Each contact has 16 input attributes and one decision attribute. The types of input attributes are different. Three attributes are binary type, seven attributes are numeric, and six attributes are categorical. The target attribute is binary type. It has two results; the client subscribes to a term deposit or not.

3.2.2 Data preprocessing

In this thesis, many steps were followed to prepare the research data for the prediction model. The data is processed and prepared into prediction classification stage because actual datasets like (Dataset - Marketing Targets and bank-additional-full) datasets might have some unsuitable structure. This stage consists of four steps include Data Cleaning (Missing values), Data Transformation (Nominal to Binary, Nominal to Numeric), Normalization (Standardization, Smote), Data Reduction. Algorithm (3.1) demonstrates this process.

```
Algorithm (3.1): Preprocessing Dataset.
Input: Two dimensional array Dataset [$n*m$] in which $n$ is number of rows and $m$ number of Colum.
Output: Processed Dataset.
Step 1: Checking the Missing Values of Dataset.
Step 2: transforming Date nominal feature to numeric feature.
Step 3: Normalization of Dataset &(Smote).
Step 4: Data Reduction.
End.
```

A-Missing values

In this Stage, the missing value is an empty cell for the missing feature in the dataset or this cell may contain the letter N, or it may contain the word null which indicates a missing value which controls the results of the calculation and gives inaccurate results. There are different ways to deal with missing values. The used datasets in this thesis are found without missing values.

B- Data Transformation

The main effectiveness of pre-pressing is preparing data for mining in proper form. It may require transforming nominal features of the two datasets (Dataset - Marketing Targets and bank-additional-full) to numerical features. Algorithm (3.2) illustrates this process.

Algorithm (3.2): Transformation of Nominal into Numbers

```
Input:  Dataset [n*m] in which n is number of rows and m number of Colum.
Output: Dataset [n*m] of Numeric Values.
for i = 1 to n
    for j = 1 to m
        N=0
            If Dataset [i, j] = Nominal _value.
                N = Numbers _value
                Dataset [i, j] = N
    end for j
end for i
End
```

C- Normalization

The normalization stage was used for all numerical features values within the dataset, which are inputs for machine learning algorithms. This process was implemented to avoid features with large values controlling the calculation results. All values are normalized to be within the constant range between zero and one using the Standardization normalization method mentioned in equation (2.1).

The Standard scaler algorithm is a common preprocessing step used in machine learning to scale numerical features of a dataset to have mean (μ)= 0 and variance(σ)= 1. This process is essential when features have different ranges, as it helps models that rely on distance calculations or gradient descent to perform better and converge faster. Algorithm (3.3) illustrates this process.

```
Algorithm (3.3): Normalization
Input: Dataset [n*m] in which n is number of rows and m number
         of Colum.
Output: Z1 after applying normalization.
For v =1 to m:
    A) Compute Mean of each input features of v (µv)
         $µv = ( \frac{\sum_{1}^{i} x}{n} )$  .......... (3.2)   // x: all values of features
                                                   // n: numbers of values
         B) Compute standard deviation of each input features of
         v ($SD_v$)
          $SDv = \sqrt{( \frac{\sum_{1}^{i}(x-\mu)2}{n-1} )}$      ............. (3.1)
         End for v
              for i = 1 to   m      // m: all features
               for j = 1 to n      // n: values of features
                 Z1 (j, i) = (x (j, i) -     µv)  /  SDv
          end for j
     end for i
  Z1   is z-score normalization
End.
```

D- SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is an algorithm used for imbalanced learning in machine learning. It is specifically designed to address the problem of imbalanced class distribution, where class A has significantly fewer instances compared to the class B. This process depicted in Algorithm (3.4).

```
Algorithm (3.4): SMOTE of Dataset
Input:   Dataset [n*m] in which n is number of rows and m number
         of Colum.
          class A: number of labels of class A (Minor).
          class B: number of labels of class B. (Major)
```

```
Output: Balanced between class A, class B
        R= Major – Minor
for k = 1 to R
  While  Major <> Minor do // compare between class A, class B
    for i = 1 to m
      for j = 1 to n
        Call  KNN class A [i, j],
        end for j
    end for i
End While.
end for k
End.
```

3.2.3 Data Reduction

This study aims to identify the most important characteristics and explore their influence on the prediction model using the reduction levels for banking dataset. There are many features in the banking dataset, but not all of them are appropriate for the prediction process because some of them decreased the prediction model's accuracy and increased the processing time. This thesis uses the feature selection approach to extract the most significant features that have an impact on banking. The reduction technique is divided into two steps: Ranking Correlation and Feature Selection.

A-Ranking correlation

The process starts by selecting a banking attribute and test its correlation coefficient with the second, third, …… till the last one and so on for other attributes. It is the first step of data reduction, so the importance of features will be calculated based on an equation (2.2) to be initialized to the next step. By identifying relevant features, this method's primary goal is to reduce the dimensionality of the dataset. The correlation coefficient for a feature calculated. Using this process, the attributes with the highest value were selected. Algorithm (3.5) illustrates this process.

Algorithm (3.5): correlation coefficient Method

```
Input: Two dimensional array Dataset [n*m] in which n is number of rows and m number of Colum.
Output: correlation coefficient for Dataset [n*m]
 for i = 1 to n   // n: values of features
     for j = 1 to m // m: all features
 Compute the correlation coefficient betweenfeature Fi and feature Fj according to the equations in section (2.2)
    end for j
 end for i
END
```

B-Feature Selection

In the Bank dataset, a feature selection approach was used to reduce feature space dimensions, select the important features for the prediction process, and improve prediction accuracy. A subset of the most useful features is selected as a result of this phase. Feature selection techniques were used first. However, a feature sequential selection (Ranking correlation) approach has been proposed to reduce the number of features and increase prediction accuracy because a very high degree of accuracy cannot be achieved. Since the dataset includes a huge number of features, the main stage in this thesis is the selection of appropriate features to reduce the curse of dimensionality. Algorithm (3.6) illustrates this selection.

```
Algorithm (3.6): Feature Selection
Input: Normalization of Dataset[n*m] in which n is number of rows and
m number of Colum.
R: array[n] of rank correlation for class label.
Output:  Reduce of features for Dataset[n*m].
 for i = 1 to m
     for j = 1 to m
          Call correlation coefficient Method R[i], R[j]
          if R[i] >R[j]
           Max = R[i]
Select and Order the features (Fi) with maximum correlation (Ri)
The selected features are sorted by the correlation coefficient
values in descending order
        end if
      end for j
  end for i
End.
```

3.2.4 Predictive modeling

Predictive modeling is a machine learning technique that would work best to predict the future outcomes. Predictive modeling represents a statistical method to analyze the patterns of the data to estimate future outcomes or events. It can be considered as an essential feature of predictive analytics. To perform a predictive modeling approach, four prediction models are utilized in this thesis (Naïve bayes, Decision tree, KNN and Logistic).

- ***Naïve bayes:***

The goal of using Naive Bayes in banking operations is to improve the accuracy and efficiency of predicting outcomes based on given input data. To apply the Naive Bayes algorithm based on the used datasets, the important step is to observe the collected data which include information such as previous, emp.var. rate, poutcome, contact, cons.price.idx, nr. employed, pdays, Duration, default, campaign, month, marital, job, age, education, day_ of_ week, housing, loan, cons.conf.idx. Using Naive Bayes theory for predictive modeling in banking operations can help to improve decision-making, reduce risk, and increase efficiency. The relationship between SMOTE, standardization and data reduction association with Naïve bayes achieved the best results. Implementing NB algorithm based a Python Software requires checking different situations. These situations are implementing the NB without activating the SMOTE, data reduction and Standardization functions. The second step is to activate one of these three function and activate different two. The final step is to activate these three functions and observe their effects on the prediction values. Table (3.1) presents the relation among SMOTE, Standardization and data reduction.

Table (3.1) Implementing NB with SMOTE, data reduction and standardization.

Algorithm NB	
SMOTE *Data Reduction* *Standardization*	*False* *Fales* *False*
SMOTE *Data Reduction*	*True* *False*

Standardization	*False*
SMOTE *Data Reduction* *Standardization*	*False* *True* *False*
SMOTE *Data Reduction* *Standardization*	*False* *False* *True*
SMOTE *Data Reduction* *Standardization*	*True* *False* *True*
SMOTE *Data Reduction* *Standardization*	*False* *True* *True*
SMOTE *Data Reduction* *Standardization*	*True* *True* *False*
SMOTE *Data Reduction* *Standardization*	*True* *True* *True*

❖ ***Decision tree***

Decision Tree (DT) algorithms are widely used in banking for various applications, to predict if the client will subscribe (yes/no) a term deposit. DTs are a supervised learning technique that uses a tree-like model to make predictions or decisions based on input features. The working principle of DT based on the Smote, standardization and data reduction. DT achieved the highest accuracy when it correlates with the important features. Table (3.2) presents the relation among SMOTE, Standardization and data reduction

Table (3.2) Implementing DT with SMOTE, data reduction and standardization

Algorithm DT	
SMOTE Data Reduction Standardization	False Fales False
SMOTE Data Reduction Standardization	True False False
SMOTE Data Reduction Standardization	False True False
SMOTE Data Reduction Standardization	False False True
SMOTE Data Reduction Standardization	True False True
SMOTE Data Reduction Standardization	False True True
SMOTE Data Reduction Standardization	True True False
SMOTE Data Reduction Standardization	True True True
SMOTE Important features Standardization	True True False
SMOTE Important features Standardization	False True True
SMOTE Important features	True True

Standardization	*True*

- ***KNN***

KNN algorithms are widely used in banking for various applications, to predict if the client will subscribe (yes/no) a term deposit. The algorithm calculates the distance between the attributes of the new applicant and those of existing customers to determine the nearest neighbors. By considering the labels (default or non-default) of the nearest neighbors, the algorithm predicts by association with Smote, standardization and data reduction. Table (3.3) presents the relation among SMOTE, Standardization and data reduction.

Table (3.3) Implementing KNN with SMOTE, data reduction and standardization

Algorithm KNN	
SMOTE *Data Reduction* *Standardization*	*False* *Fales* *False*
SMOTE *Data Reduction* *Standardization*	*True* *False* *False*
SMOTE *Data Reduction* *Standardization*	*False* *True* *False*
SMOTE *Data Reduction* *Standardization*	*False* *False* *True*
SMOTE *Data Reduction* *Standardization*	*True* *False* *True*

SMOTE *Data Reduction* *Standardization*	*False* *True* *True*
SMOTE *Data Reduction* *Standardization*	*True* *True* *False*
SMOTE *Data Reduction* *Standardization*	*True* *True* *True*

- ***Logistic Regression***

In the context of banking, logistic regression can be utilized for several purposes for customer churn prediction. To apply the logistic algorithm to the prediction model, it is necessary to relate with Smote, standardization and data reduction. Table (3.4) presents the relation among SMOTE, Standardization and data reduction.

Table (3.4) Implementing LG with SMOTE, data reduction and standardization

Algorithm LG	
SMOTE *Data Reduction* *Standardization*	*False* *Fales* *False*
SMOTE *Data Reduction* *Standardization*	*True* *False* *False*
SMOTE *Data Reduction* *Standardization*	*False* *True* *False*
SMOTE *Data Reduction* *Standardization*	*False* *False* *True*

SMOTE *Data Reduction* *Standardization*	*True* *False* *True*
SMOTE *Data Reduction* *Standardization*	*False* *True* *True*
SMOTE *Data Reduction* *Standardization*	*True* *True* *False*
SMOTE *Data Reduction* *Standardization*	*True* *True* *True*

3.2.5 Evaluate the models

Predictive modeling is the method of getting known outcomes and creating a model that can expect values for new events. It uses past data to forecast future actions. There are many dissimilar forms of predictive modeling methods in this chapter following methods are used decision trees, KNN, Logistic regression, Naïve Bayes. Choosing the accurate predictive modeling method at the beginning of a task can save time. Selecting the improper modeling technique can end in imprecise predictions and residual plots that knowledge non-constant adjustment. After defining the variables, several types of models can be formed. The greatest common ones for predicting consumer performance are: decision trees and logistic regression. Decision trees utilizes graph or typical decisions to define the conditional probability of a consequence. Logistic regression model can be created to predict the probability of occurrence of an event. Commonly used metrics are Cumulative Gains Chart, Lift Chart, and Receiver Operating Characteristics (ROC) curve. All of these provide metrics by trading off desirable outcomes. These metrics can be obtained by implementing the model on the training data set.

Chapter Four

Results and Discussion

Chapter Four

Results and Discussion

4.1 Introduction

The effectiveness of the proposed system is illustrated in chapter three, with different parameters values, and the results will be presented in this chapter. Datasets have been used to determine the behavior of the implemented models. The experimental results stages of system are described and discussed.

4.2 System Requirement

Hardware: Processor Intel(R), Core (TM), i5-6440HQ CPU, 2.60 GHz, Ram 8 GB.

Operating System windows 10, 64 bits.

Programming Language with python.

IDE Programming is done in Jupiter environment.

WEKA 3.7.11 Machine Learning Package.

4.3 Datasets

There are two sets of data used:

4.3.1 bank-additional-full

This dataset is to indicate the direct marketing movement of a banking institution in Portuguese. Its data was collected based on phone calls to access if the bank term deposit would be ('No') or ('Yes'). It was represented as csv file contains (41188) records and 20 inputs. Its data are organized from “May 2008 to November 2010”. Table (4.1) contains the detailed description of all attributes in this dataset after understanding the context from Kaggle. Table (4.2) present a sample from this dataset.

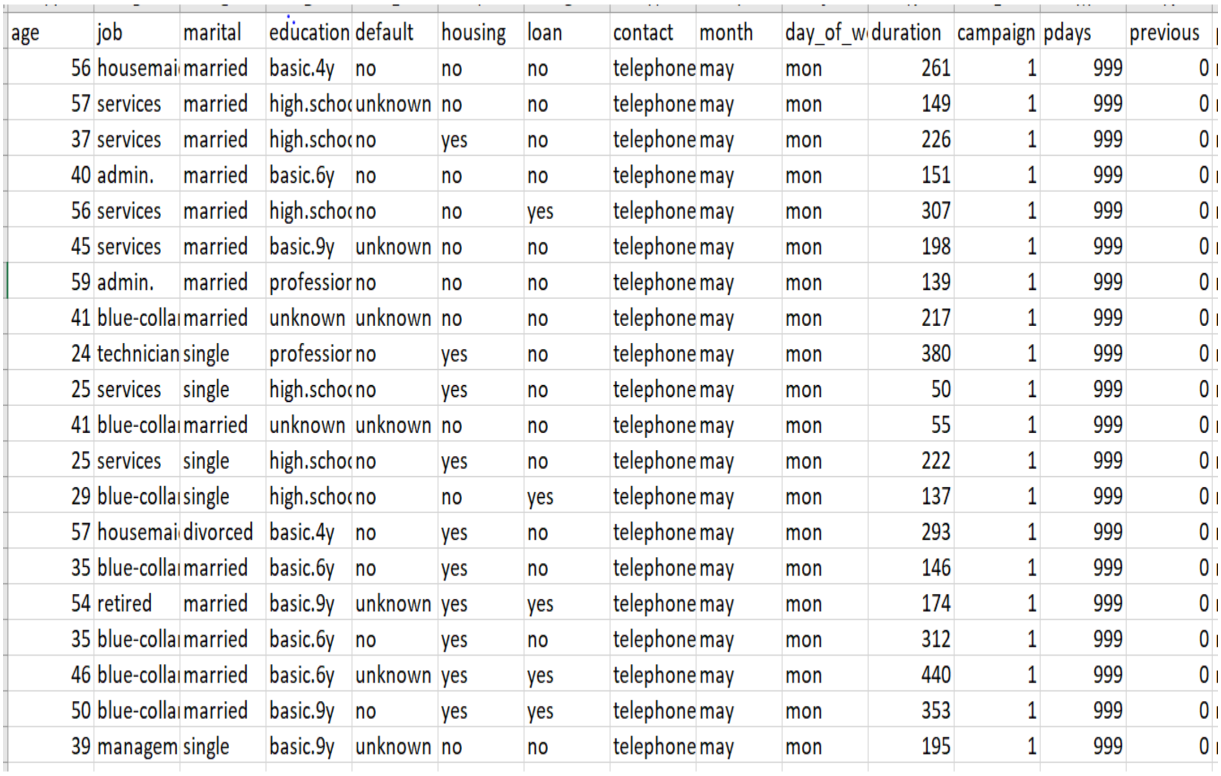
Table (4.1) Dataset Description

Variable	Variable Description
bank client data	
Age	Numeric
Job	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
Marital	marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
Education	education (categorical: 'basic.4y','basic.6y','basic.9y','high. school','illiterate','professional. course','university. degree','unknown')
Default	has credit in default? (Categorical: 'no','yes','unknown')
Housing	has housing loan? (Categorical: 'no','yes','unknown')
Loan	has personal loan? (Categorical: 'no','yes','unknown')

related with the last contact of the current campaign	
Contact	contact communication type (categorical: 'cellular’, ‘telephone')
Month	last contact month of year (categorical: 'jan', 'feb', 'mar', …, 'nov', 'dec')
Duration	last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
other attributes	
Campaign	number of contacts performed during this campaign and for this client (numeric, includes last contact)
Pdays	number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
Previous	number of contacts performed before this campaign and for this client (numeric)

Poutcome	outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
emp.var. rate	employment variation rate - quarterly indicator (numeric)
cons.price.idx	consumer price index - monthly indicator (numeric)
cons.conf.idx	consumer confidence index - monthly indicator (numeric)
- euribor3m	euribor 3-month rate - daily indicator (numeric)
nr. employed	number of employees - quarterly indicator (numeric)
Output variable (desired target)	
Y	has the client subscribed a term deposit? (Binary: 'yes','no')

Table (4.2) bank-additional-full dataset sample



age	job	marital	education	default	housing	loan	contact	month	day_of_w	duration	campaign	pdays	previous
56	housemai	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0
57	services	married	high.scho	unknown	no	no	telephone	may	mon	149	1	999	0
37	services	married	high.scho	no	yes	no	telephone	may	mon	226	1	999	0
40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0
56	services	married	high.scho	no	no	yes	telephone	may	mon	307	1	999	0
45	services	married	basic.9y	unknown	no	no	telephone	may	mon	198	1	999	0
59	admin.	married	professio	no	no	no	telephone	may	mon	139	1	999	0
41	blue-colla	married	unknown	unknown	no	no	telephone	may	mon	217	1	999	0
24	technician	single	professio	no	yes	no	telephone	may	mon	380	1	999	0
25	services	single	high.scho	no	yes	no	telephone	may	mon	50	1	999	0
41	blue-colla	married	unknown	unknown	no	no	telephone	may	mon	55	1	999	0
25	services	single	high.scho	no	yes	no	telephone	may	mon	222	1	999	0
29	blue-colla	single	high.scho	no	no	yes	telephone	may	mon	137	1	999	0
57	housemai	divorced	basic.4y	no	yes	no	telephone	may	mon	293	1	999	0
35	blue-colla	married	basic.6y	no	yes	no	telephone	may	mon	146	1	999	0
54	retired	married	basic.9y	unknown	yes	yes	telephone	may	mon	174	1	999	0
35	blue-colla	married	basic.6y	no	yes	no	telephone	may	mon	312	1	999	0
46	blue-colla	married	basic.6y	unknown	yes	yes	telephone	may	mon	440	1	999	0
50	blue-colla	married	basic.9y	no	yes	yes	telephone	may	mon	353	1	999	0
39	managem	single	basic.9y	unknown	no	no	telephone	may	mon	195	1	999	0

4.3.2 Data Preprocessing Results

The goal of this stage is to prepare the data for the prediction and the portfolio optimization stages. Steps of the preprocessing stage are applied to bank-additional-full dataset.

A- Missing Values Processing Results

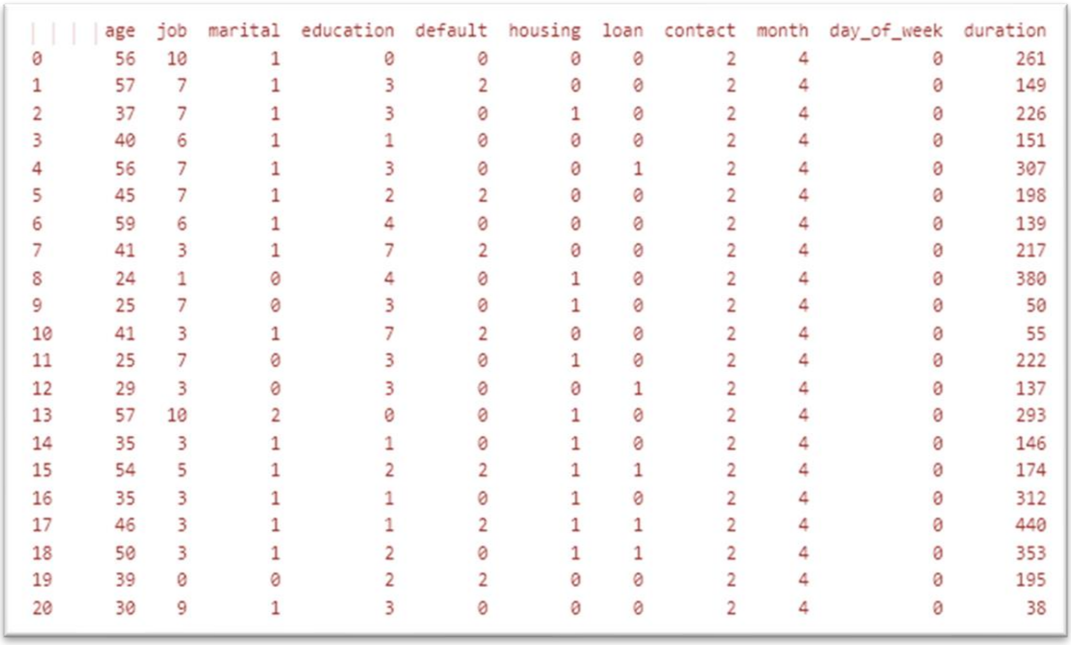
In this stage, the bank-additional-full dataset data will be tested to see if this data contains missing values (non-null) or not. After a number of tests, we noticed that the bank-additional-full dataset is free from missing values.

B- Nominal to Numeric

In this step, transform attributes to numeric form, which is performed on categorical attributes (y, contact, loan, housing, job, marital, education, default, month, poutcome, day-of-week). as shown in Table (4.3) This process converts the data type from Nominal into Numeric. This is done by converting the attributes into numeric as:

- y’: {“'no':0,'yes':1”},
- 'contact’: {“'cellular':1,'telephone':2”},
- 'loan’: {“'no':0,'yes':1,'unknown':2”},
- 'housing’: {“'no':0,'yes':1,'unknown':2”},
- 'job': {“'management':0, 'technician':1, 'entrepreneur':2, ‘blue- collar':3, 'unknown':4, 'retired':5, ‘admin.':6, ‘services':7, 'self-employed':8, 'unemployed':9, ‘housemaid':10, 'student':11”},
- 'marital’: {“'single':0,'married':1,'divorced':2,'unknown':3”},
- 'education’: {“'basic.4y':0,'basic.6y':1,'basic.9y':2,'high. school':3,'professional. course':4,'university. degree':5,'illiterate':6,'unknown':7”},
- 'default’: {“'no':0,'yes':1,'unknown':2”},
- 'month’: {“'jan':0, 'feb':1, 'mar':2, ‘apr':3, ‘may':4, ‘jun':5, ‘jul':6, ‘aug':7, 'sep':8, 'oct':9, ‘nov':10, ‘dec':11”},
- 'poutcome’: {“'failure':0,'success':1,'nonexistent':2”},
- 'Day_of_week’: {“'mon':0,'tue':1,'wed':2,'thu':3,'fri':4”}

Table (4.3) Nominal to Numeric Result

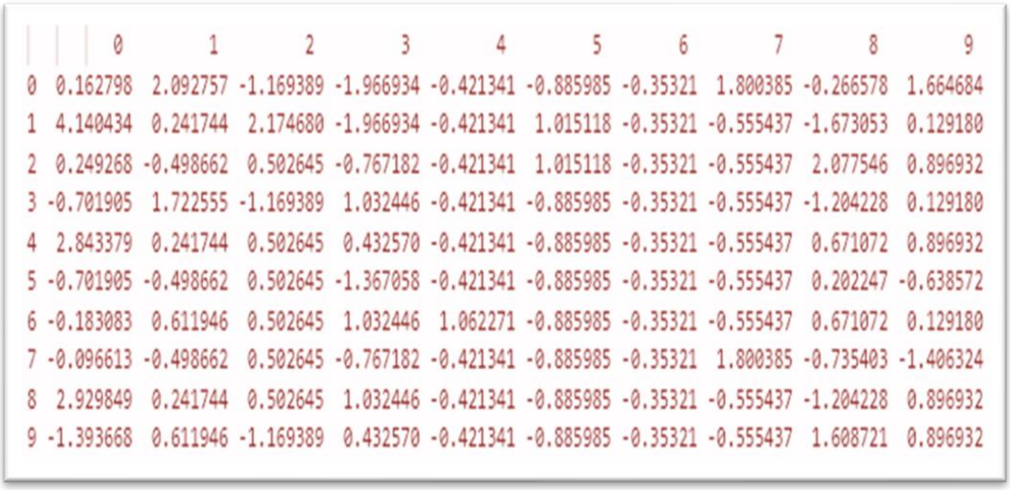


	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration
0	56	10	1	0	0	0	0	2	4	0	261
1	57	7	1	3	2	0	0	2	4	0	149
2	37	7	1	3	0	1	0	2	4	0	226
3	40	6	1	1	0	0	0	2	4	0	151
4	56	7	1	3	0	0	1	2	4	0	307
5	45	7	1	2	2	0	0	2	4	0	198
6	59	6	1	4	0	0	0	2	4	0	139
7	41	3	1	7	2	0	0	2	4	0	217
8	24	1	0	4	0	1	0	2	4	0	380
9	25	7	0	3	0	1	0	2	4	0	50
10	41	3	1	7	2	0	0	2	4	0	55
11	25	7	0	3	0	1	0	2	4	0	222
12	29	3	0	3	0	0	1	2	4	0	137
13	57	10	2	0	0	1	0	2	4	0	293
14	35	3	1	1	0	1	0	2	4	0	146
15	54	5	1	2	2	1	1	2	4	0	174
16	35	3	1	1	0	1	0	2	4	0	312
17	46	3	1	1	2	1	1	2	4	0	440
18	50	3	1	2	0	1	1	2	4	0	353
19	39	0	0	2	2	0	0	2	4	0	195
20	30	9	1	3	0	0	0	2	4	0	38

C- Normalization Process Results

Normalization is an important step that has been conducted on the banking dataset(bank-additional-full) to minimize substantial value disparities from dominating the results. The features values ranged between zero and one by using the Standardization normalization approach. Table (4.4) shows the data normalization on a sample of the banking dataset. Whereas the columns represent the attributes, while the rows represent the sequence of data.

Table (4.4) The Normalization Result

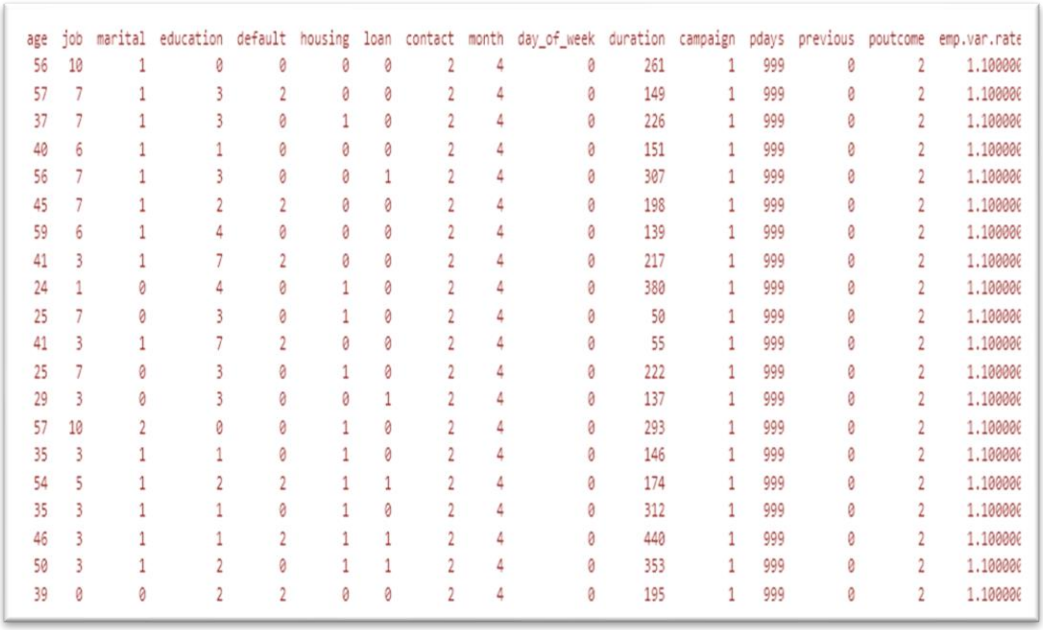


	0	1	2	3	4	5	6	7	8	9
0	0.162798	2.092757	-1.169389	-1.966934	-0.421341	-0.885985	-0.35321	1.800385	-0.266578	1.664684
1	4.140434	0.241744	2.174680	-1.966934	-0.421341	1.015118	-0.35321	-0.555437	-1.673053	0.129180
2	0.249268	-0.498662	0.502645	-0.767182	-0.421341	1.015118	-0.35321	-0.555437	2.077546	0.896932
3	-0.701905	1.722555	-1.169389	1.032446	-0.421341	-0.885985	-0.35321	-0.555437	-1.204228	0.129180
4	2.843379	0.241744	0.502645	0.432570	-0.421341	-0.885985	-0.35321	-0.555437	0.671072	0.896932
5	-0.701905	-0.498662	0.502645	-1.367058	-0.421341	-0.885985	-0.35321	-0.555437	0.202247	-0.638572
6	-0.183083	0.611946	0.502645	1.032446	1.062271	-0.885985	-0.35321	-0.555437	0.671072	0.129180
7	-0.096613	-0.498662	0.502645	-0.767182	-0.421341	-0.885985	-0.35321	1.800385	-0.735403	-1.406324
8	2.929849	0.241744	0.502645	1.032446	-0.421341	-0.885985	-0.35321	-0.555437	-1.204228	0.896932
9	-1.393668	0.611946	-1.169389	0.432570	-0.421341	-0.885985	-0.35321	-0.555437	1.608721	0.896932

4.3.3 SMOTE Process Results

A Synthetic Minority Oversampling Technique is used to generate new data to extend the Minority dataset in order to balance the class label. Table (4.5) Present a sample from the generated data in bank-additional-full dataset.

Table (4.5) illustrates this SMOTE Process Results



age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate
56	10	1	0	0	0	0	2	4	0	261	1	999	0	2	1.100000
57	7	1	3	2	0	0	2	4	0	149	1	999	0	2	1.100000
37	7	1	3	0	1	0	2	4	0	226	1	999	0	2	1.100000
40	6	1	1	0	0	0	2	4	0	151	1	999	0	2	1.100000
56	7	1	3	0	0	1	2	4	0	307	1	999	0	2	1.100000
45	7	1	2	2	0	0	2	4	0	198	1	999	0	2	1.100000
59	6	1	4	0	0	0	2	4	0	139	1	999	0	2	1.100000
41	3	1	7	2	0	0	2	4	0	217	1	999	0	2	1.100000
24	1	0	4	0	1	0	2	4	0	380	1	999	0	2	1.100000
25	7	0	3	0	1	0	2	4	0	50	1	999	0	2	1.100000
41	3	1	7	2	0	0	2	4	0	55	1	999	0	2	1.100000
25	7	0	3	0	1	0	2	4	0	222	1	999	0	2	1.100000
29	3	0	3	0	0	1	2	4	0	137	1	999	0	2	1.100000
57	10	2	0	0	1	0	2	4	0	293	1	999	0	2	1.100000
35	3	1	1	0	1	0	2	4	0	146	1	999	0	2	1.100000
54	5	1	2	2	1	1	2	4	0	174	1	999	0	2	1.100000
35	3	1	1	0	1	0	2	4	0	312	1	999	0	2	1.100000
46	3	1	1	2	1	1	2	4	0	440	1	999	0	2	1.100000
50	3	1	2	0	1	1	2	4	0	353	1	999	0	2	1.100000
39	0	0	2	2	0	0	2	4	0	195	1	999	0	2	1.100000

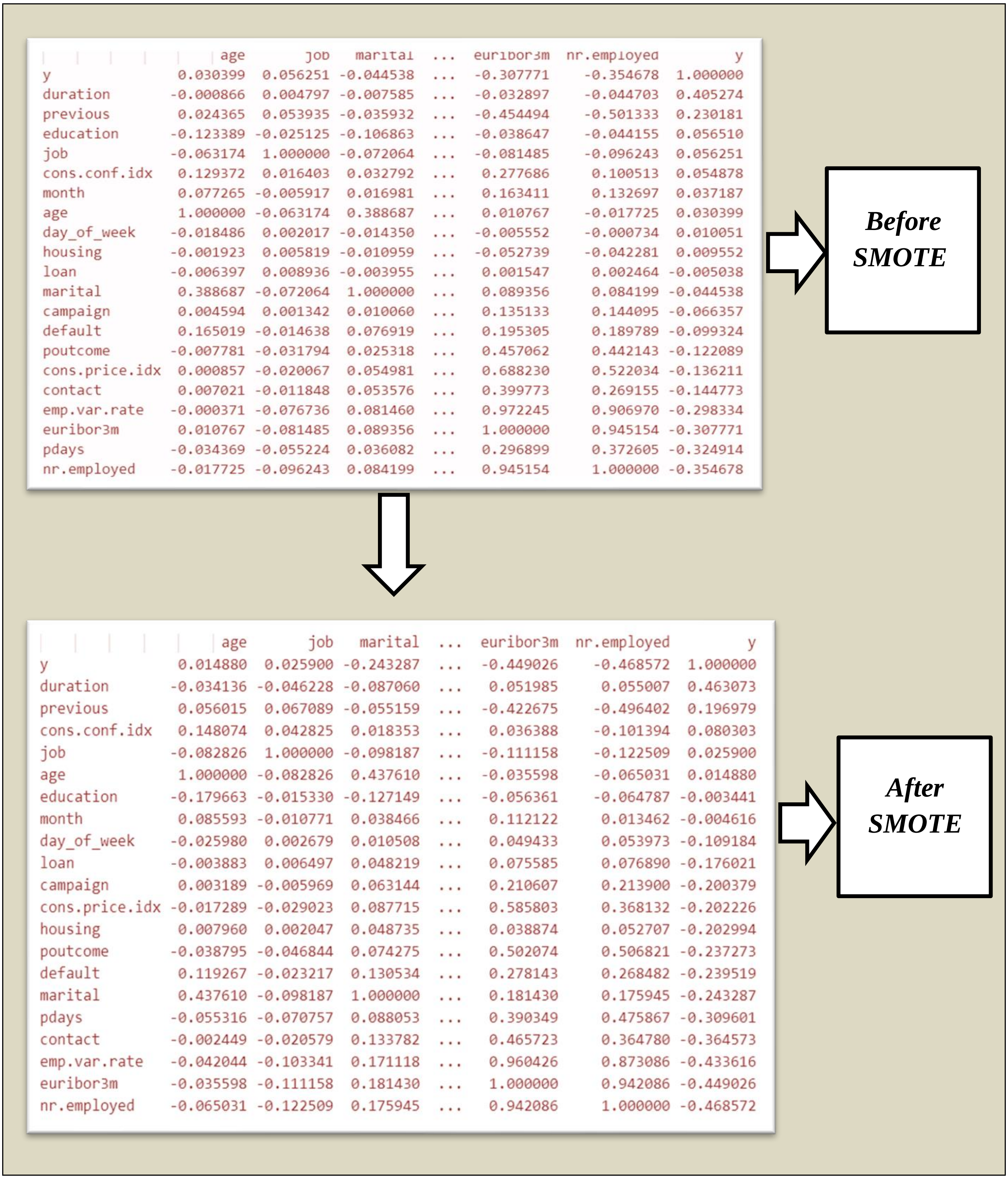
4.3.4 Data Reduction

The goal of is to use the reduction levels for banking dataset (bank-additional-full) to identify the most informative features and investigate their impact on the prediction model. Although the banking dataset (bank-additional-full) contains a large number of features, not all of them are suitable for the prediction process, as some of them reduced the prediction model's accuracy and increased the time complexity. For this reason, the feature selection system is employed in this thesis, in order to extract the most important features that affect banking. Two steps Ranking Correlation and Feature Selection make up the reduction procedure.

A- Ranking Correlation

The main goal of this method is to reduce the dimensions of the dataset (bank-additional-full) by identifying relevant features. The ranking correlation coefficient method is calculated between a feature and a class label. According to this method, the features with the highest value were selected. The first (15) features out of (21) features were selected as informational features to be used for further analysis. Table (4.6) illustrates the Ranking Correlation Result.

Table (4.6) The Ranking Correlation Result



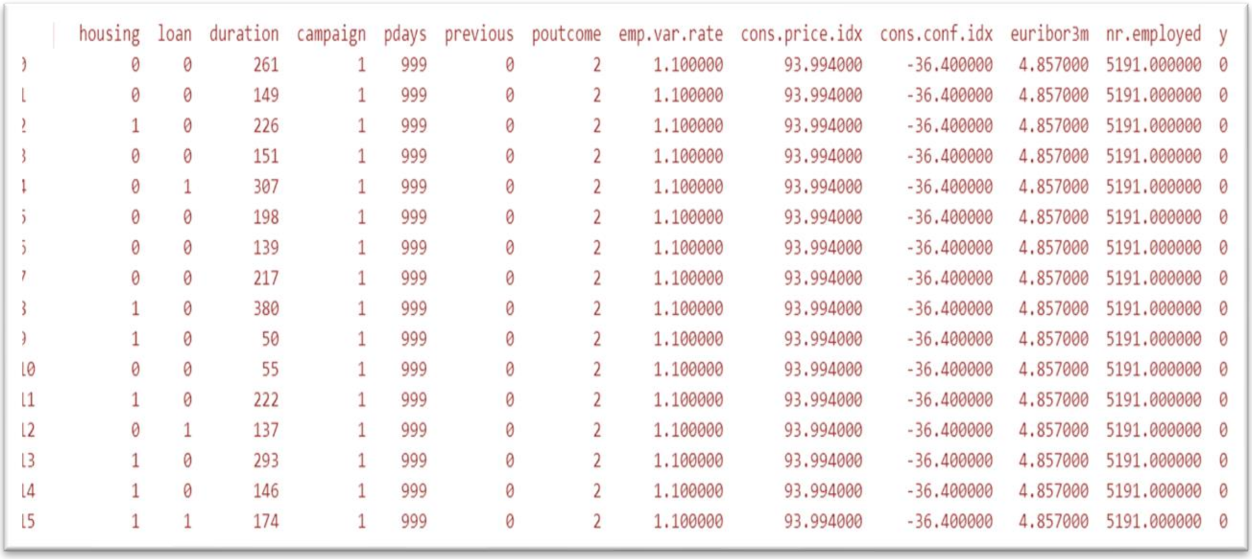
```
                   age       job   marital  ...  euribor3m  nr.employed         y
y             0.030399  0.056251 -0.044538  ...  -0.307771    -0.354678  1.000000
duration     -0.000866  0.004797 -0.007585  ...  -0.032897    -0.044703  0.405274
previous      0.024365  0.053935 -0.035932  ...  -0.454494    -0.501333  0.230181
education    -0.123389 -0.025125 -0.106863  ...  -0.038647    -0.044155  0.056510
job          -0.063174  1.000000 -0.072064  ...  -0.081485    -0.096243  0.056251
cons.conf.idx  0.129372  0.016403  0.032792  ...   0.277686     0.100513  0.054878
month         0.077265 -0.005917  0.016981  ...   0.163411     0.132697  0.037187
age           1.000000 -0.063174  0.388687  ...   0.010767    -0.017725  0.030399
day_of_week  -0.018486  0.002017 -0.014350  ...  -0.005552    -0.000734  0.010051
housing      -0.001923  0.005819 -0.010959  ...  -0.052739    -0.042281  0.009552
loan         -0.006397  0.008936 -0.003955  ...   0.001547     0.002464 -0.005038
marital       0.388687 -0.072064  1.000000  ...   0.089356     0.084199 -0.044538
campaign      0.004594  0.001342  0.010060  ...   0.135133     0.144095 -0.066357
default       0.165019 -0.014638  0.076919  ...   0.195305     0.189789 -0.099324
poutcome     -0.007781 -0.031794  0.025318  ...   0.457062     0.442143 -0.122089
cons.price.idx  0.000857 -0.020067  0.054981  ...   0.688230     0.522034 -0.136211
contact       0.007021 -0.011848  0.053576  ...   0.399773     0.269155 -0.144773
emp.var.rate -0.000371 -0.076736  0.081460  ...   0.972245     0.906970 -0.298334
euribor3m     0.010767 -0.081485  0.089356  ...   1.000000     0.945154 -0.307771
pdays        -0.034369 -0.055224  0.036082  ...   0.296899     0.372605 -0.324914
nr.employed  -0.017725 -0.096243  0.084199  ...   0.945154     1.000000 -0.354678
```

```
                   age       job   marital  ...  euribor3m  nr.employed         y
y             0.014880  0.025900 -0.243287  ...  -0.449026    -0.468572  1.000000
duration     -0.034136 -0.046228 -0.087060  ...   0.051985     0.055007  0.463073
previous      0.056015  0.067089 -0.055159  ...  -0.422675    -0.496402  0.196979
cons.conf.idx  0.148074  0.042825  0.018353  ...   0.036388    -0.101394  0.080303
job          -0.082826  1.000000 -0.098187  ...  -0.111158    -0.122509  0.025900
age           1.000000 -0.082826  0.437610  ...  -0.035598    -0.065031  0.014880
education    -0.179663 -0.015330 -0.127149  ...  -0.056361    -0.064787 -0.003441
month         0.085593 -0.010771  0.038466  ...   0.112122     0.013462 -0.004616
day_of_week  -0.025980  0.002679  0.010508  ...   0.049433     0.053973 -0.109184
loan         -0.003883  0.006497  0.048219  ...   0.075585     0.076890 -0.176021
campaign      0.003189 -0.005969  0.063144  ...   0.210607     0.213900 -0.200379
cons.price.idx -0.017289 -0.029023  0.087715  ...   0.585803     0.368132 -0.202226
housing       0.007960  0.002047  0.048735  ...   0.038874     0.052707 -0.202994
poutcome     -0.038795 -0.046844  0.074275  ...   0.502074     0.506821 -0.237273
default       0.119267 -0.023217  0.130534  ...   0.278143     0.268482 -0.239519
marital       0.437610 -0.098187  1.000000  ...   0.181430     0.175945 -0.243287
pdays        -0.055316 -0.070757  0.088053  ...   0.390349     0.475867 -0.309601
contact      -0.002449 -0.020579  0.133782  ...   0.465723     0.364780 -0.364573
emp.var.rate -0.042044 -0.103341  0.171118  ...   0.960426     0.873086 -0.433616
euribor3m    -0.035598 -0.111158  0.181430  ...   1.000000     0.942086 -0.449026
nr.employed  -0.065031 -0.122509  0.175945  ...   0.942086     1.000000 -0.468572
```

B- Feature Selection

This method aims to reduce the dataset's dimensionality, which is generated using the Feature_importances approach between the feature and the class label, by locating the relevant features. The features with the greatest Feature_importances value were chosen using this method. among (21 characteristics). Table (4.7) shows the Feature Selection of the banking dataset.

Table (4.7) The Feature Selection Result



```
     | housing  loan  duration  campaign  pdays  previous  poutcome  emp.var.rate  cons.price.idx  cons.conf.idx  euribor3m  nr.employed  y
)          0     0       261         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
l          0     0       149         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
?          1     0       226         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
3          0     0       151         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
1          0     1       307         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
5          0     0       198         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
5          0     0       139         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
7          0     0       217         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
3          1     0       380         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
)          1     0        50         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
l0         0     0        55         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
l1         1     0       222         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
l2         0     1       137         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
l3         1     0       293         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
l4         1     0       146         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
l5         1     1       174         1    999         0         2      1.100000       93.994000     -36.400000   4.857000  5191.000000  0
```

4.3.5 Predictive modeling Results

To perform a predictive modeling approach, four prediction models are utilized in this thesis (Naïve bayes, Decision tree, KNN and Logistic). Table (4.8) presents the specific measures for each of the used datasets(bank-additional-full). These measures are (precision, recall f1-score, support).

Table (4.8) presents the Predictive modeling Results

Algorithm type	*Precision*	*Recall*	*F- score*	*support*	*Class*
Naïve bayes	0.83	0.82	0.82	11012	***NO***
	0.82	0.83	0.82	10917	***YES***
	0.82	0.82	0.82	21929	***macro avg***
	0.82	0.82	0.82	21929	***weighted avg***
Decision tree	0.92	0.94	0.93	7130	***NO***
	0.94	0.93	0.93	7490	***YES***
	0.93	0.93	0.93	14620	***macro avg***
	0.93	0.93	0.93	14620	***weighted avg***
KNN	0.87	0.99	0.92	6444	***NO***
	0.99	0.88	0.93	8176	***YES***
	0.93	0.94	0.93	14620	***macro avg***
	0.94	0.93	0.93	14620	***weighted avg***
Logistic	0.97	0.93	0.95	7675	***NO***
	0.40	0.64	0.49	563	***YES***
	0.69	0.78	0.72	8238	***macro avg***
	0.93	0.91	0.92	8238	***weighted avg***

A- Naïve bayes

The initial stage in using the Naive Bayes algorithm in banking operations is to gather information about the current issue. The data gathered, for instance, might reveal particulars like earnings, age, employment history, and other pertinent details. In banking operations, applying Naive Bayes theory to predictive modeling can help to enhance decision-making, lower risk, and boost productivity. The best outcomes were obtained when smote, standardization, and data reduction were combined with Nave Bayes. Table (4.9) presents this process among SMOTE, Standardization and data reduction.

Table (4.9) Result NB with SMOTE, data reduction and standardization

Algorithm NB		***Accuracy***	***Recall***	***Precision***	***Error***
SMOTE Data Reduction Standardization	***False Fales False***	0.8606	0.5951	0.4172	**0.1393**
SMOTE Data Reduction Standardization	***True False False***	0.8245	0.8221	0.8263	**0.1754**
SMOTE Data Reduction Standardization	***False True False***	0.8545	0.5836	0.4004	**0.1454**
SMOTE Data Reduction Standardization	***False False True***	0.8478	0.5972	0.3867	**0.1521**
SMOTE Data Reduction Standardization	***True False True***	0.7946	0.7898	0.7976	**0.2053**
SMOTE Data Reduction Standardization	***False True True***	0.8481	0.5922	0.3867	**0.1518**
SMOTE Data Reduction Standardization	***True True False***	0.8076	0.8198	0.8006	**0.1923**
SMOTE Data Reduction Standardization	***True True True***	0.7814	0.7908	0.776	**0.2185**

B- Decision tree

To implement a Decision Tree (DT) algorithm on the dataset (bank-additional-full), to predict if the client will subscribe (yes/no) a term deposit. DTs are a supervised learning technique that uses a tree-like model to make predictions or decisions based on input features. The working principle of DT based on the Smote, standardization and data reduction. DT achieved the highest accuracy when it correlates with the important features. Table (4.10) presents this process among SMOTE, Standardization, important features and data reduction.

Table (4.10) Result DT with SMOTE, data reduction, important features and standardization

Algorithm DT		***Accuracy***	***Recall***	***Precision***	***Error***
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***Fales*** ***False***	0.8913	0.5508	0.5347	**0.1086**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***False***	0.9281	0.9381	0.9205	**0.0718**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***False***	0.8907	0.5518	0.532	**0.1092**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***False*** ***True***	0.8899	0.5288	0.7392	**0.11009**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***True***	0.9287	0.9381	0.9217	**0.0712**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***True***	0.8911	0.5404	0.5343	**0.10888**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***False***	0.9288	0.9391	0.9211	**0.0711**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***True***	0.9296	0.94009	0.9218	**0.0703**
SMOTE	***True***	0.9336	0.9428	0.9267	**0.0663**

Important features ***Standardization***	***True*** ***False***				
SMOTE ***Important features*** ***Standardization***	***False*** ***True*** ***True***	0.8962	0.5643	0.5556	**0.1037**
SMOTE ***Important features*** ***Standardization***	***True*** ***True*** ***True***	0.9336	0.9428	0.9267	**0.0663**

C- KNN

The prediction of whether a client would sign up for a term deposit using KNN algorithms is a common practice in banking for a variety of applications. In order to identify the closest neighbors, the algorithm evaluates the separation between the new applicant's qualities and those of the existing clientele. The algorithm predicts through association with Smote, standardization, and data reduction while taking the labels (default or non-default) of the closest neighbors into account. Table (4.11) presents this process among SMOTE, Standardization and data reduction.

Table (4.11) Result KNN with SMOTE, data reduction and standardization

Algorithm knn		***Accuracy***	***Recall***	***Precision***	***error***
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***Fales*** ***False***	0.8983	0.4754	0.5372	**0.1016**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***False***	0.9285	0.9907	0.8805	**0.0714**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***False***	0.8953	0.4754	0.5207	**0.1046**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***False*** ***True***	0.9008	0.4241	0.5580	**0.0991**

SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***True***	0.9209	0.9457	0.9001	**0.0790**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***True***	0.8991	0.4787	0.5409	**0.1008**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***False***	0.9093	0.9510	0.8770	**0.0906**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***True***	0.9154	0.9357	0.8983	**0.084**

D- ***Logistic***

For predicting client churn in the banking industry, logistic regression can be used in a variety of ways. It is important to relate with Smote, standardization, and data reduction in order to apply the logistic algorithm to the prediction model. Table (4.12) presents this process among SMOTE, Standardization and data reduction.

Table (4.12) Result Logistic with SMOTE, data reduction and standardization

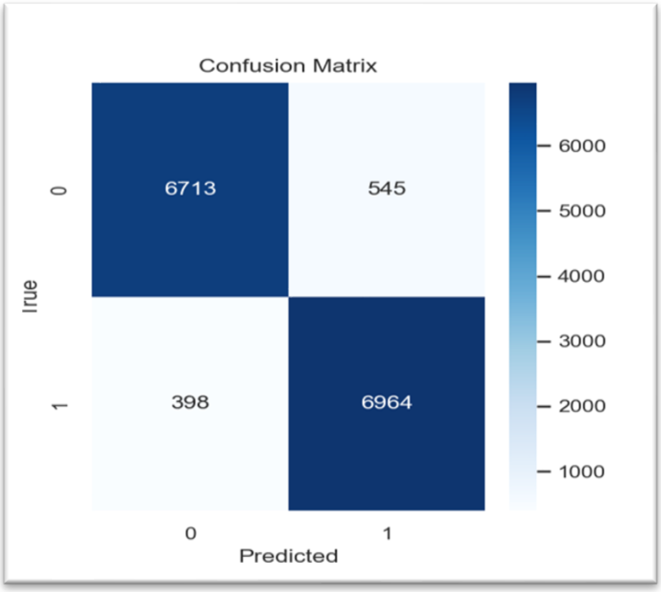
Algorithm LG		***Accuracy***	***Recall***	***Precision***	***error***
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***Fales*** ***False***	0.9081	0.3716	0.6318	**0.0918**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***False***	0.8605	0.8651	0.8558	**0.1394**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***False***	0.9061	0.3616	0.6171	**0.0938**

SMOTE ***Data Reduction*** ***Standardization***	***False*** ***False*** ***True***	0.9102	0.4017	0.6394	**0.0897**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***True***	0.8889	0.8963	0.8821	**0.1110**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***True***	0.9071	0.3727	0.6219	**0.0928**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***False***	0.8625	0.8787	0.8497	**0.1374**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***True***	0.8645	0.8645	0.8630	**0.1354**

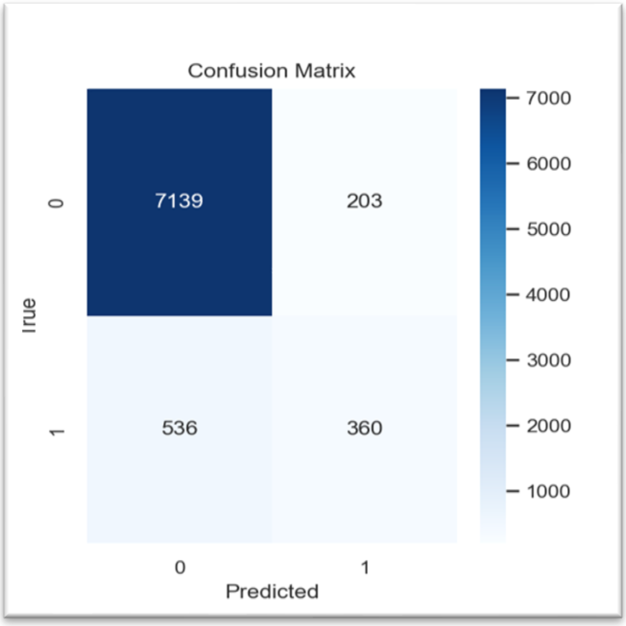
4.3.6 Evaluation Metric Results

The performance of the prediction model (Naïve bayes, Decision tree, KNN and Logistic) was evaluated. Usually, all features that have appeared are taken as influencer features. Then, they were added to the model to confirm their importance and the level of accuracy that could be achieved by these features. In fact, the use of selected features greatly improves accuracy. Confusion matrices are a way to evaluate the performance of a classification model by analyzing the number of true positives, true negatives, false positives, and false negatives. (Decision trees, Naïve bayes, KNN and Logistic) are a popular classification algorithm that can be evaluated using confusion matrices Table (4.13) Confusion Matrixes (LG, DT, NB and KNN).

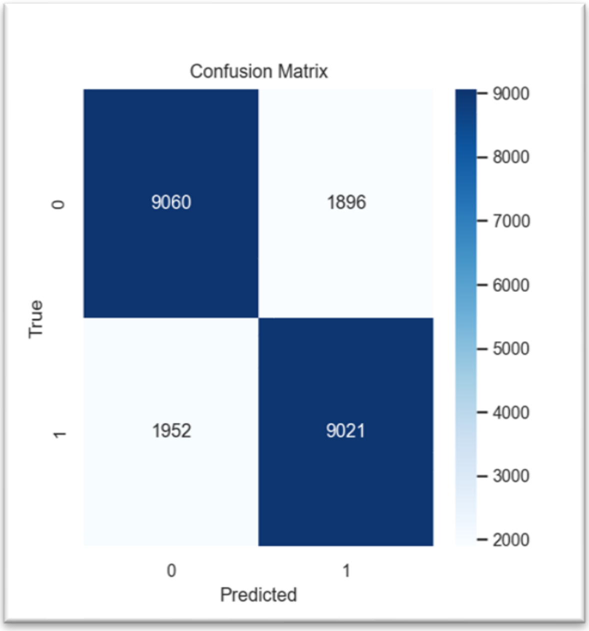
Table (4.13) Confusion Matrixes Results



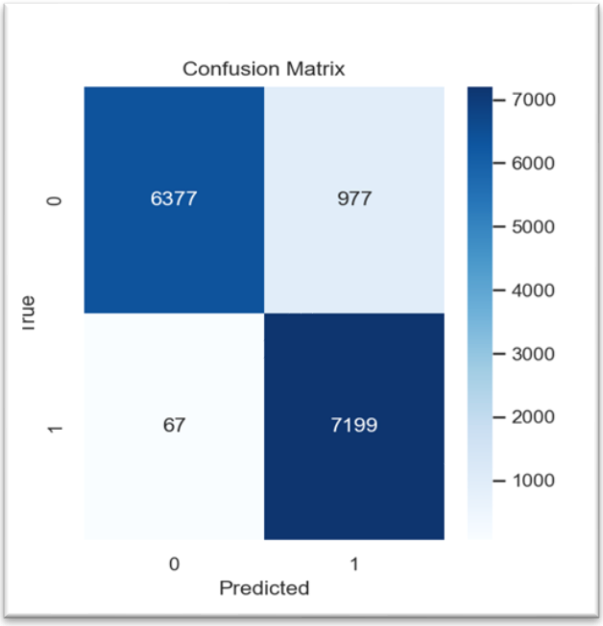
confusion matrix DT



confusion matrix LG



confusion matrix NB



confusion matrix KNN

4.4 Banking Dataset - Marketing Targets

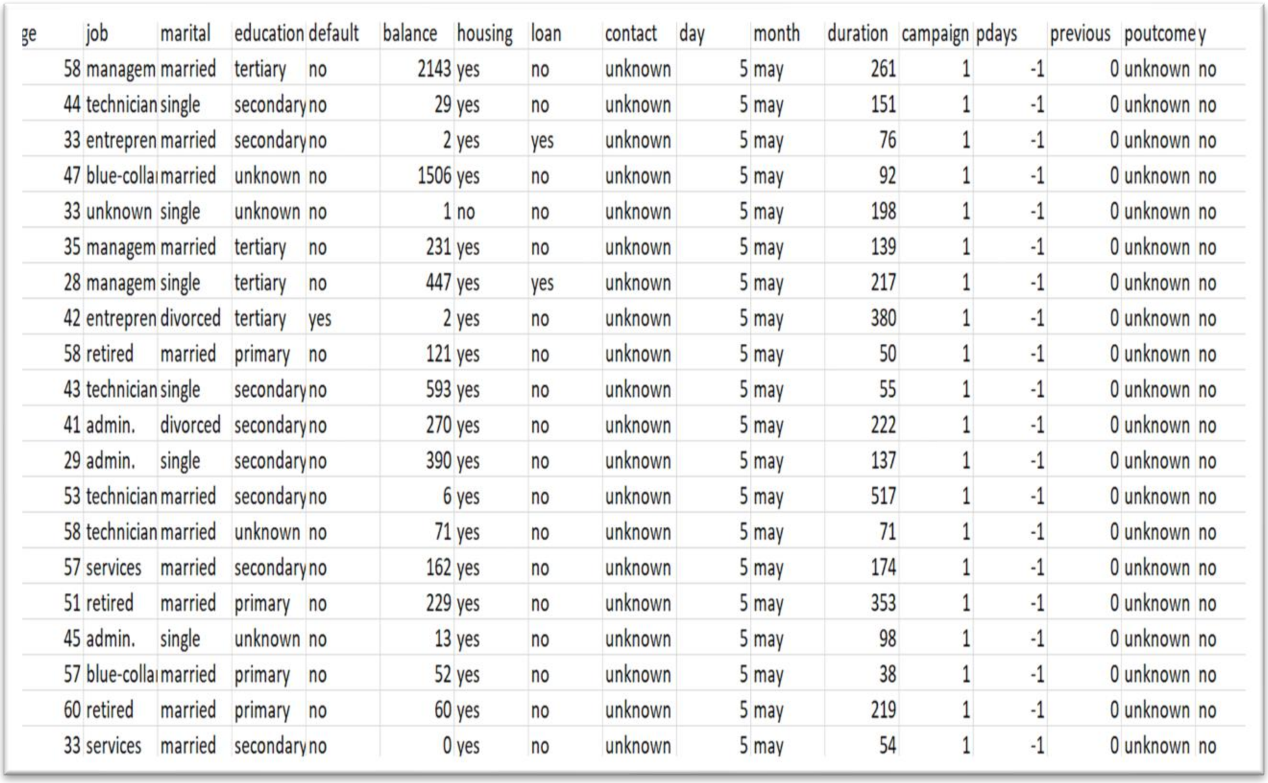
The dataset used for experiments in this paper was related with direct marketing campaigns of a Portuguese banking institution and is available at UCI Machine Learning Repository The marketing campaigns (Cortez, n.d.). Were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed results of direct bank marketing campaigns. It includes 17 campaigns of a Portuguese bank conducted between May 2008 and November 2010. The details of 17 attributes with 45,211 instances composed of numerical, nominal and binary attributes. Table (4. 14) contains the detailed description of all attributes in this dataset after understanding the context from Kaggle. Table (4.15) present sample from the dataset.

Table (4.14) Dataset Description

Variable	Variable Description
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)
balance	numeric, average yearly balance (in euros)

housing	binary, status of housing loan (yes, no)
Loan	binary, client's personal loan (yes, no)
contact	categorical, contact communication type (unknown, telephone, cellular)
Day	numeric, the last contact day of the month range (1-31)
Month	categorical, last contact month of the year
duration	numeric, last contact duration (in seconds)
campaign	numeric, number of contacts performed during this campaign
Pdays	numeric, number of days that passed by after the client was last contacted from a previous campaign
previous	numeric, number of contacts which are made before this campaign
poutcome	categorical, result or outcome of the previous marketing campaign (Unknown, other, failure, success)
Y	binary, (desired target) client subscribed a term deposit or not

Table (4.15) Banking Dataset - Marketing Targets dataset sample



ge	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
58	managem	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	entrepren	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-colla	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
35	managem	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
28	managem	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
42	entrepren	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
57	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
51	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no
45	admin.	single	unknown	no	13	yes	no	unknown	5	may	98	1	-1	0	unknown	no
57	blue-colla	married	primary	no	52	yes	no	unknown	5	may	38	1	-1	0	unknown	no
60	retired	married	primary	no	60	yes	no	unknown	5	may	219	1	-1	0	unknown	no
33	services	married	secondary	no	0	yes	no	unknown	5	may	54	1	-1	0	unknown	no

4.4.1 Data Preprocessing Results

To prepare the data for the prediction and portfolio optimization stages is the aim of this step. the whole dataset from the Banking Dataset - Marketing Targets stage of preparation is used.

A- Missing Values Processing Results

The Banking Dataset - Marketing Targets data will be examined at this stage to determine whether or not it contains missing values (nan-null). Whereas, depending on the values of other features, missing values are processed by substituting the mean of the column containing the missing values. After conducting a variety of tests, we found that the Banking Dataset - Marketing Targets does not contain any missing values.

B- Nominal to Numeric Result

This stage involves transforming categorical attributes (y, contact, loan, housing, job, marriage, education, default, month, and outcome) into numeric form. according to table (4.16). The data type is changed from nominal to numeric through this technique. This is accomplished by turning the qualities into numbers like follows:

- y': "no":0, "yes":1

- "Contact": "cellular," "telephone,"

- "loan": "no": 0, "yes": 1, "unknown": 2,

- Housing: "no," "yes," "unknown," and "2"
- "Job": "management": 0, 'technician': 1, 'entrepreneur': 2, 'blue-collar': 3, 'unknown': 4,'retired': 5, 'admin.': 6,'services': 7,'self-employed': 8, 'unemployed': 9, 'housemaid': 10, and ‘student':11

- "Marital": "single": 0; "married": 1; "divorced": 2; "unknown": 3;

- Education: "basic.4y": 0, "basic.6y": 1, "basic.9y':2, "high school," "professional course," "university degree," "illiterate," "unknown," and "7,"

- "default": "no": 0, "yes": 1, and "unknown": 2,

- 'month': 'January': 0, 'failure':0,'success':1, 'nonexistent':2,'may':4, 'jun':5, 'jul':6, 'aug':7,'sep':8, 'oct':9, 'nov':10, 'dec':11'

- "poutoute": "failure":0, "success":1, "nonexistent":2

Table (4.16) Nominal to Numeric Result

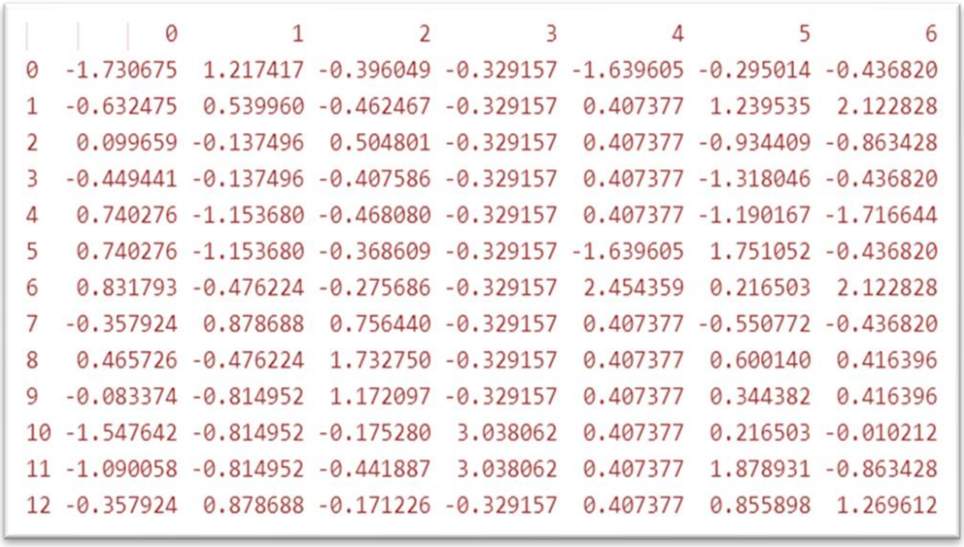
age	*job*	education	*marital*	*default*	*balance*	*housing*	contact	*loan*
58	0	0	1	0	2143	1	0	0

44	1	1	0	0	29	1	0	0
33	2	1	1	0	2	1	0	1
47	3	3	1	0	1506	1	0	0
33	4	3	0	0	1	0	0	0

C- Normalization Process Results

In order to prevent significant value differences from predominating the results, normalization is a crucial step that has been carried out on the banking dataset (Banking Dataset - Marketing Targets). Using the Standardization normalization approach, the feature values varied from zero to one. A portion of the banking dataset's data normalization is displayed in table (4.17) Whereas the columns represent the attributes, while the rows represent the sequence of data.

Table (4.17) The Normalization Result

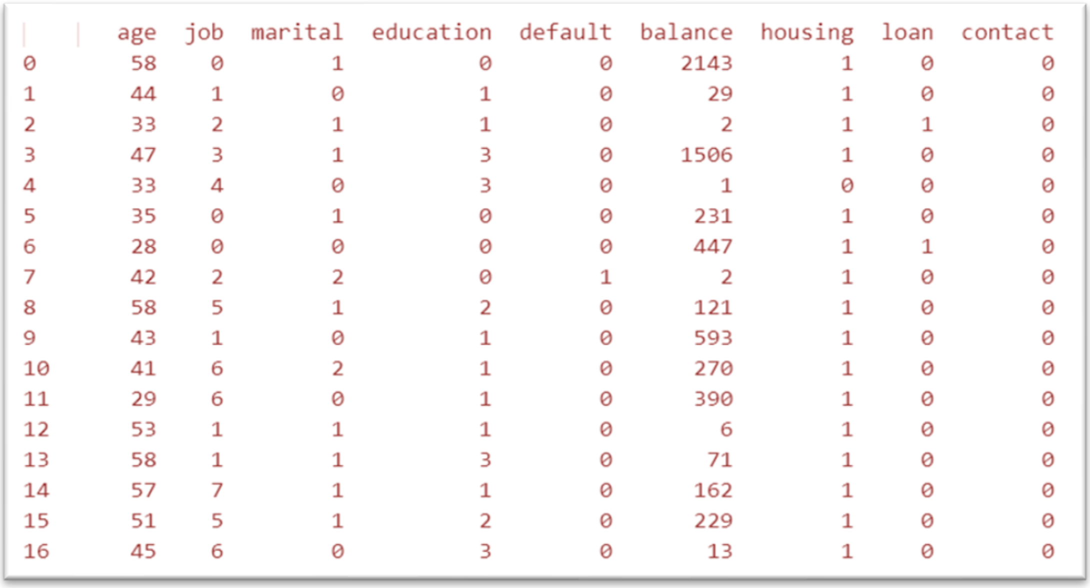


```
           0         1         2         3         4         5         6
0  -1.730675  1.217417 -0.396049 -0.329157 -1.639605 -0.295014 -0.436820
1  -0.632475  0.539960 -0.462467 -0.329157  0.407377  1.239535  2.122828
2   0.099659 -0.137496  0.504801 -0.329157  0.407377 -0.934409 -0.863428
3  -0.449441 -0.137496 -0.407586 -0.329157  0.407377 -1.318046 -0.436820
4   0.740276 -1.153680 -0.468080 -0.329157  0.407377 -1.190167 -1.716644
5   0.740276 -1.153680 -0.368609 -0.329157 -1.639605  1.751052 -0.436820
6   0.831793 -0.476224 -0.275686 -0.329157  2.454359  0.216503  2.122828
7  -0.357924  0.878688  0.756440 -0.329157  0.407377 -0.550772 -0.436820
8   0.465726 -0.476224  1.732750 -0.329157  0.407377  0.600140  0.416396
9  -0.083374 -0.814952  1.172097 -0.329157  0.407377  0.344382  0.416396
10 -1.547642 -0.814952 -0.175280  3.038062  0.407377  0.216503 -0.010212
11 -1.090058 -0.814952 -0.441887  3.038062  0.407377  1.878931 -0.863428
12 -0.357924  0.878688 -0.171226 -0.329157  0.407377  0.855898  1.269612
```

4.4.2 SMOTE Process Results

To enlarge the Minority dataset and balance the class label, a Synthetic Minority Oversampling Technique is utilized to provide fresh data. Table (4.18) Give an example of the generated data in Banking Dataset - Marketing Targets dataset.

Table (4.18) Banking Dataset - Marketing Targets dataset sample



```
      age  job  marital  education  default  balance  housing  loan  contact
0      58    0        1          0        0     2143        1     0        0
1      44    1        0          1        0       29        1     0        0
2      33    2        1          1        0        2        1     1        0
3      47    3        1          3        0     1506        1     0        0
4      33    4        0          3        0        1        0     0        0
5      35    0        1          0        0      231        1     0        0
6      28    0        0          0        0      447        1     1        0
7      42    2        2          0        1        2        1     0        0
8      58    5        1          2        0      121        1     0        0
9      43    1        0          1        0      593        1     0        0
10     41    6        2          1        0      270        1     0        0
11     29    6        0          1        0      390        1     0        0
12     53    1        1          1        0        6        1     0        0
13     58    1        1          3        0       71        1     0        0
14     57    7        1          1        0      162        1     0        0
15     51    5        1          2        0      229        1     0        0
16     45    6        0          3        0       13        1     0        0
```

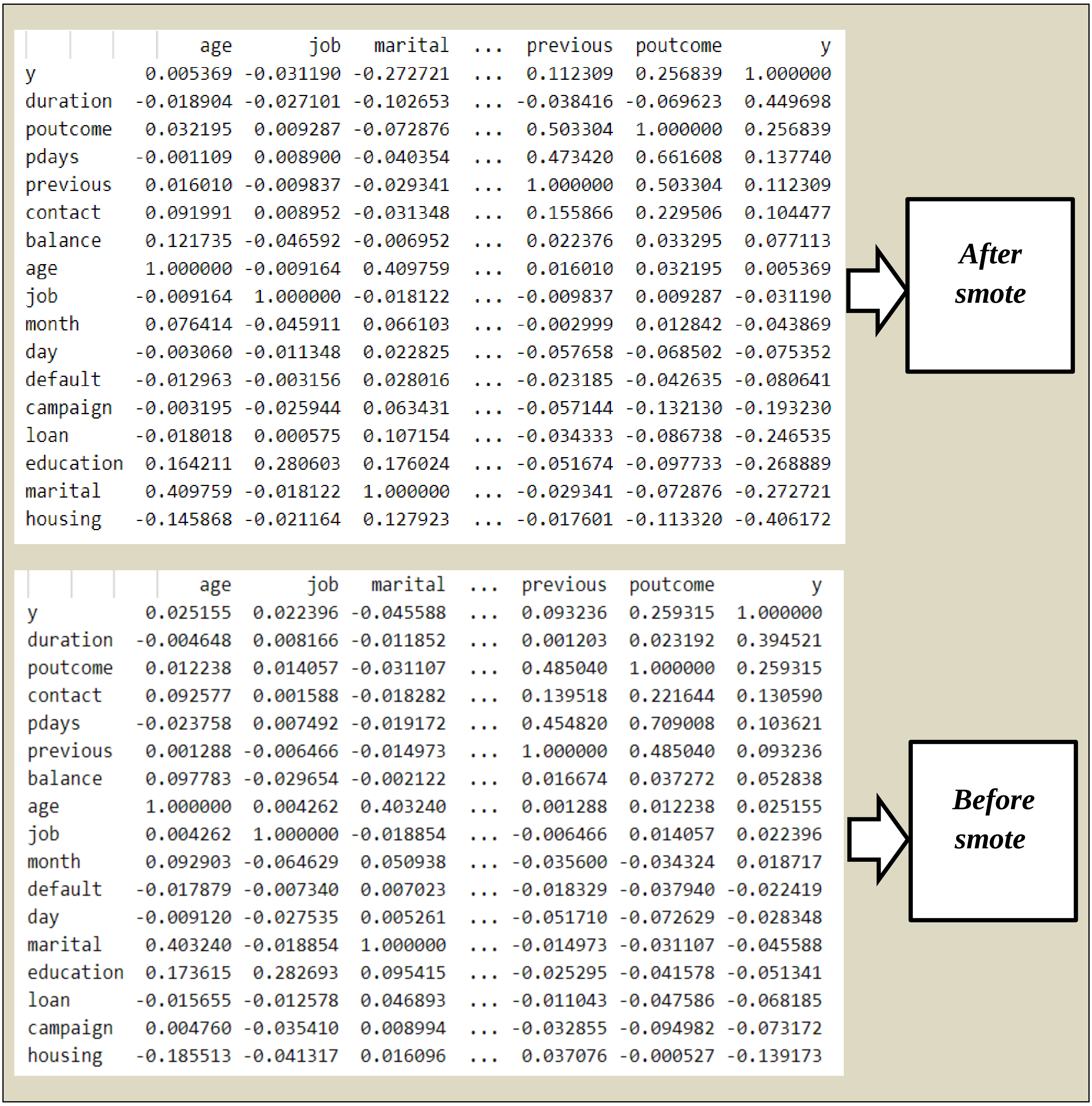
4.4.3 Data Reduction

The reduction levels for banking dataset (Banking Dataset - Marketing Targets) will be used in this study to find the most useful features and look at how they affect the prediction model. There are many features in the banking dataset (Banking Dataset - Marketing Targets), but not all of them are appropriate for the prediction process because some of them decreased the prediction model's accuracy and increased the processing time. To extract the most significant factors that have an impact on banking, the feature selection system is used in this thesis. The reduction process consists of the two processes of Feature Selection and Ranking Correlation.

A- Ranking Correlation

This technique's main objective is to shrink the dataset (Banking Dataset - Marketing Targets) dimensionality by locating pertinent features. A class label and a feature's ranking correlation coefficient are determined. The features with the highest value were chosen using this methodology. Out of a total of 17, the first (9) attributes were chosen as informational features for additional research. Table (4.19) The Ranking Correlation Result.

Table (4.19) The Ranking Correlation Result



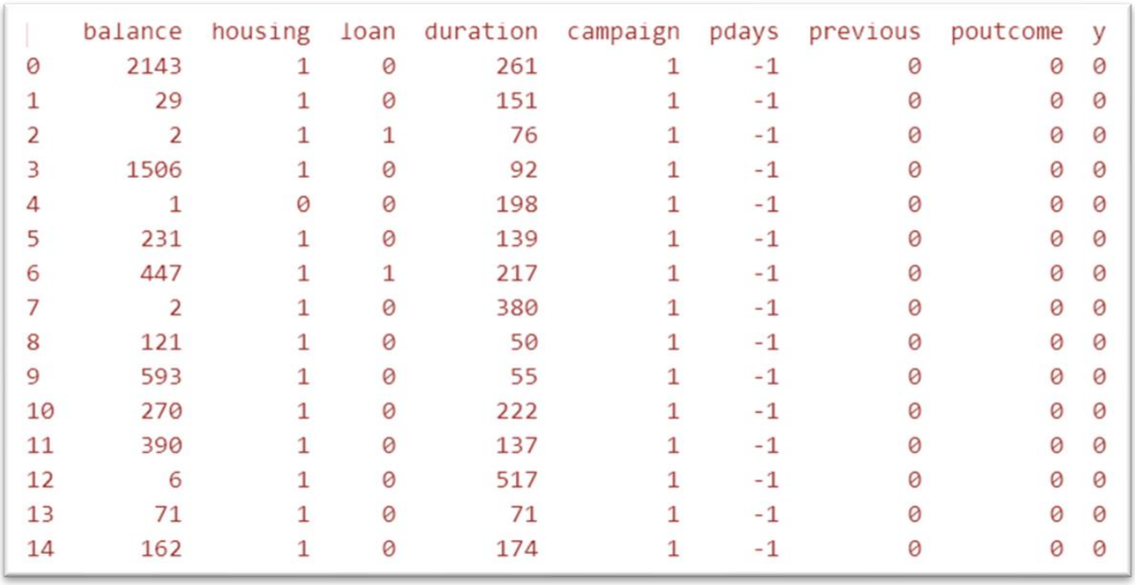
```
                age       job   marital  ...  previous  poutcome         y
y          0.005369 -0.031190 -0.272721  ...  0.112309  0.256839  1.000000
duration  -0.018904 -0.027101 -0.102653  ... -0.038416 -0.069623  0.449698
poutcome   0.032195  0.009287 -0.072876  ...  0.503304  1.000000  0.256839
pdays     -0.001109  0.008900 -0.040354  ...  0.473420  0.661608  0.137740
previous   0.016010 -0.009837 -0.029341  ...  1.000000  0.503304  0.112309
contact    0.091991  0.008952 -0.031348  ...  0.155866  0.229506  0.104477
balance    0.121735 -0.046592 -0.006952  ...  0.022376  0.033295  0.077113
age        1.000000 -0.009164  0.409759  ...  0.016010  0.032195  0.005369
job       -0.009164  1.000000 -0.018122  ... -0.009837  0.009287 -0.031190
month      0.076414 -0.045911  0.066103  ... -0.002999  0.012842 -0.043869
day       -0.003060 -0.011348  0.022825  ... -0.057658 -0.068502 -0.075352
default   -0.012963 -0.003156  0.028016  ... -0.023185 -0.042635 -0.080641
campaign  -0.003195 -0.025944  0.063431  ... -0.057144 -0.132130 -0.193230
loan      -0.018018  0.000575  0.107154  ... -0.034333 -0.086738 -0.246535
education  0.164211  0.280603  0.176024  ... -0.051674 -0.097733 -0.268889
marital    0.409759 -0.018122  1.000000  ... -0.029341 -0.072876 -0.272721
housing   -0.145868 -0.021164  0.127923  ... -0.017601 -0.113320 -0.406172
```

```
                age       job   marital  ...  previous  poutcome         y
y          0.025155  0.022396 -0.045588  ...  0.093236  0.259315  1.000000
duration  -0.004648  0.008166 -0.011852  ...  0.001203  0.023192  0.394521
poutcome   0.012238  0.014057 -0.031107  ...  0.485040  1.000000  0.259315
contact    0.092577  0.001588 -0.018282  ...  0.139518  0.221644  0.130590
pdays     -0.023758  0.007492 -0.019172  ...  0.454820  0.709008  0.103621
previous   0.001288 -0.006466 -0.014973  ...  1.000000  0.485040  0.093236
balance    0.097783 -0.029654 -0.002122  ...  0.016674  0.037272  0.052838
age        1.000000  0.004262  0.403240  ...  0.001288  0.012238  0.025155
job        0.004262  1.000000 -0.018854  ... -0.006466  0.014057  0.022396
month      0.092903 -0.064629  0.050938  ... -0.035600 -0.034324  0.018717
default   -0.017879 -0.007340  0.007023  ... -0.018329 -0.037940 -0.022419
day       -0.009120 -0.027535  0.005261  ... -0.051710 -0.072629 -0.028348
marital    0.403240 -0.018854  1.000000  ... -0.014973 -0.031107 -0.045588
education  0.173615  0.282693  0.095415  ... -0.025295 -0.041578 -0.051341
loan      -0.015655 -0.012578  0.046893  ... -0.011043 -0.047586 -0.068185
campaign   0.004760 -0.035410  0.008994  ... -0.032855 -0.094982 -0.073172
housing   -0.185513 -0.041317  0.016096  ...  0.037076 -0.000527 -0.139173
```

B- Feature Selection

This strategy aims to reduce the dataset's dimensionality, which is calculated using the Feature_importances approach between the feature and the class label. The features with the greatest Feature_importances value have been chosen as a result of this methodology. between (17 characteristics). Table (4.20) shows the Feature Selection of the banking dataset.

Table (4.20) shows the Feature Selection of the banking dataset



	balance	housing	loan	duration	campaign	pdays	previous	poutcome	y
0	2143	1	0	261	1	-1	0	0	0
1	29	1	0	151	1	-1	0	0	0
2	2	1	1	76	1	-1	0	0	0
3	1506	1	0	92	1	-1	0	0	0
4	1	0	0	198	1	-1	0	0	0
5	231	1	0	139	1	-1	0	0	0
6	447	1	1	217	1	-1	0	0	0
7	2	1	0	380	1	-1	0	0	0
8	121	1	0	50	1	-1	0	0	0
9	593	1	0	55	1	-1	0	0	0
10	270	1	0	222	1	-1	0	0	0
11	390	1	0	137	1	-1	0	0	0
12	6	1	0	517	1	-1	0	0	0
13	71	1	0	71	1	-1	0	0	0
14	162	1	0	174	1	-1	0	0	0

4.4.4 Predictive modeling

In this thesis, four prediction models Naive Bayes, Decision Tree, KNN, and Logistic are used to conduct a predictive modeling technique. For each of the used datasets (Banking Dataset - Marketing Targets), the specific measurements are shown in Table (4.21) Precision, recall, and support, F- score are the metrics in question.

Table (4.21) Precision, recall, F- score, and support Result

Algorithm type	*Precision*	*Recall*	*F- score*	*support*	*Class*
Naïve bayes	0.92	0.93	0.92	11687	***NO***
	0.53	0.46	0.49	1877	***YES***
	0.72	0.70	0.71	13564	***macro avg***
	0.86	0.87	0.87	13564	***weighted avg***
Decision tree	0.88	0.91	0.89	7631	***NO***
	0.91	0.88	0.90	8338	***YES***
	0.89	0.89	0.89	15969	***macro avg***
	0.89	0.89	0.89	15969	***weighted avg***
KNN	0.88	0.92	0.90	7631	***NO***
	0.92	0.88	0.90	8338	***YES***
	0.90	0.90	0.90	15969	***macro avg***
	0.90	0.90	0.90	15969	***weighted avg***
Logistic	0.98	0.91	0.95	8579	***NO***
	0.29	0.66	0.41	464	***YES***
	0.64	0.79	0.68	9043	***macro avg***
	0.94	0.90	0.92	9043	***weighted avg***

A- Naïve bayes

The first step in using the Naive Bayes algorithm in banking operations is to gather data relevant to the current issue. Information like salary, age, employment history, and other pertinent details may be among the data gathered, for instance. Using Naive Bayes theory for predictive modeling in banking operations can help to enhance decision-making, lower risk, and boost efficiency. The best outcomes were obtained when smote, standardization, and data reduction were combined with Naive Bayes. Table (4.22) presents this process among SMOTE, Standardization and data reduction.

Table (4.22) Result NB with SMOTE, data reduction and standardization

Algorithm NB		*Accuracy*	*Recall*	*Precision*	*error*
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***Fales*** ***False***	0.8696	0.2885	0.4554	**0.1303**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***False***	0.8403	0.8863	0.8104	**0.1596**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***False***	0.8684	0.4950	0.4558	**0.1315**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***False*** ***True***	0.8399	0.5085	0.3765	**0.1600**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***True***	0.7781	0.9305	0.7119	**0.2218**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***True***	0.8463	0.4772	0.3865	**0.1536**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***False***	0.8322	0.8788	0.8027	**0.1677**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***True***	0.8038	0.8826	0.7612	**0.1961**

B- Decision tree

to apply a Decision Tree (DT) algorithm to the dataset (Banking Dataset - Marketing Targets) in order to predict whether the client would sign up for a term deposit (yes/no). A tree-like model is used by DTs, a supervised learning technique, to produce predictions or choices based on input features. The Smote, standardization, and data reduction are the foundation of the DT's operation. When DT corresponds with the crucial traits, it has the maximum accuracy. Table (4.23) presents this process among SMOTE, Standardization, important features and data reduction.

Table (4.23) Result DT with SMOTE, data reduction, important features and standardization

Algorithm DT		*Accuracy*	*Recall*	*Precision*	*Error*
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***Fales*** ***False***	0.8763	0.4647	0.4541	**0.1236**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***False***	0.8941	0.9117	0.8828	**0.1058**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***False***	0.8763	0.4804	0.4554	**0.1236**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***False*** ***True***	0.8781	0.4804	0.4623	**0.1218**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***True***	0.8934	0.9098	0.8829	**0.1065**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***True***	0.8767	0.4657	0.4555	**0.1232**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***False***	0.8941	0.9109	0.8833	**0.1058**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***True***	0.8952	0.9104	0.8855	**0.1047**
SMOTE ***Important features*** ***Standardization***	***True*** ***True*** ***False***	0.8828	0.8960	0.8752	**0.1171**

SMOTE ***Important features*** ***Standardization***	***False*** ***True*** ***True***	0.8751	0.4657	0.4494	**0.1248**
SMOTE ***Important features*** ***Standardization***	***True*** ***True*** ***True***	0.8840	0.8982	0.8757	**0.1159**

C- KNN

In many banking applications, KNN algorithms are used to forecast whether a client would sign up for a term deposit (yes/no). The algorithm determines the closest neighbors by calculating the distance between the new applicant's qualities and those of current clients. The algorithm predicts through association with Smote, standardization, and data reduction while taking into account the nearby neighbors' labels (default or non-default). Table (4.24) presents this process among SMOTE, Standardization and data reduction.

Table (4.24) Result KNN with SMOTE, data reduction and standardization

Algorithm KNN		*Accuracy*	*Recall*	*Precision*	*Error*
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***Fales*** ***False***	0.8696	0.2885	0.4554	**0.1303**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***False***	0.8870	0.9649	0.8342	**0.1129**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***False***	0.8699	0.2965	0.4584	**0.1300**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***False*** ***True***	0.8841	0.3691	0.5449	**0.1158**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***True***	0.9001	0.9240	0.8812	**0.0998**

SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***True***	0.8732	0.3297	0.4804	**0.1267**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***False***	0.8483	0.9123	0.8079	**0.1516**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***True*** ***True***	0.8607	0.8992	0.8341	**0.1392**

D- *Logistic*

Logistic regression can be used in the banking industry for a variety of purposes, including predicting customer attrition. The logistic algorithm must be related to Smote, standardization, and data reduction in order to be used to the prediction model. Table (4.25) presents this process among SMOTE, Standardization and data reduction.

Table (4.25) Result Logistic with SMOTE, data reduction and standardization

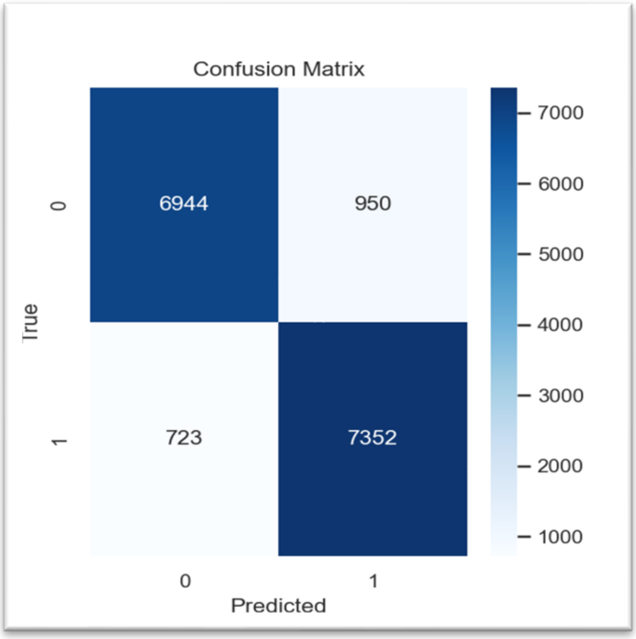
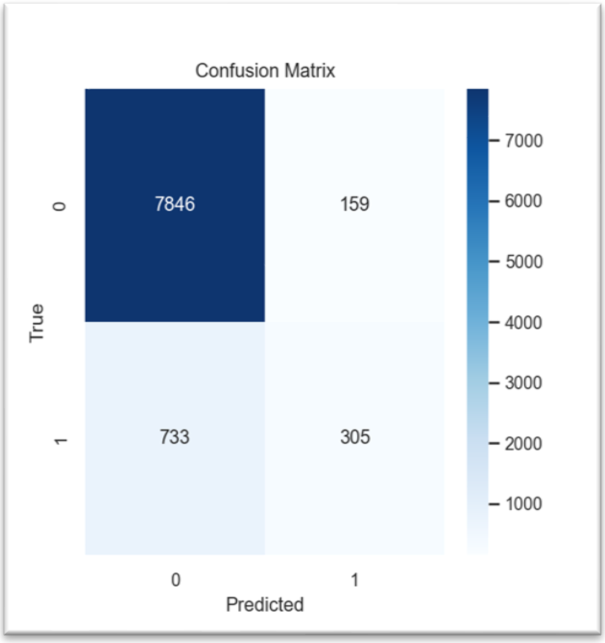
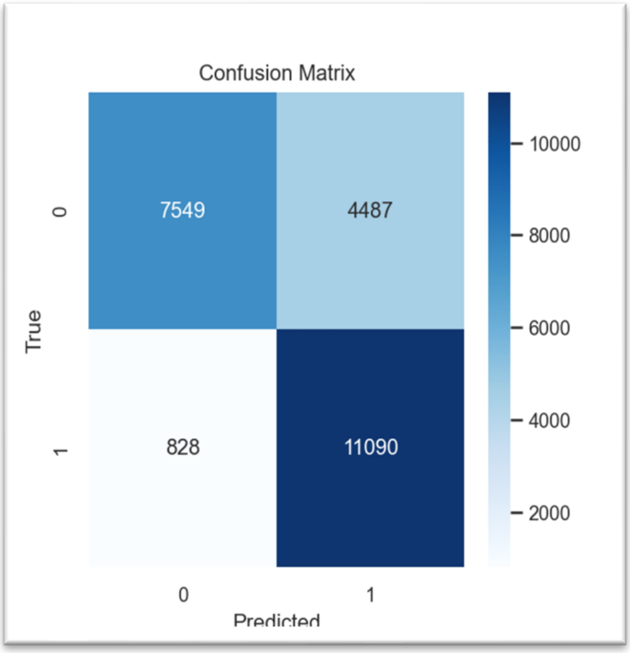
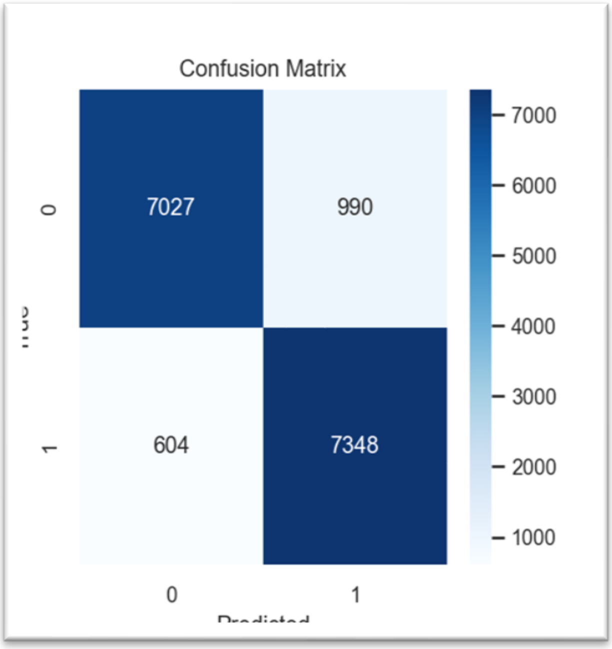
Algorithm LG		***Accuracy***	***Recall***	***Precision***	***error***
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***Fales*** ***False***	0.8906	0.2013	0.5906	**0.1093**
SMOTE ***Data Reduction*** ***Standardization***	***True*** ***False*** ***False***	0.8123	0.8036	0.8187	**0.1876**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***True*** ***False***	0.9011	0.3439	0.6263	**0.098**
SMOTE ***Data Reduction*** ***Standardization***	***False*** ***False*** ***True***	0.9013	0.3150	0.6437	**0.0986**
SMOTE	***True***	0.8540	0.8641	0.8477	**0.1459**

Data Reduction *Standardization*	*False* *True*				
SMOTE *Data Reduction* *Standardization*	*False* *True* *True*	0.9013	0.2938	0.6369	**0.0986**
SMOTE *Data Reduction* *Standardization*	*True* *True* *False*	0.7808	0.7299	0.8138	**0.2191**
SMOTE *Data Reduction* *Standardization*	*True* *True* *True*	0.7821	0.7421	0.8076	**0.2178**

4.4.5 Evaluation Metric

It was determined how well the Naive Bayes, Decision Tree, KNN, and Logistic prediction models performed. All features that have surfaced are often regarded as influencer features. After that, they were incorporated into the model to demonstrate both their significance and the level of precision that could be obtained by these attributes. In reality, using a few features really makes accuracy much better. A confusion matrix is a table that is used to evaluate the performance of a classification model by comparing the actual and predicted class labels. In the case of a (Naive Bayes, Decision Tree, KNN, and Logistic) classification models, a confusion matrix can be used to determine the accuracy of the model in predicting the class labels for a given set of input data. Table (4.26) Confusion Matrixes (NB, TD, KNN, LG).

Table (4.26) Confusion Matrixes (NB, TD, KNN, LG)

 ***confusion matrix DT***	 ***confusion matrix LG***
 ***confusion matrix NB***	 ***confusion matrix KNN***

4.5 Results Using WEKA

Two data sets are utilized:

4.5.1 Banking Dataset - Marketing Targets

The 1st dataset is Banking Dataset - Marketing Targets dataset is used to studying customer behavior in the banking industry and developing predictive models for marketing campaigns is presented in Table (4.15).

4.5.2 Attribute Ranking

A preprocessing process is performed by analyzing the 1st dataset (Banking Dataset - Marketing Targets) and observing its attributes. It composed of 17 attributes. A competition process is implemented using a WEKA software to rank these attributes based on their significance. Table (4.27) presents the sorted list of the ranked attributes.

Table (4.27) attributes ranking for a Banking Dataset - Marketing.

Rank	Attributes	Old No
1	Duration	**12**
2	pout come	**16**
3	Housing	**7**
4	Contact	**9**
5	P days	**14**
6	Previous	**15**
7	Campaign	**13**
8	Loan	**8**
9	Marital	**3**
10	Month	**11**
11	Balance	**6**

12	Education	**4**
13	Job	**2**
14	Day	**10**
15	Age	**1**
16	Default	**5**

4.5.3 Attributes Correlation

The value of the next time values for each attribute can be predicted. To reach accurate prediction, a correlation process must be performed to show the correlation coefficient between any attributes pair. Strong or weak correlation may give a good indication in ignore or perform an attribute reduction process. Table (4.28) shows the correlation coefficient matrix for the used banking dataset.

Table (4.28) correlation coefficient matrix for banking dataset

	Age	**Job**	**marital**	**Education**	**default**	**balance**	**Housing**	**loan**	**contact**	**day**	**Month**	**Duration**	**Campaign**	**P days**	**Previous**	**Pout come**	**Y**
Age	–	0.0686	0.312	0.105	0.0179	0.0978	0.186	0.0157	0.061	0.00912	0.00623	0.00465	0.00476	0.0238	0.00129	0.00332	**0.0252**
Job	0.0686	–	0.0564	0.187	0.0066	0.0376	0.08	0.0266	0.0619	0.0176	0.0462	0.00987	0.0148	0.0148	0.0124	0.015	**0.0344**
Marital	0.0321	0.0564	–	0.0537	0.0116	0.0215	0.0166	0.0387	0.0364	0.00642	0.0281	0.0201	0.0272	0.0249	0.013	0.0226	**0.0546**
Education	0.105	0.187	0.0537	—	0.0116	0.0634	0.0849	0.0537	0.0743	0.0124	0.0475	0.00191	0.0162	0.0165	0.0125	0.0125	**0.0448**
Default	0.0179	0.0066	0.0116	0.0116	–	0.0667	0.00603	0.0772	0.0134	0.00942	0.0134	0.01	0.0168	0.03	0.0183	0.0365	**0.0224**
Balance	0.0978	0.0376	0.0215	0.0634	0.0667	–	0.0688	0.0844	0.0235	0.0045	0.0497	0.0216	0.0146	0.00344	0.0167	0.0276	**0.0528**

0.139	**0.0682**	**0.132**	**0.0283**	**0.0531**	**0.395**	**0.0732**	**0.104**	**0.932**	**0.149**	_
0.0684	0.0281	0.239	0.0803	0.0772	0.0064	0.101	0.811	0.493	_	
0.0371	0.011	0.132	0.0517-	0.0418	0.0012	0.0329-	0.455	_		
0.124	0.0228	0.219	0.093-	0.0914	0.00156-	0.0886-	_			
0.0236	0.0098	0.0258	0.162	0.0796	0.0846-	_				
0.00508	0.0124	0.222	0.0302	0.0172	_					
0.208	0.0518	0.253	0.0945	_						
0.028	0.0114	0.0244	_							
0.166	0.0117	_								
0.0413	_									
_										
0.0688										
0.00603										
0.0849										
0.0166										
0.08										
0.186										
Housing	Loan	Contact	Day	Month	Duration	Campaign	Pdays	Previous	Pout come	Y

4.5.4 predictive modeling

To perform a predictive modeling approach, four prediction models are utilized in this thesis (Naïve bayes, Decision tree, KNN and Logistic). Table (4.29) presents the specific measures for each of the used datasets. These measures are (TP rate, FP rate, precision, Recall, F-Measure, MCC, ROC Area, PRC Area).

Table (4.29) some class-specific accuracy

Algorithm type	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Naïve bayes	0.975	0.861	0.440	0.932	0.927	0.937	0.472	0.927	N0
	0.445	0.861	0.440	0.507	0.528	0.488	0.073	0.528	Yes
	0.913	0.861	0.440	0.882	0.880	0.884	0.426	0.880	avg
Decision tree (J48)	0.947	0.843	0.488	0.946	0.959	0.933	0.519	0.959	NO
	0.486	0.843	0.488	0.537	0.481	0.609	0.041	0.481	Yes
	0.893	0.843	0.488	0.898	0.903	0.895	0.463	0.9.3	Avg
KNN	0.914	0.646	0.319	0.927	0.938	0.917	0.645	0.938	No
	0.237	0.646	0.319	0.389	0.355	0.431	0.062	0.355	Yes
	0.835	0.646	0.319	0.864	0.870	0.860	0.577	0.870	avg
Logistic	0.985	0.907	0.427	0.946	0.975	0.918	0.654	0.975	N0
	0.551	0.907	0.427	0.451	0.346	0.648	0.025	0.346	Yes
	0.934	0.907	0.427	0.888	0.902	0.887	0.580	0.902	avg

A- Confusion Matrixes

A classification model's performance can be assessed using confusion matrices, which count the true positives, true negatives, false positives, and false negatives. Confusion matrices can be used to assess (DT, KNN, LG and NB) a popular classification approach. Confusion Matrixes Table (4.30).

Table (4. 30) Confusion Matrixes

Confusion Matrix		Predicated Class	
		No	Yes
Naive bayes	No	2924	36998
	Yes	2791	2498

Confusion Matrix		Predicated Class	
		No	Yes
Decision tree	No	1633	38289
	Yes	2542	2747

Confusion Matrix		Predicated Class	
		No	Yes
KNN	No	2482	37440
	Yes	1879	3410

Confusion Matrix		Predicated Class	
		No	Yes
Logistic	No	990	38927
	Yes	1831	3458

B- Algorithm accuracy

accuracy is a table that compares the actual and anticipated class labels to assess the effectiveness of a classification model. accuracy can be used to evaluate how well a (NB, DT, KNN and LG) classification model predicts the class labels for a specific set of input data. Confusion Matrixes, Table (4.31).

Table (4.31) Algorithm accuracy

Algorithm type	Accuracy	Error	Time
Naïve bayes	88.0073	0.1532	**0.23**

Decision	90.3121	0.1269	**2.12**
KNN	86.9678	0.1303	**0.01**
Logistic	90.1506	0.1391	**3.13**

4.5.5 bank-additional-full

The 2nd dataset is a bank-additional-full is used in the classification goal to predict if the client will subscribe a term deposit (variable y) which is presented in Table (4.2).

4.5.6 Attribute Ranking

A preprocessing process is performed by analyzing the data and observing its attributes. It composed of 21 attributes. A competition process is implemented using a WEKA software to rank these attributes based on their significance. Table (4.32) presents the sorted list of the ranked attributes.

Table (4.32) attributes ranking for a bank-additional-full

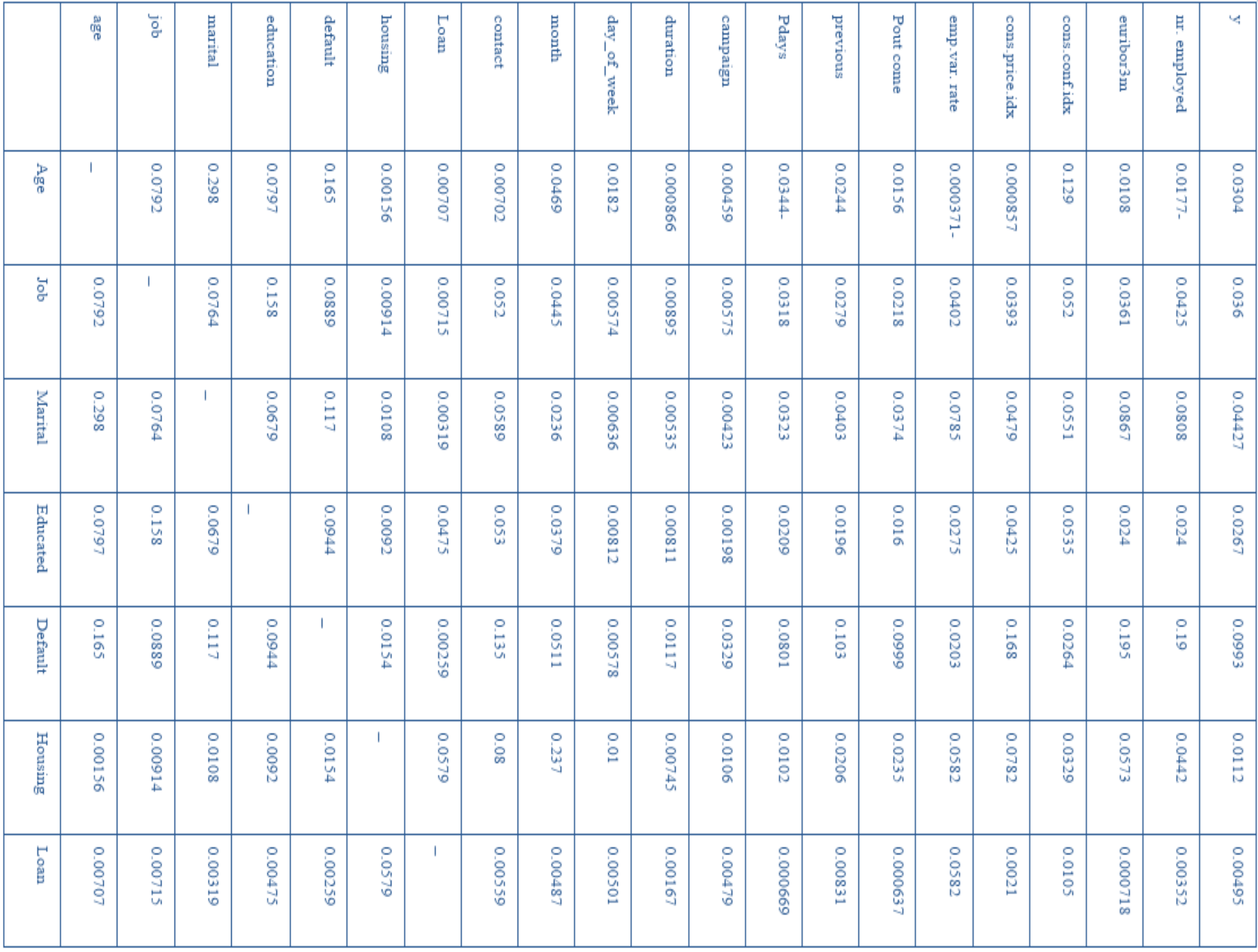
Rank	Attributes	Old No
1	Duration	**11**
2	nr. employed	**20**
3	Pdays	**13**
4	euribor3m	**19**
5	emp.var. rate	**16**
6	Previous	**14**
7	pout come	**15**
8	Contact	**8**
9	cons.price.idx	**17**
10	Default	**5**
11	Campaign	**12**
12	Month	**9**
13	cons.conf.idx	**18**
14	Marital	**3**

15	Job	**2**
16	Age	**1**
17	Education	**4**
18	day_of_week	**10**
19	Housing	**6**
20	Loan	**7**

4.5.7 Attributes Correlation

To study and analyze the relation and the effects of each attribute on the others, a correlation study is conducted. The correlation between any two attributes. The process starts by selecting a bank-additional-full attribute and test its correlation coefficient with the second, third, …… till the last one and so on for other attributes. The correlation results are recorded in a matrix of (21 x 21). The strong correlation coefficient value (which is greater or equal 0.7) is selected and attended. Table (4.33) shows a sample of the created correlation matrix.

Table (4.33) correlation coefficient matrix for bank-additional-full



	Age	Job	Marital	Educated	Default	Housing	Loan
age	_	0.0792	0.298	0.0797	0.165	0.00156	0.00707
job	0.0792	_	0.0764	0.158	0.0889	0.00914	0.00715
marital	0.298	0.0764	_	0.0679	0.117	0.0108	0.00319
education	0.0797	0.158	0.0679	_	0.0944	0.0092	0.00475
default	0.165	0.0889	0.117	0.0944	_	0.0154	0.00259
housing	0.00156	0.00914	0.0108	0.0092	0.0154	_	0.0579
Loan	0.00707	0.00715	0.00319	0.0475	0.00259	0.0579	_
contact	0.00702	0.052	0.0589	0.053	0.135	0.08	0.00559
month	0.0469	0.0445	0.0236	0.0379	0.0511	0.237	0.00487
day_of_week	0.0182	0.00574	0.00636	0.00812	0.00578	0.01	0.00501
duration	0.000866	0.00895	0.00535	0.00811	0.0117	0.00745	0.00167
campaign	0.00459	0.00575	0.00423	0.00198	0.0329	0.0106	0.00479
Pdays	0.0344-	0.0318	0.0323	0.0209	0.0801	0.0102	0.000669
previous	0.0244	0.0279	0.0403	0.0196	0.103	0.0206	0.00831
Pout come	0.0156	0.0218	0.0374	0.016	0.0999	0.0235	0.000637
emp.var. rate	0.000371-	0.0402	0.0785	0.0275	0.0203	0.0582	0.0582
cons.price.idx	0.000857	0.0393	0.0479	0.0425	0.168	0.0782	0.0021
cons.conf.idx	0.129	0.052	0.0551	0.0535	0.0264	0.0329	0.0105
euribor3m	0.0108	0.0361	0.0867	0.024	0.195	0.0573	0.000718
nr. employed	0.0177-	0.0425	0.0808	0.024	0.19	0.0442	0.00352
y	0.0304	0.036	0.04427	0.0267	0.0993	0.0112	0.00495

4.5.8 predictive modeling

To perform a predictive modeling approach, four prediction models are utilized in this thesis (Naïve bayes, Decision tree, KNN and Logistic). Table (4.34) presents the specific measures for each of the used datasets. These measures are (TP rate, FP rate, precision, Recall, F-Measure, MCC, ROC Area, PRC Area).

Table (4.34) specific accuracy by class

Algorithm type	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Naïve bayes	0.978	0.871	0.458	0.927	0.905	0.949	0.383	0.905	N0
	0.481	0.871	0.458	0.522	0.617	0.453	0.095	0.617	Yes
	0.922	0.871	0.458	0.881	0.873	0.893	0.350	0.873	avg
Decision tree (J48)	0.966	0.884	0.533	0.951	0.959	0.942	0.462	0.959	NO
	0.538	0.884	0.533	0.580	0.538	0.627	0.041	0.538	Yes
	0.918	0.884	0.533	0.909	0.912	0.907	0.414	0.912	Avg
KNN	0.918	0.645	0.317	0.929	0.938	0.920	0.645	0.938	No
	0.225	0.645	0.317	0.386	0.355	0.422	0.062	0.355	Yes
	0.840	0.645	0.317	0.868	0.873	0.864	0.580	0.873	Avg
Logistic	0.991	0.935	0.484	0.951	0.973	0.930	0.577	0.973	N0
	0.597	0.935	0.484	0.516	0.423	0.663	0.027	0.423	Yes
	0.947	0.935	0.484	0.902	0.911	0.900	0.515	0.911	Avg

A- Confusion Matrixes

A classification model's performance can be assessed using confusion matrices, which count the true positives, true negatives, false positives, and false negatives. a popular classification approaches. Confusion Matrixes Table (4.35).

Table (4.35) Confusion Matrixes (Naïve Bayes)

Confusion Matrix		Predicated Class	
		No	Yes
Naive bayes	No	3463	33085
	Yes	2865	1775

Confusion Matrix		Predicated Class	
		No	Yes
Logistic	No	1000	35548
	Yes	1963	2677

Confusion Matrix		Predicated Class	
		No	Yes
Decision tree	No	1483	35065
	Yes	2498	2142

Confusion Matrix		Predicated Class	
		No	Yes
KNN	No	2250	34298
	Yes	1646	2994

B- Algorithm accuracy

accuracy is a table that compares the actual and anticipated class labels to assess the effectiveness of a classification model. Accuracy can be used to evaluate how well a (Naive Bayes, Decision Tree, KNN, Logistic) classification model predicts the class labels for a specific set of input data. Accuracy Table (4.36).

Table (4.36) Algorithm accuracy

Algorithm type	Accuracy	Error	Time
Naïve bayes	87.2827	0.1406	**0.17**
Decision	91.1989	0.1132	**1.95**
KNN	87.2681	0.1273	**0.02**
Logistic	91.0726	0.2508	**3.88**

4.6 Comparison

Table (4.37) presents the comparison with the closely related works. This thesis results shows a superiority on the previous related works.

Table (4.37) The Comparison Result

Ref. No, Year	The used algorithm	The used Dataset	Results	Results in this thesis
2022	Decision tree	Banking Dataset - Marketing Targets	decision tree (DT) accuracy 0.784	Naïve bayes 0.87 & Recall 0.92 Decision tree 0.8952 KNN 0.9001 Logistic 0.9013
2020	KNN, linear model, logistic model	Banking Dataset - Marketing Targets	KNN is Accuracy 0.88 Linear Accuracy 0.89 Logistic Accuracy 0.89	

2021	NN, SVM, NB	Banking Dataset - Marketing Targets	NN Accuracy 94.8684 SVM Accuracy 89.8031 NB Accuracy 88.3294	
2020	Logistic Regression, KNN, SVM	A bank-additional-full	Logistic Regression Accuracy 0.848 SVM Accuracy 0.856 KNN Accuracy 0.917	Naïve bayes 0.8606 Decision tree 0.9336 KNN 0.9285 Logistic 0.9102
2019 In WEKA	NB, One-R	Banking Dataset - Marketing Targets	NB Accuracy 88.4541 One-R Algorithm 89.3875	Naïve bayes 88.0073 Decision tree 90.3121 KNN 86.9678 Logistic 90.1506

Chapter Five

Conclusions and Future Works

Chapter Five

Conclusions and Future Works

5.1 Conclusions

The following are the main conclusions gained from the results obtained utilizing the proposed system for predict if the client will subscribe (yes/no) a term deposit expression:

- ❖ The suggested system has demonstrated its effectiveness in detecting relevant features (the best features) and deleting irrelevant or harmful features. Furthermore, according to all conventional assessment metrics, this proposed system produces good outcomes in the prediction model.
- ❖ Model balancing using the Smote algorithm gave the best results for the class. Therefore, a high accuracy was obtained.
- ❖ The proposed system has successfully proved the feature selection approach according to the dataset's, with satisfying results in the prediction model.
- ❖ The comparison between the models such as DT, Logistic, Naive Bayes and KNN, proves that the proposed model gives better results for all datasets with all parameters. The best enhancement results are 0.9336, 0.9102, 0.8606, and 0.9285 in Accuracy.
- ❖ The comparison between the models such as KNN, DT, Naive Bayes, and logistic proves that the proposed model gives better results for all datasets with parameter. The better results improvement to minimize error ratio (0.071, 0.066, 0.139, 0.089).
- ❖ To identify the most important features affecting clients' deposit performance among the total characteristics, a sequential feature selection algorithm was developed.
- ❖ For each prediction period, influential features are selected by Ranking Algorithm.

5.2 The Future Works

The following is recommended for future works:

- Classifying the features related with the bank operations using a big dataset containing thousands of features.
- Apply the proposed approach to other different datasets such as education.
- Examine various models that can be compared to the CNN, SVM, Random Forest and ANN models and attempting to reduce prediction error.
- studding other feature selection methods, such as chi-squared test, ANOVA, or mutual information and others, are being investigated for picking a best subset of features and highlighting their impact on the prediction model.
- Compare the results of other classification approaches, such as SVM (Support Vector Machines), with the results of the Logistic algorithm and DT algorithm with the results of the Random Forest.

References

Reference

A. Jović, K. Brkić, and N. Bogunović, “A review of feature selection methods with applications,” 2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc., pp. 1200–1205, 2015.

Alcalde, R., Alonso de Armiño, C., & García, S. (2022). Analysis of the economic sustainability of the supply chain sector by applying the Altman Z-score predictor. Sustainability, 14(2), 851.

Ali, Q., Yaacob, H., Parveen, S., & Zaini, Z. (2021). Big data and predictive analytics to optimise social and environmental performance of Islamic banks. Environment Systems and Decisions, 41, 616-632.

Appiahene, P., Missah, Y. M., & Najim, U. (2020). Predicting bank operational efficiency using machine learning algorithm: comparative study of decision tree, random forest, and neural

Ashraf, S., GS Félix, E., & Serrasqueiro, Z. (2019). Do traditional financial distress prediction models predict the early warning signs of financial distress? Journal of Risk and Financial Management, 12(2), 55.

Ashraf, S., GS Félix, E., & Serrasqueiro, Z. (2019). Do traditional financial distress prediction models predict the early warning signs of financial distress? Journal of Risk and Financial Management, 12(2), 55.

Barga, R., Fontama, V., Tok, W. H., & Cabrera-Cordon, L. (2015). Predictive analytics with Microsoft Azure machine learning (pp. 221-241). Berkely, CA: Apress.

Bluwstein, K., Buckmann, M., Joseph, A., Kapadia, S., & Simsek, Ö. (2021). Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach.

Borugadda, P., Nandru, P., & Madhavaiah, C. (2021). Predicting the success of bank telemarketing for selling long-term deposits: An application of machine learning algorithms. *St. Theresa Journal of Humanities and Social Sciences*, *7*(1), 91-108.

Broby, D. (2022). The use of predictive analytics in finance. The Journal of Finance and Data Science, 8, 145-161.

Choi, Y., & Choi, J. (2022). How does Machine Learning Predict the Success of Bank telemarketing?

Demraoui, L., Eddamiri, S., & Hachad, L. (2022). Digital Transformation and Costumers Services in Emerging Countries: Loan Prediction Modeling in Modern Banking Transactions. In AI and IoT for Sustainable Development in Emerging Countries: Challenges and Opportunities (pp. 627-642). Cham: Springer International Publishing.

Gavurova, B., Packova, M., Misankova, M., & Smrcka, L. (2017). Predictive potential and risks of selected bankruptcy prediction models in the Slovak business environment. Journal of Business Economics and Management, 18(6), 1156-1173.

Giordana, G. A., & Schumacher, I. (2017). An empirical study on the impact of Basel III standards on banks' default risk: The case of Luxembourg. Journal of Risk and Financial Management, 10(2), 8.

Ian H. Witten, Eibe Frank, Christopher J. Pal, Data Mining Practical Machine Learning Tools and Techniques, 2017.

Journal, I. J. C. S. M. C. (2019). Mining a Marketing Campaigns Data of Bank. IJCSMC, 8(3), 285–290.

Jullum, M., Løland, A., Huseby, R. B., Ånonsen, G., & Lorentzen, J. (2020). Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control*, *23*(1), 173-186.

Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive modeling with big data: is bigger really better? Big data, 1(4), 215-226.

Kikan, D., Singh, S., & Singh, Y. (2019). Predictive analytics adoption by banking and financial services: The future perspective. International Journal of Recent Technology and Engineering, 8, 832-837.

Kinga Włodarczyk & Kingsley Success Ikani, (2020), Data Analysis of a Portuguese Marketing Campaign using Bank Marketing data Set.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.

Kumar, V., & Garg, M. L. (2018). Predictive analytics: a review of trends and techniques. International Journal of Computer Applications, 182(1), 31-37

Kumar, V., & Garg, M. L. (2018). Predictive analytics: a review of trends and techniques. International Journal of Computer Applications, 182(1), 31-37.

L. Al Shalabi and Z. Shaaban, “Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix,” Proc. Int. Conf. Dependability Comput. Syst. DepCoS-RELCOMEX 2006, pp. 207–214, 2006.

Li, J. P., Mirza, N., Rahat, B., & Xiong, D. (2020). Machine learning and credit ratings prediction in the age of fourth industrial revolution. *Technological Forecasting and Social Change*, *161*, 120309.

M. Toloo, B. Sohrabi and S.Nalchigar, "A new method for ranking discovered rules from data mining by DEA", *journal Expert Systems with Applications*, vol. 36, p 8503–8508,2008 , doi:10.1016/j.eswa.2008.10.038.

Martens, D., Provost, F., Clark, J., & de Fortuny, E. J. (2016). Mining massive fine-grained behavior data to improve predictive analytics. MIS quarterly, 40(4), 869-888.

Martens, D., Provost, F., Clark, J., & de Fortuny, E. J. (2016). Mining massive fine-grained behavior data to improve predictive analytics. *MIS quarterly*, *40*(4), 869-888.

Mhlanga, D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. *International Journal of Financial Studies*, *9*(3), 39.

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, *165*, 113986.

Niemann, M., Schmidt, J. H., & Neukirchen, M. (2008). Improving performance of corporate rating prediction models by reducing financial ratio heterogeneity. Journal of Banking & Finance, 32(3), 434-446.

Oni, J. O. (2020). Exploratory analysis of bank marketing campaign using machine learning; logistic regression, support vector machine and k-nearest neighbour (Doctoral dissertation, Dublin, National College of Ireland).

P. Tang, M. Steinbach, and V. Kumar, "Introduction to data mining", Pearson Education, Inc.,2006.

Pang-Ning Tan, Michael Steinbach, Vipin Kumar and Anuj Karpatne, Introduction to Data Mining, 2021

Patwary, M. J., Akter, S., Alam, M. B., & Karim, A. R. (2021). Bank deposit prediction using ensemble learning. *Artificial Intelligence Evolution*, 42-51.

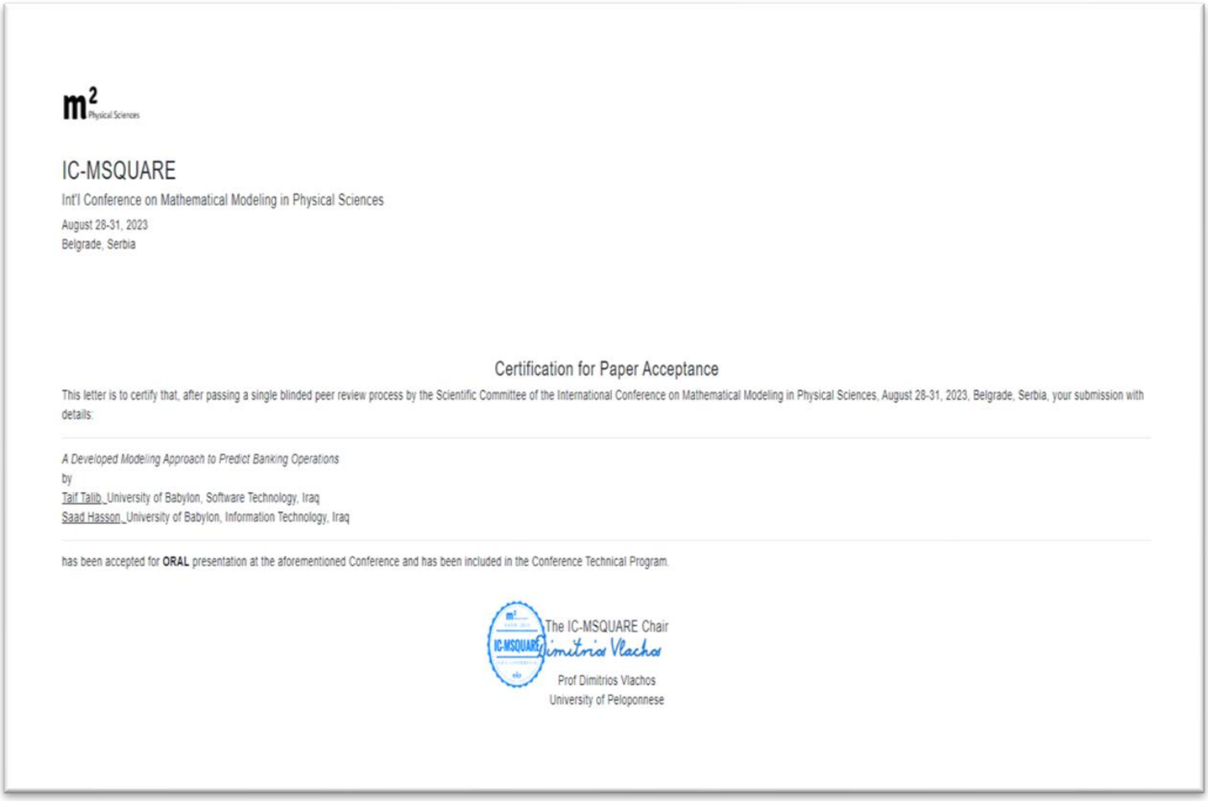
S. Jiang, A. E. Williams, K. Schenke, M. Warschauer, and D. O. Dowd, "Predicting MOOC Performance with Week 1 Behavior," Proc. 7th Int. Conf. Educ. Data Min., pp. 273–275, 2014.

S. Wang and W. Shi, Data mining and knowledge discovery. 2012.

Sravani, K., & Mahaveerakannan, R. (2023). An innovative method of loan prediction that compares decision tree algorithm accuracy with random forest. Journal of Survey in Fisheries Sciences, 10(1S), 3033-3041.

Uthayakumar, J., Metawa, N., Shankar, K., & Lakshmanaprabu, S. K. (2020). Intelligent hybrid model for financial crisis prediction using machine learning techniques. *Information Systems and e-Business Management*, *18*(4), 617-645.

Widyastuti, M., Simanjuntak, A. G. F., Hartama, D., Windarto, A. P., & Wanto, A. (2019, August). Classification Model C. 45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch. In *Journal of Physics: Conference Series* (Vol. 1255, No. 1, p. 012002). IOP Publishing.



m^2 Physical Sciences

IC-MSQUARE

Int'l Conference on Mathematical Modeling in Physical Sciences

August 28-31, 2023

Belgrade, Serbia

Certification for Paper Acceptance

This letter is to certify that, after passing a single blinded peer review process by the Scientific Committee of the International Conference on Mathematical Modeling in Physical Sciences, August 28-31, 2023, Belgrade, Serbia, your submission with details:

A Developed Modeling Approach to Predict Banking Operations

by

Taif Talib, University of Babylon, Software Technology, Iraq

Saad Hasson, University of Babylon, Information Technology, Iraq

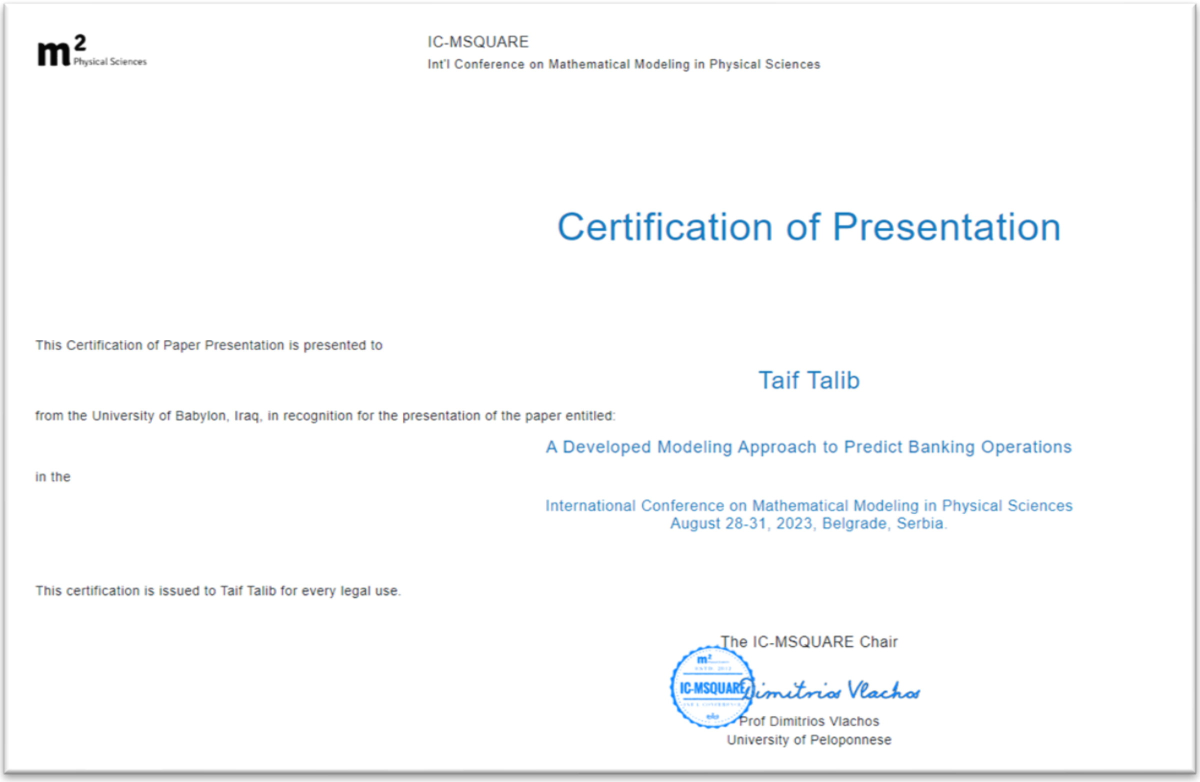
has been accepted for **ORAL** presentation at the aforementioned Conference and has been included in the Conference Technical Program.

The IC-MSQUARE Chair

Dimitrios Vlachos

Prof Dimitrios Vlachos

University of Peloponnese



m^2 Physical Sciences

IC-MSQUARE

Int'l Conference on Mathematical Modeling in Physical Sciences

Certification of Presentation

This Certification of Paper Presentation is presented to

Taif Talib

from the University of Babylon, Iraq, in recognition for the presentation of the paper entitled:

A Developed Modeling Approach to Predict Banking Operations

in the

International Conference on Mathematical Modeling in Physical Sciences
August 28-31, 2023, Belgrade, Serbia.

This certification is issued to Taif Talib for every legal use.

The IC-MSQUARE Chair

Dimitrios Vlachos

Prof Dimitrios Vlachos

University of Peloponnese

المستخلص

تعتبر النمذجة التنبؤية أداة مهمة للبنوك والمؤسسات المالية الأخرى. من أجل تحديد الاحتمالات الواقعية للنتائج المستقبلية، يجب اتباع طريقة لتقييم البيانات المصرفية والتنبؤ بالاحتمالات المحددة. النمذجة التنبؤية هي طريقة لاستخدام البيانات الحالية لإنشاء نموذج مناسب للتنبؤ بنتائج البيانات المستقبلية.

الهدف من هذه الدراسة هو اختبار مدى نجاح خوارزميات التعلم الآلي في توقع ما إذا كان العميل الجديد سيحصل على وديعة لأجل أم لا. يمكن استخدامه لمعرفة أفضل استراتيجية لاكتشاف مستهلكي الشركات المصرفية الذين يغادرون كثيرًا. الهدف الأساسي من هذه الرسالة هو إنشاء نموذج بقدرات تنبؤ دقيقة لتعزيز عمليات البنك. يمكن تحقيق هذا الهدف بأقل قدر من الخطأ عن طريق اختيار الميزات الأكثر أهمية، جنبًا إلى جنب مع (اجمع فهمًا أفضل لاحتياجات العميل بناءً على تحليل مجموعات البيانات المصرفية المختلفة، وقم بإجراء تصنيف للميزات للإشارة إلى فعالية الميزات، والإشارة إلى ارتباط الميزات لإظهار إمكانيات التخفيض).

مرحلة إعداد البيانات ومرحلة التنبؤ هما المرحلتان الرئيسيتان للنظام المقترح. يتم زيادة دقة التنبؤ بالنموذج المقترح عن طريق إجراء المعالجة المسبقة للبيانات باستخدام تنظيف البيانات (القيم المفقودة)، وتحويل البيانات (الاسمي إلى الثنائي، والاسمي إلى الرقمي)، والتطبيع (التقييس، والتخفيض)، وإجراءات تقليل البيانات. نهج معامل ارتباط الترتيب هو أحد تقنيات تقليل البيانات. من أجل تأكيد أهمية هذه الميزات والدقة التي يمكن تحقيقها، يستخدم النظام المقترح نهج معامل ارتباط الترتيب، والذي يحدد الميزات الأكثر فائدة في كل خطوة قبل دمجها في النموذج.

علاوة على ذلك. تستخدم هذه الرسالة تقنيات إحصائية وخوارزميات التعلم الآلي وتحليل الانحدار لاستخدام وتقييم مجموعتي بيانات (مجموعة بيانات بنكية إضافية كاملة ومجموعة بيانات مصرفية - مجموعات بيانات أهداف التسويق). تتضمن تقنيات النمذجة التنبؤية Naive BaysوDT وKNN وLogistic Regression. تم استخدام Accuracy, precision, recall, F-measure, and error كأساس للتقييم. أظهرت النتائج أن أداء النظام المقترح فعال، كان لدى DT **(0.93)** وKNN **(0.92)** وLogistic Regression **(0.89)** وNaive Bays **(0.86)** أعلى دقة تنبؤ.

جمهورية العراق

وزارة التعليم العالي والبحث العلمي

جامعة بابل

كلية تكنولوجيا المعلومات

قسم البرمجيات

نـهج النمذجة التنبؤية لتحسين العمليات المصرفية

رسالة مقدمة إلى
مجلس كلية تكنولوجيا المعلومات - جامعة بابل كجزء من متطلبات
نيل درجة الماجستير في تكنولوجيا المعلومات / البرمجيات

مقدمة من قبل
طيف علي طالب جواد

اشراف

الاستاذ الدكتور
سعد طالب حسون

1445هـ **2023 مـ**