CLASSIFICATION OF BANK MARKETING DATA USING SUPPORT VECTOR MACHINE

EI EI KHIN

M.C.Sc. **MAY2023**

CLASSIFICATION OF BANK MARKETING DATA USING SUPPORT VECTOR MACHINE

BY

EI EI KHIN

B.C.Sc.

A Dissertation Submitted in Partial Fulfillment of the Requirement for the Degree of

Master of Computer Science

(M.C.Sc.)

University of Computer Studies, Yangon

MAY 2023

ACKNOWLEDGEMENTS

First and foremost, I would like to thank **Prof. Dr. Mie Mie Khin,** Rector of University of Computer Studies, Yangon, from the bottom of my heart for giving the opportunity to do this dissertation and for providing me with good advice during this study.

Secondly, I would like to express my sincere gratitude to my supervisor **Dr. Tin Tin Htar**, Associate Professor, Department of Information Technology Support and Maintenance, University of Computer Studies, Yangon, for being kind to me during the writing and preparation of this dissertation and giving me many helpful suggestions and valuable support.

Thirdly, I would like to express special thanks to **Dr. Si Si Mar Win** and **Dr. Tin Zar Thaw**, Professors, Faculty of Computer Science, University of Computer Studies, Yangon, as Deans of Master's Course, for providing me with insightful advice and recommendations during the development of this thesis.

Moreover, I would like to express special thanks to **Daw Hnin Yee Aung,** Lecturer**,** Department of English, University of Computer Studies, Yangon, for support and revising my dissertation from the language point of view.

Last but not least, I would like to express my gratitude to all of the professors who guided me during my master's program and to my friends for their collaboration. I want to thank my parents in particular for their help and encouragement during my thesis.

STATEMENT OF ORIGINALITY

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

....

Date Ei Ei Khin

ABSTRACT

Nowadays, banking system plays an important role of financial sectors all over the world. The more accurate predictive modeling system is required for their services or products in the banking industry. Bank workers can make those predictive models with manually, but this process takes long time and lots of man-hours. For these reasons, machine learning techniques are useful to predict the outcomes with huge amounts of data. Classification is an important technique to analyze and to predict the data. This system will implement the classification of bank marketing data using support vector machine (SVM) to predict the probability of the customers' subscription to the term deposit whether subscribe or not. Support Vector Machine (SVM) is a supervised learning model used for classification and prediction of data. The purpose of this system is to predict the customers' response to the term 'deposit' using bank marketing data. The precision, recall, and F-Measure confusion matrix is used to gauge the system's correctness. In the first experiment when the training data is used, the accuracy without feature engineering is 86%, the accuracy with feature engineering is 83% and the accuracy with feature engineering of Correlation Matrix and Principal Component Analysis gets 96%. In the second experiment which is used the testing data, the accuracy without feature engineering gets 85%, the accuracy with feature engineering before using PCA is 83% and the accuracy after using PCA is 95%. The system shows the best results in both training data and testing data after using the Principal Component Analysis.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1
INTRODUCTION

Nowadays, data is widely used in several areas and the one who holds a huge amount of data will be the one who holds the sources of information that is necessary for their business development. The banking sector is one of the sectors that holds the customers data which can make decisions on the services of banking systems. In [16], Bank deposit is one of the bank products that saves the customers to hold an amount of money at a bank for a specific length of time. In return, the financial institution will pay the customers the relevant amount of interest, based on how much they choose to deposit and for how long.

In [10], deposit in banking system is a fixed-term investment and the main source of revenue for banks. Many banks offer different types of accounts to attract customers willing to deposit their funds. A bank can increase the number of subscribers to term deposit and can collect huge amounts of customer data through effective marketing. Bank marketing campaign can be carried out or launched in various ways using telephone, social media, emails, short message services, blogging, and others. The purpose of bank marketing campaign is to meet the targeted needs of the customers to satisfy the bank's product.

Banking industry requires more accurate prediction system because of various challenges offering products or services. Data mining is useful and important research area in banking sectors for analyzing and identifying important, useful and unknown data of banks [1]. Data mining techniques are able to be used to process and analyze data to turn raw data into useful information. For these reasons, Machine Learning techniques are needed to use to predict the result when processing huge amounts of data.

In Machine Learning techniques in data mining, classification is one of the most important techniques to analyze the raw data, to extract the model, to classify the data into the required forms of outcomes using classification methods, and to predict the categorical labels of data using the prediction models. In classification, unlabeled data is given to the model and the classification methods find the outcomes of the class to which it belongs [11]. Support Vector Machine is a supervised learning model with

related learning algorithms analyzing the data used for classification and prediction of data.

In this experiment, the system will classify the bank marketing data using Support Vector Machine and predict the customers' subscription to the term deposit whether subscribe or not. By using the proposed system, bank can achieve their organizational objectives to increase the number of subscriptions to term deposit.

1.1 Objectives of the Thesis

The primary goals are

- To classify and predict the bank marketing data to the term deposit using Support Vector Machine
- To help the banks in identifying the main factor that can increase the customers' subscription to the term deposit
- To evaluate the performance of classification of bank marketing data by using confusion matrix

1.2 Related Works

In this section, the various analysis of related works in bank marketing sectors are described below.

In paper [13], Support Vector Machine was used to predict the behaviors of bank customers toward the term deposits. The dataset was collected during a direct marketing campaign. The result came out the accuracy of 93% when predicting customer behavior with high level of predictability. This model was obtained to minimize the cost and expense of a marketing campaign for banks.

In [15], the purpose of this paper was to predict the acceptance of the bank loan offers using Support Vector Machine. The authors used to predict results with four kernels of SVM such as Linear kernel, Polynomial kernel, RBF kernel and Sigmoid kernel. The best result was obtained with Polynomial kernel as 97.2% and the lowest success rate was Sigmoid kernel as 83.3% accuracy. The study showed that Polynomial kernel was a good choice to predict loan results.

In paper [7], the supervised machine learning, Support Vector Machine and unsupervised machine learning, Support Vector Data Description were used to compare the prediction performance of bank telemarketing area based on the dataset collected

from a Portuguese banking institution. The result indicated that the machine learning was more suitable for binary classification problem with the accuracy value 0.9867. Therefore, the authors proposed that Support Vector Machine model is more suitable for binary classification problem compared to Support Vector Data Description.

1.3 Structure of the Dissertation

This dissertation is constructed into five chapters. In **Chapter-1**, Classification of Bank Marketing Data using Support Vector Machine is introduced and the chapter describes the objectives, related works, the overall framework and the structure of the dissertation.

Chapter-2 consists of the background theory and various machine learning techniques.

In **Chapter-3**, the design, overview of the system, methodology and system explanation are presented.

The architecture, the system implementation, experimental results and model comparisons are explained in **Chapter-4**.

Chapter-5 describes the conclusion of the thesis work, the limitations and further extensions of the system.

CHAPTER 2
BACKGROUND THEORY

Many banking systems use a large amount of data and data mining techniques are applied when predicting the outcomes of customers' subscriptions in their products or services. Data mining is important in many business sectors because the industries require more accurate predictive modeling system for their services or products. As a result, the industries can save lots of man hours and can predict the result exactly.

In this chapter, the related background theory of bank marketing data, data mining and various machine learning classifiers, such as Support Vector Machine, Naïve Bayes, Random Forest, K-Nearest Neighbors, Neural Network, Decision Tree, Regression Analysis, are described.

2.1 Bank Marketing Data

Bank marketing data refers to the information collected by banks to gain insights into their customers' behavior and preferences, as well as to develop effective marketing strategies. This data includes a variety of information, such as customer demographics, transaction history, account balances, credit scores, and other financial information. By analyzing this data, banks can gain a deeper understanding of their customers and tailor marketing campaigns to specific segments of customers.

One of the uses of bank marketing data is to identify potential customers for various financial products, such as deposits, loans, credit cards, and savings accounts. For example, a bank may use data analytics to identify customers who have a high credit score and a history of responsible financial behavior as potential candidates for a deposit or loan. By targeting these customers with personalized marketing messages, banks can increase the likelihood that they will sign up for these products [16].

Bank marketing data can also be used to improve customer engagement and loyalty. By analyzing customer data, banks can identify trends in customer behavior, such as the types of products they are interested in and the channels they prefer for communication. This information can be used to personalize marketing messages and improve the customer experience, which can lead to higher levels of customer satisfaction and loyalty.

Another key benefit of bank marketing data is that it can be used to optimize marketing budgets. By analyzing the data, banks can identify which marketing channels are the most effective for reaching specific customer segments. For example, if a bank finds that a particular segment of customers is more likely to respond to email marketing than direct mail, they can allocate more of their marketing budget to email campaigns for that segment. This can help banks to maximize their marketing ROI and avoid wasting resources on ineffective marketing channels [8].

Bank marketing data is a valuable resource that can help banks to understand their customers, develop effective marketing strategies, and improve customer engagement and loyalty. By analyzing this data, banks can identify potential customers for various financial products, personalize marketing messages, optimize marketing budgets, and improve the overall customer experience. However, it is important for banks to handle this data ethically and comply with privacy regulations to protect their customers' personal and financial information.

2.2 Machine Learning

Machine Learning is one of the branches of Artificial Intelligent (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Machine Learning is an important component that is growing in the data science fields. In Machine Learning, to make classifications and predictions, the algorithms are trained to use. The training data is used to determine correct data and to improve algorithms [3]. The goal of machine learning is to develop algorithms that can automatically improve their performance on a task, as more data becomes available.

Machine Learning has numerous applications, including image recognition, speech recognition, natural language processing, recommender systems, fraud detection, and many others. It has become a powerful tool for solving complex problems that would be difficult or impossible to solve using traditional programming methods as shown in Figure 2.1.

There are two subsets of learning techniques in Machine Learning and they are as follow:

1. Supervised Machine Learning
2. Unsupervised Machine Learning

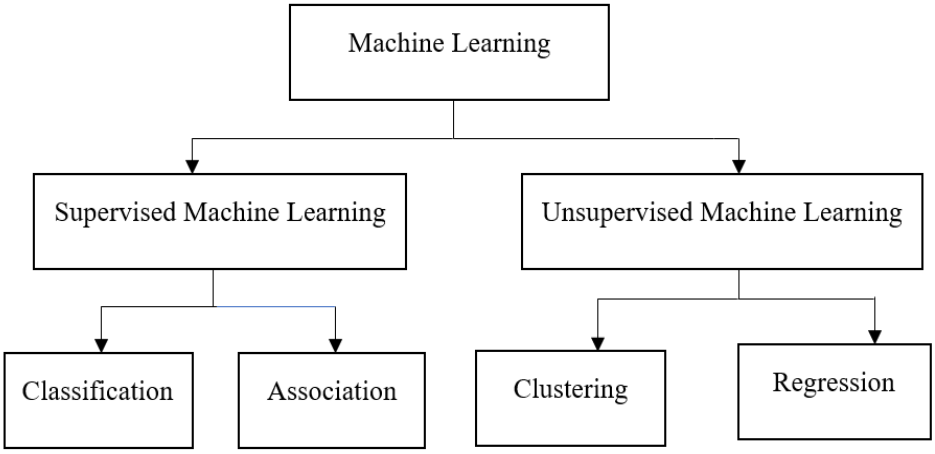


Figure 2.1 Algorithms for Machine Learning

2.2.1 Supervised Machine Learning

Supervised Machine Learning is a subcategory of machine learning that learns the relationship between input and output. It helps to solve the various varieties of real world problems. Supervised Learning is based on the training phase and during its training phase, the labeled datasets are fed in the system [12]. The labeled data is a type of data which contains both features and the targets. The features in the dataset or 'X variables' are called the inputs and the target or 'Y variables' are generally referred to output.

$$\boldsymbol{Y = f(X)} \qquad \textbf{2.1}$$

In equation 2.1, Y = the output variable, X = the input variable and f(X) = the hypothesis. The main objective of Supervised Machine Learning is to find the hypothesis as approximate. In Supervised Machine Learning, when there is new input data X, the output or target Y can be predicted. The application of the learning is to predict whether the target is Yes or No [26].

The two main categories that are included in Supervised Machine Learning are as follow:

1. **Classification**: A classification algorithm is to solve the inputs when the output is a number of categories or classes based on the labeled data. Classification algorithm can be used when the output is the categorical features such as the customer feedback is 'Positive' or 'Negative' or 'Neutral', the email is 'Spam' or 'Non-spam' and etc.

2. **Regression**: Regression is used to find the relationship between the input and output when the output variable is real value. Examples of regression models are determining how much customers will buy a certain product based on their ages, predicting real estate prices based on zip code and others.

2.2.2 Unsupervised Machine Learning

Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision. Unsupervised Machine Learning algorithms are used to analyze and cluster the unlabeled datasets. The process of Unsupervised Machine Learning is to identify the similarities and differences of the data analysis [4].

Unlike Supervised Machine Learning, Unsupervised Machine Learning is used when the input data is supported and there is no corresponding output data. Unsupervised learning is significant because it operates on unlabeled and uncategorized data. In real word, to solve the cases which have the input data but no the related output, Unsupervised Learning is needed [26]. Unsupervised Learning Algorithms, allow users to perform more complex processing tasks, compared to supervised learning [5].

The Unsupervised Learning Algorithms can be categorized in two types of problems:

1. **Clustering:** Clustering is one of the data mining methods that groups the unlabeled dataset into clusters which the objects with the most similarities are into a group and the objects with the less or no similarities are in another groups. The processes of cluster analysis are to process raw data, to find the commonalities between the data objects and to categorize them into groups as the presence and absence of structures or patterns in the information [4].
2. **Association:** An association rule is a rule-based method of Unsupervised Machine Learning for finding connection between variables in a huge dataset. These techniques are widely employed for market basket analysis and make the marketing strategies more effective. Understanding the relationships between different products and customer's habits helps the businesses to develop better cross-selling strategies. For example, people who buy X item (Bread) are also tend to purchase Y item (Butter/ Jam) [29].

2.3 Algorithms for Classification in Machine Learning

Machine learning involves the use of statistical methods to analyze data, identify patterns and relationships, and make predictions or decisions based on that data. Machine learning techniques are the field of data analysis by providing powerful tools for extracting insights from complex data sets [3].

Classification algorithms are a key component of machine learning, and are used to predict the class label or category of a given data point based on its features. Some of the most commonly used classification algorithms in machine learning will be explored in this article [5].

2.3.1 Support Vector Machine

Support Vector Machine, is a supervised machine learning model with related learning algorithms analyzing the data, which is used for classification and prediction of data. In the Support Vector Machine, each data point corresponds to a single data item in n-dimensional space. The objective of a support vector classifier is to define an optimal hyperplane to separate the two classes. The hyperplane separates the two classes to determine a plane with the largest margin. Support Vector Machine is known as the algorithm that finds a special type of linear model called the maximum margin hyperplane, which gives the maximum separation between decision classes [26].

In Support Vector Machine model as seen in Figure 2.2, the algorithm tries to find the optimal hyperplane by increasing the margin, which measures how far each class's nearest data points are from the hyperplane (called support vectors) ass. The algorithm can also use a kernel function to transform the input data into a higher-dimensional feature space, where a linear hyperplane can separate the data points more effectively.

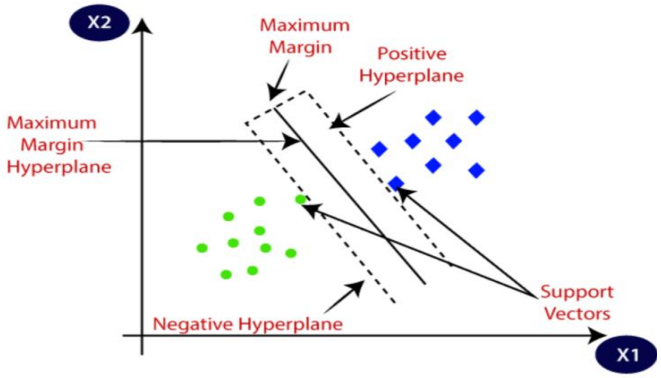


Figure 2.2 Support Vector Machine

Support Vector Machine is a powerful algorithm that can handle non-linear and high-dimensional datasets, and is robust to noise and outliers. It can also provide a measure of confidence in the predictions, through the use of the decision function or the probability estimates. Support Vector Machine is a widely used machine learning algorithm that has been successfully applied in various fields, including bioinformatics, finance, and image recognition.

2.3.2 Naïve Bayes

Naïve Bayes algorithm is based on Bayes' Theorem which is a statistical rule that calculates the probability of a hypothesis given evidence. It is mainly used for text classification problems. The algorithm learns the probability distribution of the features for each class in the training data, and then uses Bayes' Theorem to evaluate the probability of classes given features of a new instance. The formula of Bayes' Theorem states that:

$$\boldsymbol{P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}} \qquad \textbf{2.2}$$

In Equation 2.2, P(X|Y) is Posterior probability: Probability of hypothesis A on the observed event B. P(Y|X) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true. P(X) is Prior Probability: Probability of hypothesis before observing the evidence. P(Y) is Marginal Probability: Probability of Evidence [12].

2.3.3 Random Forest

Random Forest algorithm, one kind of supervised machine learning algorithms, is used for both classification and regression problems. It is an ensemble learning technique that mixes various decision trees to provide predictions that are more accurate. In a Random Forest model, a large number of decision trees are constructed, each tree is trained on a random subset of the training data and using a random subset of the features. This approach helps to reduce overfitting and increase the accuracy of the model [25].

In [21], Random Forest has several advantages, including its ability to handle large datasets with high dimensionality and its robustness to noise and outliers. It also provides feature importance measures that can help identify the most significant

features for the task at hand. Overall, Random Forest is a powerful and versatile machine learning algorithm that is widely used in various fields, including finance, healthcare, and natural language processing.

2.3.4 K-Nearest Neighbors

K-Nearest Neighbors algorithm is used for both classification and regression in Machine Learning. It is the simple and easy to implement supervised machine learning algorithm that can handle complex decision boundaries and non-linear relationships between the features and the target variable [20]. In a KNN model, the algorithm stores all the training data points in memory and when a new instance is presented, it calculates the distance between the new instance and all the training data points. The algorithm then selects the K closest neighbors (where K is a user-defined hyperparameter) and predicts the class or value of the new instance based on the most common class or the average value of the K neighbors.

However, K-Nearest Neighbors algorithm can suffer from the "curse of dimensionality" problem, where the distance metric becomes less meaningful as the number of dimensions increases. In addition, selecting the optimal value of K can be challenging and may require cross-validation. A powerful and flexible machine learning approach is K-Nearest Neighbors algorithm which is used in various applications, including image recognition, recommendation systems, and anomaly detection [28].

2.3.5 Neural Network

Neural Networks are also known as Artificial Neural Network and the classes of machine learning algorithms inspired by the structure and function of the human brain. They consist of a large number of interconnected nodes (neurons) organized into layers, which can be trained to learn complex patterns and relationships in the data [14]. In a neural network model, the input data is fed into the first layer (input layer), which then passes the information through a series of hidden layers, where the neurons apply non-linear transformations to the data. The output of the last hidden layer is then passed to the output layer, which produces the final output of the model. An example of Artificial Neural Network is as seen in Figure 2.3.

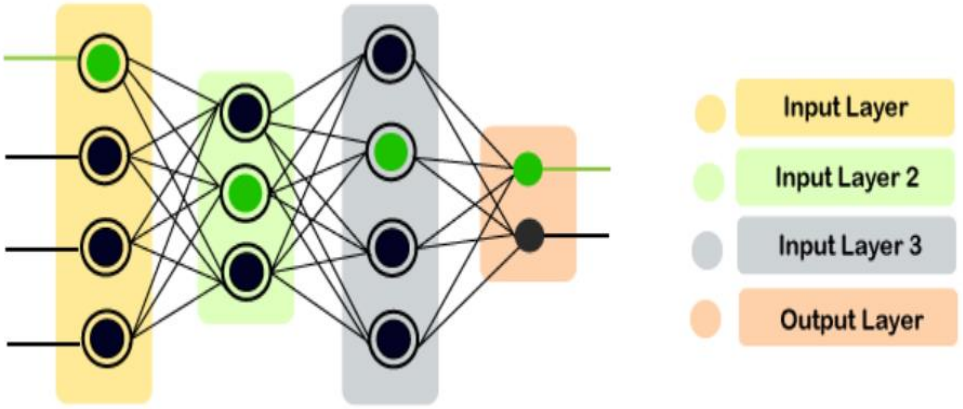


Figure 2.3 Artificial Neural Networks

The neural network learns by adjusting the weights between the neurons to minimize the difference between the predicted output and the true output (i.e., the loss function). Neural networks can be used for a variety of tasks, including classification, regression, and image and speech recognition. They are capable of learning highly non-linear relationships between the input and output variables, and can handle large and complex datasets. However, they can also be computationally expensive to train, and can suffer from overfitting if the model is too complex or the dataset is too small. Neural networks are a powerful and versatile class of machine learning algorithms that have been successfully applied in various fields, including computer vision, natural language processing, and robotics [24].

2.3.6 Decision Tree

Decision trees are well-liked machine learning approaches for classification and regression tasks. The main idea behind decision trees is to create a series of rectangular regions out of the feature space, by asking a series of yes-or-no questions based on the features of the input data. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node**.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

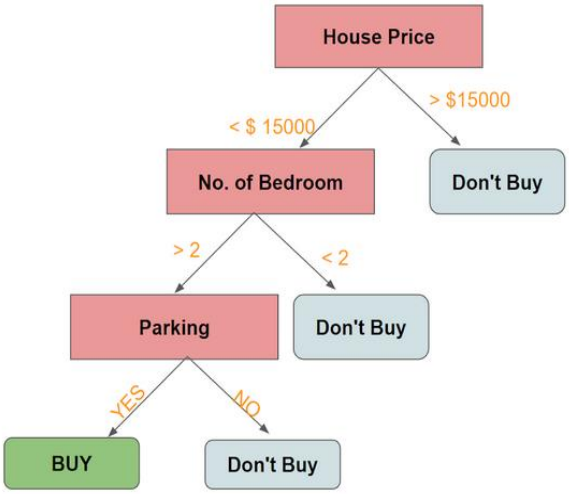


Figure 2.4 Decision Tree

In Figure 2.4, the algorithm starts with a root node that represents the entire dataset, and then recursively splits the data into smaller subsets, based on the most informative feature and threshold value that maximizes the information gain or minimize the impurity of the subsets. The algorithm stops when a stopping criterion is met, such as reaching a maximum depth, minimum number of samples per leaf, or minimum improvement in the impurity [30].

Decision trees are easy to interpret and can handle both categorical and numerical features, as well as interactions between the features. They can also handle missing values and outliers, and are applied for both binary and multi-class classification. However, decision trees are prone to overfitting if the model is too complex or the dataset is too small, and may not generalize well to unseen data. Decision trees are a simple and powerful machine learning algorithm that can be used in various applications, including medical diagnosis, credit risk assessment, and customer segmentation [19].

2.3.7 Regression Analysis

Regression is a statistical method which is used to identify and quantify the relationship between variables and helps to predict the continuous output variable based on the one or more predictor variables. It is commonly used to predict or estimate the value of the dependent variable based on the values of the independent variables [31]. The basic idea behind regression analysis is to find a mathematical formula that

describes the relationship between the dependent and independent variables. This formula is called a regression equation or model, and it is typically represented as:

$$Y = a_0 + a_1Z_1 + a_2Z_2 + \cdots + a_nZ_n + e \quad \textbf{2.3}$$

In equation 2.3, Y is the dependent variable, Z_1, Z_2, ..., Z_n are the independent variables, a_0, a_1, a_2, ..., a_n are the regression coefficients, and e is the error term. The regression coefficients show how much the related independent variable and dependent variable change by one unit each. There are different types of regression analysis, including simple linear regression, multiple linear regression, logistic regression, and nonlinear regression, among others. The type of regression analysis used depends on the nature of the data and the research question being addressed. Regression analysis is widely used in various fields such as economics, finance, psychology, and engineering, among others, to understand the relationships between variables, make predictions, and inform decision-making [2].

2.4 Data Pre-processing

In Data Mining, data pre-processing is the most important processing step before going into further process. The procedures of data pre-processing entails preparing raw data for analysis. It involves handling, cleaning and transforming data to ensure that it is accurate, complete, and suitable for analysis. The quality of the data used for analysis can significantly affect the accuracy and validity of the results obtained. Therefore, it is essential to pre-process data before analysis.

By performing data pre-processing, it can improve the accuracy and reliability of analysis, reduce the risk of errors and biases, and improve the performance of machine learning models. Additionally, data pre-processing can help to identify and address data quality issues early in the analysis process, saving time and resources [6].

2.4.1 Handle Missing Values and Remove Duplicate Values

Missing values and duplicate values can badly affect the prediction results. Missing data can be anything such as missing sequence, incomplete feature, files missing, information incomplete, data entry error, etc. Data is cleaned to remove errors, inconsistencies, missing values, duplicates values.

In [6], handling missing and duplicate values is an essential step in data pre-processing step. Missing values can be handled by deleting, imputing, or using machine learning algorithms. Duplicate values can be handled by deleting, merging, or identifying and correcting the root cause of the duplication. Therefore, firstly the missing values and duplicate values are needed to handle before going into further process.

2.4.2 Handle Outliers

An outlier is a data point that is far away from other related points in the dataset. They can be due to variability in the measurement or can show the experimental errors. Handling outliers is essential because they can significantly affect the analysis and interpretation of the data, leading to inaccurate conclusions and decisions. Outliers can significantly affect the accuracy and validity of the results obtained. Therefore, the outliners are needed to handle and remove them before processing the next steps [9]. Several approaches, can be used to handle outliers, including removing them from the dataset, transforming the data, or using robust statistical methods. The basic methods to handle outliers are box plot and Percentile methods.

2.4.3 Data Transformation

Data transformation is a fundamental step in data mining which involves the process of converting, cleansing, and structuring data into a usable format that can be analyzed to support decision making processes, and to propel the growth of an organization. Data transformation is used when data needs to be converted to match that of the destination system. It is a crucial aspect of data preparation as it helps to improve data quality, consistency, and relevance for mining purposes [32].

Data transformation involves several techniques such as normalization, aggregation, sampling, and dimensionality reduction, to name a few. Normalization is a process of scaling numerical data to a specific range, whereas aggregation is used to summarize data by grouping or merging them. Sampling helps to reduce the size of the dataset by selecting a representative subset of data, while dimensionality reduction techniques such as principal component analysis (PCA) and factor analysis help to reduce the number of variables in the dataset by identifying the most important ones.

Data transformation is necessary to ensure that the data is suitable for mining and analysis. It helps to address issues such as missing values, inconsistent data formats,

and outliers, which can adversely affect the results of the analysis. Moreover, it helps to extract valuable insights and knowledge from the data, which can be used for making informed decisions.

CHAPTER 3
DESIGN OF THE PROPOSED SYSTEM

In this Chapter, the system overview, methodology and justification are presented in details. This chapter also discusses about dataset, feature engineering, the methodologies for Support Vector Machine to accomplish the primary goal of the system which is to analyze and classify the banking marketing data using Support Vector Machine. Python programming language is employed to conduct the experiments for the thesis.

3.1 Overview of the System

The main goal of the system is to predict the bank customers whether subscribes the term 'deposit' or not using data mining and its machine learning classification techniques. The flow chart of the system is shown in Figure 3.1. In this system, data pre-processing, feature engineering, splitting dataset as training and testing, classification, model building, model evaluation are performed. The dataset can include a variety of missing values and errors. The pre-processing steps will be processed the data cleansing, handling and transforming the data to the usable format.

The used dataset has 17 attributes which includes 16 attributes is to classify the data and 1 attribute is the feature attributes. It consists of 11162 rows, two classes (Yes = 5289 rows and No = 5873 rows) that predict the deposit. The dataset is divided into 80 % and 20 % as training data and testing data respectively. In the model training stage, the classification model is trained using a portion of the data, and the model's performance is evaluated using a separate portion of the data in the evaluation stage.

Feature engineering will be completed with Principal Component Analysis (PCA) and Support Vector Machine is used for classification of data. Finally, the model's performance is measured using various metrics, such as accuracy, precision, recall, and F1-score, to assess its effectiveness in predicting customer behavior. The results will be compared the accuracy of the predicted system with feature engineering and without feature engineering.

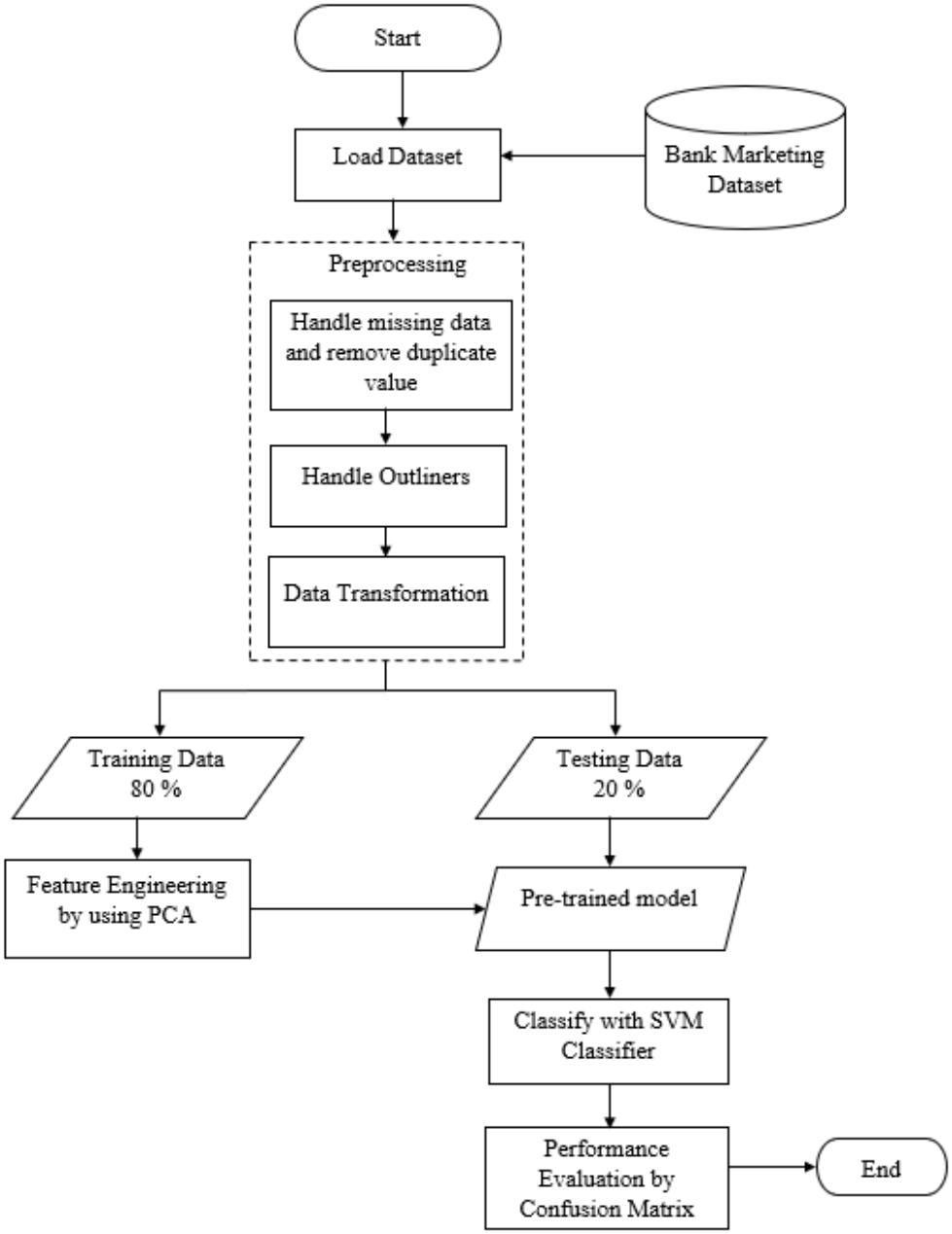


Figure 3.1 System Flow Diagram

3.2 The Explanation of the System

The goal of the proposed system is to build a model to classify the bank marketing data to the term 'deposit' whether the customers subscribe or not using feature engineering, Principal Component Analysis (PCA) method and classification techniques- Support Vector Machine (SVM). The implementation of the suggested system is run using the Python programming language.

3.2.1 Collection of Dataset

The dataset of bank marketing data is extracted from the website [29], Kaggle. The dataset contains 11162 rows, 17 columns and two classes (Yes or No). The attributes are both categorical values and numeric values. Using the train-split method, 80 % serve as training data and the remainders are testing data.

3.2.2 Dataset Information

The below Table 3.1 provides contents and details regarding the attribute "deposit" and the bank marketing dataset.

Table 3.1 Bank Marketing Data and Contents of its Description

No	Attributes	Information	Values
1	Age	Age of the customer	Numeric value
2	Job	Types of job	Admin, Blue-collar, Entrepreneur, Housemaid, Management, Retired, Self-employed, Services, Student, Technician, Unemployed, Unknown
3	Marital	Marital Status	Married, Single, Divorced
4	Education	Level of education of customer	Primary, Secondary, Tertiary, Unknown
5	Default	Does the customer have credit default?	Yes, No
6	Balance	Bank Balance	Numeric value
7	Housing	Does the customer have housing?	Yes, No
8	Loan	Does the customer have loan?	Yes, No
9	Contact	The type of contact	Cellular, Telephone, Unknown
10	Day	Last contact day of the month	Numeric value
11	Month	Last contact month	Jan, Feb, Mar, April, May, Jun, Jul, Aug, Sept, Nov, Dec
12	Duration	Contact time in second	Numeric value
13	Campaign	Number of contacts during this campaign	Numeric value

14	Pdays	Number of days that the customer last contacted from previous campaign	Numeric value
15	Previous	Number of contacts performed before this campaign	Numeric value
16	Poutcome	Outcome of previous marketing campaign	Success, Failure, Other, Unknown
17	Deposit	Does the customer subscribe the term deposit?	Yes, No

3.2.3 Splitting Data into Training and Testing Data

Splitting data into training and testing sets is a crucial step in building effective machine learning models. This involves dividing the available data into two separate sets: the training set and the testing set. The training set is used to train the model, while the testing set is used to evaluate the performance of the model on new, unseen data. The goal of splitting the data is to prevent the model from overfitting, which occurs when the model is too closely fitted to the training data and does not generalize well to new data. By evaluating the model's performance on the testing set, the data can estimate how well it will perform on new, unseen data.

Typically, the data is randomly split into training and testing sets, with a certain percentage of the data allocated to each set and it can vary depending on the size and complexity of the dataset. In this proposed system, the data is splitted into training data and testing data as 80% and 20% respectively as shown in Figure 3.2.

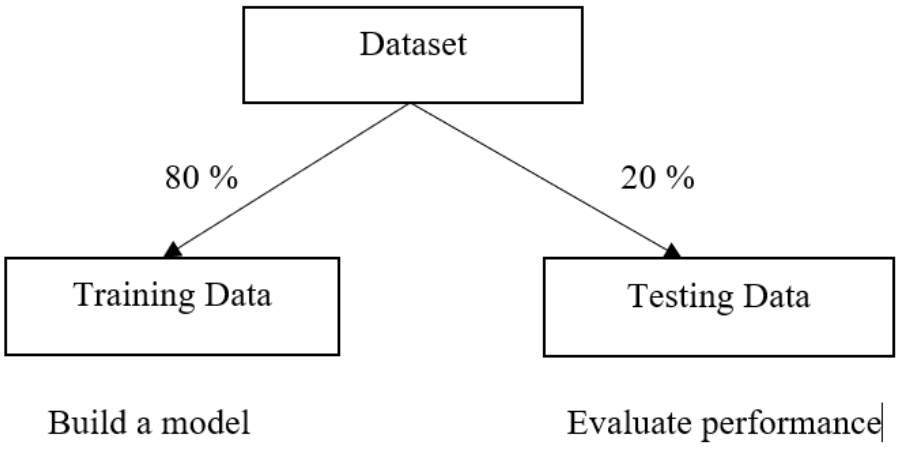


Figure 3.2 Splitting Data into Training Data and Testing Data

3.2.4 Data Pre-processing

In machine learning, data pre-processing is one of the most important processes to transform raw data to clean data. Data pre-processing stage requires filling the missing values with necessary values, deleting duplicates from the dataset and identifying and removing the features that are needed to train before processing the next processes. In this proposed system, handling missing values, removing duplicate values, handling outliers and data transformation are carried out.

```
RangeIndex: 11162 entries, 0 to 11161
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        11162 non-null  int64
 1   job        11162 non-null  object
 2   marital    11162 non-null  object
 3   education  11162 non-null  object
 4   default    11162 non-null  object
 5   balance    11162 non-null  int64
 6   housing    11162 non-null  object
 7   loan       11162 non-null  object
 8   contact    11162 non-null  object
 9   day        11162 non-null  int64
 10  month      11162 non-null  object
 11  duration   11162 non-null  int64
 12  campaign   11162 non-null  int64
 13  pdays      11162 non-null  int64
 14  previous   11162 non-null  int64
 15  poutcome   11162 non-null  object
 16  deposit    11162 non-null  object
dtypes: int64(7), object(10)
memory usage: 1.4+ MB
```

Figure 3.3 Observation of the Data Frame

The raw data is always incomplete and needed to check the values of each column. At the beginning of data pre-processing, data are needed to check which features are highly effect with the target variable, deposit. The method of "pandas.DataFrame.info" is used to extract the related details of the data. This method displays an observation of dataset that consists of the columns, non-null count, the data types and memory usages. The dataset has 16 attributes, 1 class labeled and 11162 observations as seen in the above Figure 3.3.

The number of missing values and duplicate values can affect the process of classification. Therefore, these values are needed to check and it is better to remove them to go further processes. There is no missing values and duplicate values as seen in the following Figure 3.4 and Figure 3.5.

```
age            0
job            0
marital        0
education      0
default        0
balance        0
housing        0
loan           0
contact        0
day            0
month          0
duration       0
campaign       0
pdays          0
previous       0
poutcome       0
deposit        0
dtype: int64
```

Figure 3.4 Handling Missing Values

```
False    11162
dtype: int64
```

Figure 3.5 Handling Duplicate Values

Handling outliers is one of the most important steps in the data pre-processing stage. They can affect the accuracy of the output of system. Therefore, they are needed to handle before going any other processes. There are two basic methods to handle the outliers, percentile and box plot. In Figure 3.6, there are a couple of outliers which are needed to fix.

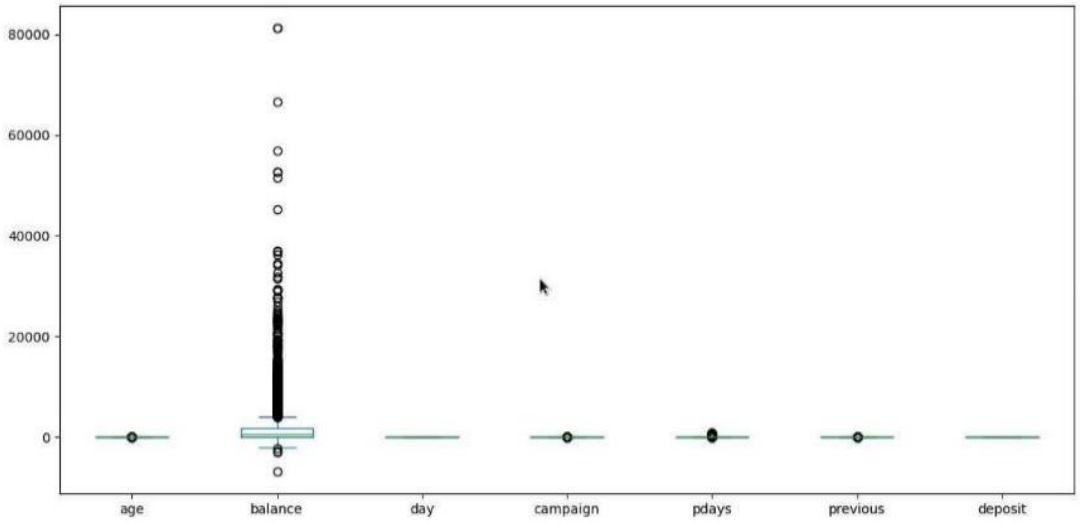


Figure 3.6 Checking Outliers

As an example of handling outliers, Campaign is having an outlier with a value of more than 40 as seen in the following Figure 3.7. The other values which are higher than the upper quartile range cannot be considered as outliers since they are recognized as anomalies. Therefore, they are needed to remove them.

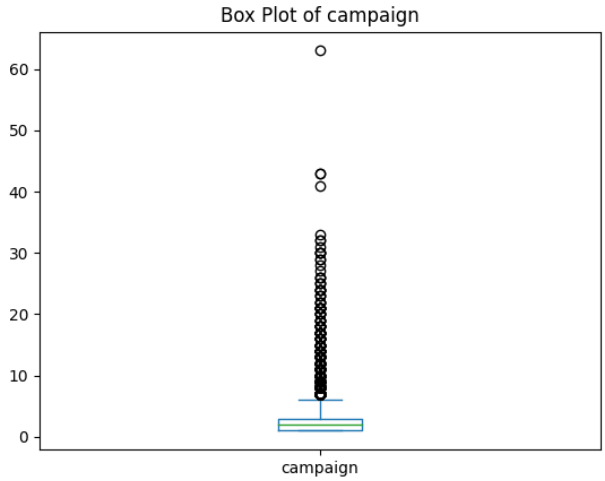


Figure 3.7 Before removing the outliers of " Campaign" Attribute

The result of removing campaign having outlier with a value of more than 40 is shown in the following Figure 3.8.

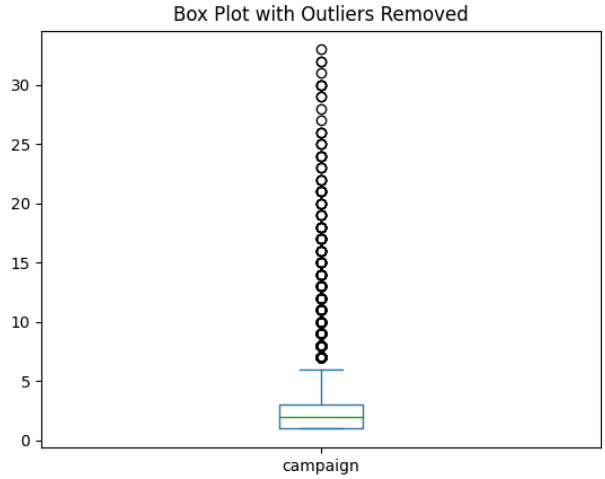
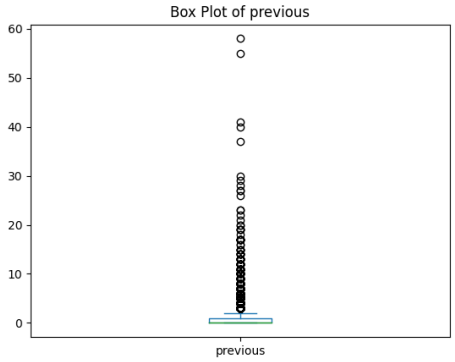
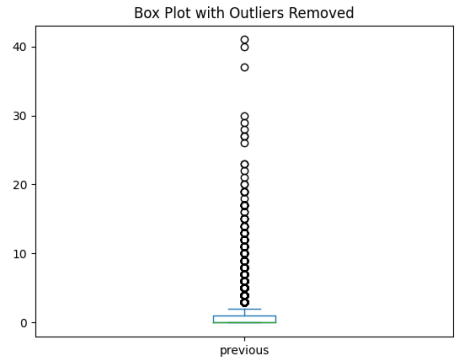


Figure 3.8 After removing the outliers of " Campaign" Attribute

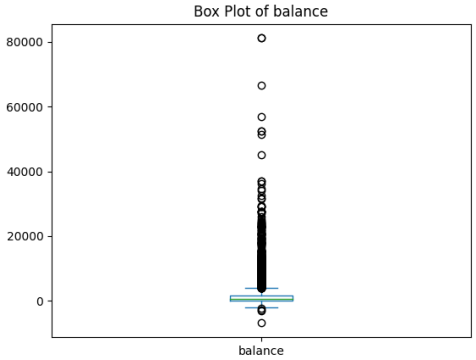
The other features which are needed to fix outliers, such as Balance and Previous, are also fixed. The result can be seen in the following Figure 3.9.



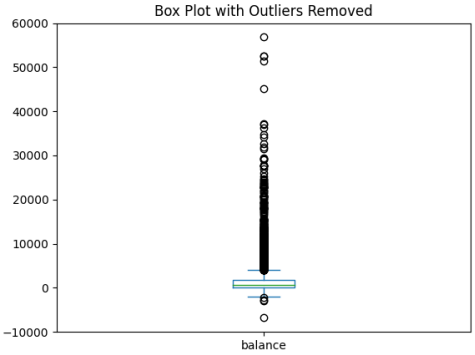
Before removing the outliers of "Previous" Attribute



Before removing the outliers of "Previous" Attribute



Before removing the outliers of "Balance" Attribute



Before removing the outliers of "Balance" Attribute

Figure 3.9 Handling Outliers

3.2.5 Label Encoding

Label encoding is a process of transforming categorical data into numerical data by assigning a unique integer to each category. It is a simple and efficient technique to transform the categorical data into numerical data. In this technique, each category is assigned a label starting from 0, 1, 2, and so on based on its order of appearance in the dataset. In this system, label encoding is used to transform the data into the usable format to process the system.

Table 3.2 Label Encoding

Education	Education (Label Encoding)
Primary	1
Secondary	2
Tertiary	3
Unknown	4

3.3 Feature Engineering

Feature engineering is the process of selecting and transforming the raw data variables to create new features that improve the performance of machine learning models. It involves selecting, transforming, and creating new features from the raw data, with the goal of improving the accuracy and efficiency of the model, identifying the relevant variables and manipulating them to extract meaningful information that can help the model learn patterns and make accurate predictions. The process typically involves analyzing the data to identify which features are the most relevant, and then applying various techniques to extract meaningful information from those features.

The goal is to extract the most important and informative features from the data, to increase the model's accuracy and predictive power. Proper feature engineering can significantly improve the performance of machine learning models, especially in cases where the raw data is noisy or incomplete.

3.3.1 Correlation Matrix between Features

The correlation coefficients between several variables are shown in a table called a correlation matrix. Correlation coefficients are a statistical measure that indicates how strongly two variables are related to each other. The correlation matrix provides a visual representation of the strength and direction of the relationships between the variables. The correlation matrix of the transformed data is used to identify the dependent and independent features. If the two features are strongly correlated to each other, the two variables are highly dependent on each other. Correlation matrices are commonly used in various fields to help identify patterns and relationships between variables. The followings Table 3.3 is shown the different correlations between the features of the dataset.

Table 3.3 Different Correlations between the Features

Correlation	Interpretation
Near to +1	Strong positive correlation between variables
Near to -1	Strong negative correlation between variables
Near to 0	No correlation between variables

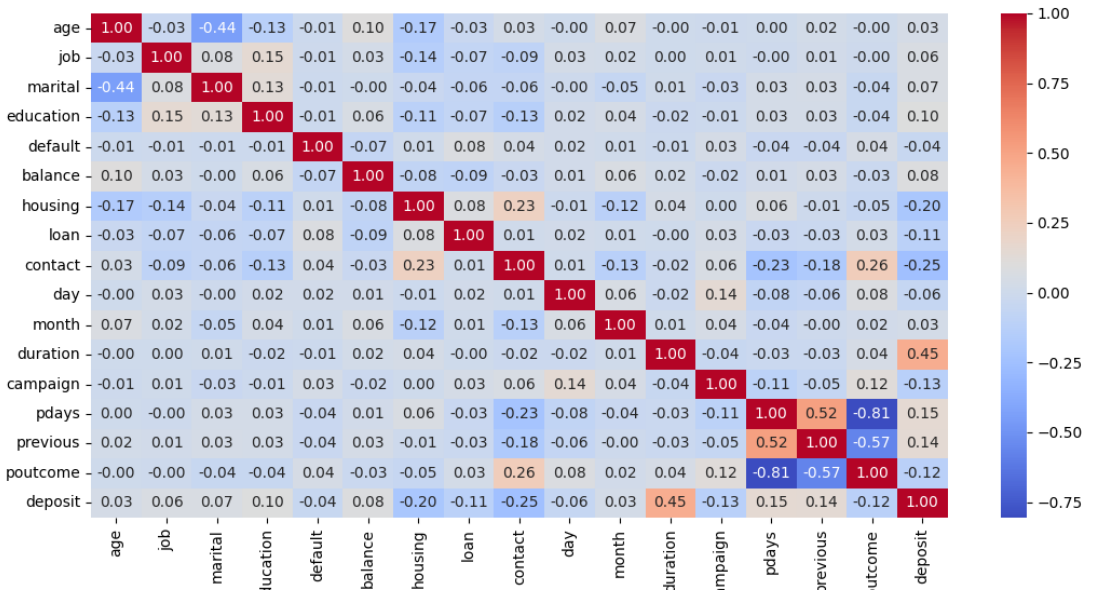


Figure 3.10 Correlation Matrix between 17 Features

In Figure 3.10, the correlation matrix shows that "Previous" is strongly positive correlated with "Pdays". Moreover, "Poutcome" is strongly negative correlated with "Pdays" and "Previous". The 3 features are highly dependent on each other. Therefore, the correlations between these 3 features and the target variable can be seen as following Table 3.4. According to the value of correlations, Pdays is the highest correlation with the target variable, Deposit. The other variables are weak relationship with the target. According to these values, Pdays is kept and the other 2 variables, "Previous" and Poutcome, are dropped from the dataset.

Table 3.4 Correlations between 3 Dependent Features and Deposit

No	Features	Correlation with Deposit
1	Pdays	0.15
2	Previous	0.14
3	Poutcome	-0.12

Removing highly dependent features from the correlation matrix is to avoid the impact of multicollinearity and improve the stability and reliability of our statistical analyses and machine learning models. This can ultimately lead to better predictions and more accurate insights from our data.

3.3.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a popular technique used in machine learning and data science for dimensionality reduction. It is used to transform a high-dimensional dataset into a lower-dimensional dataset while retaining as much of the original information as possible. The resulting dataset is referred to as the principal components. The process of PCA is simply reducing the number of variables of a dataset, while preserving as much information as possible.

PCA works by identifying the directions of maximum variance in the original data and projecting the data onto those directions. This is done by finding the eigenvectors of the covariance matrix of the data, which correspond to the directions of maximum variance. The corresponding eigenvalues indicate the amount of variance in the data that is accounted for by each eigenvector.

In [32], PCA can be useful for several reasons. First, it can simplify the data and make it easier to analyze. By reducing the number of features, it can make the data more interpretable and easier to visualize. Second, it can improve the performance of machine learning algorithms by reducing the noise and redundancy in the data. Third, it can address the curse of dimensionality, which can occur when there are too many features relative to the size of the dataset, by reducing the number of features while retaining as much information as possible.

After dropping the 2 features - "Previous" and "Poutcome", "Contact" and "Housing" are positively correlated with the value of 0.23. Therefore, these 2 variables are selected for Principal Component Analysis (PCA). PC1 (the first Principal Component) has the better relationship with deposit than "Contact" and "Housing" as seen in the following Table 3.5. Therefore, these 2 variables, "Contact" and "Housing", are replaced with PC1.

Table 3.5 Correlations between Features After Using PCA

No	Features	Correlation with Deposit
1	PC1	-0.27
2	Contact	-0.26
3	Housing	-0.20

3.4 The Proposed System's Methodology

At this point of the system, the pre-processed is used to develop the machine learning model to predict the deposit of bank customers. Support Vector Machine (SVM) classifier is employed to calculate the accuracy and performance of the system,. The goal is to evaluate the performance and effectiveness of Support Vector classifier in predicting the bank marketing data. The proposed system uses the bank marketing dataset as an input and to build a prediction model using Support Vector Machine. The dataset is splitted into training and testing data and the training data is used to create the prediction model and the testing data is used to evaluate the performance of the model.

3.4.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a popular supervised machine learning algorithm used for classification and regression tasks. SVM aims to find the best hyperplane that separates the data into different classes by maximizing the margin between the hyperplane and the data points. In the case of classification, an SVM finds the best hyperplane that separates the different classes of data. The hyperplane is chosen such that the distance between the closest data points of each class and the hyperplane is maximized. These closest data points are known as support vectors, and they define the decision boundary of the model.

In order to make predictions for new data points, the SVM calculates the distance between the new data point and the hyperplane. If the distance is positive, the new data point is classified as belonging to one class, and if the distance is negative, it is classified as belonging to the other class. The decision boundary of an SVM can be linear or non-linear, depending on the kernel used. A kernel is a function that maps the original data into a higher-dimensional space, where a linear boundary can be found.
The SVM model can be formulated as follows:

For a given training set $\{(x_i, y_i)\}$ with i = 1,2,...,n, where x_i is a vector of input features and y_i is the corresponding class label, the SVM finds the hyperplane that maximizes the margin:

$$w^T x + b = 0 \qquad \textbf{3.1}$$

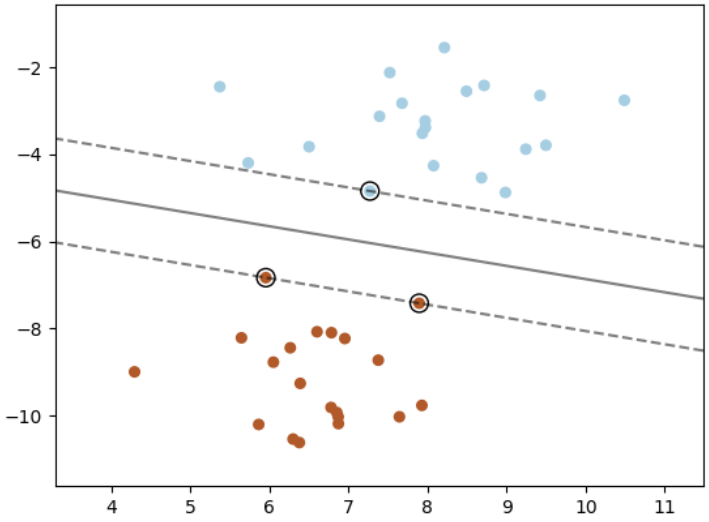


Figure 3.11 Support Vector Machine

In the above Figure 3.11, the solid line represents the hyperplane found by the SVM, while the dashed lines represent the margin. The support vectors are the points closest to the hyperplane and are circled in red. The SVM aims to find the hyperplane that maximizes the margin between the support vectors of each class. In SVM, three main lines can be found as follow:

$w^Tx + b = 0$ –for Decision boundary

$w^Tx + b = (-1)$ – for Class-1 boundary

$w^Tx + b = (+1)$ – for Class-2 boundary

1. Positive samples - $w^Tx + b \geq (+1)$
2. Negative samples - $w^Tx + b \leq (-1)$

Then by inducing the variable Y as following, a conditional statement can be generated as follow:

$$y(w^Tx + b) \geq (+1) \quad \textbf{3.2}$$

Therefore, the support vector is

$$y_i(w^Tx_i + b) = 1 \quad \textbf{3.3}$$

When this condition is satisfied, all the positive and negative data points will be behind the boundary lines. When applying SVM, the maximum width between the boundaries are needed.

$$\text{Width} = (x_{i+} - xi_{-}) \frac{w^T}{||w||}$$

$$= \frac{(1-b) - (b-1)}{||w||}$$

$$= \frac{2}{||w||} \quad \textbf{3.4}$$

Where, $y_i = 1$ is for positive samples and $y_i = -1$ is negative samples.

To find a hyperplane with the maximum margin, which can be expressed as an optimization problem shown as:

$$\text{Minimize: } \frac{1}{2} ||w||^2 \quad \textbf{3.5}$$

Subject to: $y_i(w^T x + b) \geq 1$, i = 1,2,…,n

Where,

w = the weight vector

b = the bias term,

y_i = either +1 or -1 depending on the class label of the i^{th} data point.

Support Vector Machine (SVM) is a powerful machine learning algorithm with several advantages. They are as follows:

- Can handle both linear and nonlinear data.
- Work well with high-dimensional data, and have a strong theoretical foundation.
- Support Vector Machine has good generalization performance.
- Effectively handle classification and regression tasks.

Some of the disadvantages of Support Vector Machine (SVM) can be included and they are as follow:

- Sensitive to the choice of kernel function.
- High computational requirements for large datasets.
- Difficult in interpreting the model's results.
- Support Vector Machine does not handle noisy data well.

CHAPTER 4
IMPLEMENTATION OF THE PROPOSED SYSTEM

A detail explanation of the implementation of the predicted system is described in this part. The comparison of the accuracy of the model by using with feature engineering and without feature engineering will be evaluated. The performance of Support Vector classifier is evaluated to predict the customers' deposit in bank marketing data. The accuracy, precision, recall and f-measure are calculated by using Confusion Matrix. The analysis results of the comparison are shown figures and these results will indicate the performance of the model.

4.1 Performance Evaluation

Performance evaluation is the process of measuring the effectiveness and efficiency of a system, process, or model. In machine learning, performance evaluation involves measuring the accuracy and effectiveness of a model by comparing its predicted outputs with the actual outputs. Various metrics, such as accuracy, precision, recall, and F1 score, can be used to evaluate the performance of the model. The performance evaluation helps in identifying the strengths and weaknesses of the model and improving its accuracy and effectiveness. It is an essential step in machine learning and helps in making informed decisions based on the model's performance.

The performance of a model is often evaluated using a confusion matrix. A confusion matrix is a table that is used to evaluate the performance of a classification model by comparing the actual and predicted values. It is an effective tool for analyzing the accuracy of a model and identifying where the model is making errors. The confusion matrix is constructed by comparing the actual values (ground truth) of the target variable with the predicted values generated by the model. The matrix is composed of four parts: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

- **True Positives (TP)** refer to the total number of deposit customers who are actually deposit and the model made the correct predictions as deposit customers.

- **True Negatives (TN)** refer to the total number of deposit customers who are actually non-deposit and the model made the correct predictions as non-deposit customers.
- **False Positives (FP)** refer to the total number of deposit customers who are non-deposit but the system incorrectly predicted as deposit customers.
- **False Negatives (FN)** refer to the total number of deposit customers who deposit but the system incorrectly predicted as non-deposit customers.

Table 4.1 Binary Classifier's Confusion Matrix

	Classes Prediction		
Actual Classes	Yes	TP (True Positive)	FN (False Negative)
	No	FP (False Positive)	TN (True Negative)

Accuracy

The percentage of all classes that were correctly predicted is determined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \textbf{4.1}$$

Precision

The percentage of consumers who were expected to make deposits actually did so is determined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \textbf{4.2}$$

Recall

The following formula is used to assess the percentage of actual deposit customers who are correctly identified as such:

$$\text{Recall} = \frac{TP}{TP + FN} \quad \textbf{4.3}$$

F-Measure

A precision and recall weighted average, where the best value is 1 and the worst is 0. The relative contributions of memory and precision to the F-1 score are equal. The calculation is as follows:

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad \textbf{4.4}$$

4.2 Experimental Setup

The objective of this part is to describe the system implementation, design, performance of the model and the experimental outcomes. The client deposit prediction system can help retain valuable customers who are anticipated to leave. The programming language, Python is used to run the system.

4.3 System Implementation

The "Welcome" page, as seen in following Figure 4.1, is found when the proposed system is started. To start the system, "Start" button is clicked.

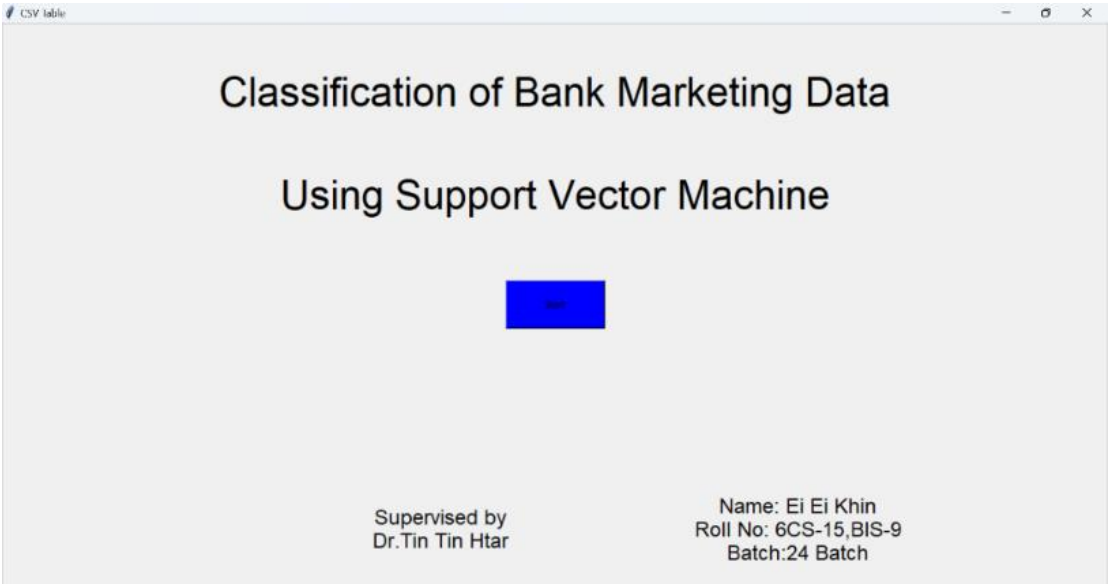


Figure 4.1 The Start Page of the System

By pressing the "Start" button, the system's "Home" screen appears as seen in Figure 4.2. The "Load Dataset," "Data Preprocessing," "Import," and "Comparison" buttons are located on this page.

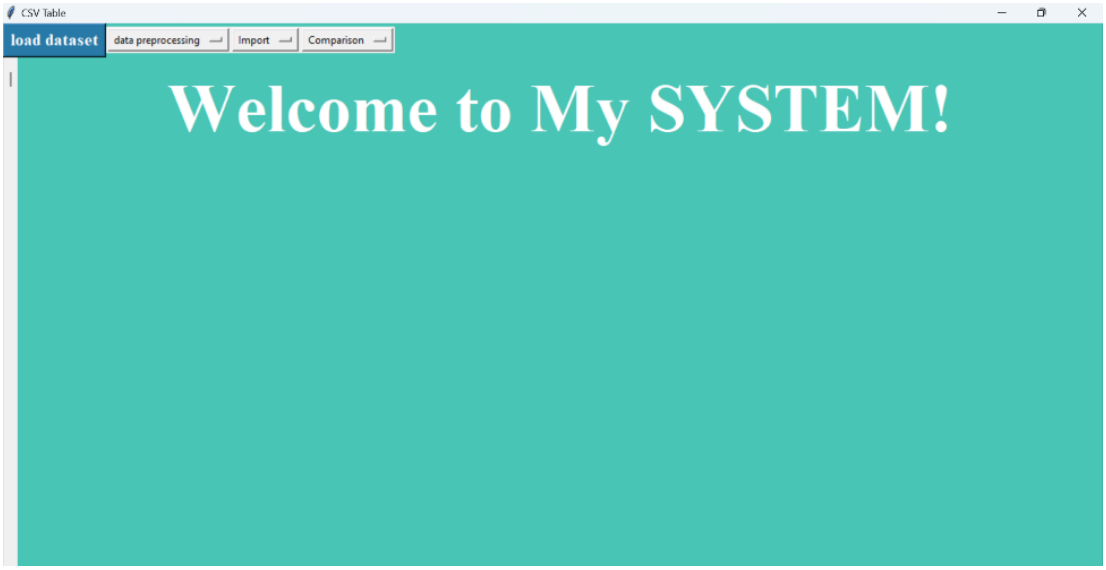
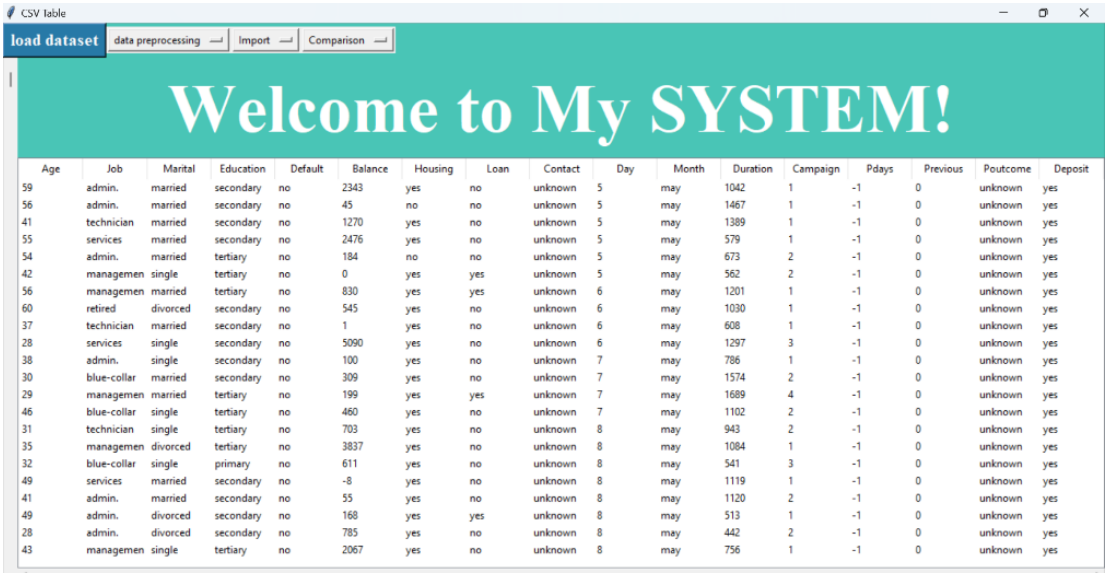


Figure 4.2 Home Screen of the System

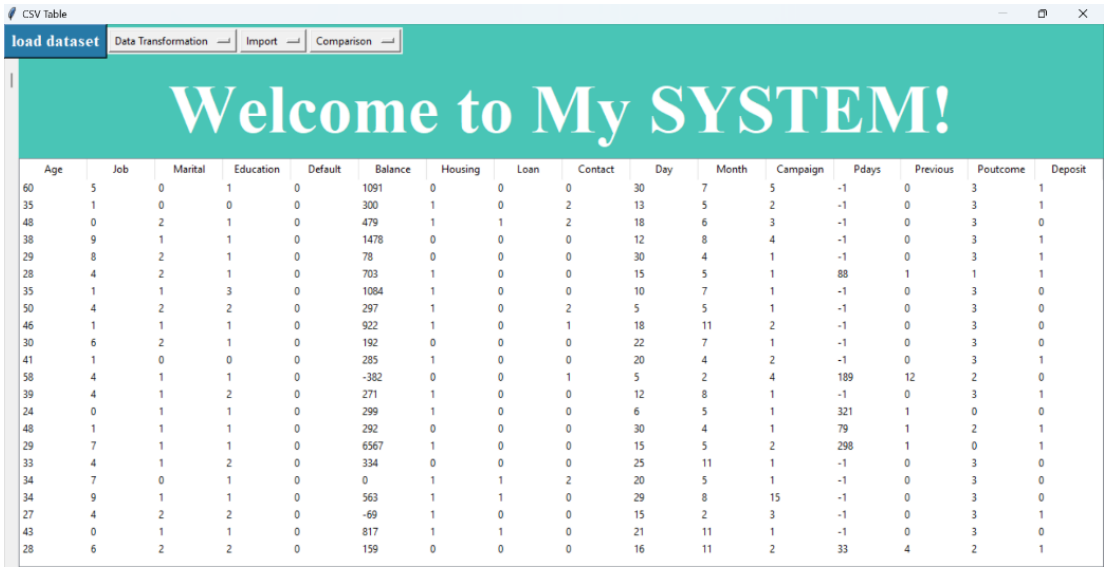
By using "Load Dataset" option, one may observe the bank marketing dataset that is utilized in the suggested system. Figure 4.3 displays the outcome of pressing this button.



Age	Job	Marital	Education	Default	Balance	Housing	Loan	Contact	Day	Month	Duration	Campaign	Pdays	Previous	Poutcome	Deposit
59	admin.	married	secondary	no	2343	yes	no	unknown	5	may	1042	1	-1	0	unknown	yes
56	admin.	married	secondary	no	45	no	no	unknown	5	may	1467	1	-1	0	unknown	yes
41	technician	married	secondary	no	1270	yes	no	unknown	5	may	1389	1	-1	0	unknown	yes
55	services	married	secondary	no	2476	yes	no	unknown	5	may	579	1	-1	0	unknown	yes
54	admin.	married	tertiary	no	184	no	no	unknown	5	may	673	2	-1	0	unknown	yes
42	managemen	single	tertiary	no	0	yes	yes	unknown	5	may	562	2	-1	0	unknown	yes
56	managemen	married	tertiary	no	830	yes	yes	unknown	6	may	1201	1	-1	0	unknown	yes
60	retired	divorced	secondary	no	545	yes	no	unknown	6	may	1030	1	-1	0	unknown	yes
37	technician	married	secondary	no	1	yes	no	unknown	6	may	608	1	-1	0	unknown	yes
28	services	single	secondary	no	5090	yes	no	unknown	6	may	1297	3	-1	0	unknown	yes
38	admin.	single	secondary	no	100	yes	no	unknown	7	may	786	1	-1	0	unknown	yes
30	blue-collar	married	secondary	no	309	yes	no	unknown	7	may	1574	2	-1	0	unknown	yes
29	managemen	married	tertiary	no	199	yes	yes	unknown	7	may	1689	4	-1	0	unknown	yes
46	blue-collar	single	tertiary	no	460	yes	no	unknown	7	may	1102	2	-1	0	unknown	yes
31	technician	single	tertiary	no	703	yes	no	unknown	8	may	943	2	-1	0	unknown	yes
35	managemen	divorced	tertiary	no	3837	yes	no	unknown	8	may	1084	1	-1	0	unknown	yes
32	blue-collar	single	primary	no	611	yes	no	unknown	8	may	541	3	-1	0	unknown	yes
49	services	married	secondary	no	-8	yes	no	unknown	8	may	1119	1	-1	0	unknown	yes
41	admin.	married	secondary	no	55	yes	no	unknown	8	may	1120	2	-1	0	unknown	yes
49	admin.	divorced	secondary	no	168	yes	yes	unknown	8	may	513	1	-1	0	unknown	yes
28	admin.	divorced	secondary	no	785	yes	no	unknown	8	may	442	2	-1	0	unknown	yes
43	managemen	single	tertiary	no	2067	yes	no	unknown	8	may	756	1	-1	0	unknown	yes

Figure 4.3 Bank Marketing Dataset

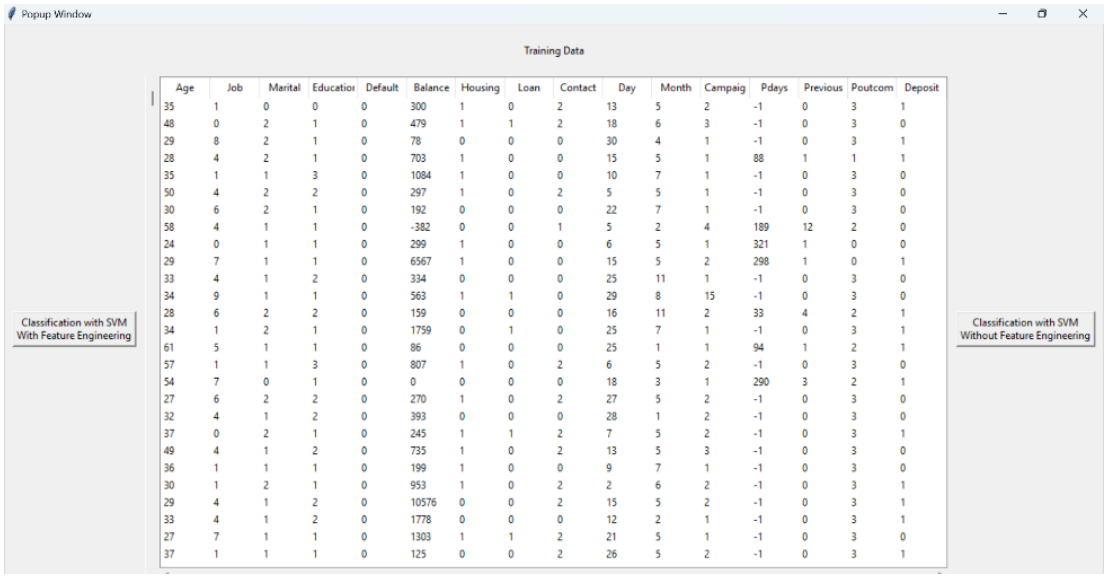
Data Pre-processing step can be complete by clicking "Data Preprocessing" button. In this stage, "Handle Missing Value and Duplicate Value", "Handle Outlier" and "Data Transformation" are carried out. Figure 4.4 displays the outcome of the preprocessing of the data.



Age	Job	Marital	Education	Default	Balance	Housing	Loan	Contact	Day	Month	Campaign	Pdays	Previous	Poutcome	Deposit
60	5	0	1	0	1091	0	0	0	30	7	5	-1	0	3	1
35	1	0	0	0	300	1	0	2	13	5	2	-1	0	3	1
48	0	2	1	0	479	1	1	2	18	6	3	-1	0	3	0
38	9	1	1	0	1478	0	0	0	12	8	4	-1	0	3	1
29	8	2	1	0	78	0	0	0	30	4	1	-1	0	3	1
28	4	2	1	0	703	1	0	0	15	5	1	88	1	1	1
35	1	1	3	0	1084	1	0	0	10	7	1	-1	0	3	0
50	4	2	2	0	297	1	0	2	5	5	1	-1	0	3	0
46	1	1	1	0	922	1	0	1	18	11	2	-1	0	3	0
30	6	2	1	0	192	0	0	0	22	7	1	-1	0	3	0
41	1	0	0	0	285	1	0	0	20	4	2	-1	0	3	1
58	4	1	1	0	-382	0	0	1	5	2	4	189	12	2	0
39	4	1	2	0	271	1	0	0	12	8	1	-1	0	3	1
24	0	1	1	0	299	1	0	0	6	5	1	321	1	0	0
48	1	1	1	0	292	0	0	0	30	4	1	79	1	2	1
29	7	1	1	0	6567	1	0	0	15	5	2	298	1	0	1
33	4	1	2	0	334	0	0	0	25	11	1	-1	0	3	0
34	7	0	1	0	0	1	1	2	20	5	1	-1	0	3	0
34	9	1	1	0	563	1	1	0	29	8	15	-1	0	3	0
27	4	2	2	0	-69	1	0	0	15	2	3	-1	0	3	1
43	0	1	1	0	817	1	1	0	21	11	1	-1	0	3	0
28	6	2	2	0	159	0	0	0	16	11	2	33	4	2	1

Figure 4.4 Bank Marketing Dataset after Data Pre-processing

By clicking "Import" button, "Training Data" and "Testing Data" can be selected. If the "Training Data" is selected, the 80% of the original dataset will be imported as the training data. In this page, "Classification with SVM with Feature Engineering" and "Classification with SVM without Feature Engineering" buttons can be seen. The result of clicking "Training Data" button are observed in the Figure 4.5.



Age	Job	Marital	Educatior	Default	Balance	Housing	Loan	Contact	Day	Month	Campaig	Pdays	Previous	Poutcom	Deposit
35	1	0	0	0	300	1	0	2	13	5	2	-1	0	3	1
48	0	2	1	0	479	1	1	2	18	6	3	-1	0	3	0
29	8	2	1	0	78	0	0	0	30	4	1	-1	0	3	1
28	4	2	1	0	703	1	0	0	15	5	1	88	1	1	1
35	1	1	3	0	1084	1	0	0	10	7	1	-1	0	3	0
50	4	2	2	0	297	1	0	2	5	5	1	-1	0	3	0
30	6	2	1	0	192	0	0	0	22	7	1	-1	0	3	0
58	4	1	1	0	-382	0	0	1	5	2	4	189	12	2	0
24	0	1	1	0	299	1	0	0	6	5	1	321	1	0	0
29	7	1	1	0	6567	1	0	0	15	5	2	298	1	0	1
33	4	1	2	0	334	0	0	0	25	11	1	-1	0	3	0
34	9	1	1	0	563	1	1	0	29	8	15	-1	0	3	0
28	6	2	2	0	159	0	0	0	16	11	2	33	4	2	1
34	1	2	1	0	1759	0	1	0	25	7	1	-1	0	3	1
61	5	1	1	0	86	0	0	0	25	1	1	94	1	2	1
57	1	1	3	0	807	1	0	2	6	5	2	-1	0	3	0
54	7	0	1	0	0	0	0	0	18	3	1	290	3	2	1
27	6	2	2	0	270	1	0	2	27	5	2	-1	0	3	0
32	4	1	2	0	393	0	0	0	28	1	2	-1	0	3	0
37	0	2	1	0	245	1	1	2	7	5	2	-1	0	3	1
49	4	1	2	0	735	1	0	2	13	5	3	-1	0	3	0
36	1	1	1	0	199	1	0	0	9	7	1	-1	0	3	0
30	1	2	1	0	953	1	0	2	2	6	2	-1	0	3	1
29	4	1	2	0	10576	0	0	2	15	5	2	-1	0	3	1
33	4	1	2	0	1778	0	0	0	12	2	1	-1	0	3	1
27	7	1	1	0	1303	1	1	2	21	5	1	-1	0	3	0
37	1	1	1	0	125	0	0	2	26	5	2	-1	0	3	1

Figure 4.5 Importing the Training Data of the System

If "Classification with SVM with Feature Engineering" button is clicked, the page of "PCA" will be uploaded. In this page, "Before PCA" and "After PCA" buttons are included. The outcome is as seen in Figure 4.6.

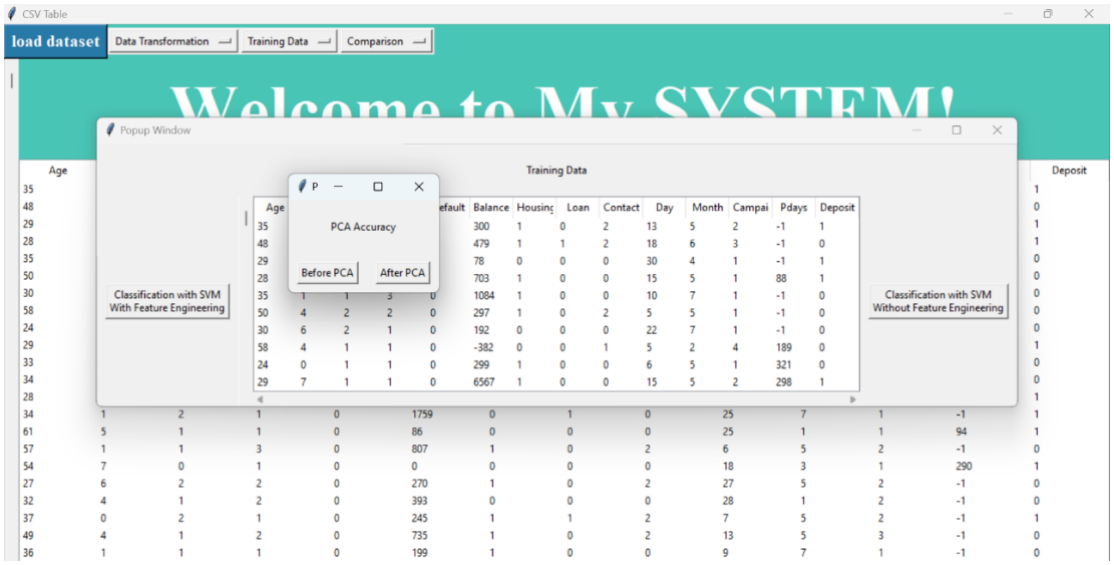


Figure 4.6 Evaluating PCA Results

By clicking "Before PCA", the correlation matrix before using PCA and the accuracy result are shown as seen in Figure 4.7.

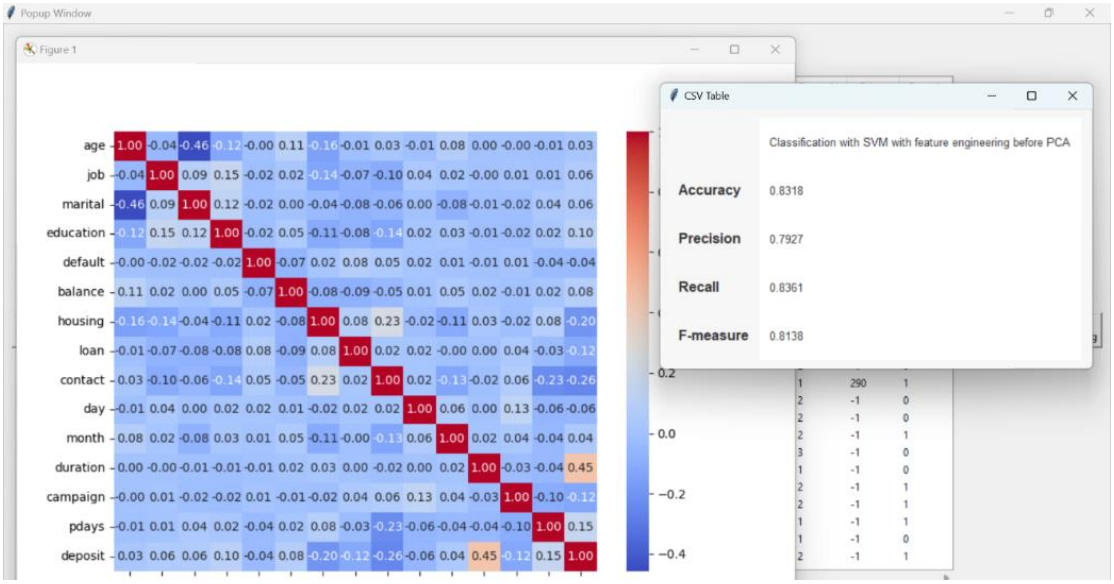


Figure 4.7 The Result of the System with Feature Engineering Before using PCA

The confusion matrix of the system with feature engineering before using Principal Component Analysis is displayed in the Figure 4.8 below.

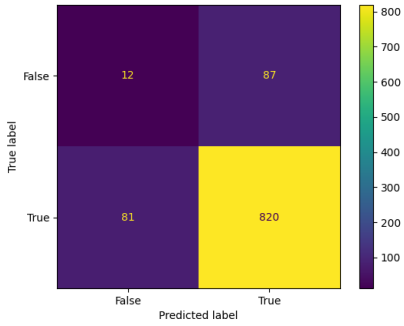


Figure 4.8 Confusion Matrix of the System with Feature Engineering Before using PCA

If "After PCA" buttons is clicked, the correlation matrix after using PCA and the accuracy result are shown as seen in Figure 4.9.

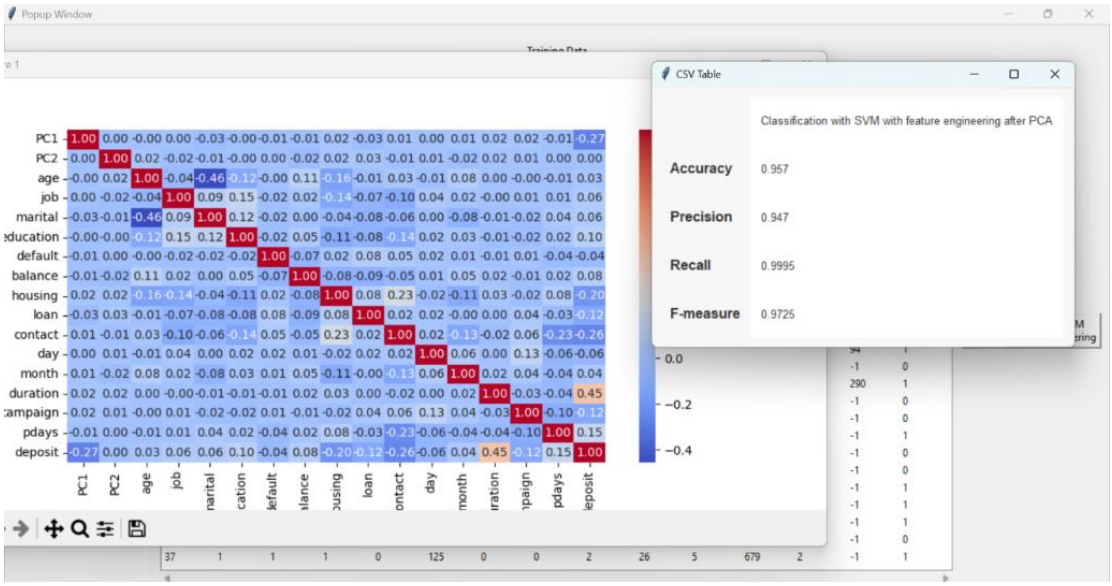


Figure 4.9 The Result of the System with Feature Engineering After using PCA

The confusion matrix of the system with feature engineering after using Principal Component Analysis is displayed in the Figure 4.10.

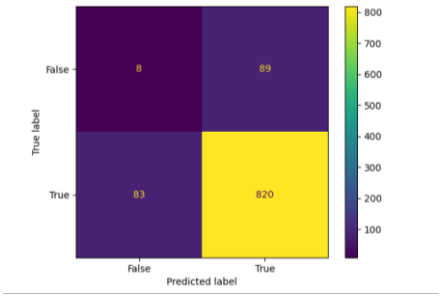


Figure 4.10 Confusion Matrix of the System with Feature Engineering After using PCA

By clicking "Classification with SVM without Feature Engineering", the accuracy of the system without feature engineering is shown as seen in Figure 4.11.

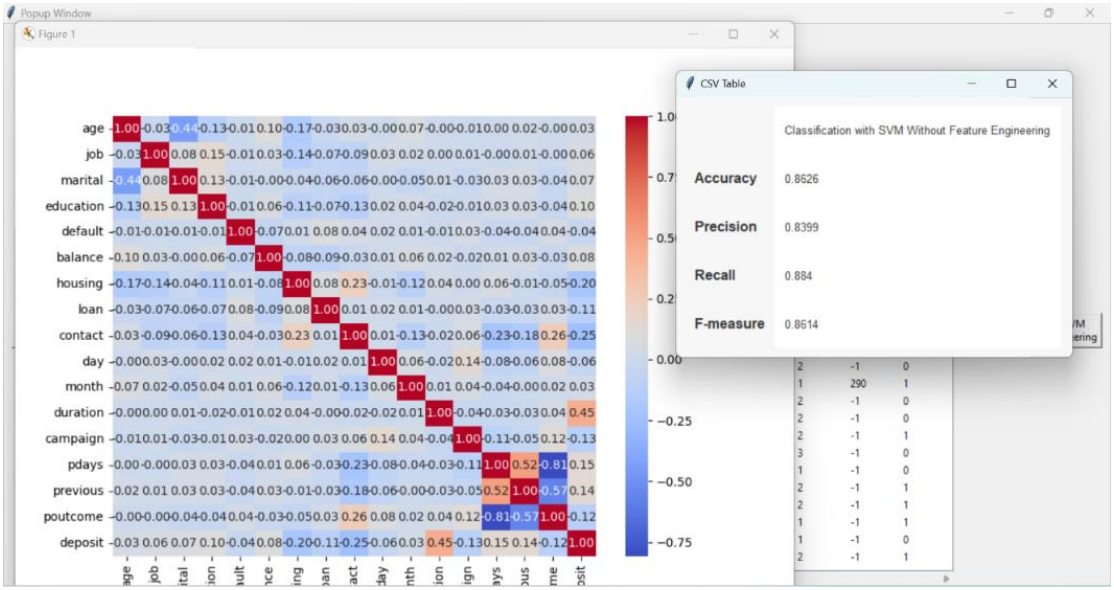


Figure 4.11 The Result of the System without Feature Engineering

The confusion matrix of the system without feature engineering is displayed in the Figure 4.12.

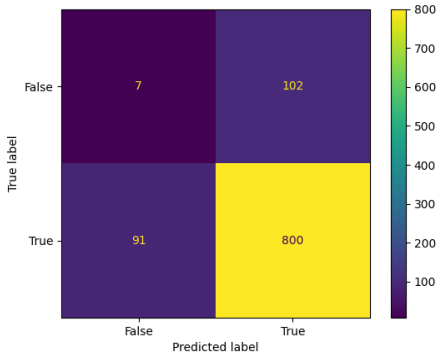


Figure 4.12 Confusion Matrix of the System without Feature Engineering

The comparison of three experimental results, the accuracy without feature engineering, the accuracy with feature engineering before using PCA and the accuracy with feature engineering after using PCA, are displayed when "Comparison" button is clicked.

CSV Table

	With Feature engineering				Without Feature engineering	
	Before PCA		After PCA			
	training	testing	training	testing	training	testing
Accuracy	0.8318	0.8271	0.957	0.954	0.8626	0.8498
Precision	0.7927	0.8251	0.947	0.9536	0.8399	0.8195
Recall	0.8361	0.8324	0.9995	1.0	0.884	0.8613
F-measure	0.8138	0.8287	0.9725	0.9762	0.8614	0.8399

Figure 4.13 The Comparison Result of Three Experiments

4.4 Experimental Result

The experiment is conducted on the bank marketing dataset to ascertain how well the supervised machine learning model predicts bank deposit customers. The proposed system is constructed by using the Support Vector Machine Classifier. Three datasets - one without feature engineering, one with feature engineering, and one with feature engineering that includes Principal Component Analysis - were employed in this investigation.

In order to remove the missing values, duplicate values and outliers, the system employs the Support Vector Machine approach. The accuracy of system with training data without feature engineering is 86% and the accuracy of the system after using the Principal Component is 96% and the accuracy before is 83%. The accuracy of system with testing data without feature engineering is 85% and the accuracy of the system after using the Principal Component is 95% and the accuracy before is 83%.

4.5 Model Comparison

The accuracy of the original model without feature engineering, the model with feature engineering, and the suggested model using principal component analysis are compared on the training data and testing data. The following Table 4.2 compares and contrasts the performances of the original model without feature engineering, the model

with feature engineering, and the final model. The final model with feature engineering after using PCA shows the best result in both training data and testing data.

Table 4.2 Accuracy, Precision, Recall and F-measure based on three models

Comparison	Model without Feature Engineering		Model with Feature Engineering Before PCA		Model with Feature Engineering After PCA	
	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Accuracy	0.8626	0.8498	0.8318	0.8271	0.9570	0.9540
Precision	0.8399	0.8195	0.7927	0.8251	0.9470	0.9536
Recall	0.8840	0.8613	0.8361	0.8324	0.9995	1.0
F-measure	0.8614	0.8399	0.8138	0.8287	0.9725	0.9762

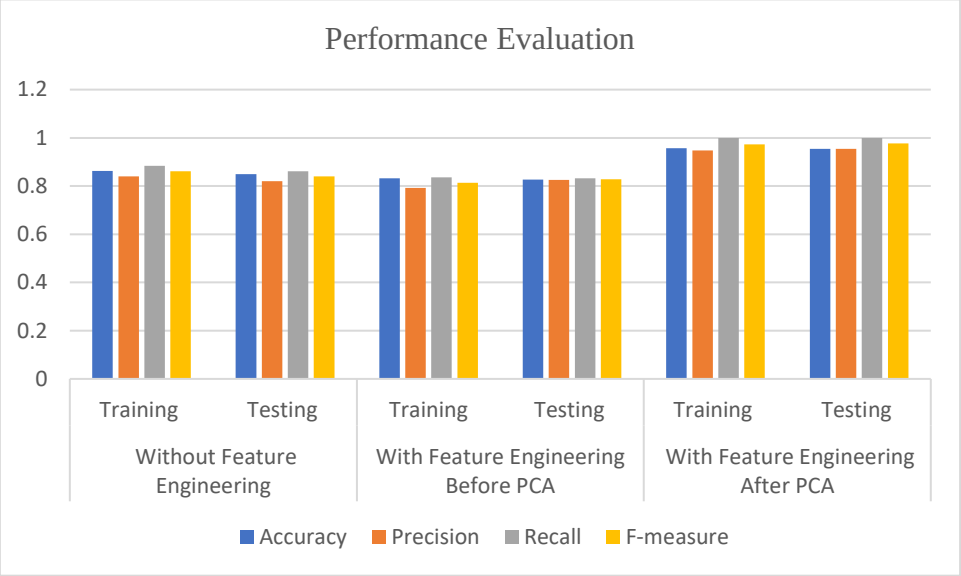


Figure 4.14 Performance Evaluation of the System

CHAPTER 5
CONCLUSION

In this system, the main objective is to classify the data of the banking industry using Data Mining and Machine Learning Technique. Support Vector Machine (SVM) classifier is used to compare the accuracy, precision, recall and F-measure of the predicted system with feature engineering and without feature engineering. Python is used as the programming language to implement the system. The conclusion, advantages, limitations and further extensions are presented in this part.

The analysis of the performance of the Machine Learning Technique, Support Vector Machine (SVM) is described. The bank marketing dataset, that is used in this system, is extracted from Kaggle. The three experimental results: without feature engineering, with feature engineering and with feature engineering (feature engineering of Correlation Matrix and Principal Component Analysis (PCA)) are carried out by using Support Vector Machine (SVM).

According to the experiment, the performance of the system is implemented with Support Vector Machine (SVM). In the first experiment when the training data is used, the accuracy without feature engineering is 86%, the accuracy with feature engineering is 83% and the accuracy with feature engineering of Correlation Matrix and Principal Component Analysis gets 96%. In the second experiment which is used the testing data, the accuracy without feature engineering gets 85%, the accuracy with feature engineering before using PCA is 83% and the accuracy after using PCA is 95%. Therefore, the proposed system (with feature engineering of Correlation Matrix and Principal Component Analysis) shows the best result of the experiments by comparing and contrasting the accuracies of the results.

5.1 Advantages

By using this system, many banks can easily know and get both the deposit customers and non-deposit customers.. This system also saves time for the customers who buy their product such as deposit and can know which features are highly dependent on the customers' subscription to the term deposit. Bank can achieve their organizational objectives to increase the number of subscriptions to deposit.

5.2 Limitations and Further Extensions

The system suggests some prospective future projects that could be undertaken. In this study, a single classification algorithm and a single portion of the dataset are employed, investigated, analyzed, and examined. In future, the bank marketing dataset will be analyzed in another area. In this work, the machine learning approach is predicted and tested using the bank marketing dataset. The dataset for bank marketing can be investigated using the other supplementary approaches and techniques. It is feasible to evaluate different machine learning techniques and compare and contrast the performance of these machine learning methods.

PUBLICATION

[1] Ei Ei Khin, Tin Tin Htar, "Classification of Bank Marketing Data Using Support Vector Machine", University of Computer Studies, Yangon, Myanmar, 2023.

REFERENCES

[1] A. Nachev, T. Teodosiev, "Using Support Vector Machines for Direct Marketing Models", April, 2015.

[2] Alchemer, "What is regression analysis and why should I use it?", June, 8, 2021.

[3] Aggarwal, C. C, "Data Mining: The Textbook". Springer, 2015.

[4] Bishop, C. M, "Pattern Recognition and Machine Learning", Springer, 2006.

[5] Daniel Johnson, "Unsupervised Machine Learning: Algorithms, Types with Example", March, 2023.

[6] Deepak_jain, abhishekolympics, mukuljain1092, avinashrat55252, "Data Processing in Data Mining".

[7] Han Gao*; Pei Shan Fam; Heng Chin Low, "A comparative study between support vector machine and support vector data description in bank telemarketing", 2021.

[8] Henrique Ap. Laureano, "Bank Marketing Dataset: An overview of classification algorithms", Spring Semester, 2018.

[9] Harika Bonthu, "Detecting and Treating Outliers | Treating the odd one out!", April, 14, 2023.

[10] Jamiu Olalekan Oni, "Exploratory analysis of bank marketing campaign using machine learning; logistic regression, support vector machine and k-nearest neighbor", 2020.

[11] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques".

[12] Karim, Md Rezaul, Ali, Ahmed Wasif, Khan, Asifullah, "A novel instance selection technique using a Naïve Bayes classifier for text categorization", Journal of Computational Science, 44, 101164, 2020.

[13] Karim Amzile, Rajaa Amzile, "Using SVM for Smart Direct Marketing (SDM): A case of predicting bank customers interested in the Term Deposits", 2021.

[14] Larry Hardesty, "Explained: Neural Networks", April, 14, 2017.

[15] Mehmet Furkan Akça, Onur Sevli b, "Predicting acceptance of the bank loan offers by using support vector machines", 2022.

[16] María José Lovera**,** "What are Bank Deposits?".

[17] Müller, A. C., & Guido, S, "Introduction to Machine Learning with Python: A Guide for Data Scientists", 2016.

[18] Nachev, T. Teodosiev, " Using Support Vector Machines for Direct Marketing Models", 2015.

[19] Nagesh Singh Chauhan, "Decision Tree Algorithm, Explained", February 9, 2022.

[20] Onel Harrison, "Machine Learning Basics with the K-Nearest Neighbors Algorithm", Sep, 11, 2018.

[21] Prasad, Anurag, Charoenkitkarn, Nopasit, Srihari, Rajeev K., Sankhya, "Random Forests for Regression and Multiclass Classification", The Indian Journal of Statistics, 68(3), 579-601, 2006.

[22] Rohith Gandhi, "Support Vector Machine- Introduction to Machine Learning Algorithm", June, 2018.

[23] Shalev-Shwartz, S., "Understanding Machine Learning: From Theory to Algorithms", Ben-David, S, 2014.

[24] Schmidhuber, J, "Deep learning in neural networks: An overview", Neural Networks, 61, 85-117, 2015.

[25] Tony Yiu, "Understanding Random Forest, June, 12, 2019.

[26] Thomas Wood, "What is Unsupervised Learning".

[27] Zakaria Jaadi, "A step by step explanation of Principal Component Analysis (PCA)", 2022.

[28] Zhang, Shichao, Zhang, "K-Nearest Neighbors Based on Improved Distance Metric and Optimal K Value for Text Classification", Yuanyuan Mathematical Problems in Engineering, 2021.

[29] Janio Martinez Bachmann, "Bank Marketing Dataset" Kaggle, from https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset

[30] "Javat Point- Decision Tree Classification Algorithm", from https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[31] "Javat Point- Regression Analysis in Machine Learning", from https://www.javatpoint.com/regression-analysis-in-machine-learning

[32] "What is data transformation?", from https://www.tibco.com/reference-center/what-is-data-transformation