Arab American University – Palestine

Faculty of Graduate Studies

Loan Risk Prediction Model Using Machine Learning

By

Ramzi Assaf Assaf

Supervisor

Dr. Mustafa Abusalah

This thesis was submitted in partial fulfillment of the requirements for the Master's degree in Data Science and Business Analytics

June 2021

Loan Risk Prediction Model Using Machine Learning

By

Ramzi Assaf Assaf

This thesis was defended successfully on 16-07-2021 and approved by:

Committee members	Signature
1. Supervisor Name: Mustafa Abusalah	
2. Internal Examiner Name: Mohammed Maree	Alaa Armiti
3. External Examiner Name: Alaa Armiti-Juber	

Abstract

With the increase in the number of Credit applications and current expansion in loan and financing services, processing these applications manually has become time-consuming. Customers may apply for loans, such as establishing startups, marriage, buying new or old cars, mortgage, education, new construction projects, etc.

However, only qualified customers can be granted loans to ensure the funds are going to be returned.

Many banks around the world still use manual lending decision-making procedures to make judgements. These decisions are made by bank officers based on personal interaction with loan applicants and customer transactional lending profiles. Bank loan officers make their credit decisions based on information about applicants if they are qualified, such as personal impressions about loan applicants and other information not existent in financial records. Making decisions requires a lot of efforts, time, as well as deep knowledge about customers and their loan plans.

Bank loan officers also use transactional lending practice to choose their quantitative information, such as cash flows, annual reports, and liquidity measures. They combine personal information and quantitative information to set the bank officers' final decision or recommendation.

This study aims to utilize Machine Learning (ML) to create an automated decision support system to identify customers who are eligible to receive loans based on quantitative information.

Supervised Machine Learning methods will be used to judge whether the customer will pay back. These methods will use training datasets from existing customer records to build a classification model; the training dataset is manually judged. The subject of this study is the Bank of Palestine, which provided us with the necessary training data to carry out this research.

The outcome of the research demonstrated promising decisions based on quantitative information; these judgement results accompanied with Bank officers' personal interactions with the client will definitely reduce the time spent in loan decision making with high accuracy similar to the manual workflow implemented by the Bank of Palestine.

Declaration

I certify that this thesis submitted for the degree of Master in Data Science and Business Analytics, is the result of my research, except where otherwise acknowledged, and that this study has not been submitted for a higher degree to any other university or institution.

Name:Ramzi Assaf Assaf

Signature:

Date:16-07-2021

Acknowledgement

My praises and thanks to God, the Almighty, for all the blessings and strength he bestowed upon me to complete this research successfully.

My sincere thanks go to everyone who played a role in my accomplishment, my supervisor, and the master’s committee members.

I would also like to acknowledge Bank of Palestine opportunity to work and learn in their organization.

Contents

List of Figure

List of Tables

List of Abbreviations

Machine Learning	ML
Decision Support System	DSS
Support Vector Machines	SVM
Digital Transformation	DT
Data Science	DS
K-Nearest Neighbor	KNN
Bank of Palestine	BOP
Customer Number	Cust.NM
Date Warehouse	DWH
Customer identification	Cust.Wid
Credit transaction	Cr.trans
Type	Tp
True positive	TP
False-positive	FP
False Negative	FN
True Negative	TN
Area Under the Curve	AUC
Receiver Operating Characteristics	ROC
True Positive Rate	TPR
False Positive Rate	FPR
Log Average Salary	Log AVE Salary
Log Credit Transaction	Log CR
Palestine Monetary Authority	PMA
Employee	EMP
Non-Governmental Organization	NGO
Small and Medium-sized Enterprises	SME

Chapter One: Introduction

1.1 Preface

Financial organizations offer credit in the form of loans to meet the demands of customers. Those customers shall repay the loan with interest by a specific deadline. Some clients fail to pay back their loans on time, raising financial risks on lending institutions, while the applicants themselves suffer greater losses as a result of having to let go of their assets to facilitate payment (Soni P. M. & Varghese P., 2019). As such, it is important that only qualified customers will be granted loans, ensuring that the funds will be returned. After customers apply for a loan, the bank loan officer validates the customer's eligibility by identifying her/his personal and quantitative information, in addition to subjective judgements through social interactions between the officer and the customer. The difficulties stem from financial data for customer, the strength of risk models and length of the processes. The objective of this research is to utilize Machine Learning (ML) methods to build a prediction model to assist in judging if the customer is eligible for a loan or not based on previous judgements of similar cases.

The implemented Loan Risk Prediction Model in this research will be compared to the current manual loan decision-making procedure at the Bank of Palestine.

1.2 Lending Decision Making in Bank of Palestine

This section outlines the lending process for small and medium products programs covering all aspects of the credit lending decision-making workflow.

Branch Loan Officer:

The loan officer in the branch meets with the customer, ensures that all the required documents are submitted according to the program, and asks the customer to bring any missing documents to process his application, then he issues an inquiry to the Palestinian Monetary Authority. After that, the loan officer initiates the “Lending decision-making” workflow by submitting the lending case to the branch manager, accompanied with his recommendation, after ascertaining the customer’s right to borrow according to the internal systems.

Branch Manager:

The branch manager reviews the documents and sends them through the “Lending decision making” workflow to the Financing Department, accompanied by his recommendation.

The Lending decision-making at the branch typically takes one working day in case all documents are ready.

The Financing Department:

After the branch sends the application form, accompanied with the branch recommendation, whether acceptance or rejection, to the Financing Department, the Financing Department Credit Officer will:

- Validate the integrity of the application and required documents. If anything is missing, she/he will communicate with the branch.
- Verifies that the loan meets the eligibility criteria set by the Financing Department.

- If the application does not meet the eligibility criteria, she/he has to return the application to the Head of the Financing Department to take the necessary actions, and if the application meets the criteria, she/he has to continue the loan process described in the next steps.
- The officer verifies the applicant's financial records from the Palestinian Monetary Authority; the officer will inquire records about the main applicant, loan shareholders if any, and guarantors.
- If the feedback from the Palestinian Monetary Authority is positive, meaning that whether the main applicant, loan shareholders, if any, or guarantors have deficit records, then the officer will reject the application. Otherwise, the process continues to the next step.
- Check the bank financial records and credit records of the main applicant, loan shareholders, if any, or guarantors for other deposits or loans, etc. This is called Credit Scoring; please check section 2.4 Credit Scoring.
- Verifies that the applicant guarantees liabilities according to the bank's credit policy manual.
- Prepare cash flow analysis, financial statements, and bank account statements within the bank and other banks, if any for the main applicant, loan shareholders, if any, or guarantors.
- Provides her/his recommendation to the head of the Financing Department.
- The head of the Financing Department receives the report from the Credit Officer and will:

- Approve, reject, or suggest amendments to the application. Possible adjustments might include increasing/decreasing the loan amount, asking for additional guarantees, providing personal guarantees, etc.
- If there are amendments, she/he updates the application according to the bank procedure and updates the “Lending decision making” workflow for the top management to be able to make the final recommendation.
- After making updates to the workflow, the status of the application will be set to “Approved” or “Rejected”.

The processing time in the Financing Department typically takes up to two working days.

The Applicant Branch:

- Once the decision is made, the bank branch which initiated the workflow is notified, and the branch loan officer will:
- Communicate with the client to inform her/him about the outcome of the application.
- If the application is rejected, the status of the request will be changed to “Closed”.
- If the application is approved, the officer will finalize the legal terms with the customer and grant him the loan.
- The decision made by the bank is based on the bank officer’s interaction with the loan applicant and the transactional lending profile of the applicant. This process does not consider the history of similar financial cases of other existing customers, which could assist in the lending decision-making process.

1.2 Objectives of the Study

A Decision Support System (DSS) is an automated software that assists in decision making, assessments, judgements, and actions, mainly in financial institutes. Decision Support System in financial institutes improves internal control, enhances long term profitability, advance the quality of decision making and boosts response to change. A DSS examines large volumes of data and analyses detailed knowledge that could be used in problem-solving and decision-making (Ignatius et al., 2016). In this study, a "Loans Risk Prediction" model is developed using supervised machine learning methods, which will act as a DSS to determine whether the applicant worth's credit or not. (Leo M., Sharma S. & Maddulety K., 2019).

The supervised machine learning classification model will use training dataset from actual customer records, which are manually labeled. This dataset is provided by the Bank of Palestine for the purpose of this research.

Feature engineering, data cleaning, and resampling are applied to the dataset; then different supervised machine learning algorithms are utilized. Then the outcome is evaluated and compared; also a voting classifier, which combines multiple classifiers, is used to come up with the best optimal solution.

The research result will attempt to answer whether the use of a supervised machine learning model will improve or at least show similar judgement results compared to the baseline manual judgement workflow. This model will use Dataset provided by the Bank of Palestine and evaluated based on accurate customer data considering customers' privacy and security.

1.3 Significance of the Study

The proposed Loans Risk Prediction Model tackles a classical decision-making process in which an individual's loan eligibility is determined by forecasting possible future earnings. (gnatius J., Hatami-Marbini A., Rahman A., Dhamotharan L. & Khoshnevis P., 2018). Palestinian banks still measure credit risks for loans manually, which depends on the bank officer or committee depending on the type and amount of the loan to assess the submitted loan application (Alfarra, et. al, 2016).

With the latest trend in Digital Transformation (DT), which has been driven by the Covid-19 Pandemic lockdown, many companies have begun to shift employees to work from home and develop their digital services. Furthermore, with the massive increase in data, it has become essential to apply Data Science (DS) and Machine Learning (ML) in a variety of industries and fields. Especially for banks and financial sectors, which need to apply financial transactions on big data of customers and credits. This research will contribute to DT by utilizing DS and ML to develop an automated DSS to aid in credit decision-making. This will enhance decision-making, and in the future, reduce the bank officers' time and resources.

In Supervised Learning, a training dataset that contains both features and labels is used to train the classification model. The label may consist of continued data such as price or discrete data such as Boolean or a few categories. In this study, labels contain decisions made to whether accept or reject the loan. The classifier uses this dataset to predict the class label based on feature characteristics. (Goyal A. & Kaur R., 2016). Classification using supervised machine learning is the approach of the proposed model.

In this thesis, we apply various supervised machine learning methods for estimating the loan status prediction to investigate whether the ML model can distinguish between customers who are eligible for a loan and those who are not.. Furthermore. We will apply evaluation metrics to compare between different classification approaches.

In this context, we apply feature engineering extract relevant features from data and enhance the training dataset, i.e., resampling datasets and cleaning, to create an automatic DSS that can replace or at least assist in the manual workflow currently being used by several banks. The study may also recommend combining various techniques to accomplish this objective.

1.4 Organization of the Thesis

This thesis is organized into four chapters: Chapter one contains the introduction part which discusses the research problem, research objectives, significance, and limitations of the research. The second chapter discusses the related literature about using ML in loan risk prediction. Chapter three focuses on the research design and methodology to evaluate the loan risk prediction problem and explores applying machine learning methods for estimating the loan status prediction. Finally, chapter four covers the conclusions of the findings and presents the recommendations.

Chapter Two: Literature Review

This chapter summarizes the available literature in loan risk prediction models that uses data science and machine learning techniques to assist in making loan decisions. It will review theories of loan risk prediction models in different research studies.

2.1 Theoretical Review

The theoretical review aims to contribute a better understanding of the loan risk prediction process and the relevant theory behind the chosen classification methods.

2.2 The Concepts of Loans

A loan is the lending of money to other persons or organizations through banks or other financial institutes for a specific amount, for an agreed period time. In general, loans are defined as the main component of the total assets of each bank.

"A bank loan is a contractual agreement between a lender and a borrower in which the lender typically agrees to grant a sum of money to a borrower who intends to repay the lender in installment payments over a given period time" (Berhano Y., 2017).

2.3 Problems of Loan Default

Default occurs when a debtor has not met his or her legal obligations according to the debt contract. For example, a debtor may have missed a payment or breached a loan condition in the debt contract. Failure to repay a loan is referred to as a default.

Furthermore, Pearson and Greeff defined default as a risk threshold that defines a borrower's repayment history in which he or she has missed at least three payments in 24 months. This is a period in time and a behavioral indicator in which there is a proven increase in the chance that the borrower would eventually default by stopping all payments. The definition complied with international standards and was required because consistent analysis needed a single definition (Addae-Korankye A., 2014).

2.4 Credit Scoring

Credit scoring is the most common method for determining a person's creditworthiness. It is calculated by converting loan parameters into numeric values using statistical methods as a guide to make loan decisions (Abdou H. A. & Pointon J., 2011).

Companies proceed by credit-awarding laws and regulations developed by experts. Non-experts used credit standards to help businesses with the loan decision-making processes. Since banks offer such a broad range of services, the number of credit applications has risen dramatically. As a result of the increased demand in the banking sector, more credit applications have been submitted.

With such a large variety of financial applications, applying manual methods to each of the evaluated requests (such as credit cards, home loans, and small business loans) one at a time has become extremely time consuming. Banks began using credit rating in credit applications evaluation because both credit seekers and banks wanted to test applications quickly.

Loan risk forecasting is one of the most critical challenges facing bankers worldwide, as banks focus on loan status prediction techniques for clients that show the difference between developing good and poor credit and actions that increase clients' profits. (Alfarra A. N., Hui X. F. & Ahmed J., 2016).

The credit score system uses past credit history to review loan applications for acceptance or rejection. Furthermore, due to the short amount of time available to complete the loan application. A credit score system can also be used as a method for evaluating customer performance.

There are many statistical methods for assessing credit risk, but since they focus on fixed features and numerical assumptions such as gender, age, marital status, dependents, having a telephone, educational level and occupation, they have drawbacks in modeling complex economic systems. When machine learning techniques are used and compared to these statistical techniques and methods, the results showed that machine learning techniques are more effective (Abdou H. A. & Pointon J., 2011).

The selection of the variables varies from study to another based on the nature of the data, and also on what cultural or economic variables may affect the quality of the model.

2.5 Machine Learning and Default Predictions

The global financial crisis of 2007–2008 caused mainly by loan defaults was the most severe since the Great Depression of the 1930s. Some of the world's largest financial institutions collapsed or were nationalized. The crisis of 2007–2008 began in the United States with the bursting of a housing bubble and the growth of mortgage defaults due to less creditworthy borrowers (CM Reinhart & KS Rogoff, 2009). It was an epic budgetary and financial collapse that cost many ordinary people their jobs, their life savings and their homes. (Helleiner E., 2011).

Risk management in banks has become much more critical since the global economic crisis, and there has been a continuous emphasis on how risks are identified, measured, published, and handled.

Machine learning demonstrated a growing impact in enterprise applications, with several solutions still in place and many others are investigated.

The existing and emerging banking and risk management issues have been the subject of extensive studies in academia and industry.

The use of data mining techniques in banking risk management, such as credit and market risk, operational and liquidity risk, has been analyzed; however, the present rate of industry emphasis on both risk management and machine learning doesn’t really appear to be comparable. There are still many areas of bank risk management that might gain from research into how machine learning methods can be used to solve financial problems (Leo M., Sharma S. & Maddulety K., 2019).

Machine learning can be applied to different domains including but not limited to: science, engineering, economics, politics, finance, education, agriculture, climate and environment.

It employs techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, and information science. It was illustrated as a method that would be used to solve real-world problems, particularly those requiring data analysis and implementation.

Developing a machine learning application is a process that is even more iterative and exploratory than software development. Machine learning is used to solve problems that are complex for humans to solve.

It is very challenging to give clear and precise specifications on the implementation stage when met with the necessity of extracting useful information from big data and the resulting difficulty of patterns to be examined.

This challenge can be addressed by machine learning, which “endows algorithms” with the ability to “learn and adapt.” Machine learning can be used with these problems because Machine learning systems continuously learn and evolve (Goyal A. & Kaur R., 2016).

2.5.1 Machine Learning Categories

Machine learning techniques follow different approaches on how machine learning may be used. Machine learning techniques are decomposed mainly into three categories: Supervised Learning, Unsupervised Learning and Reinforcement Learning.

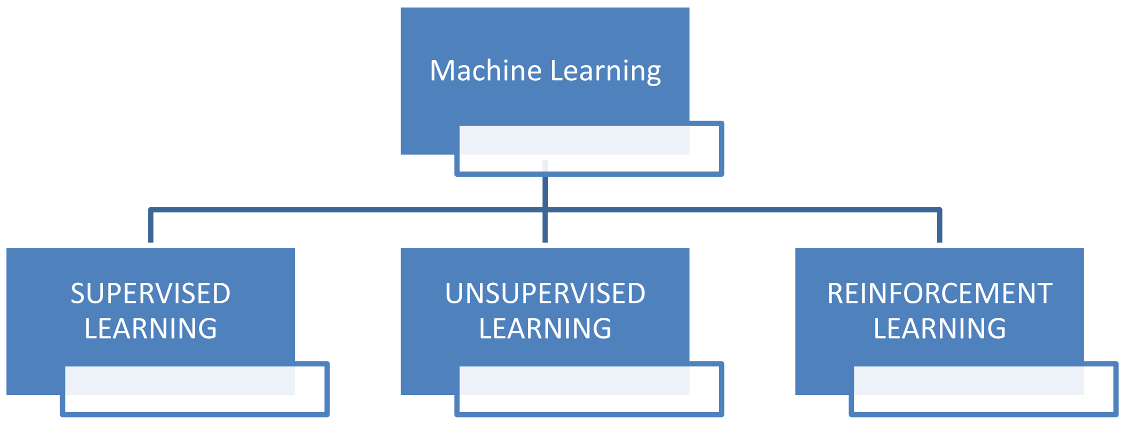

Figure 2. 1: Machine Learning Categories

Supervised learning: is widely used machine learning technique; It is named supervised because the learning happens under the supervision of label observation variables. The aim is to construct a classifier that can predict the class label of an object using a set of features.

Supervised learning is decomposed into classification and regression. Given such input variables, the classification job is predicting the class value. Regression is the method of estimating the value of a continuously changing variable while input variables are given (Goyal A. & Kaur R., 2016).

Unsupervised Learning: when dataset attributes have no label, we look for correlations between the objects. This type can be used to display the pattern in the data. This requires tasks such as dimensionality reduction, clustering, association, etc (Goyal A. & Kaur R., 2016).

Reinforcement Learning: is directly inspired by how people learn from data in their daily lives. It includes a trial-and-error algorithm that improves upon itself and learns from different scenarios. Positive outcomes are rewarded or 'reinforced,' while negative outcomes are discouraged or 'punished'. (Muih, et. al, 2016)

2.6 Data Mining

"Data Mining is a multi-stage procedure in which raw data from various sources is processed and knowledge extracted" (Soni P. M. & Varghese P., 2019). Figure 2. 2 illustrates different steps to extract information from unstructured data.

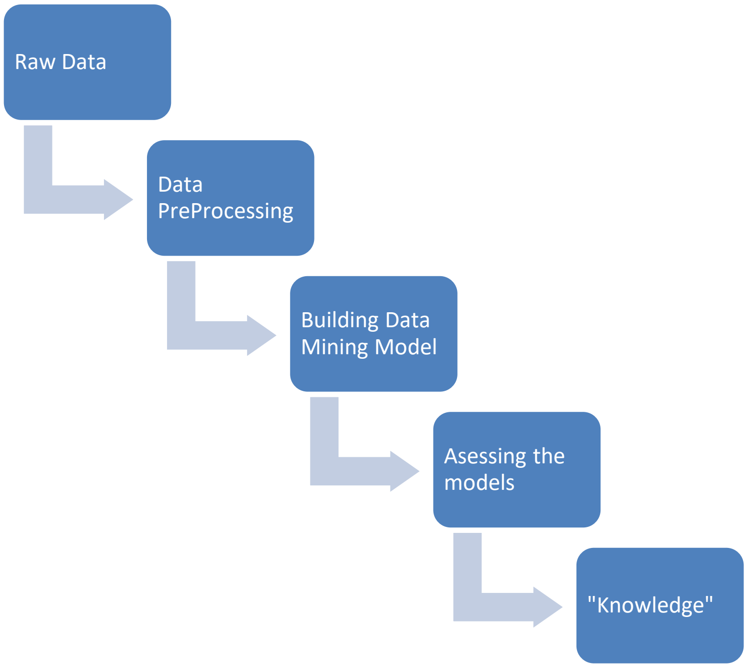

Figure 2. 2: Steps in Knowledge Extraction

Classification, clustering, association mining, prediction, and regression are examples of data mining techniques. The most widely used data mining method is classification, which uses a collection of pre-classified samples to create a model that classify a large volume of data (Sivasree M. S. & Rekha Sunny T., 2015).

The classification method is highly utilized in credit risk applications. Training data is used to construct the classifier model which used for learning, organize the collected data into correct classes, and test data is used to assess the performance of the classifier.

2.7 Data Mining in Banking

Analysis and transformation of data into meaningful knowledge has become a task beyond human capabilities because of the huge increase in data that the banking industry interacts with. Data mining techniques can solve this problem by identifying associations and correlations between various variables hidden in business information models and user profiles stored in financial databases.

Data mining techniques can be used to forecast customer behavior and improve model accuracy by predicting which customers are more likely to accept future deals, which customers are more likely to default on a loan, and how to keep customer engagement more profitable.

Data mining is focused on analyzing current and historical data to assist banks in predicting creditworthy customers and preventing fraud. Data mining is used effectively by the financial and banking sectors in various areas, including marketing, industry dynamics, stocks, risk management, loans, and customer relationship management (Sivasree M. S. & Rekha Sunny T., 2015).

2.8 Credit Scoring System

Credit rating models are designed to give good or poor credit to loan customers or to forecast bad lenders. As a result, a credit scoring model's goal is to distinguish between loan customers. Sex, age, marital status, dependent family members, phone, level of education, job, current address, and using the credit card are all factors usually used in the credit scoring system. In some cases, the variables list has included personal information about the client, such as bank account, income, and other details (Abdou H. A. & Pointon J., 2011).

2.9 Statistical Techniques for Credit Scoring System

Statistical techniques used to derive the mathematical relationships between dependent variables (factors whose value is influenced by other factors) and independent variables which refer to that variable that measures the effect of the independent variable(s).

Credit scoring models are constructed using a variety of statistical methods,

such as weight of evidence measure, correlation analysis, regression analysis, discriminant analysis, probability analysis, logistic regression, linear programming and non-parametric techniques such as support vector machines, decision trees, neural networks and k-nearest-neighbor.

Most of these statistical models can be employed to create an efficient and effective credit scoring system that can be used for prediction reasons.

Credit analysts, researchers, and computer software use techniques together with the activity of proof weight, regression analysis, logistical regression, vector support machines, k-nearest-neighbor random forest naïve bays, and decision trees programming to construct credit scoring models.

Statisticians use various types of tests, two of them are parametric and non-parametric tests, in parametric we assume underlying statistical distribution in the data where in nom-parametric we don’t really in any distribution.

In this research we depend on parametric test which is more powerful for the sample size used in experiments.

The aim of this research is to investigate traditional statics approaches to see whether they will be useful to build the DSS. (Abdou H. A. & Pointon J., 2011)

2.10 Classification Methods

This section discusses the classification methods used in this study in detail.

2.10.1 Logistic Regression

Logistic regression is the most used supervised learning algorithm that can be used for classification tasks, whereas linear regression is used for regression, or to predict continuous values.

The goal of logistic regression is to categorize observations based on their modeled maximum likelihood of belonging to a specific class. With a given input, the logistic function can be used to calculate the likelihood of a customer being in the default class. The difference between logistic regression and linear regression models is that the output variable in logistic regression is binary (a 0/1 result), while in linear regression the value is continues. This distinction can be seen in selecting a parametric model as well as the assumptions used in logistic and linear regression (Abdou H. A. & Pointon J., 2011).

Logistic regression predicts the probability of occurrence of a binary event utilizing a logit function.

Linear Regression Equation:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots. + \beta nXn \qquad (1)$$

Where y is the dependent variable and X1, X2 ..., and Xn are explanatory variables.

Sigmoid Function:

$$\text{P} = 1/(1 + e^{-y}) \qquad (2)$$

Apply Sigmoid function on linear regression:

$$\text{P} = 1/(1 + e^{-(\beta 0+\beta 1X1+\beta 2X2+\ldots\ldots+\beta nXn)}) \tag{3}$$

The sigmoid function, also known as the logistic function, produces 'S' shaped curve that can be used to transfer any real-valued number to a value between 0 and 1. if the y values approach positive infinity, then the probability P tends to 1. While as the y values approach negative infinity, then the probability P tends to 0.

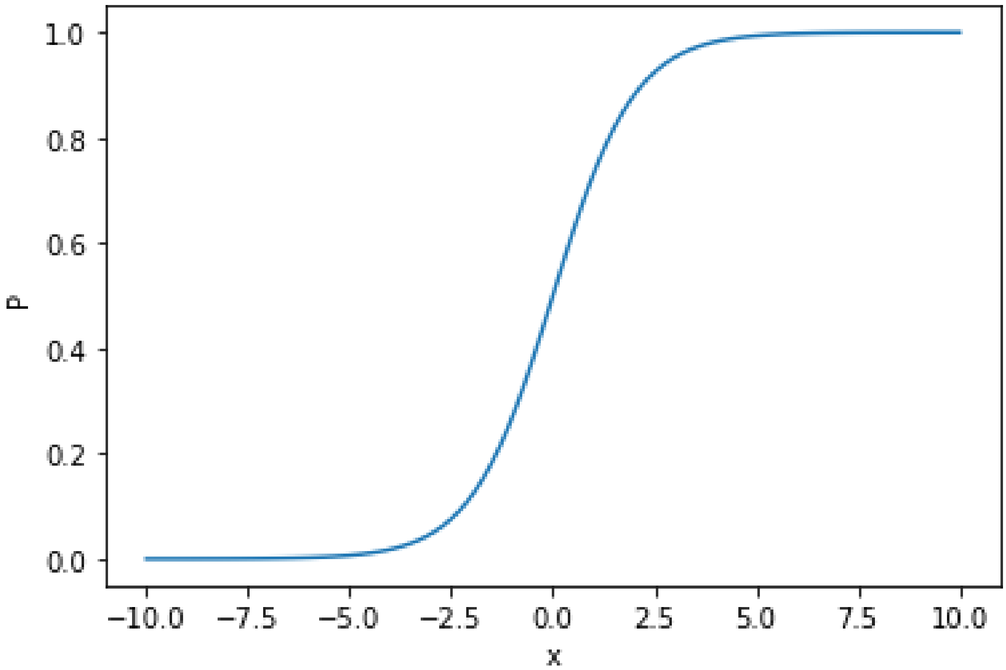

Figure 2. 3 Sigmoid Function

We have three types of logistic regression:

1. Binary logistic regression: The target variable has only two possible outcomes
2. Multinomial Logistic Regression: when the target variable contains three or more nominal categories.
3. Ordinal Logistic Regression is used when the target variable has three or more ordinal categories.

Advantages of logistic regression:

1. Logistic regression is easier to implement.
2. Classifies unknown records quickly with good accuracy for simple data.
3. It easily extends for multiple classes.

Disadvantages of logistic regression:

1. Overfitting issue when the number of observations is less than the number of features.
2. It can only be used as a discrete function.
3. Non-linear problem cannot be solved.

2.10.2 Decision Tree

The Decision Tree is an algorithm that transforms a branch object's elements into a target value in tree leaves. In this learning process, each inner node or non-leaf node with an input function is labeled (A. Salihu & V. Shehu, 2020). The Decision Tree belongs to supervised family methods.

One common approach is to use an entropy measure to calculate the information gain (IG)

IG = E(ParentNode) − AverageE(ChildNodes)

where Entropy (E) is defined as:

$$\mathrm{E(S)} = -\sum_{i=1}^{n} \mathrm{p}_i \log_2 \mathrm{p}_i \tag{4}$$

Where S is dataset and 'Pi' is simply the frequentist probability of an element/class 'i' in our data.

Advantages of Decision Tree:

1. Decision Tree is easier to implement.
2. Data does not require scaling or normalizing.
3. Easy to explain.

Disadvantages of Decision Tree:

1. Decision Tree often needs higher time to train the model.
2. A small change in data can cause a significant difference in the structure.

2.10.3 K-Nearest Neighbor (KNN)

KNN is a method for non-parametric classification or regression. It consists of both a positive and a negative training set. KNN is also known as the lazy algorithm; there is no pre-compilation training phase in KNN ("lazy" learning); you just store the training data, and the data is loading during the compilation. The mean of the most related K cases is used to predict the outcome of KNN regression. When KNN is being used for classification, it can be measured as the class with the highest frequency in most cases. (A. Salihu & V. Shehu, 2020).

The k-nearest neighbor classifier fundamentally relies on a distance metric. The better that metric reflects label similarity, the better the classification will be. The most common distance methods are Euclidean, Manhattan, and Minkowski. Figure 2.4 explains the Knn classifier.

KNN stands for the number of closest neighbors. The number of neighbors (K) is the most important factor. If the number of classes is 2, K is usually an odd number.

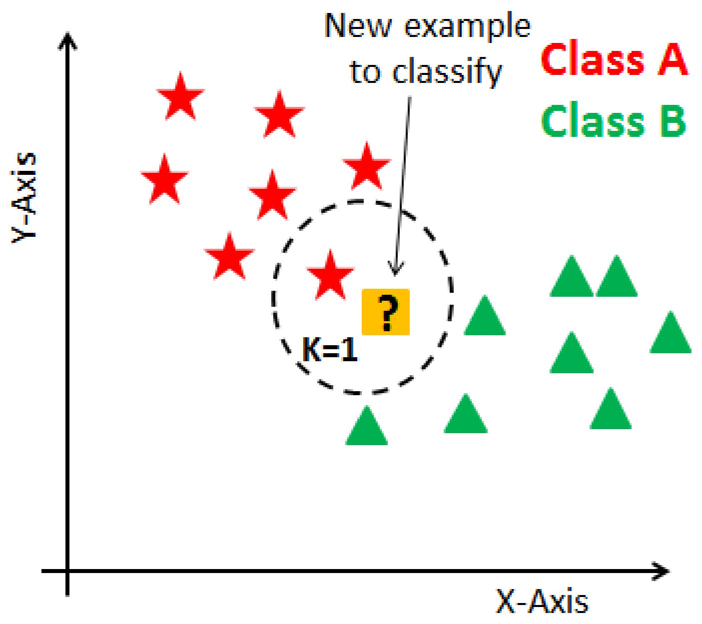

Figure 2. 4 Avinash Navlani .(2018).Number of K neighbors in KNN. https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

The time complexity of KNN is linear O(n×m), where n is the number of training examples and m is the number of dimensions in the training set.

Advantages of K-Nearest Neighbor:

1. KNN is very easy to implement; we need to know K and distance function.
2. New training data has no effect on accuracy.

Disadvantages of K-Nearest Neighbor:

1. Large datasets will affect the performance; it will be degraded.
2. Curse of dimensionality, High dimension of data will affect the complexity and difficulty to calculate the distance between neighbor points.
3. Noise data, missing value, and the outlier will affect the algorithm sensitivity.

2.10.4 Support Vector Machine

Support Vector Machine (SVM) is another supervised learning method that uses a learning algorithm to analyze data for classification and regression. In a high or infinite-dimensional space, SVM creates a hyperplane or set of hyper-planes that can be used for classification, regression, and other tasks

The decision function f(x) is given by:

$$f(x)=sgn(w,\, x+b) \qquad (5)$$

Where w is a weight vector, x is input vector, b is bias and sgn is signed of function.

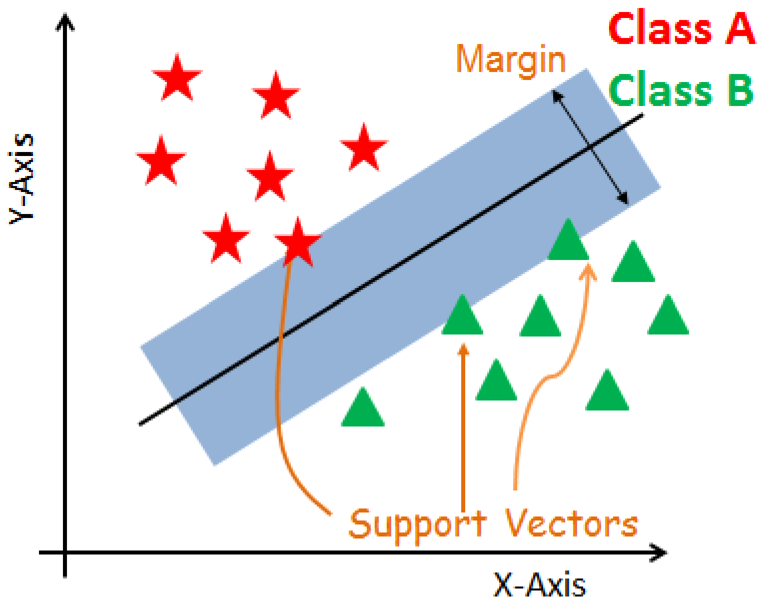

Figure 2. 5 Avinash Navlani .(2019). Support Vector Machines. https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

SVM is a classification algorithm that analyses and classifies data. This model transforms data using a technique known as a kernel function, and then it finds the optimal results based on data transformations. SVMs are classifiers defined by separating the feature space with a hyperplane. (Goyal A. & Kaur R., 2016).

SVM can handle two or three dimensions, it may be visualized as a plane separating the data, but it becomes more difficult to visualize with more dimensions.

Advantages of SVM:

1. SVM works well when there is a clear margin between classes.
2. SVM is relatively memory efficient because it is constructed as a linear combination of all the training data.

Disadvantages of SVM:

1. Noise data will affect the algorithm sensitivity.
2. SVM algorithm is not suitable for large data sets because the training complexity of SVM is highly dependent on the size of data set.

2.10.5 Random Forest Classifier

An ensemble of highly randomized decision trees is known as a Random Forest. The Random Forest algorithm assembles various option trees of different samples of the dataset; for portioning the data at each node of the tree, random subgroups of the usable factors are assumed. In the classification step, a simple dominant vote is then used for prediction (and the average for regression). Random Forest is often used for large datasets and, in particular, many data variables (hundreds or even a thousand factors), especially to deal with the overfitting problem in the dataset since it repeatedly subsets the usable factors.

Thus, Random Forest models are differentiated from Decision Trees in two aspects:

1. Several trees are generated and trained on different samples.

2. The variables chosen in each node are not chosen based on discriminatory power but randomly.

In generating the Random Forest model, at least two parameters must be specified, the number of trees to grow and the number of variables to consider at each node (Marte A., 2019).

The main advantage of Random Forest Classifier is that it is less vulnerable to overfitting than the other algorithms.

Disadvantages of Random Forest Classifier:

1. Small change in data will cause a large change in the structure.
2. Random Forest algorithm computations are more complex compared to different algorithms.

2.10.6 Naïve Bayes Classifiers

In statistics, naive Bayes classifiers are a form of “probabilistic classifier” based on Bayes’ theorem and strong (naive) independence assumptions between features. The naıve Bayes model assigns a probability to a class variable Cn given a vector of independent variables (x1, ..., xn) with the assumption that the x's are mutually independent. They are one of the most basic Bayesian network models, but when combined with kernel density estimation, they can achieve higher levels of accuracy.

It is based on the Bayes theorem, which states that:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad (6)$$

Where P(A|B) and P(B|A) is the conditional probabilities of occurrence of an event A given that event B is true and vice versa. A, P(A), P(A|B), and P(B|A) is called a proposition, prior probability, posterior probability, and likelihood, respectively (Assiri A. S., Nazir S. & Velastin S. A, 2020).

Advantages of Naive Bayes:

1. Very fast algorithm and can easily predict the class
2. Better than other models to perform the classification with less training data

Disadvantages of Naive Bayes:

1. Low accuracy for prediction values
2. You should select independent features in the experiment which is too hard

2.10.7 AdaBoost Classifier

An AdaBoost stands for adaptive boosting is a meta-estimator algorithm that starts by applying a classification model on the data set, then fits multiple copies of the algorithm on the same dataset but adjusts the weights of wrongly classified instances so that subsequent classifiers concentrate more on complex cases (A. Salihu & V. Shehu, 2020). The models are sequentially arranged in the ensemble. This means that at each step, we try to boost our weak learners (base model) based on the mistakes of our previous models, so together, they are one robust ensemble model.

Advantages of Adaboost:

1. AdaBoost is not prone to overfitting, which is good
2. AdaBoost can be used to improve the accuracy

Disadvantages of Adaboost:

1. Extremely sensitive to Noisy data and outliers
2. Features variable should be independent without correlation in the experiment which is not always achievable.

2.10.8 Voting Classifier

A Voting Classifier is an algorithm of machine learning that learns from an ensemble of variables and predicts an output (class) based on the highest likelihood of the output being the chosen class.

It simply adds up the results of each classifier supplied into the Voting Classifier and predicts the output class based on the most votes. Instead of developing separate specialized models and determining their performance, we develop a single model that trains on various models and predicts output based on each output class's cumulative majority of votes.

The Voting Classifier supports two types of voting, hard voting, and soft voting (Assiri A. S., Nazir S. & Velastin S. A., 2020).

In hard voting, the voting classifier counts each class instance and then assigns to a test instance a class that was voted by the majority of the classifiers.

In majority voting (hard voting), the class label y is predicted via majority (plurality) voting of each classifier C:

$$y = mode\{\, \mathrm{C1(x)}, \mathrm{C2(x)}, .., \mathrm{Cn(x)}\} \qquad (7)$$

y is the prediction class which has highest probability averaged by each classifier, single quantity representing the majority of votes.

In soft voting, we give different weights to each classifier to bias the outcome based on the classifier's accuracy. Predictions are chosen by summing the prediction probabilities from the individual models, then we take the average of probabilities then use it to classify the test instance.

The probability of class labels assigned by the classifier C to input x is defined as:

Average of Probabilities Voting:

$$y = \text{AVERAGE}\,\{\, \text{C1}(\text{x}), \text{C2}(\text{x}), \ldots\ldots, \text{Cn}(\text{x}))\} \qquad (8)$$

Product of Probabilities Voting:

$$y = \text{PROD}\,\{\, \text{C1}(\text{x}), \text{C2}(\text{x}), \ldots\ldots, \text{Cn}(\text{x})\,\} \qquad (9)$$

Minimum of Probabilities Voting:

$$y = \text{MIN}\,\{\text{C1}(\text{x}), \text{C2}(\text{x}), \ldots\ldots, \text{Cn}(\text{x})\,\} \qquad (10)$$

Maximum of Probabilities Voting:

$$y = \text{MAX}\,\{\, \text{C1}(\text{x}), \text{C2}(\text{x}), \ldots\ldots, \text{Cn}(\text{x})\,\} \qquad (11)$$

2.11 Classification in Loan Defaults

Many studies in the financial and banking sectors have used machine learning and data mining techniques. This section present previous work that have been carried out to assess the risk of loans.

2.11.1 Review of Previous Works

In (Sivasree M. S. & Rekha Sunny T., 2015), a prediction model for banks was introduced. The model assists in predicting the most trustworthy customers who have qualified for a loan. Decision Trees were used to predict the attributes that are essential in determining credibility. A prototype of the proposed model was developed for companies to assist the decision about whether to accept or reject a customer's loan application. The dataset used in this work is considered to be small (4520) records and they were reviewed and reduced to 3271 records with 17 features. The accuracy of the model reached was 75%.

In (Hamid A. J. & Ahmed T. M., 2016), data mining was used to propose a model for predicting loan risk in the banking by the consumers to a good or poor loan by looking at consumer actions and past credit repayment history. The proposed model was created using information from the banking industry to predict the status of loans, and the following algorithms were used: j48, naiveBayes, and bayesNet. The proposed model was implemented using the Weka tool. J48 was chosen as the best algorithm of instances were 78.3% of instances were correctly classified instances when compared with the other algorithms.

In (Goyal A. & Kaur R., 2016), the authors examined the accuracy of several machine learning models using R language and tested them to determine the best model for forecasting an institution's financial status. The results showed that the Tree model form Genetic Algorithm - this algorithm is usually based on theory of natural selection and survival of fittest - is the best algorithm for predicting customer finance, as 80.7% on average correctly classified after researcher run experiment five times on the same dataset.

In (Gahlaut A. & Singh P. K., July 2017), the authors investigated if data mining methods can be used to forecast and identify a customer's credit rating as good or as bad in order to avoid potential risks of lending to people who can't pay back their loans. The authors used a bank's historical data for predictive modeling that banks may use to improve the overall performance of their credit system. The authors presented a loan's credibility forecasting mechanism that can assist financial institutions in making the correct decision to accept or reject a customer's proposal for a loan dependent on family history, occupation, financial situation, marital status, and a variety of other factors. Rattle is used to execute the model. There are 1000 cases and 18 variables in this dataset; dataset is divided into two parts, with 700 records in the Training Dataset and 300 records in the Validation Dataset.

The Random Forest algorithm was found to be the efficient technique for risky credit classification 0.7968 ROC value for validation on the dataset. There is a chance of incorporating the most relevant features that assess the credit riskiness of clients into a bank's risk automated framework using Random Forest and SVM.

In (Chen, Ya-Qi, Jianjun Zhang & Wing WY Ng, 2018), the authors focused on addressing the imbalance data issue to improve loan default prediction. The method used in this research is a hybrid under-sampling to validate the output of their process; the authors have used real loan defaults dataset from a Chinese P2P business. The results of the experiments showed that the proposed method produces better results. The dataset contains 25504 applicants, 1309 applicants to whom credit should be accepted, and 24195 instances correspond to applicants to whom credit should be refused

The authors demonstrated that the Diversified sensitivity under-sampling method of clustering to maintain information distribution is efficient; however, the clustering algorithms used in this process are k-means, which are inefficient. As a result, a comparison of the effectiveness of the Diversified Sensitivity Under sampling method (DSUS) with clustering methods is needed.

In (Chen, Ya-Qi, Jianjun Zhang & Wing WY Ng, 2018), the authors applied different classification techniques to choose the high accuracy algorithm in default prediction. Different machine learning methods were studied in this research. The method of oversampling is referred to as Synthetic minority Oversampling technique (SMOTE) was introduced to address the variable's response imbalance between classes. The results revealed that XGBoost without SMOTE achieved the best result in terms of the F1 Score evaluation metric. The big limitation of this study is applying only F1 Score as evaluation metrics; there is a need to use different evaluation metrics such as Precision, Recall, ROC Curve, and AUC as we did in our research.

In (Nabende P. & Senfum S., 2019), the authors used a digital loan application set of data from the Ugandan financial institution to apply supervised machine learning methods to estimate the loan status predictive model. The data set was created by manually converting 1000 loan application documents from hard copy to digital form. Since the dataset included some unique characteristics and obtained features, a specific comprehensive application, and evaluation of supervised learning methods were applied. The random forest method generated the most accurate classification results; where the model's accuracy was 94% (Nabende P. & Senfum S., 2019).

(Soni P. M. & Varghese P., 2019) study aims to introduce a new methodology that can help the banking sector to develop. Wrapper model and fisher score are used in a hybrid feature selection method to forecast a customer's ability to pay back a loan in a cost-effective manner. This research aims to show that the current hybrid model outperforms the standard random forest algorithm in terms of accuracy. Ten thousand loan transaction records were used in this work. Manual data preprocessing was used to delete some of the fields. The experiments were carried out with the help of software such as Weka, R, and Python. The results showed that the proposed algorithm, compared to other classifiers, produces improved classification accuracy in determining a customer's potential. Wrapper-Fisher algorithm with random forest always produces better accuracy; it gave 99% accuracy using the original dataset.

Chapter Three: Research Methodology

This chapter will discuss the proposed implementation method, which starts with data collection, preprocessing and cleaning, feature engineering, and then implementing classification methods and evaluating their performance as shown in Figure 3. 1. The research subject to this study is Quantitative, where quantitative research is characterized as a systematic analysis of phenomena through the collection of quantifiable data and the application of statistical, mathematical, and computational methods. Quantitative research gathers information from current customers, the results of which can be represented in the form of numeric figures.

We have also used Descriptive analysis, Correlational research, and Experimental research. Descriptive analysis aims to explain the current state of a variable that has been described. The aim of correlational research is to assess the strength of the relationship between two or more variables. Experimental analysis, also known as true testing, employs the empirical method to determine the cause-and-effect relationship between a set of variables in a sample.

In this study, we used primary data that has been collected from the Bank of Palestine. We will describe the data in more detail in the data collection section.

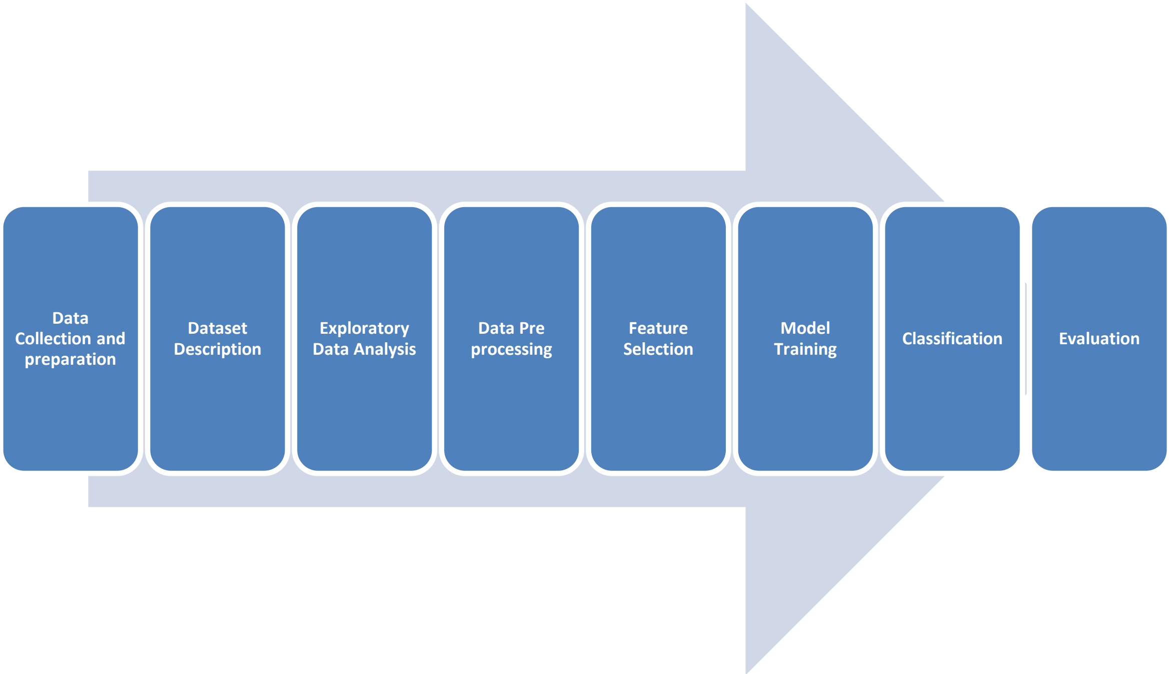

Figure 3. 1:The proposed methodology steps

3.1 Data Collection and Data Preparation

The Bank of Palestine (BoP) was founded in 1960 as a financial bank to improve payment systems in Palestine, funding multiple projects and responding to the finance and business requirements of various socio-economic sectors. BoP is considered one of the biggest national banks in Palestine with the largest organization of branches, workplaces, and Computerized Teller Machines. Through its qualified staff, BOP serves over 850,000 customers, contributes to the building and development of the economy, and remains side by side with the latest technologies.

In this study, we have collected 307,113 loan records queried from different database tables. Those records consist of information about customers from all branches, including retail and SMEs (Small and medium-sized enterprises).

The dataset was preprocessed and cleaned to create a cleaned dataset, which then will be used to train and test the ML model. The preprocessing, cleaning and feature engineering steps are described in the next sections of this chapter.

3.2 Dataset Description

3.2.1 Customer Attributes

Bank of Palestine officers identified 25 attributes related to the loan applicants, which are important for the behavior of loan applicants in the future. Table 3.1 describes each attribute and it's datatype.

Table 3. 1:Customer Attributes

Attribute	Description	Data type
Branch	Customer Original Branch	Nominal
Cust. Nm	Customer Unique ID	Numeric
Cust. Wid	Customer DWH ID	Nominal
Cust. Class	Customer Business Class	Nominal
Salary Flag	Gets Salary within the last 6 months	Nominal
Time Dep. Balance	Time Deposit Balance	Nominal
CR. Trans.	Credit Transactions	Numeric
Nm. Of Employees	Number of Employees within the Org.	Numeric
Public Sector Employee	Public Sector Flag	Nominal
Merchant	Has POS Flag	Nominal
Last salary DT.	Last salary Dates	Nominal
Avg Salary USD	Avg Salary evaluated in USD within the last 6 months	Numeric
Facility ID	Facility WID	Nominal
Facility TP	Facility Type (cards, Loans … etc)	Nominal
Facility Sub TP	Facility SubType (Overdraft, Credit Cards, Easy life, Loans … etc)	Nominal
Currency	Currency Type	Nominal
Facility Class	Special Assets Class	Nominal

Days Past Due	Number Days that Customer Delayed the Payments	Numeric
Outstanding Balance USD	Current Balance for Facility in USD	Numeric
Due Balance USD	Due Balance USD	Numeric
Collaterals USD	Collaterals Value	Numeric
Cleared Cheques	Cleared Cheques	Numeric
Rejected Cheques	Rejected Cheques	Numeric
Gender	Gender of Customer	Nominal
Age	Age of Customer	Numeric
location	location	Nominal
Date Account Open	Date Account Open	Nominal

3.2.2 Customer Class

The Customer Class aims to classify customers based on their nature and size. It is a classification based upon the basic customer information such as whether the customer is a legal entity or a person and the size of the customer's activities both as a business/person or the size of the customer's activity with the bank.

The Customer Class aims to segment customer based on their value to the bank (simply how much profit they generate for the Bank) or Customer Market Segment, which segments the customer based on their stage in the Lifecycle (individuals) or Sector (companies), as well as the customer's banking, needs as shown in the table below.

Table 3. 2:Customer Class

Class	Subclass
General Retail	Public Sector Employees
General Retail	Private & NP Sector Employee
General Retail	BOP Group Employees
General Retail	General Retail Other
Premier	Premier Public Sector EMP
Premier	Premier Private Sector EMP
Premier	Premier BOP Group EMP
Premier	Premier Other
Small Business	Small Business
SME	Medium SME

Corporate	Corporates & NGO's
Corporate	Large Corporations & NGO's
Public_Sector	Ministry of Finance
Public_Sector	Central-Government
Public_Sector	Local-Government
Public_Sector	Other Public Sector
Financial Institutions	PMA
Financial Institutions	Foreign Central Banks
Financial Institutions	Banks Operating in Palestine
Financial Institutions	Foreign & Correspondent Banks
Financial Institutions	Other Fin Institution / Settlement A/C
Unclassified	Unclassified

In this research, Data is grouped based on four main types based on customer class, General Retail, Premier, Small business, and Small and medium-sized enterprises (SME).

3.2.3 Facility Attributes

Bank of Palestine is a main economic organization devoted to selling retail and SME. It has four primary facilities: Credit Cards, Loans, Overdrafts & Overdrawn, and Indirect credit Facilities like (letters of credit), and every facility is divided into multiple subtypes, as shown in the table below.

Table 3. 3:Facility Type

Facility Type	Facility Subtype
Credit Cards	Credit Cards
Credit Cards	Easy Life
Loans	Car Loans
Loans	Restructured Loans
Loans	Suspended Loans
Overdrafts & Overdrawn	Public Sector Overdrafts

Overdrafts & Overdrawn	Project Finance Overdrafts
Overdrafts & Overdrawn	Indirect Facility Overdrafts
Loans	Other Business Loans
Loans	Commercial Deals
Loans	Other Student Loans
Loans	Micro Finance Loans
Loans	Employment Fund Loans
Indirect Facilities	Outward Letters of Guarantee
Loans	Alquds Uni Student Loans
Loans	LGF III
Loans	Disability Loans
Overdrafts & Overdrawn	Revolving Overdrafts
Overdrafts & Overdrawn	Suspended Overdrafts
Loans	Mortgages
Loans	Ariz II
Loans	Palestinian Manufacturing Association Loans
Loans	Murabaha
Indirect Facilities	Outward Letters of Credit
Overdrafts & Overdrawn	Palpay Overdrafts
Overdrafts & Overdrawn	Temporary Business Overdraft
Overdrafts & Overdrawn	Overdrawn Accounts
Loans	Consumer Loans
Loans	LGF
Loans	Equipment Loans
Loans	Sunref Loans
Loans	Istidameh Loans
Loans	Arab Fund Loans
Overdrafts & Overdrawn	Corporate/Business Overdraft
Overdrafts & Overdrawn	Overdrawn Loan Settlement
Loans	Housing Loans
Loans	Falastiniya
Loans	SIDA
Loans	EPCGF
Loans	Italian Grant Loans
Loans	Other Green Loans

Loans	Ariz I
Loans	French Grant Loans
Loans	Other Program Loans
Loans	Corporate Loans
Indirect Facilities	Other Off-balance Sheet

3.2.4 Repayment Rating Categories

There are three categories for loan repayment commitment at the Bank of Palestine as follows: performing, non-performing, and watch list. A performing loan is a one in which a customer has decided or able to make all of her/his payments on time. A loan becomes non-performing once the bank considers that the customer is unlikely to repay or when she/he is ninety days late on a payment. When the customer fails to repay the loan installment on time but did not yet exceed the 90 days, he is added to the watch list.

3.3 Exploratory Analysis

The ratio of the performing customers to non-performing ones is imbalanced 95% to 5%, and this will definitely bias our ML predictors; we have avoided this bias by applying a resampling technique.

3.3.1 Data Type

In machine learning during the feature engineering phase, it is important to convert categorical features in datasets to numerical as machine learning algorithms can only read numerical values. Our dataset is decomposed into:

1. Numerical: any feature that consists of numbers other than dates.

2. Categorical: any other features that contain objects, text, string, date, or any other data.
3. Null: features that have an empty value.

The provided data by the Bank of Palestine contains categorical variables that consist of date, which will be converted to a numerical value by calculating the duration in months. For example, the date account open variable is calculated in months from the open day of the account. Also, the last salary date variable is converted to a numerical value using the same method, by calculating the past last month in which the customer received a salary.

3.3.2 Dataset Overview

We use the heat map below to present the overview of the dataset and discover if any attribute in the dataset is empty or has a value. If any attribute has empty values, this means the dataset contains missing data, and we have to deal with these missing data. We will explain how to deal with missing data in the next section.

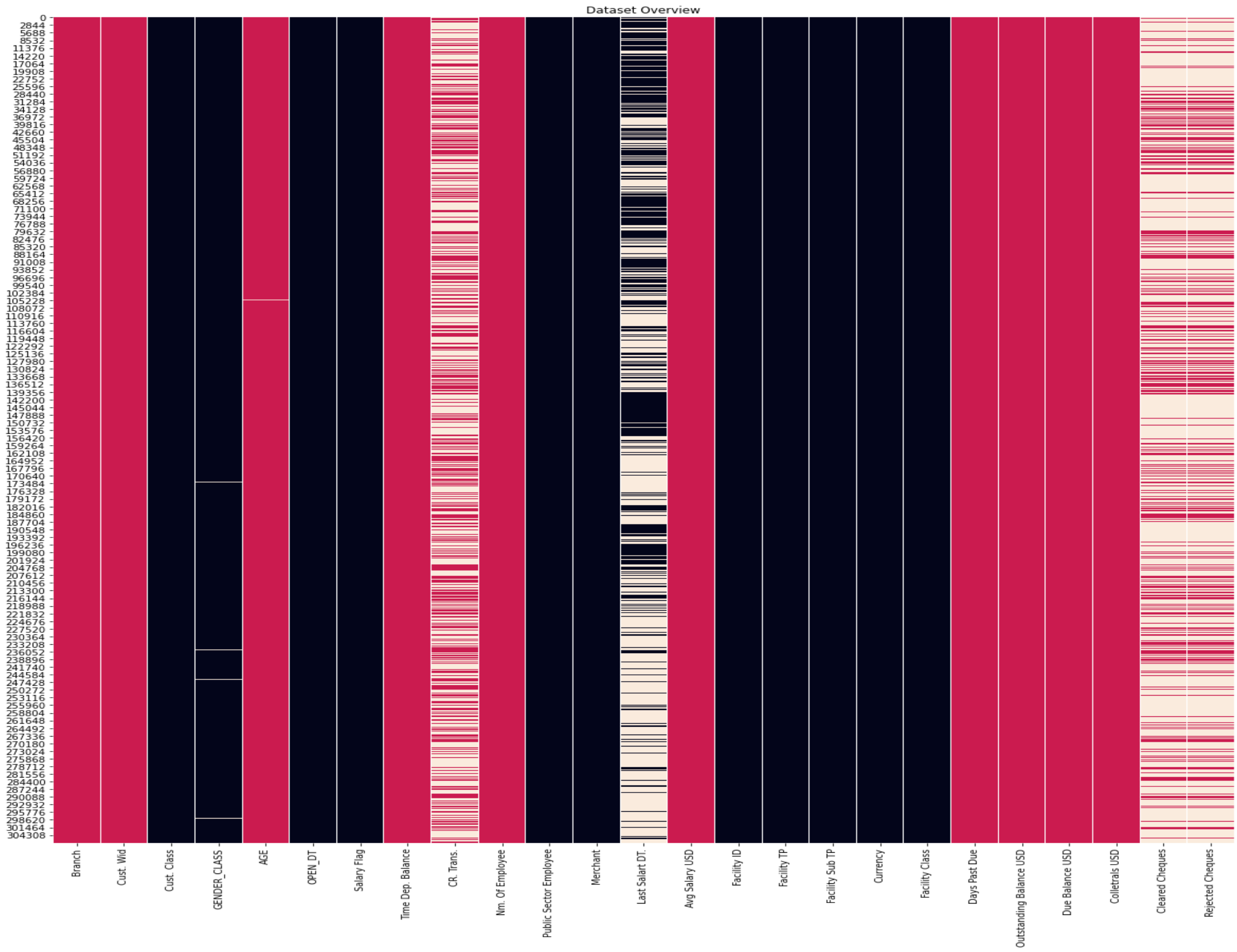

Figure 3. 2:Dataset Overview Before Preprocessing steps. Red color means numerical variable while black color means categorical variable.

3.4 Preprocessing Steps

Data preprocessing is an essential step in any Machine Learning model, as the quality of data that can be derived from it highly affects the ability of the model to learn; therefore, it is extremely important that we preprocess our data before feeding it into our model. The preprocessing started by removing any duplicate records if any.

3.4.1 Missing Values

There are three known approaches to handle missing values. First, when the feature is not so important, we may drop the whole column.

Or we can calculate the mean value along with the feature and annotate it to the missing value. This method is called Imputation.

The last method is to use imputation but also to add a new column(s) referring that the value was initially been null, to assist the machine learning model.

When dealing with categorical attributes, replacing missing data with the highest value in the table is a safe choice. In this research, categorical attributes were checked one by one, and each of the attributes replaced the missing data with the highest value that occurred in the same column to avoid removing it or dropping the whole row (Badr W., 2019).

On the other hand, we checked numerical values in the dataset that we have, in case if we encounter this issue in the future. We have missing values in numerical columns; we will remove the data by dropping it from the dataset.

In this research data is grouped into four main types based on customer class, General R etail, Premier, Small business, and Small and medium-sized enterprises (SME).

3.4.2 Transform Data

A categorical variable is the one that can only have one of a few, typically fixed, values, such as Sex (Male/Female) OR Single (Yes/No), and so on.

Machines, on the other hand, do not understand characters; everything on the machine is a number, thus we must transform these values to numbers.

Any categorical information must be encoded. The Label-Encoder (every specific label is converted to an integer) and the One-Hot-Encoder (every label is converted to a binary vector) are the two most popular encoders.

We don’t really need The Label-Encoder in our experiment because it’s only good for data with ordinality. We have used a binary vector (One-Hot-Encoder) which is more suitable.

Log transformation was also used to achieve linearity; this can help us obtain further insights into the dataset that may not be obvious at first by making highly skewed distributions less skewed.

We applied log transformation to two numerical. The first one was that the average salary in US Dollars, which was log-transformed in a new column named log average salary in dollar. The other numerical is credit transaction, which was log-transformed to a new column named log credit transaction Fig 3.6 and Fig 3.7 present distribution for credit transaction before and after applied Log transformed less skewed.

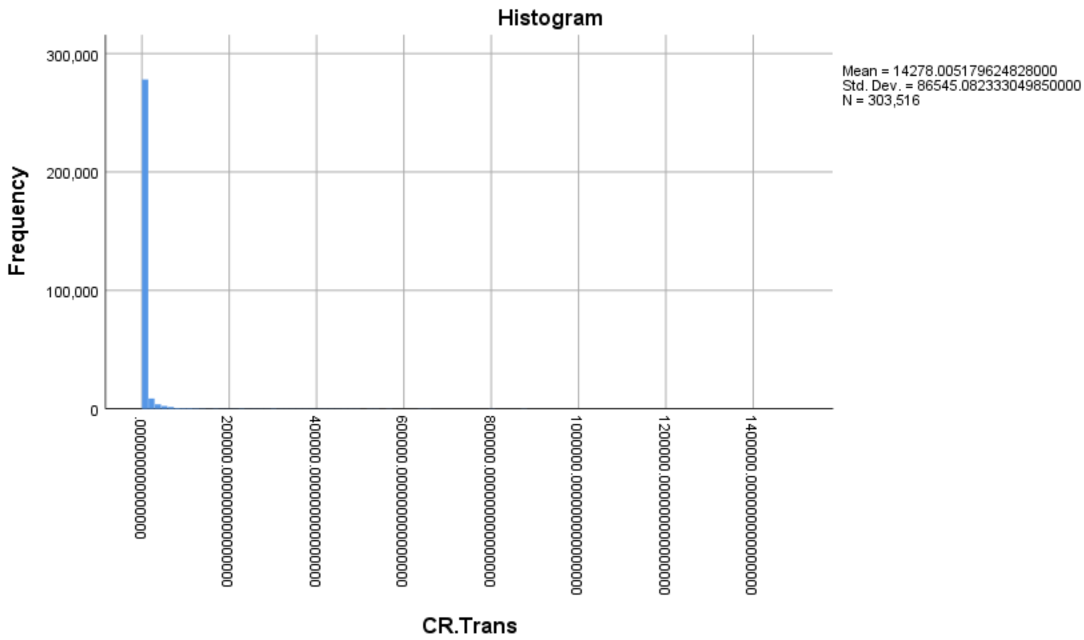

Figure 3.6: Credit Transaction Distribution

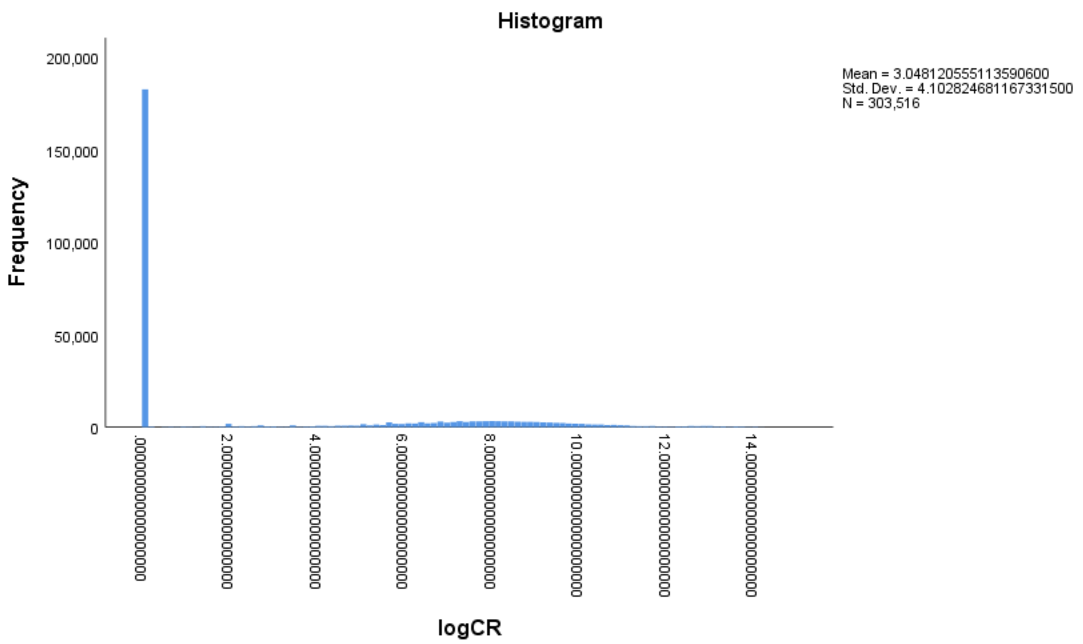

Figure 3.7: Log Credit Transaction Distribution

3.4.3 Outlier Detection

An outlier is an observation that lies in an abnormal distance from other values in a random sample in a dataset. Outliers are detected using Standard Deviation σ; and mean μ then we compute $\mu + 3\sigma$. Any values greater than $\mu + 3\sigma$ are considered outliers. Returning to the mean, compute $\mu - 3\sigma$. Outliers are then any values less than $\mu - 3\sigma$.

A credit transaction is a numeric variable that calculates the volume of transactions denominated in US dollars for a customer. Each class of customer must be in the same scope with respect to the volume of transactions accordingly, and which defined by bank before.

We used outlier equation for each group class depending on credit transaction for each customer on each group class. We had dropped 3,597 records from all group classes. Table 3.4 describes in detail the number of records dropped in each class.

Table 3. 4: Outlier Detection

Customer Class	# of record before	# of record after	# of record dropped
General Retail	265171	262798	2373
Small Business	28431	27483	948
Premier	1463	1453	10
SME	12048	11782	266
ALL	307113	303516	3,597

3.5 Feature Selection

Feature selection is the process of selecting a subset of relevant attributes to build a machine learning model. It makes the model easier to interpret and reduces overfitting. To reduce overfitting, we removed the highly correlated features by applying Pearson correlation matrix.

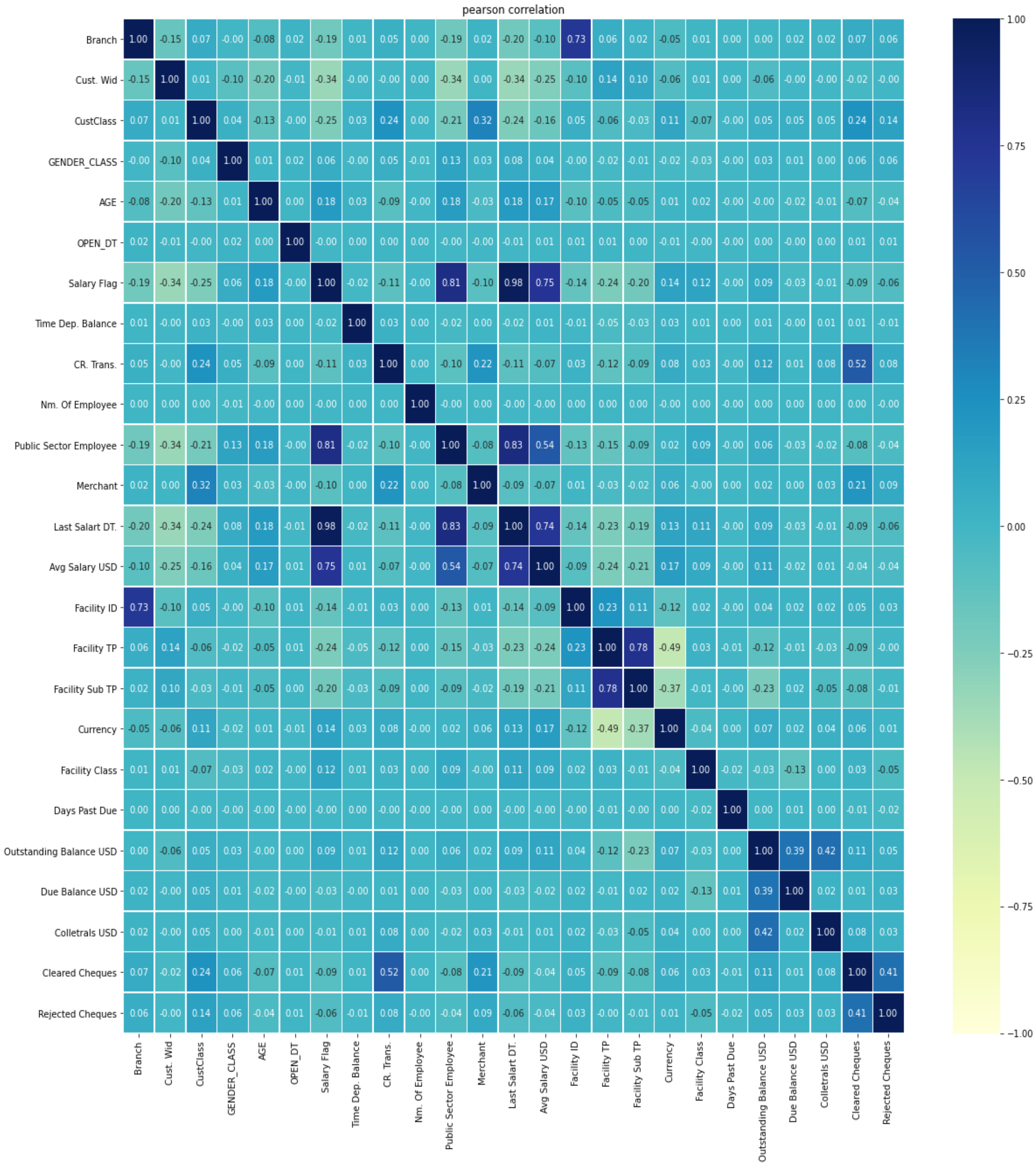

Figure 3. 3:Pearson correlation

The correlation coefficient is a metric for determining how strong a relationship exists between two variables. Pearson’s correlation is a linear regression correlation coefficient (Sedgwick P., 2012). By using Pearson’s correlation coefficient, we identify pairs of features that correlate higher than 0.8 or lower than -0.8 from the matrix; there are two columns to remove out of 25 columns are Public Sector Employee has correlation 0.81 with salary flag and last salary date has correlation 0.98 with salary flag.

3.6 Classification Model Design

Machine learning models should be able to give accurate predictions to create real value for a given business or use case. After the dataset was cleaned and tuned by applying feature engineering, features reductions, outlier detection, and linearly transforming few features to achieve a better sense of data, we have applied the ML predictors. Experimental details are discussed in Chapter Four: Experiments and Result.

3.7 Evaluation

Evaluating the performance is an integral component of building an effective machine learning model. There are several evaluation metrics that we have used to estimate the accuracy of different models on the test dataset. Evaluation methods include Confusion matrix, Jaccard index, Precision, Recall and AUC-ROC curve.

Different evaluation metrics are used for different kinds of problems; an important aspect of evaluation metrics is their capability to distinguish between model results.

3.7.1 Confusion Matrix

The confusion matrix is used to explain how well a classification model performs on collecting test data with known true values.

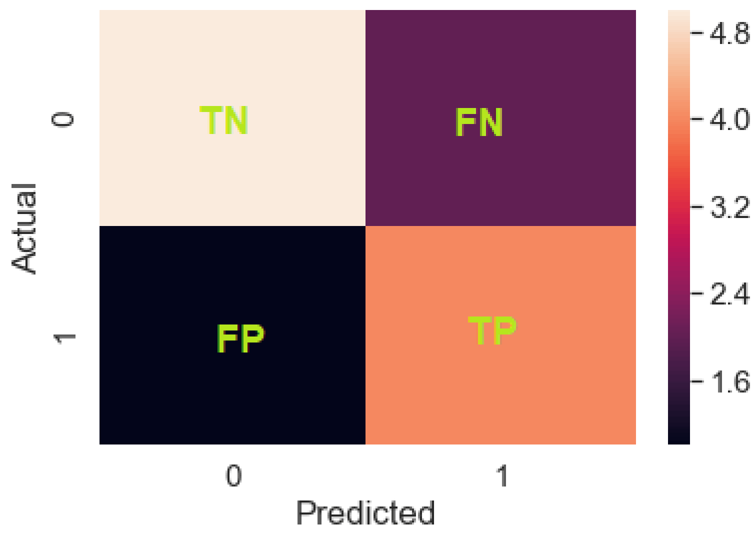

Figure 3. 4:Confusion Matrix for Binary Classification

where:

(TP) True positive: “it demonstrates that a model correctly predicted positive cases as positive cases”.

(FP) False positive: “it demonstrates that a model predicted negative cases as positive incorrectly”.

(FN) False Negative: “it demonstrates that a model incorrectly predicted positive cases as negative cases”.

(TN) True Negative: “it demonstrates that a model correctly predicted negative cases as negative cases”.

3.7.2 Jaccard Index

“The Jaccard index measures similarities between sets, the volume of the intersection, divided by the two-set union size”. We computed the Jaccard Index in the same manner as performed by (Oduor George, 2019). The Jaccard index can be calculated as:

$$j(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (12)$$

Where A is set 1 and B is set 2.

3.7.3 Precision

A confusion matrix can be used to calculate the Precision and the recall as follows;

Given that a class label has been expected, precision is a measure of accuracy. We computed the Precision in the same manner as performed by (Abusalah, 2008).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (13)$$

3.7.4 Recall

Recall is the percentage of all positive cases that the classifier accurately identified as positive, we computed recall in the same manner as performed by (Abusalah, 2008).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (14)$$

3.7.5 F1 Score

The F1 score is calculated using the accuracy and recall metrics. F1 is computed as the Precision weighted average as well as the recall scores. The F1 score ranges from one to zero, with one being the best and zero being the worst. It’s a great metric to demonstrate that a classifier has strong recall and precision, We computed F1 in the same manner as performed by (Oduor George, 2019).

$$F1\ Score = 2\left(\frac{Precision\ \times\ Recall}{Precsion\ +\ Recall}\right) \quad (14)$$

3.7.6 Accuracy

Accuracy refers to the overall accuracy or the percentage of total cases correctly identified by the classification model (Granström D. & Abrahamsson J., 2019).

$$\text{accuarcy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \tag{16}$$

3.7.7 Log Loss

Log loss is the most important classification metric based on probabilities; a lower log loss value means better predictions. The next equation can be used to measure log loss with each row in the datasets (Oduor George, 2019).

$$L = -(y \times \log(\text{pi}) + (1 - y) \times \log(1 - \text{pi}) \tag{17}$$

where pi is the probability of class 1, and (1-pi) is the probability of class 0.

The equation simply calculates the distance between each expected likelihood and the real label. The optimal value for log loss is the sum of the log losses from all of the rows.

$$Lloss = -\frac{1}{n}\sum_{i=1}^{N}(y \times \log(\text{pi}) + (1 - y) \times \log(1 - \text{pi})) \tag{18}$$

3.7.8 ROC Curve and AUC

AUC - ROC curve is an output metric for classification tasks at different threshold conditions; AUC represents the degree or metric of separability, while ROC represents a probability curve (Soni P. M. & Varghese P., 2019). It indicates how well the model can differentiate between classes.

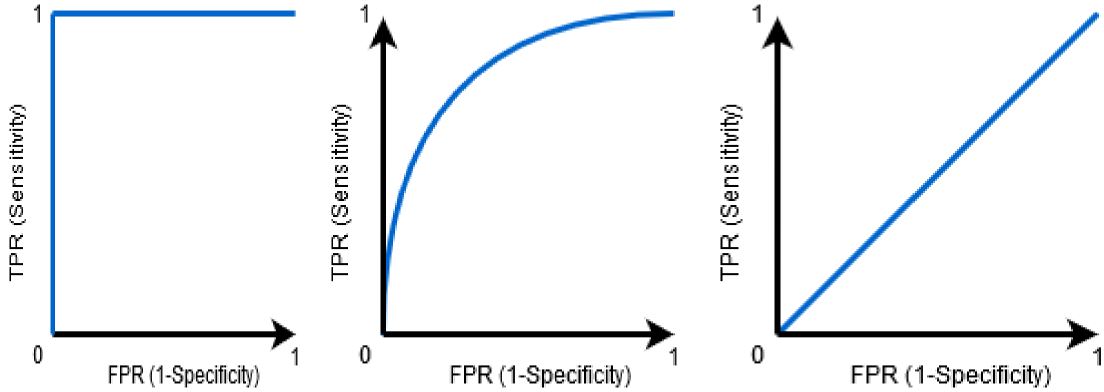

Figure 3:5 Syed Sadat Nazrul .(2018). ROC Curve. https://towardsdatascience.com/receiver-operating-characteristic-curves-demystified-in-python-bd531a4364d0

The ROC graph is constructed with (TPR) True Positive Rate on the y-axis and (FPR) False Positive Rate on the x-axis. The AUC indicates how well the model predicts 0s as 0s and 1s as 1s.

TPR/recall tells us what proportion of the positive class got correctly classified.

FPR tells us what proportion of the negative class got incorrectly classified by the classifier.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \tag{15}$$

If AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly.

When 0.5<AUC<1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values.

If AUC=0.5, then the classifier is not able to distinguish between Positive and Negative class points.

3.7.9 Feature Importance

The feature importance attribute of the model can be used to obtain the feature importance of each feature in the dataset.

Feature importance assigns a value to each of the data's features; the greater the score, the more essential or relevant the feature is to your output variable.

The fitted attribute provides feature importance attributes; importance is computed as the mean and standard deviation of accumulation of the impurity decrease within each tree.

3.7.10 Optimal Value of K in KNN

K value is the count of the closest neighbors. Distances between test and trained label locations should be calculated.

The challenge is how to select optimal K value; we should derive a plot between error rate and K value, and we choose the minimum error rate to pick optimal K using Elbow method.

The dataset was first classified as a whole dataset and then divided into four datasets based on the classification of the customers to calculate optimal K-related minimum error for each class.

1. Optimal K for the Whole Dataset without classification

   Minimum error: 0.03144936651990295 at K = 13

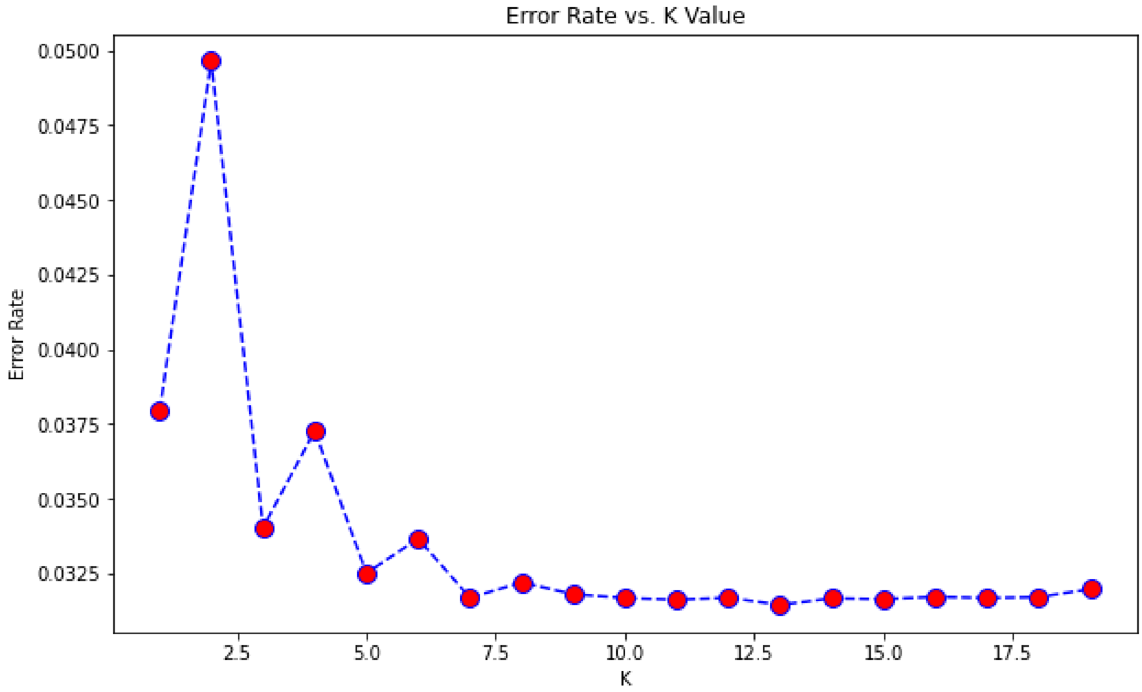

Figure 4. 1 Optimal K for the Whole Dataset

2. Optimal K for sub dataset General Retail class

Minimum error: 0.021332041879987086 at K = 9 GM

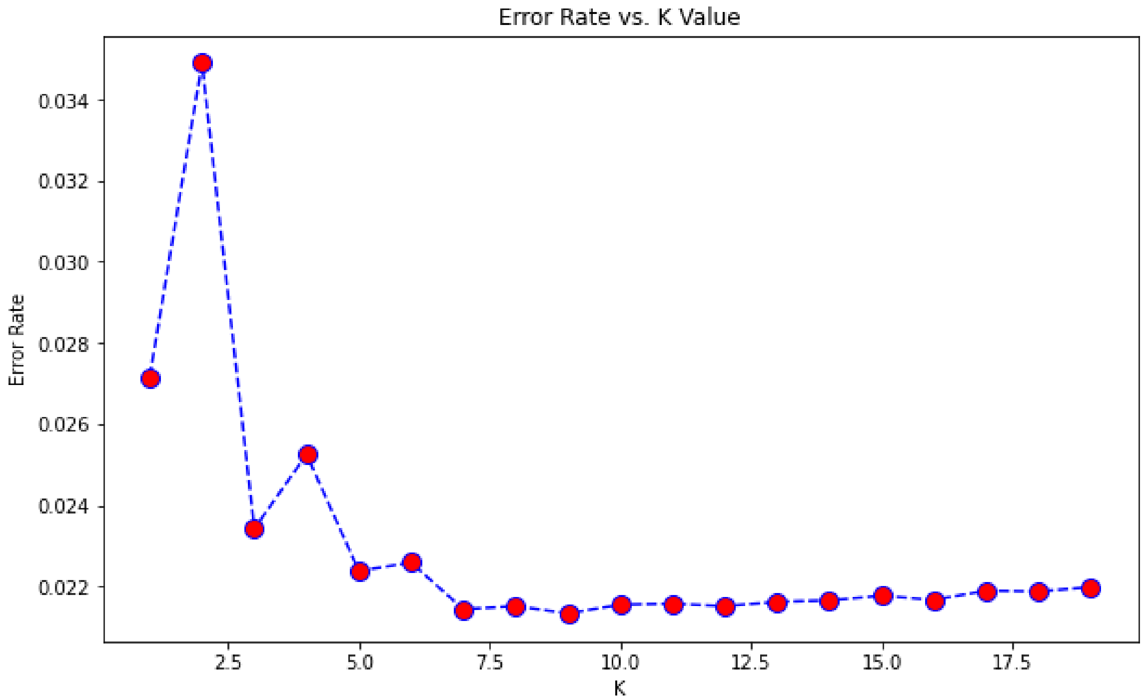

Figure 4. 2 Optimal K for sub dataset General Retail

3. Optimal K for sub dataset Premier class

Minimum error: 0.014583333333333334 at K = 7

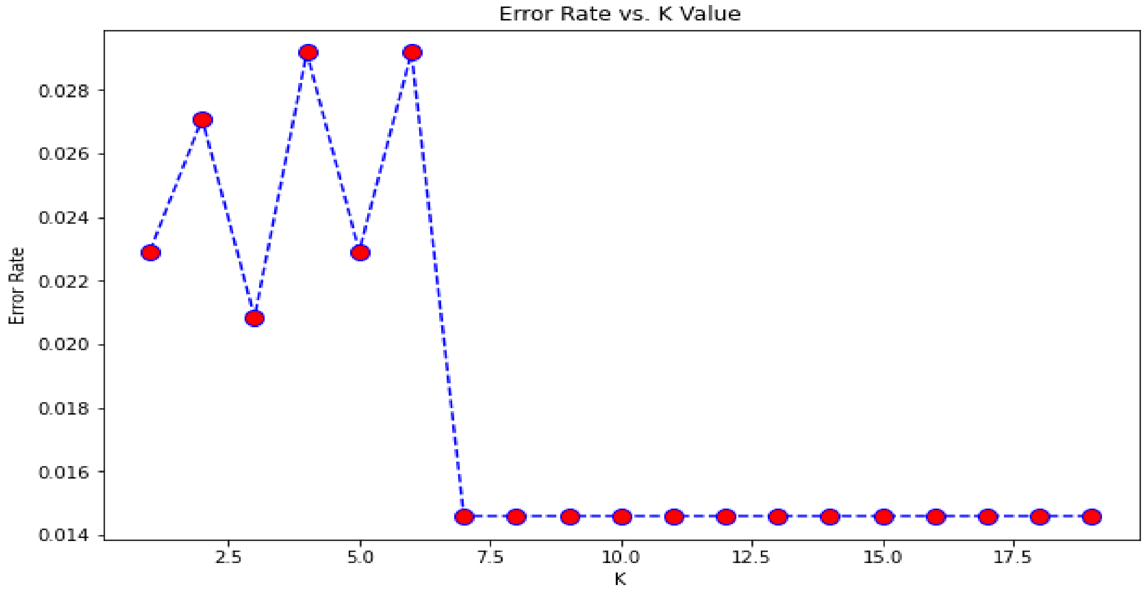

Figure 4. 3 Optimal K for sub dataset Premier class

4. Optimal K for sub dataset SME class

Minimum error: 0.08999742864489586 at K = 1

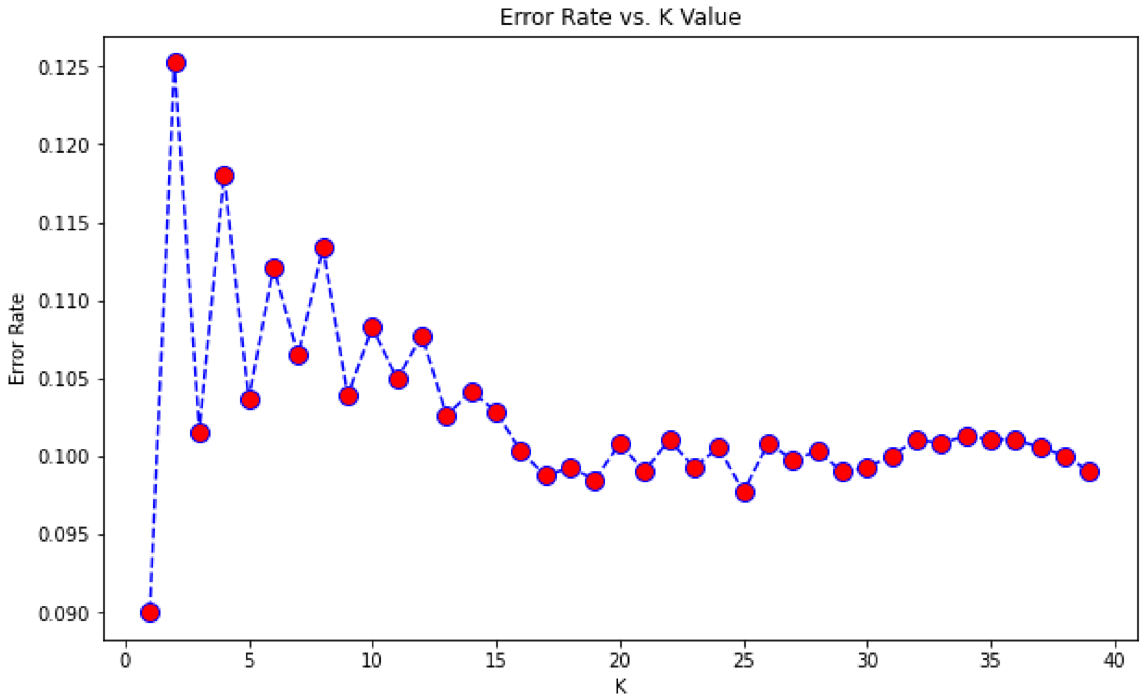

Figure 4. 4 Optimal K for sub dataset SME class

5. Optimal K for sub dataset small business class

Minimum error: 0.10584343991179714 at K = 13

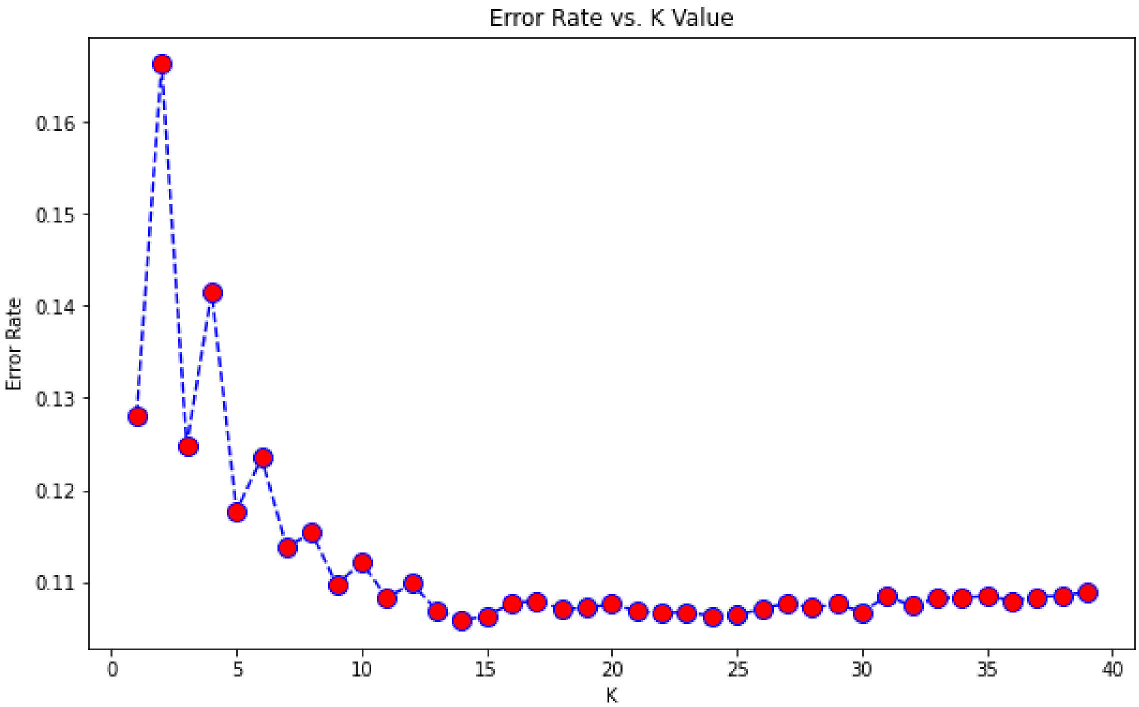

Figure 4. 5 Optimal K for sub dataset small business class

Chapter Four: Experiments and Results

This chapter presents the supervised machined learning implementation using eight classification algorithms; the chapter will also explain the design of the model and optimizations. The results of those classifiers will be described and compared.

4.1 Experiment Configuration

For building the loan prediction model, we used jupyter notebook based on Python, a powerful programming language version 3.7, which is rich with data science and machine learning libraries such as sklearn. Python also makes it easier to construct prediction models by allowing you to work easily and implement processes more efficiently.

The dataset was first classified as a whole dataset and then divided into four datasets based on the classification of the customers to see if this will improve the prediction model.

4.2 Measuring Classification Models Performance on the Whole Dataset

The first experimental work started by training the supervised machine learning classifiers on 67% of the dataset and the 33% was used to test the models.

In all experiments, we applied eight classification methods with default parameters for each classifier except the soft voting classifier and the K in the KNN classifier depends on the optimal error for each class.

The below table demonstrates the results of the evaluation of the classifiers' performance for all test samples of the dataset.

Table 4. 1: Evaluation of the classifiers' performance for all samples

Classifier	Precision	Recall	F1 score	log_ loss	jcard	auc	Accuracy (%)	processing time (in seconds)
Logistic Regression	0.957	0.998	0.956	1.487	0.956	**0.76**	95.69	11.5
K-nearest neighbors	0.975	0.989	0.965	1.175	0.965	**0.48**	96.60	230.893
Decision Tree	0.981	0.979	0.962	1.284	0.961	**0.81**	96.28	5.881
SVC	0.973	0.812	0.801	6.869	0.795	**0.80**	80.11	77.171
Random Forest	0.978	0.996	0.975	0.848	0.974	**0.96**	97.54	52.532
Bernoulli NB	0.978	0.952	0.934	2.255	0.932	**0.91**	93.47	0.558
Voting Classifier	0.991	0.998	0.990	0.315	0.990	**1.00**	99.09	411.43
AdaBoost	0.978	0.952	0.934	2.255	0.932	**0.96**	96.8	27.89

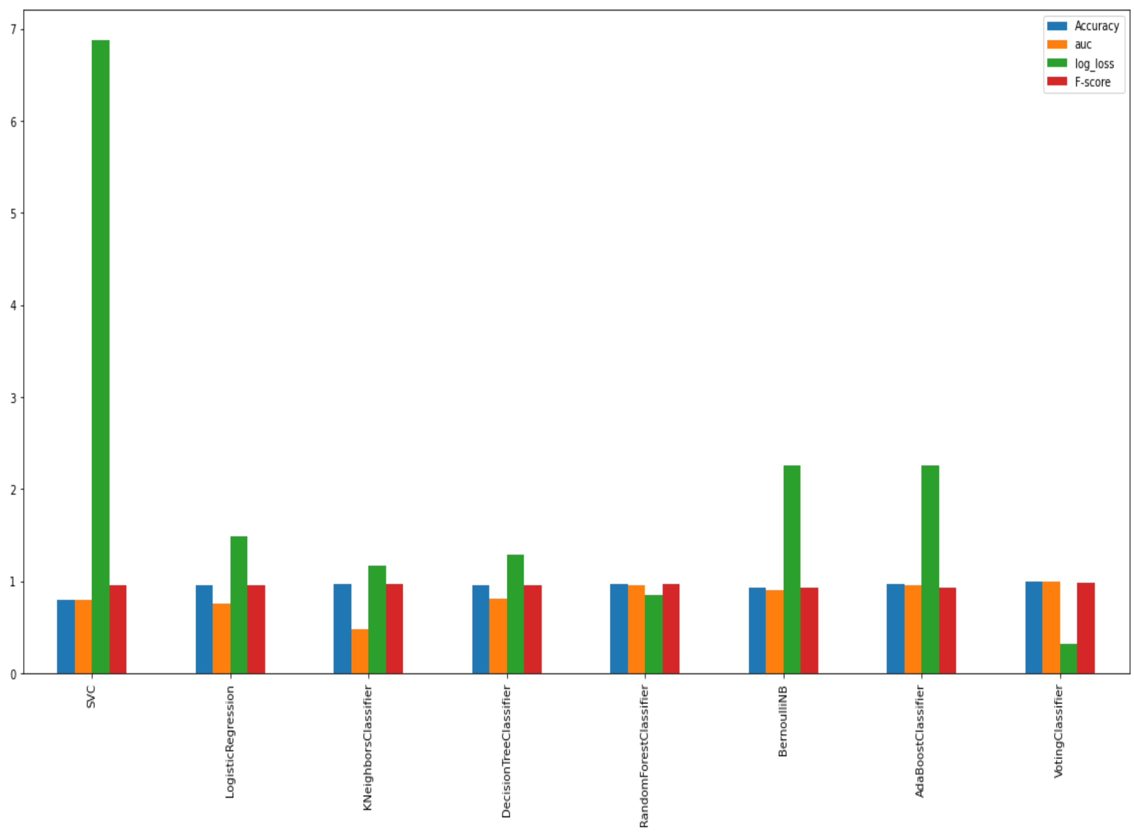

Figure 4. 6 Evaluating of the classifiers Performance for all sample

From the previous figure and table, for all samples, all classifiers performed very well and produce high classification results, except SVC had achieved low F1 score and AUC values with 0.801 and 0.80. In respect to F1 score and AUC, the Voting Classifier and Random Forest Classifier achieved the highest results with F1 Score 0.990 and 0.975, respectively. And for AUC, the Voting Classifier and Random Forest Classifier achieved the highest results with 1 and 0.96, respectively. But for the log loss, the Voting Classifier comes first with 0.315, then comes the random forest 0.848.

Table 4. 2: Feature Ranking for all samples

Feature ranking:	Feature Desc:
1. feature 7 (0.120326)	Due Balance USD
2. feature 5 (0.098297)	Days Past Due
3. feature 6 (0.092248)	Outstanding Balance USD
4. feature 106 (0.076008)	passed_months
5. feature 0 (0.066069)	AGE
6. feature 12 (0.063004)	logAvgSalaryUSD
7. feature 11 (0.048726)	logCR
8. feature 92 (0.043109)	Suspended Overdrafts
9. feature 10 (0.033762)	Rejected Cheques
10. feature 2 (0.032738)	CR. Trans.
11. feature 3 (0.028938)	Nm. Of Employee
12. feature 8 (0.027539)	Collaterals USD
13. feature 9 (0.026722)	Cleared Cheques
14. feature 91 (0.025487)	Suspended Loans
15. feature 82 (0.016228)	Overdrawn Accounts

In the previous table, the most important features in the Small Business class are; Due Balance USD, Days Past Due, Outstanding Balance USD, passed_ months, AGE, and log Avg Salary USD.

4.3 Measuring Classification Models Performance on Sub Datasets

In all experiments we applied eight classification methods on the Sub datasets with default parameters for each classifier except the voting classifier where we used soft voting, and KNN classifier where we used K that we get from optimal error for each class dataset. The below table demonstrates the results for evaluation of the classifiers' performance for General Retail class.

Table 4.3: Evaluating of the classifier's performance for General Retail class

Classifier	Precision	Recall	F1 Score	log_l oss	jcard	auc	Accuracy (%)	process ing time (in seconds)
Logistic Regression	0.973	0.997	0.971	0.994	0.970	**0.85**	97.12	10.3
K-nearest neighbors	0.982	0.993	0.976	0.808	0.976	**0.84**	97.66	203.73 0
Decision Tree	0.986	0.984	0.972	0.955	0.971	**0.80**	97.23	4.557
SVC	0.980	0.970	0.953	1.608	0.952	**0.82**	95.34	71.3
Random Forest	0.984	0.998	0.983	0.581	0.982	**0.96**	98.32	46.59
BernoulliNB	0.982	0.986	0.969	1.029	0.970	**0.92**	97.02	6.35
AdaBoost	0.982	0.986	0.970	1.029	0.969	**0.97**	97.8	33.848
Voting Classifier	0.993	0.999	0.992	0.247	0.992	**1.00**	99.28	368.43

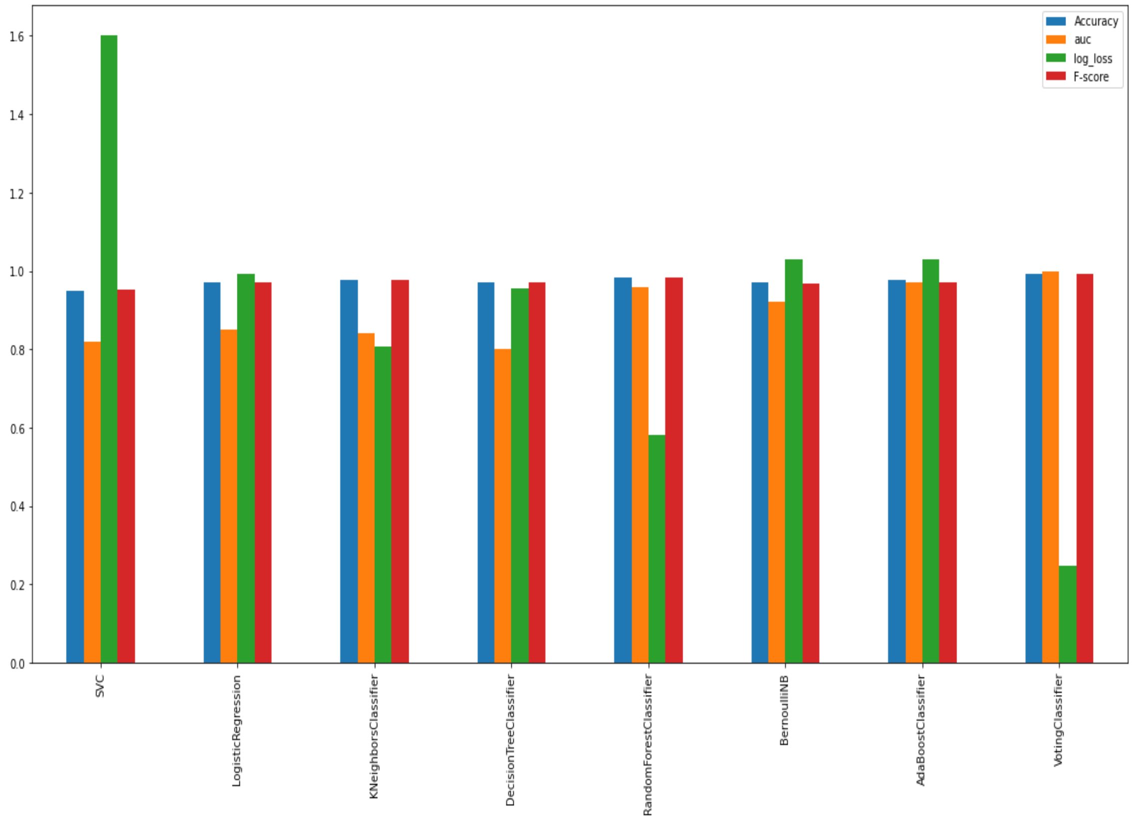

Figure 4. 7:Evaluating of the classifiers' performance for General Retail class

From the previous figure and table, for the General Retail class, all classifiers perform very well and produce high classification results. With respect to F1 Score and AUC, the Voting Classifier, Random Forest Classifier, AdaBoost Classifier, and naïve Bayes achieved the highest results with F1 Score 0.992, 0.983, 0.970, and 0.969, respectively. And for AUC, the Voting Classifier, AdaBoost Classifier, Random Forest Classifier, and the naïve Bayes achieved the highest results with 1, 0.97, 0.9, 0.92, respectively. But for the log loss the Voting Classifier came first with 0.247, then Random Forest Classifier with 0.581, then AdaBoost Classifier came with 1.029, and finally, the naïve Bayes came with 1.029.

Table 4. 3:Feature Ranking for General Retail class

Feature ranking:	Feature Desc:
1. feature 7 (0.139287)	Due Balance USD
2. feature 6 (0.105959)	Outstanding Balance USD
3. feature 5 (0.099288)	Days Past Due
4. feature 106 (0.084657)	passed_months
5. feature 12 (0.082300)	logAvgSalaryUSD
6. feature 0 (0.071690)	AGE
7. feature 92 (0.046665)	Suspended Overdrafts
8. feature 11 (0.035558)	logCR
9. feature 10 (0.028611)	Rejected Cheques
10. feature 2 (0.027905)	CR. Trans.
11. feature 8 (0.025619)	Collaterals USD
12. feature 4 (0.025574)	Avg Salary USD
13. feature 91 (0.025123)	Suspended Loans
14. feature 9 (0.023980)	Cleared Cheques
15. feature 82 (0.023390)	Overdrawn Accounts

From Table 4.3, the most important features in General Retail class are; Due Balance USD, Outstanding Balance USD, Days Past Due, passed months, log Avg Salary USD and AGE.

The below table demonstrates the results of evaluating classifiers' performance for Premier class.

Table 4. 4:Evaluating of the classifiers' performance for Premier class.

Classifier	Precision	Recall	F1 Score	log_ loss	jcard	auc	Accuracy (%)	processing time (in seconds)
Logistic Regression	0.985	0.989	0.975	0.863	0.975	**0.96**	97.50	11.75
K-nearest neighbors	0.987	0.991	0.979	0.719	0.979	**0.56**	97.92	7.1
Decision Tree	0.995	0.997	0.993	0.215	0.993	**0.86**	99.38	1.8
SVC	0.984	0.961	0.947	1.798	0.947	**0.78**	94.79	8.8
Random Forest	0.993	**1**	0.993	0.215	0.993	**0.96**	99.38	9.59
BernoulliNB	0.989	0.985	0.975	0.863	0.974	**0.93**	97.50	5.0
AdaBoost	0.989	0.985	0.975	0.863	0.974	**0.94**	098.7	1.95
Voting Classifier	1.0	0.997	0.997	0.071	0.997	**1.00**	99.79	43.5

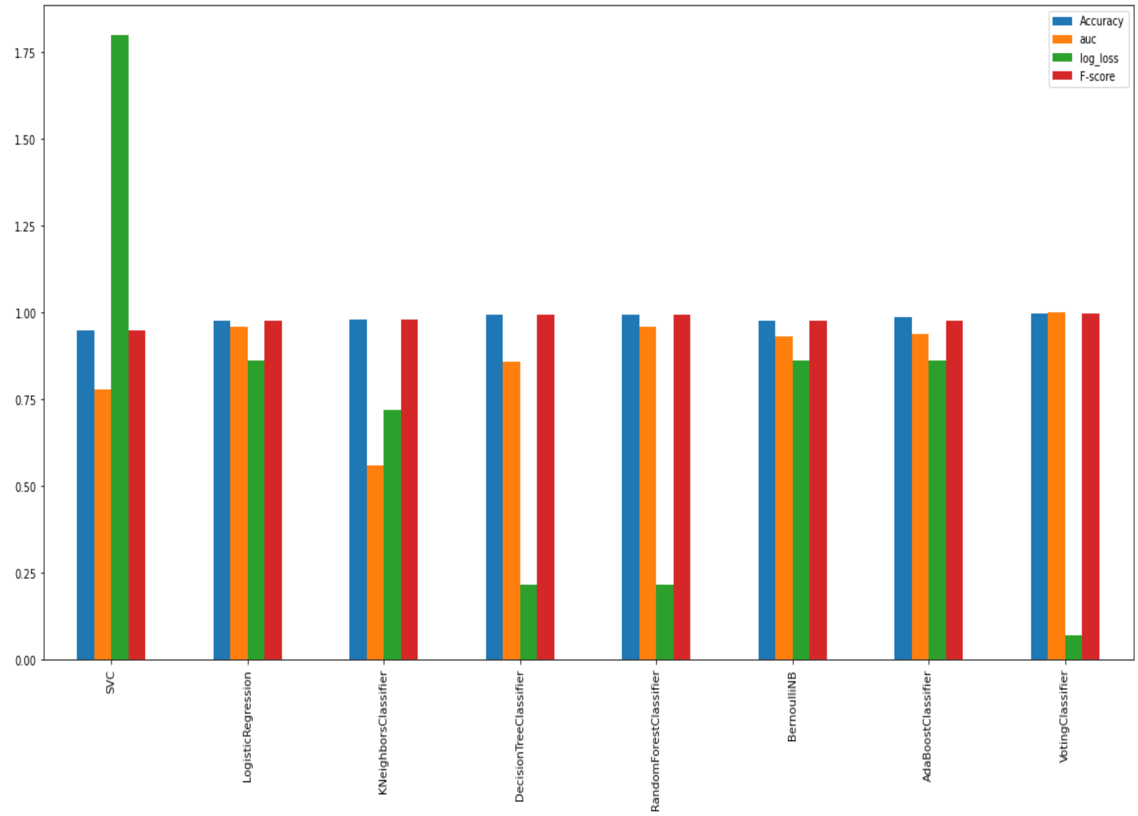

Figure 4. 8:Evaluating of the classifiers' performance for Premier class.

From the previous figure and table, for the Premier class, all classifiers perform very well and produce high classification results except for K-nearest neighbors Classifier, and SVC had achieved low AUC values 0.56 and 0.78. in respect to F1 Score and AUC, the Voting Classifier, Random Forest Classifier, Logistic Regression, AdaBoost Classifier, and naïve Bayes achieved the highest results with F1 Score 0.997, 0.993, 0.975, 0.975, and 0.975, respectively. And for AUC, the Voting Classifier, Random Forest Classifier, Logistic Regression AdaBoost Classifier, and naïve Bayes achieved the highest results with 1, 0.96, 0.96, 0.94, and 0.93, respectively. But for the log loss the Voting Classifier came first with 0.071, then Random Forest Classifier with 0.215, Logistic Regression, AdaBoost Classifier, and naïve Bayes came with the same value = 0.863.

Table 4. 5:Evaluating of the classifiers' performance for Premier class

Feature ranking:	Feature Desc:
1. feature 7 (0.086300)	Due Balance USD
2. feature 5 (0.085687)	Days Past Due
3. feature 10 (0.080716)	Rejected Cheques
4. feature 0 (0.076899)	AGE
5. feature 106 (0.073518)	passed_months
6. feature 11 (0.070595)	logCR
7. feature 2 (0.066091)	CR. Trans.
8. feature 6 (0.049848)	Outstanding Balance USD
9. feature 3 (0.039682)	Nm. Of Employee
10. feature 8 (0.036880)	Collaterals USD
11. feature 1 (0.027118)	Time Dep. Balance
12. feature 39 (0.026380)	481
13. feature 9 (0.025921)	Cleared Cheques
14. feature 101 (0.022079)	JOD
15. feature 33 (0.021580)	469

From Table 4.5, the most important features in Premier class are; Due Balance USD, Days Past Due, Rejected Cheques, AGE, passed months, logCR, and CR Trans.

The below table demonstrates the results for evaluation of the classifiers' performance for SME class.

Table 4. 6:Evaluating of the classifiers' performance for SME

Classifier	Precision	Recall	F1 Score	log_ loss	jcard	auc	Accuracy (%)	processing time (in seconds)
Logistic Regression	0.891	0.997	0.892	3.730	0.889	**0.82**	89.20	16.85
K-nearest neighbors	0.927	0.957	0.898	3.508	0.891	**0.48**	89.84	1.144
Decision Tree	0.956	0.952	0.920	2.744	0.912	**0.83**	92.05	5.05
SVC	0.936	0.522	0.554	15.391	0.504	**0.70**	55.44	4.738
Random Forest	0.940	0.984	0.932	2.318	0.927	**0.97**	93.29	3.779
BernoulliNB	0.931	0.894	0.851	5.124	0.839	**0.85**	85.16	9.4
Voting Classifier	0.987	0.997	0.986	0.461	0.984	**1.00**	98.66	45.9
AdaBoost	0.931	0.894	0.851	5.124	0.839	**0.97**	93.3	4.379

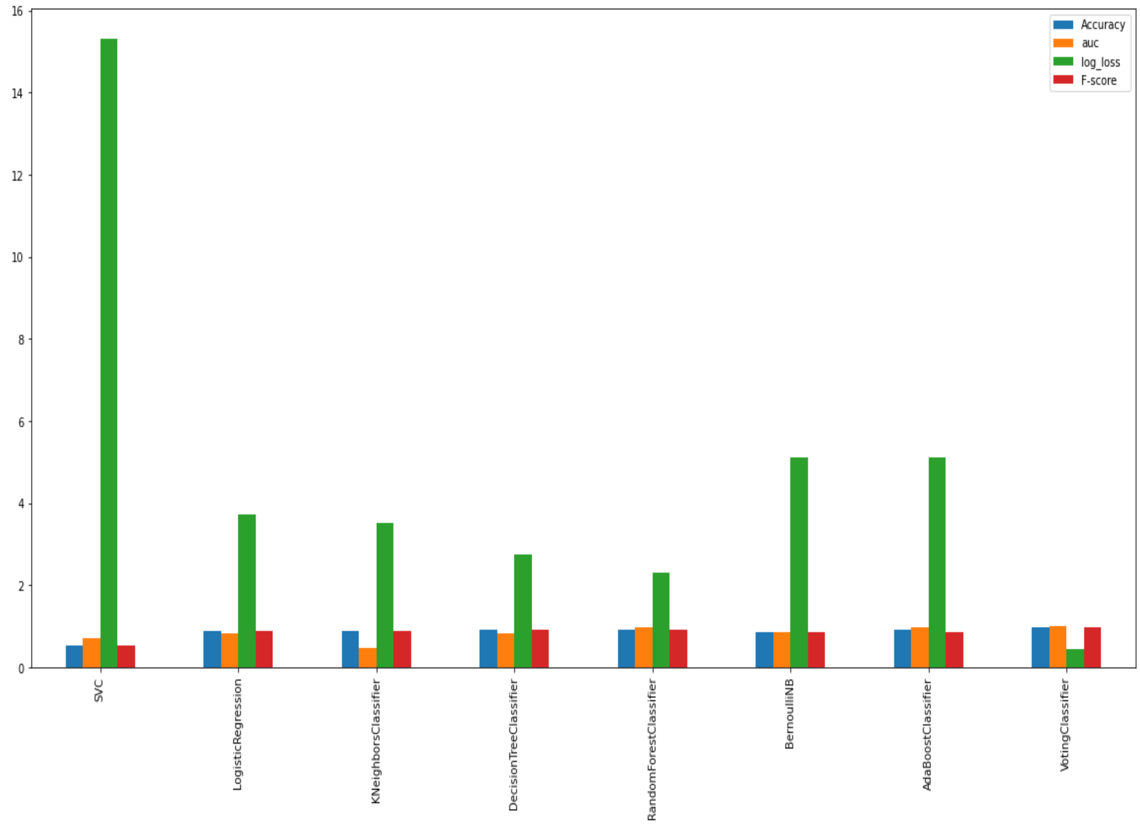

Figure 4. 9:Evaluating of the classifiers' performance for SME Class.

From the previous figure and table, for SME class, all classifiers perform very well and produce high classification results except SVC had achieved low F1 Score and AUC values with 0.554 and 0.70. with respect to F1 Score and AUC, the Voting Classifier, Random Forest Classifier, and Decision Tree Classifier achieved the highest results with F1 Score 0.986, 0.932, and 0.920, respectively. And for AUC, the Voting Classifier, Random Forest Classifier, and Decision Tree Classifier achieved the highest results with 1, 0.97, and 0.83, respectively. But for the log loss the Voting Classifier came first with 0.461, and the other algorithm achieved noticed high log loss.

Table 4. 7:Feature Ranking for SME Class

Feature ranking:	Feature Desc:
1. feature 11 (0.134581)	logCR
2. feature 7 (0.076943)	Due Balance USD
3. feature 5 (0.075814)	Days Past Due
4. feature 2 (0.060931)	CR. Trans.
5. feature 6 (0.054312)	Outstanding Balance USD
6. feature 106 (0.053481)	passed_months
7. feature 0 (0.052271)	AGE
8. feature 10 (0.051579)	Rejected Cheques
9. feature 3 (0.048461)	Nm. Of Employee
10. feature 9 (0.040551)	Cleared Cheques
11. feature 92 (0.038776)	Suspended Overdrafts
12. feature 8 (0.034999)	Collaterals USD
13. feature 91 (0.033726)	Suspended Loans
14. feature 87 (0.015109)	Restructured Loans
15. feature 35 (0.010358)	471

From Table 4.7, we can observe that the most important features in SME class are; logCR, Due Balance USD, Days Past Due, CR. Trans, Outstanding Balance USD, passed_ months, AGE, and Rejected Cheques.

The below table demonstrates the results of evaluating the classifiers' performance for Small Business class.

Table 4. 8:Evaluating of the classifiers' performance for Small Business.

Classifier	Precision	Recall	F1 Score	log_ loss	jcard	auc	Accuracy (%)	processing time (in seconds)
Logistic Regression	0.871	0.991	0.869	4.524	0.865	**0.74**	86.90	6.389
K-nearest neighbors	0.912	0.943	0.875	4.310	0.864	**0.82**	87.52	4.316
Decision Tree	0.937	0.927	0.885	3.945	0.873	**0.79**	88.58	9.10
SVC	0.847	0.993	0.843	5.407	0.843	**0.51**	84.34	12.532
Random Forest	0.927	0.985	0.921	2.699	0.914	**0.93**	92.18	8.809
BernoulliNB	0.913	0.929	0.865	4.645	0.854	**0.84**	86.55	1.50
Voting Classifier	0.972	0.997	0.974	0.894	0.970	**1.00**	97.41	33.24
AdaBoost	0.913	0.929	0.865	4.645	0.854	**0.94**	92.0	13.304

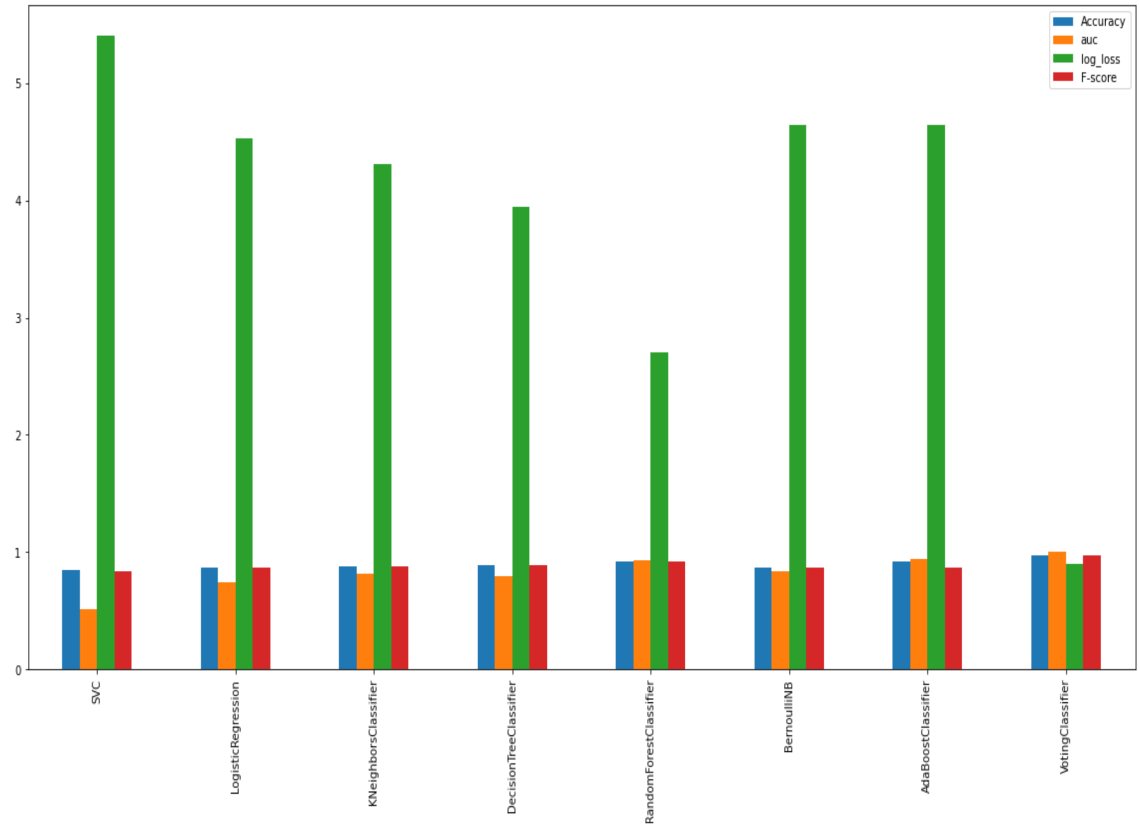

Figure 4. 10:Evaluating of the classifiers' performance for SME Class

From the previous figure and table, for Small Business class, all classifiers perform very well and produce high classification results, except SVC had achieved low F1 Score and AUC values with 0.843 and 0.51. in respect to F1 Score and AUC, the Voting Classifier and Random Forest Classifier achieved the highest results with F1 Score 0.974 and 0.921, respectively. And for AUC, the Voting Classifier and Random Forest Classifier achieved the highest results 1 and 0.93, respectively. But for the log loss the Voting Classifier came first with 0.894, and the other algorithm achieved noticed high log loss.

Table 4. 9:Feature Ranking for Small Business Class

Feature ranking:	Feature Desc:
1. feature 7 (0.122238)	Due Balance USD
2. feature 5 (0.106037)	Days Past Due
3. feature 6 (0.088470)	Outstanding Balance USD
4. feature 106 (0.070313)	passed_months
5. feature 0 (0.065247)	AGE
6. feature 11 (0.053539)	logCR
7. feature 3 (0.048541)	Nm. Of Employee
8. feature 10 (0.043213)	Rejected Cheques
9. feature 2 (0.041170)	CR. Trans.
10. feature 8 (0.032045)	Collaterals USD
11. feature 9 (0.031959)	Cleared Cheques
12. feature 12 (0.030478)	logAvgSalaryUSD
13. feature 50 (0.021445)	Loans
14. feature 92 (0.018806)	Suspended Overdrafts
15. feature 82 (0.013720)	Overdrawn Accounts

From Table 4.9, the most important features in Small Business class are; Due Balance USD, Days Past Due, Outstanding Balance USD, passed months, AGE, and log CR.

4.4 Findings

For all experiments, the best algorithm is Random Forest Classifier which had achieved high F1 Score and AUC values with low log loss; Random Forest Classifier provides an effective method for a big dataset, which was used in this research. As Random Forest Classifier is an Ensemble model, it is robust to variance and bias.

SVC Classifier achieved the lowest F1 Score and AUC values with high log loss; SVC Classifier was less effective because of the nature of the dataset, which has more than 300,000 records, which increases the complexity to find the best fit and need long time to run. SVC best works on smaller dataset.

We have achieved high accuracy which was beyond 99%, this result is very promising when compared to previous studies. In (Soni P. M. & Varghese P., 2019) the accuracy reached 99%, but the researchers used a much smaller dataset in the experimental work. Our high accuracy results were empowered by the use of Ensemble learning and voting classifier which combined seven classifies and applied soft voting method.

The result analysis and comparison of accuracy performance obtained using different models in the original dataset and the subsets (General Retail class, SME class, Small Business class, Premier class) showed that the best result was achieved on experimental work based on the Premier class subset (Accuracy 99.7%). The nature of customers in the Premier class makes the classification easy as normally they share common characteristics.

For all experimental work, the most important features which has effect on the results are Due Balance USD, Days Past Due, Outstanding Balance USD, passed_ months, AGE, and logCR for reset features have low effect in model but we couldn’t ignore it to reach accurate implemented model.

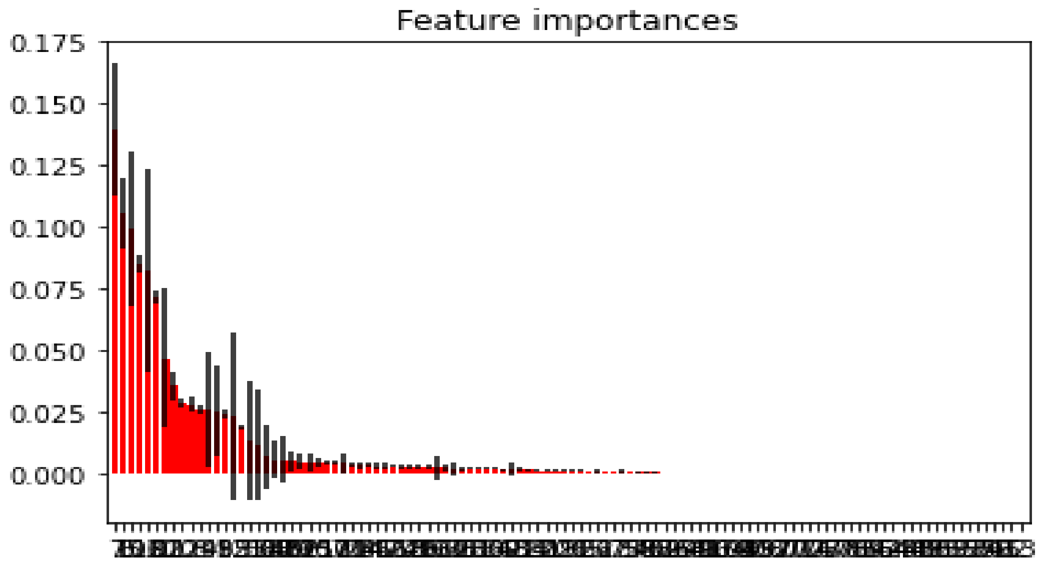

Figure 4. 11 Features Importance

Figure 4.12 demonstrates the results of the Confusion matrix for the Voting Classifier for All Classes.

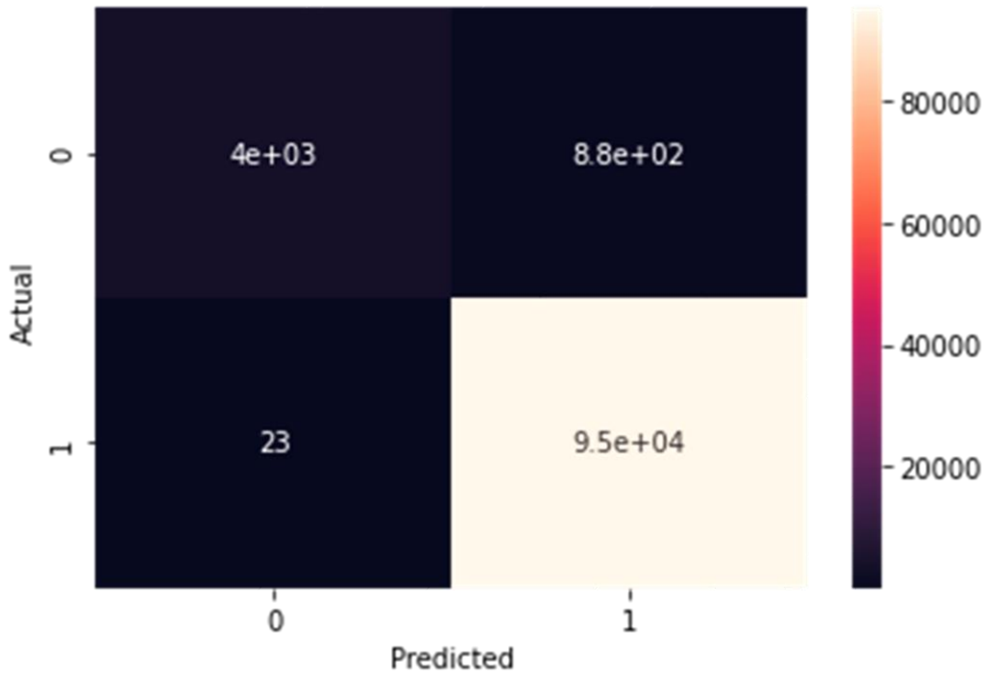

Figure 4. 12:Confusion matrix for the best classifier Voting Classifier All Classes

- True positives (TP): 95,000 customers were predicted to perform, and they are performing customer.
- True negatives (TN): 4,000 customers predicted to be non-performing, and they are nonperforming customers.
- False positives (FP): 23 customers were predicted to perform, but they don't perform. (Also known as a "Type I error.").
- False negatives (FN): 880 customers were predicted to be nonperforming, but they are performing. (Also known as a "Type II error.").

4.5 Measuring Classification Models Performance Excluding the Features (Due balance USD, Days past due)

The below table demonstrates the results of the evaluation of the classifiers' performance excluding the feature (Due balance USD, Days past Due).

Figure 4. 13 Evaluating of the Voting classifiers' performance

	Precision	Recall	F1 Score	log_ loss	jcard	Accuracy (%)	processing time (in seconds)
All samples	0.990	0.999	0.990	0.326	0.990	99.06	332.270
Premier	0.997	1.0	0.997	0.071	0.997	99.79	7.1
SME	0.991	0.993	0.986	0.470	0.984	98.64	8.101
Small Business	0.970	0.984	0.961	1.33	0.955	96.14	21.83
General Retail	0.990	0.999	0.990	0.336	0.990	99.03	289.41

From the previous table, all experiments perform very well and produce high classification results, in respect to F1 Score and Accuracy, the Premier Class achieved the highest results with F1 Score 0.997 and 0.997, respectively. And for log-loss, the Premier Class achieved the lowest results .071.

Figure 4. 14 Feature Ranking for All Sample Class Exclude Due Balance & Days Past

Feature ranking:	Feature Desc:
1. feature 5 (0.172745)	Outstanding Balance USD
2. feature 104 (0.104836)	passed_months
3. feature 0 (0.091263)	AGE
4. feature 10 (0.096677)	logAvgSalaryUSD
5. feature 9 (0.060622)	logCR
6. feature 90 (0.043139)	Suspended Overdrafts
7. feature 8 (0.041844)	Rejected Cheques
8. feature 2 (0.041775)	CR. Trans.
9. feature 3 (0.036371)	Nm. Of Employee
10. feature 6 (0.035192)	Colletrals USD
11. feature 7 (0.032812)	Cleared Cheques
12. feature 89 (0.028373)	Suspended Loans
13. feature 4 (0.020623)	Avg Salary USD
14. feature 80 (0.018714)	Overdrawn Accounts
15. feature 48 (0.016278)	Loans

From Table 4.14 the most important features in All Class are Outstanding Balance USD, passed months, AGE, logAvgSalaryUSD and logCR.

The below table demonstrates the results of feature ranking for Premier class.

Figure 4. 15 Feature Ranking for All Premier Class Exclude Due Balance & Days Past

Feature ranking:	Feature Desc:
1. feature 0 (0.091373)	AGE
2. feature 104 (0.088953)	passed_months
3. feature 9 (0.088905)	logCR
4. feature 8 (0.088578)	Rejected Cheques
5. feature 2 (0.080640)	CR. Trans.
6. feature 5 (0.066626)	Outstanding Balance USD
7. feature 6 (0.051714)	Colletrals USD
8. feature 3 (0.048832)	Nm. Of Employee
9. feature 7 (0.031258)	Cleared Cheques
10. feature 37 (0.031150)	Branch number 481
11. feature 1 (0.029861)	Time Dep. Balance
12. feature 39 (0.026380)	Branch number 469
13. feature 9 (0.025921)	Branch number 466
14. feature 101 (0.022079)	JOD
15. feature 33 (0.021580)	logAvgSalaryUSD

From Table 4.15 the most important features in Premier Class are AGE, passed months, AGE, passed month, logCR, Rejected Cheques and CR. Trans.

The below table demonstrates the results of feature ranking for SME class.

Figure 4. 16 Feature Ranking for SME Class Exclude Due Balance & Days Past

Feature ranking:	Feature Desc:
1. feature 9 (0.146689)	logCR
2. feature 5 (0.079055)	Outstanding Balance USD
3. feature 2 (0.073366)	CR. Trans.
4. feature 104 (0.069355)	passed_months
5. feature 0 (0.066682)	AGE.
6. feature 3 (0.063024)	Nm. Of Employee
7. feature 8 (0.060653)	Rejected Cheques
8. feature 7 (0.049884)	Cleared Cheques
9. feature 6 (0.043436)	Colletrals USD
10. feature 90 (0.040455)	Suspended Overdrafts
11. feature 89 (0.039479)	Suspended Loans
12. feature 85 (0.015820)	Restructured Loans
13. feature 33 (0.010483)	Branch number 471
14. feature 80 (0.010224)	Overdrawn Accounts
15. feature 48 (0.010035)	Loans

From Table 4.16 the most important features in SME Class are logCR, Outstanding Balance USD, passed month, AGE and CR. Trans.

The below table demonstrates the results of feature ranking for Small Business class.

Figure 4. 17 Feature Ranking for Small Business Class Exclude Due Balance & Days Past

Feature ranking:	Feature Desc:
1. feature 5 (0.154739)	Outstanding Balance USD
2. feature 104 (0.101841)	passed_months
3. feature 0 (0.093600)	AGE
4. feature 9 (0.067031)	logCR
5. feature 3 (0.062477)	Nm. Of Employee.
6. feature 8 (0.053994)	Rejected Cheques
7. feature 2 (0.051320)	CR. Trans.
8. feature 6 (0.041703)	Colletrals USD
9. feature 7 (0.039619)	Cleared Cheques
10. feature 10 (0.034138)	logAvgSalaryUSD
11. feature 48 (0.023349)	Loans
12. feature 90 (0.022138)	Suspended Overdrafts
13. feature 89 (0.015075)	Suspended Loans
14. feature 80 (0.013943)	Overdrawn Accounts
15. feature 49 (0.011595)	Overdrafts & Overdrawn

From Table 4.17 the most important features in SME Class are Outstanding Balance USD, passed months, AGE, logCR, Rejected Cheques and Nm of Employee.

The below table demonstrates the results of feature ranking for General Retail class.

Figure 4. 18 Feature Ranking for General Retail Class Exclude Due Balance & Days Past

Feature ranking:	Feature Desc:
1. feature 5 (0.172745)	Outstanding Balance USD
2. feature 104 (0.104836)	passed_months
3. feature 0 (0.091263)	AGE
4. feature 10 (0.096677)	logAvgSalaryUSD
5. feature 9 (0.060622)	logCR
6. feature 90 (0.043139)	Suspended Overdrafts
7. feature 8 (0.041844)	Rejected Cheques
8. feature 2 (0.041775)	CR. Trans.
9. feature 3 (0.036371)	Nm. Of Employee
10. feature 6 (0.035192)	Colletrals USD
11. feature 7 (0.032812)	Cleared Cheques
12. feature 89 (0.028373)	Suspended Loans
13. feature 4 (0.020623)	Avg Salary USD
14. feature 80 (0.018714)	Overdrawn Accounts
15. feature 48 (0.016278)	Loans

From Table 4.18 the most important features in SME Class are Outstanding Balance USD, passed months, AGE, logCR, and logAvgSalaryUSD.

From the previous results we find most of classes have the same ranking for feature importance except Primer class which Outstanding Balance USD feature changed from first to the rank 6 because most of primer class customers have no issue with amount of loan.

Figure 4.19 demonstrates the results of the Confusion matrix for the Voting Classifier for All Classes excluding (Due balance USD, Days past due).

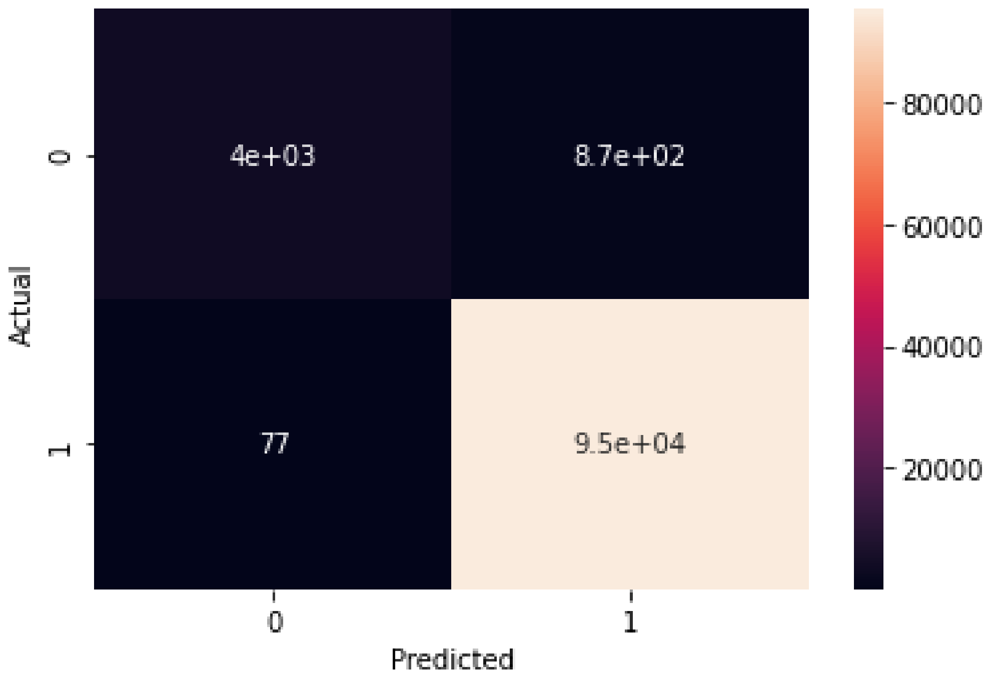

Figure 4. 19 Confusion matrix for the Voting Classifier All Classes Excluding

- True positives (TP): 95,000 customers were predicted to perform, and they are performing customer.
- True negatives (TN): 4,000 customers predicted to be non-performing, and they are nonperforming customers.
- False positives (FP): 77 customers were predicted to perform, but they don't perform. (Also known as a "Type I error.").
- False negatives (FN): 870 customers were predicted to be nonperforming, but they are performing. (Also known as a "Type II error.").

4.6 Conclusion and Future Work

In this research, seven algorithms Logistic Regression, K-nearest neighbors, Decision tree, SVM, Random Forest, AdaBoost, and naïve Bayes algorithms, were used to predict and classify the loan applicants in terms of her/his ability to repay the loan.

We have used the Voting classifier after applying seven classification models, and Those were aggregated to the voting classifier to come up with the final prediction.

Voting classifiers showed promising results. We believe that the nature of the dataset could have a major role in the improvement in prediction accuracy 99% as it contains only customers who were granted loans.

These promising results showed that our DSS could assist in the future to assist in the replace manual work by bank officers and reduce the time spent in loan decision making with high accuracy.

There are so many algorithms that could be used to do classification, and in this work seven of the most common classification algorithms were implemented, evaluated, and compared.

The potential future work for this project will be a further development to make a study concerning a different customer class in the bank of Palestine, it should be considered that there are classes not include in the scope of the dataset that we have; more loan applications should be generated and used for modeling loan risk prediction in future.

It would also be interesting to make a study concerning different Banks in Palestine, and we need to test our model on a different dataset from other banks in the same region.

There are various improvements we could apply to this study in the future.

The outlier problem, if there are outliers in the dataset, the prediction model's results will be less accurate.

References

A. Salihu & V. Shehu. (2020). A Review of Algorithms for Credit Risk Analysis. *In Proceedings of the ENTRENOVA-ENTerprise REsearch InNOVAtion Conference, Virtual Conference,* Vol. 6, pp. 134-146.

Abdou H. A. & Pointon J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting*, 59-88.

Abusalah M. (2008). *Cross language information retrieval using ontologies*. Sunderland: University of Sunderland.

Addae-Korankye A. (2014). Causes and control of loan default/delinquency in microfinance institutions in Ghana. *merican International Journal of Contemporary Research*, 4(12), 36-45.

Alfarra A. N., Hui X. F. & Ahmed J. (2016). Palestine Banks Sector Ability to Apply Credit Risk Management According to Basel III. *DEStech Transactions on Social Science, Education and Human Science , (seme).*

Alfarra, et. al. (2016). Palestine Banks Sector Ability to Apply Credit Risk Management According to Basel III. *DEStech Transactions on Social Science, Education and Human Science, (seme).*

Assiri A. S., Nazir S. & Velastin S. A. (2020). Breast tumor classification using an ensemble machine learning method. *Journal of Imaging*, 6(6), 39.

Assiri A. S., Nazir S. & Velastin S. A. (2020). Breast tumor classification using an ensemble machine learning method. *Journal of Imaging*, 6(6), 39.

Badr W. (2019, aug 5). *towardsdatascienc*. Retrieved from 6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779

Berger A. N., & Udell. (2006). A more conceptual framework. *Journal of Banking & Finance*, 2945–2966.

Berhano Y. (2017). *ASSESSMENT OF CREDIT MANAGEMENT POLICY AND LOAN COLLECTION PROBLEM IN CASE OF UNITED BANK SC.* (Doctoral dissertation, St. Mary's University).

Boot. (2000). Relationship banking: What do we know? *A. W. A*, 7–25.

Caruana R., & Niculescu-Mizil A. (2005). 1.6 A PREDICTING GOOD PROBABILITIES WITH SUPERVISED LEARNING. *In Proceedings of the 22nd international conference on Machine learning*, (pp. 625-632).

Chen Ya-Qi, Jianjun Zhang & Wing WY Ng. (2018). Loan Default Prediction Using Diversified Sensitivity Undersampling. *IEEE*, Vol. 1.

CM Reinhart & KS Rogoff. (2009). *This Time Is Different: Eight Centuries of Financial Folly*. Princeton University Press.

Gahlaut A. & Singh P. K. (July 2017). Prediction analysis of risky credit using Data mining classification models. *8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-7.

gnatius J., Hatami-Marbini A., Rahman A., Dhamotharan L. & Khoshnevis P. (2018). A fuzzy decision support system for credit scoring. *Neural Computing and Applications*, 29(10), 921-937.

Goyal A. & Kaur R. (2016). Accuracy Prediction for Loan Risk Using Machine Learning Models. *Int. J. Comput. Sci. Trends Technol*, 4(1), 52-57.

Granström D. & Abrahamsson J. (2019). Loan default prediction using supervised machine learning algorithms.

Hamid A. J. & Ahmed T. M. (2016). Developing prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal (MLAIJ)*, Vol, 3(1).

Helleiner E. (2011). Understanding the 2007–2008 global financial crisis: Lessons for scholars of international political economy. *Annual review of political science*, 14, 67-87.

Ignatius, J., Hatami-Marbini, A., Rahman, A., Dhamotharan, L., & Khoshnevis, P. (2016). A fuzzy decision support system for credit scoring. *Neural Computing and Applications*,, 29(10), 921-937.

Leo M., Sharma S. & Maddulety K. (2019). Machine learning in banking risk management: A literature review. *Risks*, 7(1), 29.

Marte A. (2019). *Machine learning in default Prediction: the incremental power of machine learning techniques in mortgage default prediction (Master's thesis).* Norwegian School of Economics.

Muih, et. al. (2016). Asynchronous methods for deep reinforcement learning. *In International conference on machine learning,* (pp. 1928-1937). PMLR.

Nabende P. & Senfum S. (2019). A study of machine learning models for predicting loan status from Ugandan loan applications. *In Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pp. 462-468.

Oduor George. (2019, May 8). *Common Classification Model Evaluation metrics.* Retrieved from towards data science: https://towardsdatascience.com/common-classification-model-evaluation-metrics-2ba0a7a7436e

Powers, D. M. ((2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv*, preprint arXiv:2010.16061.

Sadatrasoul, S. M., Gholamian, M., Siami, M., & Hajimohammadi, Z. (2013). Credit scoring in banks and financial institutions via data mining techniques: A literature review. *Journal of AI and Data Mining*, 1(2), 119-129.

Saravana, N., & Gayathri, D. V. (2018). Performance and Classification Evaluation of J48 Algorithm and Kendall’s Based J48 Algorithm (KNJ48). *International Journal of Computer Trends and Technology (IJCTT)*, Volume, 59.

Sedgwick P. (2012). *Pearson’s correlation coefficient.* Bmj: 345.

Sivasree M. S. & Rekha Sunny T. (2015). Loan Credibility Prediction System Based on Decision Tree Algorithm. *International Journal of Engineering Research & Technology (IJERT)*.

Soni P. M. & Varghese P. (2019). Algorithm For the Loan Credibility Prediction System. *International Journal of Recent Technology and Engineering (IJRTE)*, 1080-1087.

الملخص

مع زيادة عدد طلبات القروض والتمويل في البنوك والمؤسسات المصرفية أصبح معالجة هذه الطلبات يدوياً يستغرق وقتاً وجهداً مُضنياً.

يمكن للعملاء التقدم بطلب للحصول على قروض تشغيليه متعددة منها الزواج، وشراء سيارات جديدة أو قديمة، والرهن العقاري، والتعليم، ومشاريع البناء الجديدة، وما إلى ذلك.

يتم منح العملاء المؤهلين فقط وفق معايير محدده لضمان إعادة تلك الأموال الى البنك.

لا تزال العديد من البنوك في جميع أنحاء العالم تستخدم إجراءات منح او رفض القروض يدويا لإصدار الأحكام ويتم اتخاذ هذه القرارات من قبل موظفي البنك بناءً على بيانات العملاء المتوفرة في قاعدة بيانات البنك، أو من مصادر أخرى يستطيع البنك الوصول إليها كالإقرار الضريبي وغيره.

إن اتخاذ مثل هذا القرارات يتطلب معرفةً عميقةً بالعملاء وخطط قروضهم لذا يستخدم مسؤولو القروض التحليل المالي لكل عميل، مثل التدفقات النقدية والتقارير السنوية ومقاييس السيولة فهي تجمع بين المعلومات الشخصية والبيانات المالية لتحديد القرار أو التوصية النهائية لموظفي البنك.

تهدف هذه الدراسة إلى الاستفادة من التَّعلم الآلي لإنشاء نظام مؤتمت لتحديد العملاء المؤهلين لمنحهم القروض وتحديد ما إذا كان العميل سوف يسدد الالتزامات المترتبة عليه أم لا في حالة منحه القرض وفق البيانات المزودة للنظام.

موضوع هذه الدراسة هو بنك فلسطين الذي قدم لنا بيانات حقيقيه للعملاء المتوفرين في قاعدة بينات البنك اللازمة لإجراء هذا البحث.

أظهر البحث نتائج واعدة وممتازة ستؤدي الى تقليل الوقت الذي يقضيه الموظفين في اتخاذ قرارات منح القروض بدقة عالية مماثلة لسير العمل اليدوي الحالي الذي ينفذه بنك فلسطين بجهد أقل.