CREDIT RISK ANALYSIS USING MACHINE LEARNING ALGORITHMS

BY

DEPAYAN BANERJEE
ID: 221-25-088

This Report Presented in Partial Fulfilment of the Requirements for the Degree of Master of Science in Computer Science and Engineering (Major in Data Science)

Supervised By

Fahad Faisal
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Abdus Sattar
Assistant Professor & Coordinator – MSCSE
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Thesis titled "**Credit Risk Analysis Using Machine Learning Algorithms**", submitted by Depayan Banerjee, ID No: 221-25-088 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 17-01-2023.

BOARD OF EXAMINERS

Dr. S M Aminul Haque, PhD **Chairman**
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Ms. Naznin Sultana **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Md. Sadekur Rahman **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin, PhD **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that this thesis paper has been done by me under the supervision of **Fahad Faisal, Assistant Professor, Department of Computer Science and Engineering,** Daffodil International University. I also declare that neither this paper nor any part of this paper has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Faisal

Fahad Faisal
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Abdus Sattar
Assistant Professor & Coordinator – MSCSE.
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Depayan Banerjee

Depayan Banerjee
ID: 221-25-088
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENTS

First, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible for me to complete the thesis paper successfully.

I really am grateful and wish my profound indebtedness to **Fahad Faisal, Assistant Professor**, Department of Computer Science and Engineering, Daffodil International University, Dhaka. The deep knowledge & keen interest of my supervisor in the field of **"*Machine Learning*"** has enthused me greatly to carry out this thesis paper. His endless patience, scholarly guidance, continual encouragement, energetic supervision, constructive criticism, valuable advice, and patience in reading many inferior drafts and correcting them at all stages have made it possible for me to complete this paper.

I would like to express my heartiest gratitude to **Dr. Touhid Bhuiyan**, Professor and Head, Department of Computer Science and Engineering, for his kind help in completing my thesis paper. I am also grateful to other faculty members and the staff of the Computer Science and Engineering department of Daffodil International University for their kind support and help.

I would like to thank my entire course mate at Daffodil International University, who took part in many discussions and help me to complete this paper.

Finally, I must acknowledge, with due respect, the constant support, encouragement, and patience of my parents.

ABSTRACT

Almost every financial institution, for instance, credit card companies and banks heavily rely on credit risk grade systems to determine whether to issue a loan to the probable debtor. They put the applicants into 8 categories like Superior, Good, Acceptable, Marginal, Special Mention, Substandard, Doubtful, and Bad. They generally depend on traditional judgmental techniques to approve the application which takes a longer period of time. The process can be quickened by applying machine learning algorithms where the models learn from data by analyzing the pattern and then providing us with insight. Credit risk must be handled properly and it is very important for banking institutions, as loss can appear when the debtor is unable to pay back the owed money. In this study, the dataset will be analyzed where people are applying for a loan will be my research subject. Various popular machine learning algorithms such as Random Forest, Decision Tree, Naïve Bayes, KNN, Logistic Regression, and SVM will be applied to train different models and try to predict the outcome of an application being risky to grant a loan or not. The results like accuracy, precision, recall, and F1-Score, the training, and the testing time of each model trained by the mentioned machine learning algorithms will be compared. Finally, the result of each model will be evaluated by applying K-Fold Cross-validation, confusion matrix, and AUC-ROC Cure technique to find the best machine learning model among the mentioned models. In this study, it has been observed that Random Forest is overall the best model with an accuracy of 97.35%, precision of 99.84%, recall of 94.80%, F1-Score of 96.77%, AUC Value of 96.8%, while logistic regression is the second-best algorithm to tackle this problem with 96.59% of accuracy rate.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

Credit risk is a term that means the possibility of a financial institution's loss resulting from a person who lends money and fails to pay back the debt or meet the contractual commitment. Traditionally, credit risk refers to the risk that a debtor may not get the lent money and interest back, which in the future results in the cash flows being interrupted and costs being increased for collection.

Bank credit sits on top of economic development. Without ample finance, it will be hard for the economy to grow. Financial institutional lending is a vital part of the economy in the sense that it can concurrently finance all of the sub-categories of the financial field, which encompass the commercial, agricultural, and industrial activities of a country [1]. So, a financial institution is expected to issue its loanable money among economic agent-in-deficit in a way that will bring upon adequate income for it and simultaneously benefit the person who is borrowing the money to control his/her deficit.

Although it's not possible to know exactly who will not repay the money, it is necessary for any financial organization to properly assess and manage credit risk to lessen the severity of a credit loss. Banks use judgmental techniques and/or credit scoring models to either accept or reject a client’s loan application. A flawless credit risk model enables financial institutions to decline the applications that seem to not repay the loan and approved the applications of those applicants who would pay their debts with interest in time. This model will ensure that the institution will not go through financial loss rather the profit will increase. Thus, a good credit risk model will ensure an increase in cash flow and reduce the risk of a loss. However, building such a model is very complex, and it takes extensive knowledge of statistical analysis and modeling and previous data of the applicants. This knowledge and data are used to predict the outcome of each application for a loan grant. Algorithms for Machine learning are applied to the statistical methods to accelerate the process.

Algorithms for Machine learning are popularly used to filter the applicant's applications based on their previous pattern of credit usage and thus provide a result that will help to make decisions about whether to grant a loan to an applicant or not. Nowadays, algorithms are seen to be utilized in many domains with predefined purposes. For example, algorithms are utilized in enterprises to sign up persons befitting for the profile proposed. Algorithms of machine learning can make daily life easy by simplifying day-to-day tasks, making them faster and more fluid, etc. However, algorithms are predefined codes with specific tasks to complete a certain purpose. For instance, recruiting people for an organization can introduce biases or a certain profile, and then, "format" the individuals working in the organization. The same thing happens in loan provisions cases, from a financial organization to an organization, where the decision is dependent on the algorithm that is being used. So, it is crucial for decision-makers to keep aware of these kinds of biases and to detect ways to handle the use of powerful algorithms. In this paper, the fact that there are several machine learning algorithms available that can be utilized in parallel to answering a query, for example, to issue a loan to a company will be illustrated. The fact that there exist many strategies to get to the goal of discovering the choice of the features, also known as variables or attributes, the criteria, and the machine learning algorithm that provides a proper solution to the asked question will be observed.

In the case of banking, the algorithms help to evaluate the dataset's accuracy to classify the applicants who are applying for a loan into good and bad classes. Those applicants who fall in the good classes have a high chance of returning the borrowed money to the financial institution. Those applicants who fall in the bad classes have a low chance of repaying the credit to the financial institution so, they are known as the defaulters of the bank loans. To reduce the bad loan rate in the credit dataset, many types of credit risk evaluation procedures are utilized [2], [3], [4]. Sometimes enormous losses can be minimized even with a little bit of improvement in the credit evaluation accuracy. The advantages of the dependable credit risk dataset are it minimizes the cost of credit scoring, flawless decision-making in a very short amount of time, and avoid the risk associated with the collection of the issued loan. while the evaluation of credit risk plays a very important role in the financial field and it is very vital with big challenges faced by financial organizations, accuracy plays a very impactful role in the credit data classification to keep financial loss in control. The increase in bad loan rates in the dataset of credit risk which is not dependable gives motivation towards this area [5]. However, in Bangladesh, few loans get classified.

Hence, this study is intended to classify the loan risk and discover the best algorithm that performs with the best accuracy to predict whether a grant in the loan will safely bring the cash back with interest. In this research, all challenges will be tackled during the data collection stage, filtering the dataset for important attributes, and selecting a dependable classification algorithm of machine learning to discover the best possibility of a loan being repaid in due time. In the later stages of this study, with evaluation metrics, the data will be evaluated and then the best solution for the prepared dataset for this topic will be tried to be discovered.

1.2 Motivation

In modern times, most lender organizations do not classify loans rather are using the judgmental approach to determine whether a customer is credit-worthy or not. Also, since the loans are not getting classified properly, the high risk of loss exists due to the approval of applications of customers that in the future will not repay the money that was lent, and one more problem is potential customers are getting a lesser score. As such, a dynamic credit risk model created by machine learning algorithms can at a very early stage discover clients that hold a higher risk of future delinquency. It could be a valuable instrument for the lender's company. So, it is necessary to build a predictive model by using popular algorithms to classify clients' accounts as either risky or not risky. The machine learning models will be built using various algorithms trained on historical credit datasets. ML is a sub-section of computer science that has the potential to discover patterns, and insights and learn without being manually programmed after its creation. Machine learning tools are very dependable for a problem like this. Although some similar projects and research had been done, the lender's organization was not motivated to use those technologies. So, multiple new models will be built and one model will be selected from them that will provide higher accuracy and performance than the existing models to encourage the creditors to use them.

1.3 Rationale of the Study

Many studies regarding credit risk governance to predict credit-worthy applicants have been conducted over the years. The decisions were taken by examining the socio-economic and demographic profiles of a debtor. There are various restrictions in these techniques, in particular, authentic comparisons are impossible to perform between the existing client's behavior with a new client. In this process, a few features get missing from the comparison for which bad loan rates are rising over the years. This possess a big challenge for financial organizations and other banking companies, as they are now in need of a robust risk prediction model built with the help of computer science to generalize the economic behavior of their existing and potential future customers. In this paper, ML methods were utilized to make a predictive model in which various attributes are identified as important, and also the models were used to predict the behavior of customers by predicting whether they are risky or not risky to issue a loan.

1.4 Research Questions

At the beginning of every research, some questions are needed to be considered. This will help in completing the research successfully. Without, considering the research question, it will be tough to complete any research. The questions that were considered before working on this research are mentioned below:

1. What methodology will be used for this research?
2. How the data should be collected?
3. What kind of machine learning algorithm should be used?
4. What is the prefect programming language to implement the solutions?
5. Which features are important to predict the outcome perfectly?
6. Does every algorithm work perfectly (yes/no)?
7. Is it possible to get a good accuracy score from the built models?
8. How do different machine learning algorithms perform on credit risk datasets to get a prediction?
9. Is the time enough for completing the study?

1.5 Expected Output

ML and AI have made daily life easy by providing automated predictive results which are close to 100% accurate. It is needed to train the computers, based on the previous behavior of people, which comes as datasets, and apply their learning in different domains from medical to financial sectors. This paper will center around various modeling approaches and model performance of machine learning to identify dependable algorithms for credit risk systems after selecting the accurate features and data related to those features. In this research, whether an applicant is risky or not to give a loan will be predicted. This is a binary classification problem. So, the result will be in simple “Yes” or “No form. It is expected that the accuracy, precision, recall, F1-Score, and AUC value of the ML models trained with various ML algorithms will increase and the scores will be better than the scores of most of the research.

1.6 Report Layout

The summary of the entire report has been mentioned in this section where what chapter will contain what type of information will be mentioned in short. This will be helpful in finding important pieces of information from a particular chapter quite easily.

Chapter 1: Introduction

This chapter provides the introduction, motivation, rationale of the study, questions regarding the research topic, and the outcome of this research that was expected. By giving a gentle introduction to this research, those topics were discussed briefly.

Chapter 2: Background

This chapter consists of some basic terminologies, what has already been accomplished via means of the preceding workers/researchers on this topic, what the findings and problems of their findings are, and why the mentioned approach in this paper is more accurate compared to theirs. The scope of the research problem will be elaborated and the challenges will be pointed out that were tackled during this research.

Chapter 3: Research Methodology

This chapter shows the methodology that was proposed. But before that, the research subject and instruments required will be discussed and then some essential methods like data collection, data visualization, data augmentation, data pre-processing, etc. will be mentioned. Some general information about ML and its algorithms will be mentioned so that the concept can be understood by everybody. The statistical analysis of the topic will be discussed. Some implementation requirements for the data analysis stage will be pointed out at the end of this chapter.

Chapter 4: Experimental Results and Discussion

This chapter elaborates on the performance comparison of the "Success rate" of this paper's trained ML models with some well-known ML algorithms such as Decision Tree(DT), Random Forest(RF), Naïve Bayes(NB), K-Nearest Neighbor(KNN), Support Vector Machine(SVM), and Logistic Regression(LR). After that, the complete result analysis of all models will be discussed and then the result of the models will be evaluated using a confusion matrix, AUC-ROC Curve, and Cross-Validation to select a suitable model for this topic.

Chapter 5: Impact on Society, Environment and Sustainability

This is the chapter where bank loans' impact on society and the environment has been mentioned. The ethical aspects that need to be followed during the research work have been described. Finally, a sustainability plan of the machine learning technique has been detailed so that financial institutions can upgrade their policy, and following the plan, they can improve their performance.

Chapter 6: Summary, Conclusion, Recommendation, and Implication for Future Research

This is the chapter where a conclusion was made by summing up the entire work. Some of the works for the future regarding this topic will be discussed. Some recommendations were also provided that can be used by future researchers in their studies.

CHAPTER 2
BACKGROUND

2.1 Introduction

In this chapter, the reasons for studying this field of ML algorithms, and how the algorithms are utilized to automate the process of providing risk-free loans to borrowers, especially new clients will be discussed. Before moving toward the work, here the work that had been done which is related to the area of prediction of success rate will be mentioned. Here, various study papers, their procedures with their plan of action, and the problems with their works will be discussed. A comparative analysis of what was discovered from other research papers will be made. The current decision-making method for approving or denying the loan-related application will also be mentioned.

2.2 Related Works

There are many studies available for credit risk assessment. However, in Bangladesh, judgmental techniques are applicable for detecting risky applicants. Firstly, the techniques used to filter applications in Bangladesh and then the topic where machine learning is used to predict loan risk will be discussed to make the application process faster.

In banking, bank authorities gather information about an applicant who wishes to get a loan from there. They use this information to evaluate the chances of the loan and interest not being repaid. They typically evaluate the data in a subjective manner and typically judgmental technique is used. Typically, the bank manager assesses the information on a particular applicant. This assessment involves a detailed study of the 6 aspects which are mentioned in the next page:

- Character – credit officer must be satisfied with the borrower's intention for applying for a bank loan and a consequential motive to repay.
- Capacity – credit officer must be certain that the client has the authorization to apply for credit and the applicant must have legal standing to put pen to paper on a binding credit agreement.
- Cash – credit officer must have the answer to the question: Does the applicant have the capability to produce cash to repay the debt?
- Collateral – credit officer must inquire, does the applicant have enough net worth or adequate quality properties to give enough support for the debt.
- Conditions – both credit officers and credit analysts must be keeping their eyes open on the current trends in the applicant's line of work and how changing financial conditions may affect the debt.
- Control – it centers on such questions as whether modifications in rules and regulations could have an adverse impact on the applicant and if the debt plea fulfills the bank's and the regulatory authorities' standards for debt quality.

After assessing the aspects, the bank authority has a good idea of whether an applicant is credit-worthy or not. Then they use the credit risk grading system suggested by Bangladesh Bank. This system has 8 categories:

TABLE 2.1: CREDIT RISK GRADE SYSTEM

Grading	Short Name	Number
Superior	SUP	1
Good	GD	2
Acceptable	ACCPT	3
Marginal/Watchlist	MG/WL	4
Special Mention	SM	5
Sub standard	SS	6
Doubtful	DF	7
Bad & Loss	BL	8

An explicit definition of the various groups of Credit Risk Grading is mentioned in the next page:

- Superior (SUP) – Loan facilities, that are considered completely secured meaning cash is entirely covered, completely covered by the guarantee of the government, and completely covered by an international bank guarantee which is top-tier.
- Good (GD) – The debtor has strong refundability, great liquidity, low leverage, a fixed amount of market share, and good managerial expertise and skill. Credit facilities are completely covered by top-tier local bank guarantees and a score of an aggregate of 85 or more than that on the basis of the Risk Grade Score Sheet(RGSS).
- Acceptable (ACCPT) – This category of debtors is not as strong as people who fall under GOOD Grade and the grade above that grade, however still display consistent wages and cash flow. They also have an up-to-the-mark track record. Debtors have enough liquidity, cash flow and wages, allowable management, and parent/sister organization guarantee. Loans in this category would generally be retained by allowable collateral and a score of an aggregate between 75 – 84 on the basis of the RGSS.
- Marginal/Watch list (MG/WL) – This category of debtors have an above-average risk because of worn liquidity, which is higher than any normal leverage, thin money flow and/or not consistent wages, weaker business money, and early signals of warning of emerging companies loan detected. The debtor sees a loss, credit refund daily fall past due, conducting account is not good, or other unmentioned components are available. This kind of loan needs attention and a score of an aggregate of 65 – 74 on the basis of the RGSS.
- Special Mention (SM) – The people who fall in this category has possible flaws that need the authority's attention. If it is left not corrected, these flaws may bring the outcome of deterioration of the refund prospects of the debtor. A score of an aggregate of 55 – 64 is on the basis of the RGSS.
- Substandard (SS) – Economical position is weak and the capacity to refund the loan is in doubt of the authority. These flaws endanger the full loan settlement. A score of an aggregate 45-54 on the basis of the RGSS.
- Doubtful (DF) – Complete refund of interest and principal is not likely and the probability of loss is pretty high. But, because of specifically identifiable unresolved factors like liquidation procedures, capital injection or litigation, the

property is so far not been classified as Bad and Loss. A score of an aggregate 35 – 44 is on the basis of the RGSS.

- Bad & Loss (BL) – The loan of this group has zero advancements in getting a refund or is on the edge of wind-up or liquidation. The recovery chance is very low and legal choices have been taken. This group shows that it is yet to be desirable or practical to defer writing off this fundamentally not valued property even though part of the loan recovered may be affected in the future. A score of an aggregate of lesser than 35 is on the basis of the RGSS.

After this, the bank authority finalizes the loan agreement [6].

Trilok Nath Pandey and co. carried out research on predicting credit risk using ML methods. The main purpose of their study shows that it is possible to predict risky classes using machine learning. They used 9 ML algorithms (Bayesian Classifier, Naïve Bayes, Decision Tree, K-Nearest Neighbors, K-Means, Multilayer Perceptron, Extreme Learning Machine(ELM), Support Vector Machine(SVM), Artificial Neural Network(ANN)) to predict the risky applicants. They used German and Australian datasets, which had 20 and 14 attributes, in their research. It is seen ELM, which is a feedback single hidden layer feedforward neural network(NN), gives the highest accuracy (96.33 for the Australian Dataset and 96.32 for the German Dataset) for both datasets [7].

Huang and his team suggested a hybrid SVM-based loan scoring ML model, which explores the optimal ML model parameters and attribute subset to enhance the loan scoring accuracy [8].

The authors Kulkarni and Purohit, for the loan evaluation ML model, differentiated logistic regression(LR), multilayer perceptron(MLP) ML model, radial basis neural network(NN), SVM, and DT and they found that SVM, DT, and LR are the best ML models for predicting outcomes for classifying the credit applications [9].

NB, NN, and DT were utilized in Ahmed and Hamid's research for predicting loan risk. The output of this work indicated that the DT algorithm is the best ML algorithm on the basis of accuracy score [10].

Fisnik Doko and his team conducted research on the central bank dataset using five different ML algorithms (SVM, DT, RF, NN, and LR). They found that the ML model trained with the decision tree algorithm had the best accuracy while keeping the data imbalanced and without attribute scaling. RF and LR next two best algorithms based on accuracy [11].

2.3 Comparative Analysis and Summary

In this section, the result of this paper's model will be compared with other researchers' models. This will give an overview of the performances of all ML models to everyone. In summary, the best model that was found at the end of this research was trained using the random forest algorithm. If this model is compared with other models, it is quite visible that the score of the accuracy rate of the ML models in this paper is better than other researchers' models that have been mentioned in the table below:

TABLE 2.2: COMPARATIVE ANALYSIS

Author	Algorithm (Model)	Accuracy (In Percent)
Trilok Nath Pandey [7]	Decision Tree	90.72%
	Naïve Bayes	78.26%
	K-Nearest Neighbor	89.10%
	Logistic Regression	90.72%
	Support Vector Machine	85.94%
	Extreme Learning Machine (Best Model)	96.33%
Cheng-Lung Huang [8]	Support Vector Machine	80.00%
Seema Purohit [9]	Decision Tree (Best Model)	94.80%
	Logistic Regression	88.70%
	Support Vector Machine	88.80%
Aboobyda Jafar Hamid [10]	Naïve Bayes	73.87%
	j48 (Best Model)	78.38%

Fisnik Doko [11]	Random Forest (Best Model)	92.15%
	Decision Tree	92.05%
	Logistic Regression	92.00%
	Support Vector Machine	91.50%
My Result	Random Forest	97.35%
	Decision Tree	95.27%
	Naïve Bayes	91.29%
	K-Nearest Neighbor	89.77%
	Logistic Regression	96.59%
	Support Vector Machine	90.15%

2.4 Scope of the Problem

The first priority of any financial institution is to evaluate borrowers' applications to ensure that the applicant does return the money they took as debt. The safe return of the money depends on this step. However, this process takes a longer period of time. Different types of filtering and verification are done to thoroughly examine the application of the debtor. There are lots of complexity arises when financial institutions include many dimensions during the credit risk analysis. These measurements typically comprise financial data such as behavioral data of the person, or liquidity ratio such as loan payment behavior. Summarizing all of these different aspects into one score is very challenging, but tools of ML can help in achieving this objective.

The main goal of ML techniques and statistical learning methods is to learn from data and provide an outcome. The training dataset contains thousands of data to train the ML models and the testing dataset is utilized to verify how flawlessly the ML models can give a prediction. Both of the methods have the same aim and that is to analyze the underlying relationships by using a training dataset. Typically, statistical learning techniques presume formal relationships between features in mathematical equation forms, while machine learning models can learn from given data without needing rules-based programming. Because of this flexibility, machine learning techniques can discover patterns from data by maintaining high accuracy. Thus, the prediction that was found from the machine learning models is way faster than manual sorting, and based on the accuracy and selecting the best algorithm, risky borrowers can be detected easily and very quickly at the same time.

2.5 Challenges

The biggest challenge tackled during this study is finding relevant data. In Bangladesh, financial institutions tend to keep their data confidential. So, borrowers' real-time data could not be collected. Thus, this research entirely depended on the dataset found online [12].

The next problem that was tackled was selecting the appropriate dataset. Some features of this research were listed. However, it was hard to get a dataset with the attributes that needed to be used. Fortunately, a dataset that had the selected features was found. The dataset was found in Kaggle which was perfect for this research. It had lots of attributes in it. So, the unnecessary features need to be deleted. The quality of the dataset was also good and it contains enough data for this research.

During building the model, some challenges were tackled as well. As machine learning is a new topic, it was required to learn about the technology from the beginning. So, coding was quite difficult at times. Some problems were encountered during the feature selection stage. There was confusion while removing identical weighted attributes. Finally, selecting the best algorithm for this topic was also a difficult task to handle. Although some of the challenges were very critical, the problems were sorted out to conduct this research. In summary, the problems that were tackled during this study:

- Finding the loophole in the available topic
- Collecting relevant data in support of this research topic
- Cleaning the large dataset
- Selecting appropriate features
- Learning about the machine learning algorithms that would work best for this research
- Learning about the coding aspect of this research
- Choosing the algorithms and comparing their accuracy

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, the research subject who were targeted to collect data from and the way of collecting the data, and then the use of research instruments like Google Colaboratory, Python and its libraries, algorithms, etc. on the gathered data to make a dataset, will be elaborated. Statistical analysis on the prepared dataset will be done and then a methodology will be mentioned and used to progress in this research. The ways of detecting errors in the dataset and the ways of handling the errors will be mentioned in this chapter. The features that were selected will be described and the ways of filtering the features to get more important features among them will be explained. A summarization of the type of results that were searched for and the ways of evaluating the models will be mentioned in this chapter. At the end of this chapter, the implementation requirements of this research topic will be described.

3.2 Research Subject and Instrumentation

Research subjects and instruments are important parts of research. the research subject and research instrument will be discussed in two subsections.

3.2.1 Research Subject

A research subject is an entity that takes part in the research. Data is gathered from or about the subject to help solve the question under study. Credit risk is a problem that is related to financial institutions. Loans are given to borrowers by banks after agreeing to some terms and conditions. To get a loan, borrowers must apply for it. The banks go through some steps to validate the applications. The application form contains data about the borrowers. The borrowers are asked by the bank authority to submit more financial documents which they also validate. The final outcome is that the bank authority either approves or declines the application of the debtor depending on their financial condition and the amount of the loan they applied for. So, the data of the borrowers is the key factor. Without the data, the proper evaluation steps cannot take place. So, in this topic, the subject of the research is the borrowers/debtor who applies for a loan from a financial institution.

3.2.2 Research Instrumentation

A Research Instrument is a tool utilized to gather, measure, visualize, and analyze the data that are related to the study interests. In this section, the instruments like python, and its libraries (Pandas, NumPy, Seaborn, etc.), which were used in this research to clean, analyze and visualize data, and also to build the model, and evaluate the result, will be discussed. The algorithms (RF, DT, NB, KNN, LR, and SVM) will be discussed in this section.

Google Colaboratory: Google Colaboratory, "Colab" in short, is a web IDE for python programming language from Google Research. Colab allows everyone to write and execute python language code through any browser just by signing in to a Google account and it is mainly used for machine learning purposes, data analysis, etc. Colab is a very good instrument for data scientists who want to run ML and Deep Learning projects in the cloud [13].

Python: Python is a high-level programming language. Python is utilized to build websites and both desktop and phone software. Automating tasks can also be done using python. It is also used to carry out data analysis. Python is also known as a general-purpose language that has the power to create various programs and it is not specialized for any specific tasks. Nowadays, it is a must-learn language to do machine learning-related tasks. It is enriched with libraries that are used in data manipulation, data augmentation, data visualization, model training, and so on [14].

Pandas: Pandas is a python library for data manipulation and data analysis. Pandas have been one of the most popularly utilized instruments for Data Science and ML. Pandas is also the best instrument for handling messy data in the real world [15].

NumPy: NumPy is an open-source python library. NumPy can be utilized to perform different operations of mathematics on arrays. It adds data structures that are very to Python that are efficient in arrays and matrices calculation. It also has a big library that consists of mathematical functions of high-level that operate on arrays and matrices which is helpful in ML problem-solving [16].

Scikit-learn: Scikit-learn, also known as sklearn is one of the most useful python libraries that is utilized to do machine learning tasks. This library possesses a lot of effective instruments for both statistic modeling and ML including classification, regression, clustering, and dimensionality reduction. It can be utilized in data preprocessing. It is also used to train models using different algorithms after splitting the dataset into two sets (Training Data and Testing Data) [17].

Matplotlib: Matplotlib is a library of python for graphical plotting and data visualization that is a cross-platform library. It is the numerical extension NumPy. It can be a viable open-source alternative to MATLAB. The pyplot functions of Matplotlib can be used to make a change to figures: like creating graphs, creating a plotting area in a graph, and even plotting s various lines in the plotting area. It can also be utilized to decorate the plot with suitable labels, etc. [18].

Pydotplus: Pydotplus is also a library of python for graphical plotting and data visualization. It delivers a Python Interface to GDL (Graphviz's Dot Language) [19].

Seaborn: Seaborn(SNS) is also a library of python that is used for drawing statistical graphics. It is built on top of matplotlib. It also combines closely with the pandas library’s data structures. SNS helps to visualize and better understand the pattern of the data. It is also used to visualize random distributions [20].

Missingno: Missingno is a very good and very simple Python library that gives a series of data visualization functions that can be worked with to better understand the presence and distribution of data that are missing from the dataset. This missing value can be represented either in a barplot, matrix plot, heatmap, or a dendrogram [21].

Imbalanced-learn: Imbalanced-learn, also known as imblearn, is an open-source python library that provides functions to deal with the classification of imbalanced classes [22]. Imbalanced classes in machine learning are those classes where the number of instances of the values of the target variable is not the same. Suppose, in a binary classification problem, let's say, in this research's problem, it is necessary to find a risky applicant. So, the values of the target variable are either risky or not risky. It can be either said "Yes" or "No". Now, consider the instances of "Yes" are 50 not "No" are 10. This is an imbalanced dataset with an imbalanced target variable. To balance the dataset, oversampling, under-sampling, and SMOTE techniques are used to balance the dataset.

SMOTE: SMOTE means Synthetic Minority Oversampling Technique is a popular oversampling system that is utilized to handle imbalance classes. It has only one main task and that is to balance the class distribution by expanding minority class examples randomly by copying them. SMOTE incorporates new minority cases between existing minority cases [23].

Machine Learning: Machine learning, or ML, in short, is a sub-section of AI, which is defined as a way that a machine can imitate intelligent human behavior. AI systems are utilized to perform very complex jobs in a way that is very similar to how human beings can solve real-world problems. In data science, ML is utilized to learn the pattern from dataset data and give insights from that data after analyzing it. Some well-known algorithms are used to train a ML model [24]. In this study, algorithms will be used to train a model which will be used on a binary classification problem. The model will predict whether an application is risky or not depending on the data that was used to train it and the data that is sent to get a prediction.

Random Forest: Random Forest, or RF, in short, is a viral ML algorithm that is part of the supervised learning method. RF is utilized for both Regression and Classification problems in computer science. RF is mainly based on the ensemble learning concept, which is a procedure of combining more than one classifier to get a solution to a very complex problem and improve the accuracy rate thus improving the performance of the trained model of ML. RF is a classifier that has many decision trees on multiple subsets of the given dataset and it considers the average of DT data and improves the predictive accuracy and performance of that dataset. It does not rely on one DT. The RF algorithm takes prediction from each and every DT and on the basis of the majority votes of predictions, it gives the final outcome as a prediction. The higher the number of DT in the forest the higher accuracy and performance can be acquired. It also disallows the overfitting problem [25]. The diagram below shows the working of the RF algorithm:

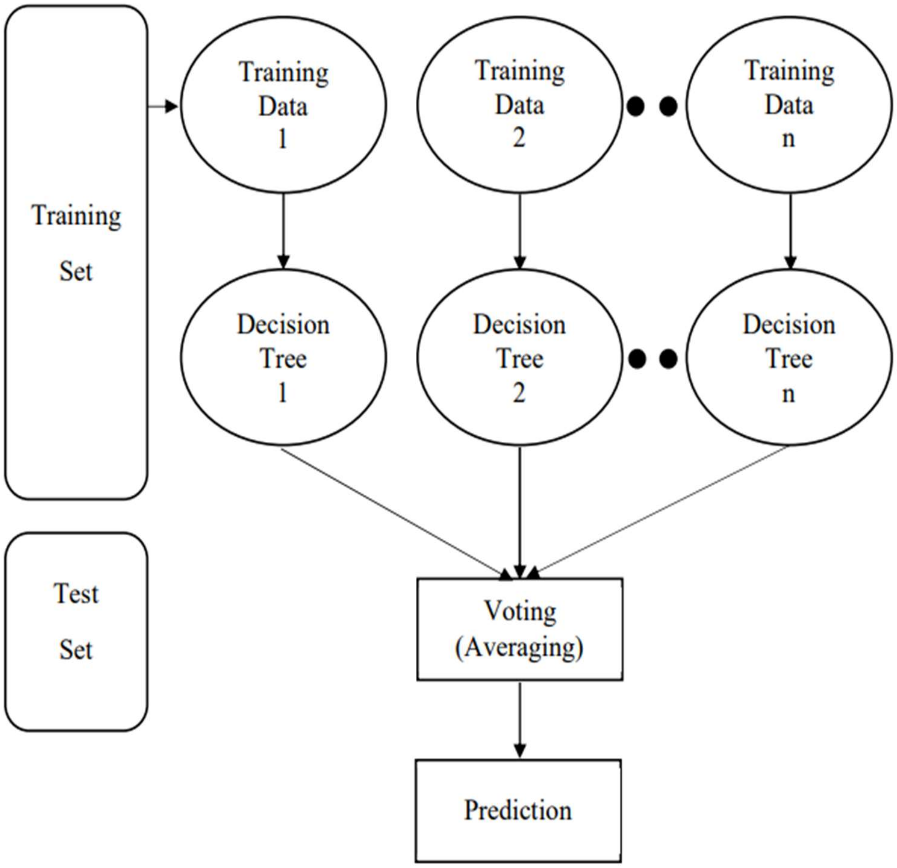


Figure 3.1: Random Forest

Decision Tree: Decision Tree, DT in short form, is also a Supervised learning(SL) method that can be utilized for both regression and classification problems; however, DT is mostly utilized for answering problems regarding classification. As the name suggests, the structure of this algorithm is just like a tree, where all internal nodes represent the attributes of a given dataset. The branches of the tree point to the decision rules and each and every leaf node of the tree points to different outcomes. In a DT, there are basically two nodes available, which are called the Decision Node, and the other one is called the Leaf Node. The decision nodes are utilized to make any kind of decision and it has more than one branch, on the other side, Leaf nodes are only the outputs that were acquired from the decisions and they do not have any other branch(es). The decisions are implemented on the basis of the attributes of the dataset. It is represented in graphical form for acquiring all the outcomes that are possible for a decision on the basis of the given conditions. The DT begins with the root/top node, which has branches and they construct a tree-like structure. A DT simply has a query and on the basis of the solution (Y/N), it further split the remaining branches into subtrees [26]. The diagram below shows the basic structure of a DT:

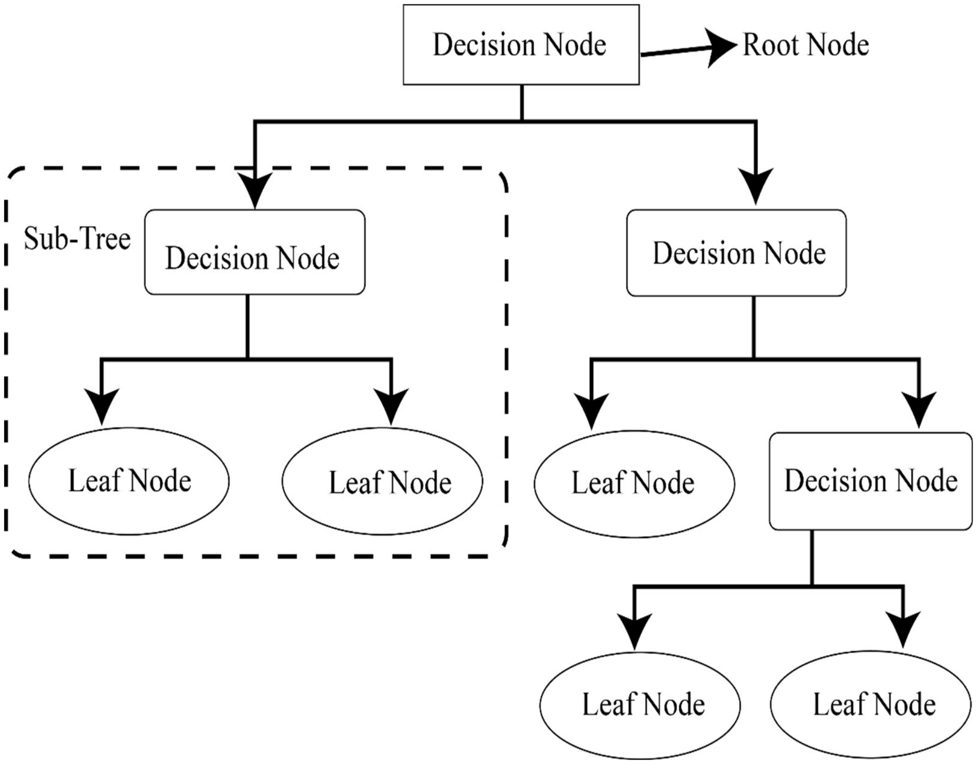


Figure 3.2: Decision Tree

Naïve Bayes: Naïve Bayes, NB is short, is an SL algorithm, which is made using the Bayes theorem and is utilized for answering problems related to classification. NB is a very simple and powerful algorithm for classification problems which helps in building faster ML models that can certainly make quicker predictions. It is known as a classifier of probabilistic because it gives prediction based on the probability of an object [27].

K-Nearest Neighbor: K-Nearest Neighbor, KNN in short, is also a simple ML algorithm that is again based on the SL method. KNN ML algorithm presumes the likeness between the new data and the data cases that are available and places the new case data into the group that is like the available groups. KNN ML algorithm keeps every data point that are available and classifies a new point of data on the basis of the similarity. This simply means when the appearance of new data then it can simply be classified into a well-suited group by utilizing the KNN ML algorithm. KNN ML algorithm can be utilized for both problems related to regression and classification. However, it is utilized for problems related to classification. KNN is a non-parametric ML algorithm. It means it does not assume underlying data. It doesn't immediately learn from the training dataset, instead, it keeps the entire dataset. At the time of classification, it carries out an action on that dataset, that is why it's called a lazy learner ML algorithm. At the training stage, the KNN ML algorithm just keeps the dataset and when data that is new appears, it classifies that new data point then and there into an existing group that is much like the new point of data [28].

Logistic Regression: Logistic regression, LR in short, is another most utilized ML algorithm, that comes under the ML SL method. It is utilized for predicting the categorical not independent feature by utilizing a given set of independent features. LR makes a prediction of the result of a categorical not independent feature. So, the result must be either a categorical or a discrete value. It is also can only be Y or N, in binary, 1 or 0, or Boolean value, True or False, etc. However, instead of giving the required value as 1 and 0, it gives the values of probabilistic. That value lies between 1 and 0. LR is utilized for answering problems related to classification. In LR, a logistic function that is shaped like "S" is fitted instead of fitting a line of regression, which makes a prediction of two maximum values (1 or 0). The curve from the logistic function tells the probability of something such as whether the cells are cancerous or not. LR is an outstanding ML algorithm. It has the capability to give predictions and classify new point of data by utilizing continuous and discrete datasets. LR can be utilized to classify the observations by utilizing various data types and can very easily determine the most useful features utilized for the classification [29]. The image below is displaying the logistic function:

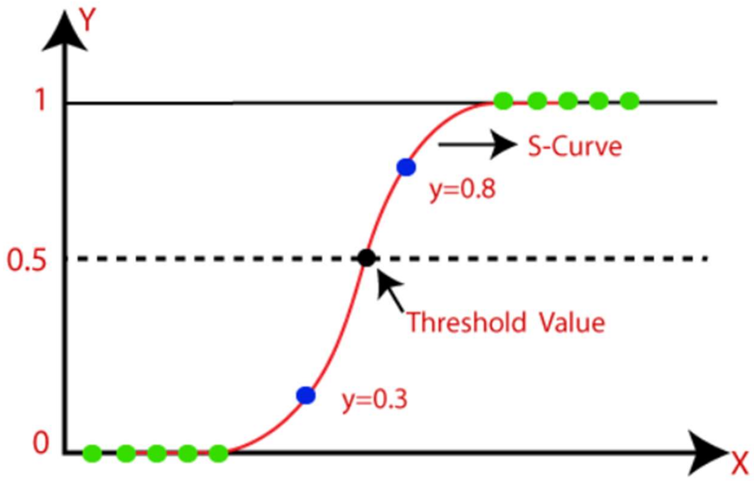


Figure 3.3: Logistic Function

Support Vector Machine: Support Vector Machine, SVM in short, is also a popular SL algorithm, which is utilized for the problems of regression and classification. But SVM has mainly utilized problems related to classification in ML. The goal of SVM ML is to graph the line that is best that can segregate n-dimensional space into more than one class so that it can be easy to put newer data points in the actual group in the time ahead. This line that is best is known as a hyperplane. The SVM algorithm chooses the extreme vectors that assist to graph that hyperplane. These points are popularly known as support vectors, and that is why the algorithm is called SVM [30].

3.3 Data Collection Procedure

Data collection is an essential part of any research. The data is analyzed to find something that is undiscovered. The case is the same for machine learning-related problems. Data is used to train ML models using training data. Some data is kept for testing purposes. Secondary data was collected from Kaggle for this research. It was not possible to get primary data from financial institutions. They follow the rule of keeping their client's data confidential strictly. After collecting the data, a dataset consisting of features was formed and then get it ready for preprocessing stage. Again, in this paper's case, a dataset was not required to be made because the dataset was already acquired from Kaggle. Although the dataset was collected from Kaggle, some questions were considered before selecting the dataset. The questions that were considered are mentioned in the next page:

1. What is their annual income?
2. What is the amount of loan they are applying for?
3. What is their purpose for wanting to take a loan?
4. For how long do they want to take a loan?
5. For how many years they have been working?
6. Is their income low, medium, or high with respect to the applied loan?
7. What is their home status?
8. What is the interest rate of the applied loan?
9. Is the interest amount of the applied loan high or low according to their salary?
10. Which grade does the applicant fall?
11. What is the debt-to-income ratio for an applicant?
12. What is the monthly installment amount?
13. What is the amount that is to be repaid at the termination date?
14. What is the total recovery principal amount?
15. What is the amount that was recovered at the termination date?

Exploratory Data Analysis: Exploratory analysis is the step of reading the data from the dataset and then exploring the features. The knowledge is very important about the total number of attributes and the total number of records the dataset holds, what type of data they hold, and the range of values of each variable they take on. In the selected dataset, there were 1343 records with 24 features. The features will be described in detail. The features are:

id: This feature holds unique numerical values which identify an individual client.

emp_length_int: This feature also holds the numerical values of the length of employment for clients. For simplicity, A value of 10 for 10 years of experience and 0.5 for 5 months of experience was taken. Every record for this variable is converted this way. It does not hold unique values as id.

home_ownership: This feature holds categorical values of home_ownership for each client. The values are RENT, OWN, and MORTGAGE. The values for all records are not unique.

home_ownership_cat: The feature holds the numerical values which were acquired from converting the values of home_ownership. RENT = 1, OWN = 2, and MORTGAGE = 3 were taken. The values for each record are not unique which means they are seen more than once.

income_category: This feature holds categorical values of the income category of each client. The values are Low, Medium, and High. It indicates in which category the client falls. The values for all records are not unique.

annual_inc: This feature holds numerical values that indicate the annual income of the client. In this dataset, the values are within the range of 6000 – 1782000. The values for all records are not unique.

income_cat: This feature holds numerical values which were acquired from converting the values of income_category. Low = 1, Medium = 2, and High = 3 were taken. The values for each record are not unique.

loan_amount: This feature holds numerical values of the amount of loan the clients applied for. In this dataset, the values are within the range of 1000 – 35000. The values for all records are not unique.

term: This feature holds categorical values of the time for which the clients want to take the loan. In this paper's dataset, there are two values for this feature. The values are 36 month and 60 month. The values for all records are not unique.

term_cat: This feature holds numerical values which were acquired by converting the values of the term. 36 month = 1, 60 month = 2 was taken. The records are not unique.
purpose: This feature holds the categorical value of the reason why the clients borrow the money. The values are credit card, debt consolidation, car, other, medical, home improvement, purpose, moving, major purchase, vacation, small business, house, wedding, and renewable energy. The values for each record are not unique.

purpose_cat: This feature holds numerical values which were acquired from converting the values of purpose. credit_card = 1, car = 2, small_business = 3, other = 4, wedding = 5, debt_consolidation = 6, home_improvement_purpose = 7, major_purchase = 8, medical = 9, moving = 10, vacation = 11, house = 12, and neweable_energy = 13 was taken. The values for each record are not unique.

interest_payments: This feature holds categorical values of how high or low the amount of the loan is. The values are Low and High.

interest_payment_cat: This feature holds numerical values which were acquired from converting the values of interest_payment. Low = 1, and High = 2 were taken. The values for each record are not unique.

interest_rate: This feature holds numerical values that indicate the rate of interest on the loanable amount of money. In this dataset, the values are within the range of 5.42 – 24.59. The values for all records are not unique.

grade: This feature holds categorical values that indicate the grade in which the client falls. Grade mainly divides the client into groups according to the interest rate. The grades are A, B, C, D, E, F, and G. The values for all records are not unique.

grade_cat: This feature holds numerical values which were acquired from converting the values of grade_cat. A = 1, B = 2, C = 3, D = 4, E = 5, F = 6, and G = 7 was taken. The values for each record are not unique.

dti: DTI means Debt-To-Income ratio which indicates the amount of money a client owes to a financial institution each month according to the amount of money he/she earns every month. This feature holds numerical values. In this dataset, the values are within the range of 0.00 – 29.99. The values for all records are not unique.

total_pymnt: This feature holds numerical values which indicate the amount of money that needs to be repaid at the end of the loan term. The repaid amount will be the money that was taken as a loan and the interest on that same amount. In this dataset, the values are within the range of 0.00 – 50110.74. The values for all records are not unique.

total_rec_prncp: This feature holds numerical values which indicate the amount of money that will be received by the financial institution after principal loss in respect of an asset of the borrower. In this dataset, the values are within the range of 4725.00 – 35000.03. The values for all records are not unique.

recoveries: This feature holds numerical values which indicate the amount that was received from the borrower, by the financial institution, prior to the termination date. In this dataset, the values are within the range of 0.00 – 22943.37. The values for all records are not unique.

installment: This feature holds numerical values which indicate the monthly payment that needs to be paid by the clients as part of the installment policy of the financial institution. In this dataset, the values are within the range of 32.23 – 1283.50. The values for all records are not unique.

loan_condition: This feature holds categorical values which indicate whether the client is eligible for a loan grant. This is the target feature or dependent feature for this research. Two types of values are held by this attribute. They are Good Loan and Bad Loan. This attribute will not be used during the programming part. The values will be converted into numerical values for simplicity.

loan_condition_cat: This feature holds numerical values which were acquired from converting the values of loan_condition. Good Loan = 0, and Bad Loan = 1 was taken. The records are not unique. This attribute will be used during the programming part.

Removing Unnecessary Attributes: In this dataset, 24 attributes were present. 22 of them are independent variables and 2 of them are dependent variables. Although there were 24 attributes, some of the attributes are conversions from categorical values to numerical values of some other attributes in the dataset. The home_ownership_cat is the conversion into the numerical value of the home_ownership attribute which holds categorical values. Similar attributes are income_category (categorical) and income_cat (numerical), the term (categorical) and term_cat (numerical), purpose (categorical) and purpose_cat (numerical), interest_payments (categorical) and interest_payment_cat (numerical), grade (categorical) and grade_cat (numerical). They are independent attributes. Even the dependent attribute loan_condition has categorical values which were converted into numerical values and that attribute is loan_condition_cat. The attributes that hold categorical values will not be used for programming cases. So, home_ownership, income_category, term, purpose, interest_payments, grade, and loan_condition attributes will be dropped. Also, the id attribute will be dropped as well since it will make no sense to use the id of an applicant. Therefore, 16 attributes will be used 15 of which are independent features (emp_length_int, home_ownership_cat, annual_inc, income_cat, loan_amount, term_cat, purpose_cat, interest_payment_cat, interest_rate, grade_cat, dti, total_pymnt, total_rec_prncp, recoveries, and installment) and 1 dependent feature (loan_condition_cat).

The description of all attributes is summarized in the table below:

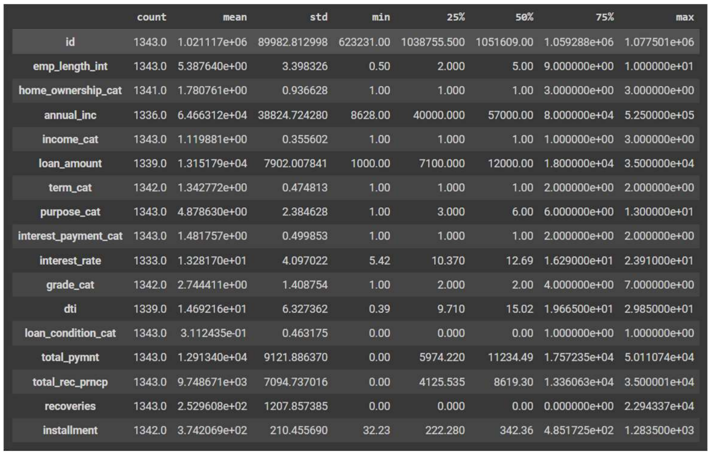
TABLE 3.1: SELECTED ATTRIBUTES' DESCRIPTION

Attribute	**Description**	**Type**
Employment Length (emp_length_int)	Applicant's employment status over the years	Numeric
Home Ownership (home_ownership)	Values: Rent, Own, Mortgage	Categorical
Income Category (income_category)	Values: High, Medium, Low	Categorial
Annual Income (annual_inc)	Applicant's yearly salary	Numeric
Loan Amount (loan_amount)	Applicant's applied loan amount	Numeric
Term (term)	36 Months (Short Term), 60 Months (Long Term)	Categorical
Purpose (purpose)	Applicant's reason for loan application	Categorical
Interest Payments (interest_payments)	Values: Low, High	Categorical
Interest Rate (interest_rate)	Interest on applied loan	Numeric

Grade (grade)	Clients are grouped into different categories according to interest rate Value: A, B, C, D, E, F, and G	Categorical
DTI (dti)	The ratio of how much an applicant owes to how much he earns	Numeric
Total Payment (total_pymnt)	Total payable amount with interest at the end of the agreement	Numeric
Total Recovery Principal (total_rec_prncp)	Principal payment received in respect of an Asset after a Principal Loss	Numeric
Recoveries (recoveries)	The amount that was recovered at the end of the contract	Numeric
Installment (installment)	The monthly amount to be paid	Numeric
Loan Condition (loan_condition_cat) (Dependent Feature)	Indicates whether a client is risky or not to issue a loan Values: 0 (Good Loan), 1 (Bad Loan)	Numeric

3.4 Statistical Analysis

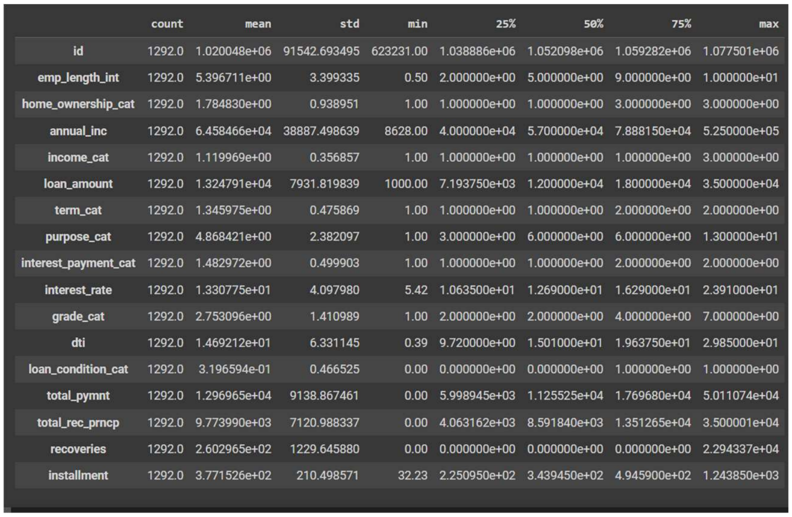
The dataset will be visualized and insight from each feature will be shown.



	count	mean	std	min	25%	50%	75%	max
id	1343.0	1.021117e+06	89982.812998	623231.00	1038755.500	1051609.00	1.059288e+06	1.077501e+06
emp_length_int	1343.0	5.387640e+00	3.398326	0.50	2.000	5.00	9.000000e+00	1.000000e+01
home_ownership_cat	1341.0	1.780761e+00	0.936628	1.00	1.000	1.00	3.000000e+00	3.000000e+00
annual_inc	1336.0	6.466312e+04	38824.724280	8628.00	40000.000	57000.00	8.000000e+04	5.250000e+05
income_cat	1343.0	1.119881e+00	0.355602	1.00	1.000	1.00	1.000000e+00	3.000000e+00
loan_amount	1339.0	1.315179e+04	7902.007841	1000.00	7100.000	12000.00	1.800000e+04	3.500000e+04
term_cat	1342.0	1.342772e+00	0.474813	1.00	1.000	1.00	2.000000e+00	2.000000e+00
purpose_cat	1343.0	4.878630e+00	2.384628	1.00	3.000	6.00	6.000000e+00	1.300000e+01
interest_payment_cat	1343.0	1.481757e+00	0.499853	1.00	1.000	1.00	2.000000e+00	2.000000e+00
interest_rate	1333.0	1.328170e+01	4.097022	5.42	10.370	12.69	1.629000e+01	2.391000e+01
grade_cat	1342.0	2.744411e+00	1.408754	1.00	2.000	2.00	4.000000e+00	7.000000e+00
dti	1339.0	1.469216e+01	6.327362	0.39	9.710	15.02	1.966500e+01	2.985000e+01
loan_condition_cat	1343.0	3.112435e-01	0.463175	0.00	0.000	0.00	1.000000e+00	1.000000e+00
total_pymnt	1343.0	1.291340e+04	9121.886370	0.00	5974.220	11234.49	1.757235e+04	5.011074e+04
total_rec_prncp	1343.0	9.748671e+03	7094.737016	0.00	4125.535	8619.30	1.336063e+04	3.500001e+04
recoveries	1343.0	2.529608e+02	1207.857385	0.00	0.000	0.00	0.000000e+00	2.294337e+04
installment	1342.0	3.742069e+02	210.455690	32.23	222.280	342.36	4.851725e+02	1.283500e+03

Figure 3.4: Dataset Description (Before data cleaning)

From figure 3.4, it can be seen that the total values within every attribute, the average values of all records of the features along with the standard deviation of the attributes. The highest and the lowest values within every attribute can also be seen. This may provide some misconceptions because the dataset is not cleaned and there may be missing values, duplicate values, and outliers present in the dataset.



	count	mean	std	min	25%	50%	75%	max
id	1292.0	1.020048e+06	91542.693495	623231.00	1.038886e+06	1.052098e+06	1.059282e+06	1.077501e+06
emp_length_int	1292.0	5.396711e+00	3.399335	0.50	2.000000e+00	5.000000e+00	9.000000e+00	1.000000e+01
home_ownership_cat	1292.0	1.784830e+00	0.938951	1.00	1.000000e+00	1.000000e+00	3.000000e+00	3.000000e+00
annual_inc	1292.0	6.458466e+04	38887.498639	8628.00	4.000000e+04	5.700000e+04	7.888150e+04	5.250000e+05
income_cat	1292.0	1.119969e+00	0.356857	1.00	1.000000e+00	1.000000e+00	1.000000e+00	3.000000e+00
loan_amount	1292.0	1.324791e+04	7931.819839	1000.00	7.193750e+03	1.200000e+04	1.800000e+04	3.500000e+04
term_cat	1292.0	1.345975e+00	0.475869	1.00	1.000000e+00	1.000000e+00	2.000000e+00	2.000000e+00
purpose_cat	1292.0	4.868421e+00	2.382097	1.00	3.000000e+00	6.000000e+00	6.000000e+00	1.300000e+01
interest_payment_cat	1292.0	1.482972e+00	0.499903	1.00	1.000000e+00	1.000000e+00	2.000000e+00	2.000000e+00
interest_rate	1292.0	1.330775e+01	4.097980	5.42	1.063500e+01	1.269000e+01	1.629000e+01	2.391000e+01
grade_cat	1292.0	2.753096e+00	1.410989	1.00	2.000000e+00	2.000000e+00	4.000000e+00	7.000000e+00
dti	1292.0	1.469212e+01	6.331145	0.39	9.720000e+00	1.501000e+01	1.963750e+01	2.985000e+01
loan_condition_cat	1292.0	3.196594e-01	0.466525	0.00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
total_pymnt	1292.0	1.296965e+04	9138.867461	0.00	5.998945e+03	1.125525e+04	1.769680e+04	5.011074e+04
total_rec_prncp	1292.0	9.773990e+03	7120.988337	0.00	4.063162e+03	8.591840e+03	1.351265e+04	3.500001e+04
recoveries	1292.0	2.602965e+02	1229.645880	0.00	0.000000e+00	0.000000e+00	0.000000e+00	2.294337e+04
installment	1292.0	3.771526e+02	210.498571	32.23	2.250950e+02	3.439450e+02	4.945900e+02	1.243850e+03

Figure 3.5: Dataset Description (After data cleaning)

From figure 3.5, it can be seen that the total values within every attribute, the average values of all records of the features along with the standard deviation of the attributes. The highest and the lowest values within every attribute can also be seen. It can be followed to detect all the missing values, duplicate values, and outliers properly.

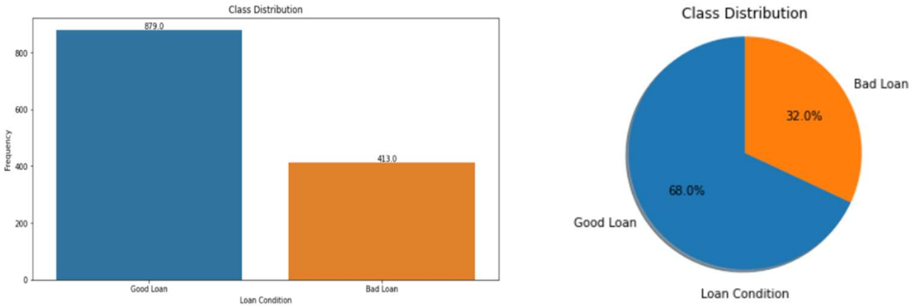


Figure 3.6: Class Distribution

From figure 3.6, it can be seen that the good loan class has 879 instances and the bad loan class has 413 instances in number. So, 68% of the total instances are classified as good loans and 32% of the total instances are classified as bad loans.

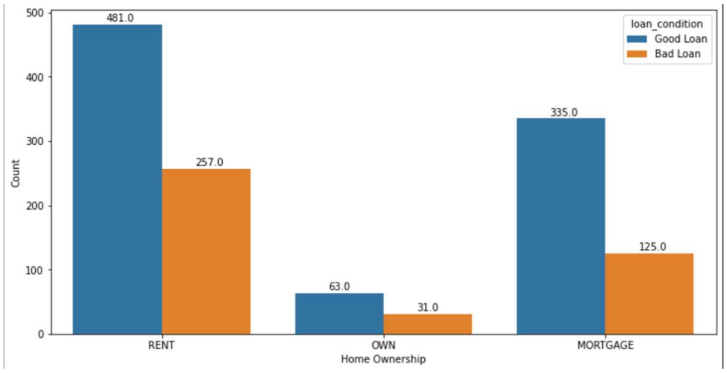


Figure 3.7: Risky Loan by Home Ownership

From figure 3.7, it can be seen that people who live in rented houses take more loan than those who own a house and takes loan by agreeing to give property to the financial institution if they fail to repay the money in time. From the graph, it can be seen that 57.12% of people who rent a house, 7.28% of people who owns a house, and 35.60% of people who mortgage their house take a loan. It can also be seen that the chances of people who live in a rented house returning the loan are 65.18% and not returning the loan is 34.82%. For, people who own a house, the good loan percentage is 67.02% and the bad loan percentage is 32.98%. People who fail in the mortgage category have a 73.82% chance of returning the loan and a 27.17% chance of not returning the loan. So, people who mortgage a home are more likely to return the money compared to people who rent a house and who own a house.

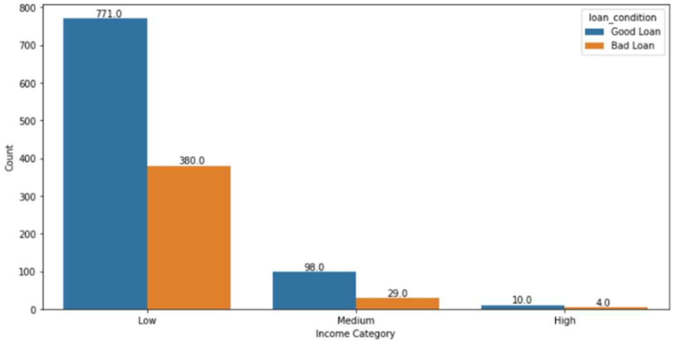


Figure 3.8: Risky Loan by Income Category

From figure 3.8, it can be seen that people with low income tend to apply for a loan more than those who earn a moderate salary and whose salary is high. From the graph, it can be seen that 89.09% of people with low income take a loan. The percentage is 9.83% for people with medium salaries, and 1.08% for people with high income. It can also be seen that the chances of people with low income who take a loan and return the loan are 66.99% and not returning the loan are 33.01%. People with medium income have a 77.17% of returning the loan and those not return the loan percentage is 22.83%. People with high income don't generally take a loan, but those who do take a loan are 71.43% likely to return the loan and have only a 28.57% of probability of not returning the loan.

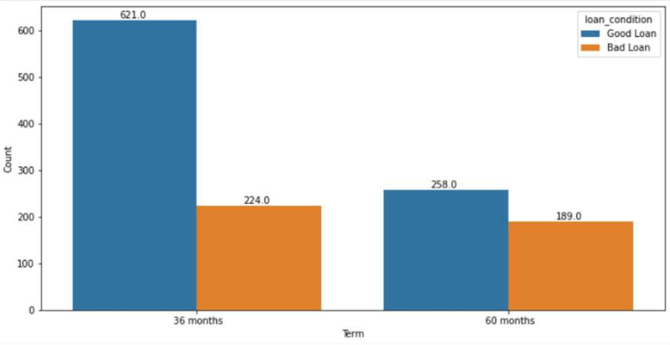


Figure 3.9: Risky Loan by Term

From figure 3.9, it can be seen that people take loans for a shorter amount of time. The percentage of short-term loans is 65.40% and 34.60% for longer loans. People who take short loans have a 73.49% chance of returning the loan and a 26.51% chance of not returning the loan. People taking long-term loans have a 57.72% chance of returning the loan and a 42.28% chance of not returning the loan.

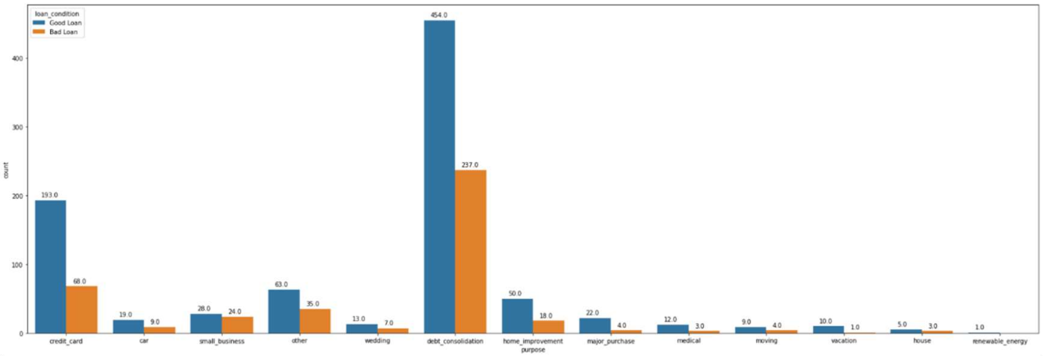


Figure 3.10: Risky Loan by Purpose

From figure 3.10, it can be seen that people take more loans for debt consolidation and fewer loans for renewable energy. People also take a great number of loans on credit cards.

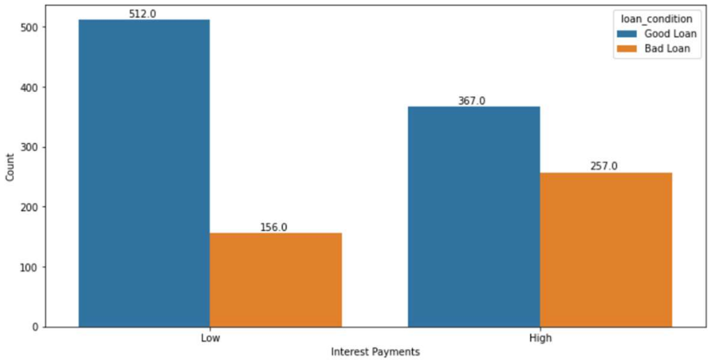


Figure 3.11: Risky Loan by Interest Payment

From figure 3.11, it can be seen people when the interest is low the chances of people taking a loan are 51.70% and when the interest is high the chances are 48.30%. When the interest is low 76.65% of the people repay the loan and 23.35% of the people do not repay the loan. For high-interest rates, 58.81% of people repay the loan, and 41.19% of people don't repay the loan. So, when the interest rate is low the chances are high of people repaying the loan and chances are low when the interest rate is high.

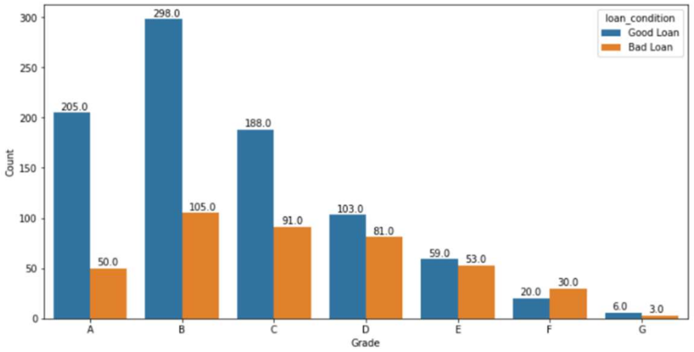


Figure 3.12: Risky Loan by Grade

From figure 3.12, it can be seen people who are in B grade takes loan more than other grades. The percentage for Grade A is 19.74, Grade B is 31.19, Grade C is 21.59, Grade D is 14.24, Grade E is 8.67, Grade F is 3.87, and Grade G is 0.70. The chances of the people who are Grade A repaying the loan are 80.39% and those not repaying the loan are 19.61%. The chances of the people who are Grade B repaying the loan are 73.95% and those not repaying the loan are 26.05%. The chances of the people being Grade C repaying the loan are 67.38% and those not repaying the loan are 32.62%. The chances of the people are Grade D repaying the loan are 55.98% and those not repaying the loan are 44.02%. The chances of the people who are Grade E repaying the loan are 52.68%

and not repaying the loan is 47.32%. The chances of the people being Grade F repaying the loan are 40.00% and those not repaying the loan are 60.00%. The chances of the people are Grade G repaying the loan are 66.67% and those not repaying the loan are 33.33%. So, People who fall into Grade A have a very high chance of paying back the loan, and people who fall into Grade F have a very high chance of not paying back the loan.

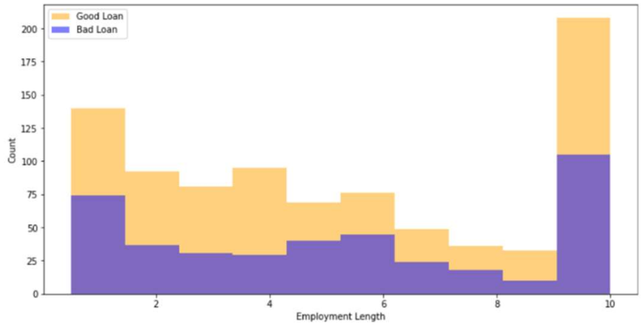


Figure 3.13: Risky Loan by Employment Length

From figure 3.13, it is seen that people employed for less than 4.5 and more than 9 years have a high rate of taking loans. However, people that are employed for more than 4.5 and less than 9 years have a high chance of not repaying the loan. People who employed for more than 1.5 to less than 4.5 years have a high chance of returning the loan.

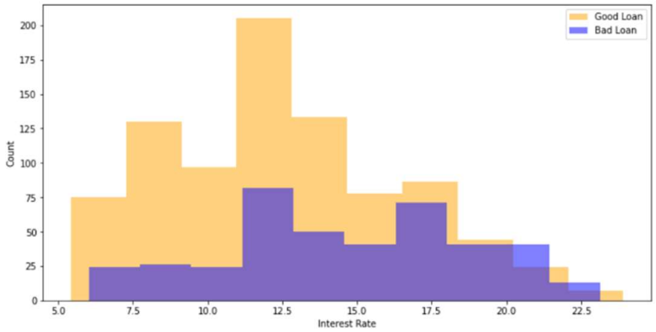


Figure 3.14: Risky Loan by Interest Rate

From figure 3.14, it is seen that people like taking loans with less interest rates, and the chances of the loan being repaid are loan. People take loans when the interest rate is between 5.8 – 17.8. The chances of the loan being not repaid are high when the interest rate is between 20 – 21 and 22 – 22.5. However, there are some instances when the interest rate is over 23, people repay the loan successfully.

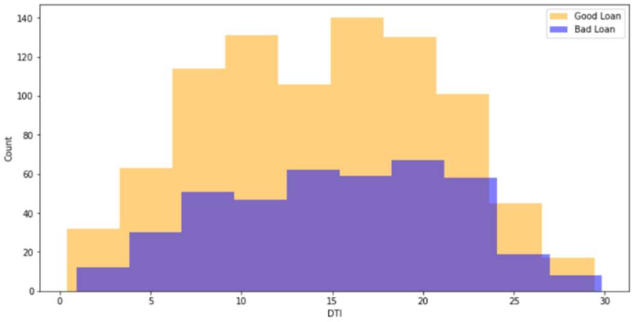


Figure 3.15: Risky Loan by Debt-To-Income

From figure 3.15, it is seen that for people who take a loan from a financial institution the debt-to-income is between 7 and 24 people. The chances of the loan being repaid are high. However, when the debt-to-income gets over 28, there is a risk that the loan might not be repaid and at that, there are not many instances of people taking a loan.

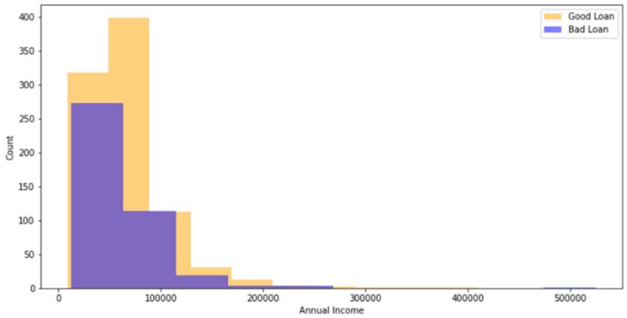


Figure 3.16: Risky Loan by Annual Income

From figure 3.16, it is seen that when the annual income of a person is 100000 or less that people like to take loans. People with higher salaries don’t take loans. The chances of the loan being repaid by low-income people are quite good. However, people who have a salary of over 200000 have a little tendency to not repay the loan.

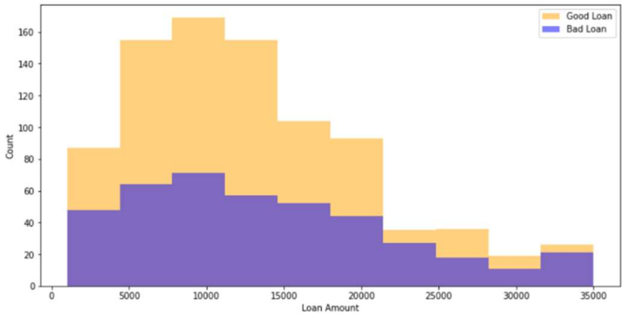


Figure 3.17: Risky Loan by Loan Amount

From figure 3.17, it is seen that when the loan amount is high the probability of a bad loan decreases. The chances of a bad loan are very high when the total payment amount is very low. The chances of a bad loan decrease when the total principal amount enters 12000 and it almost disappears when the amount is over 30000. People like to take a small amount of loan as the graph suggests when the loan amount increases people are less likely to take a loan.

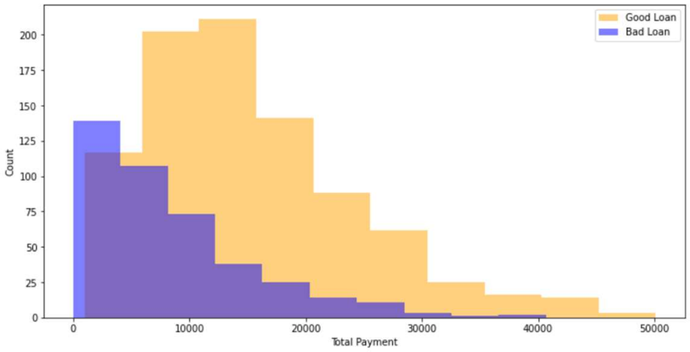


Figure 3.18: Risky Loan by Total Payment

From figure 3.18, it is seen that when the total payment amount is high the probability of a bad loan decreases. The chances of a bad loan are very high when the total payment amount is very low. The chances of a bad loan decrease when the total principal amount enters 12000 and it almost disappears when the amount is over 30000. People like to take loans when the total payment amount is high.

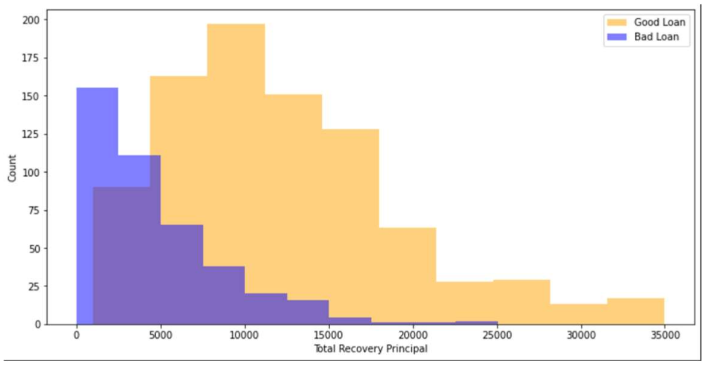


Figure 3.19: Risky Loan by Total Recovery Principal

From figure 3.19, it can be seen that when the total recovery principal amount is high the probability of a bad loan decreases. The chances of a bad loan are very high when the total recovery principal amount is very low. The chances of a bad loan decrease when the total principal amount enters 2800 and it almost disappears when the amount is over 15000.

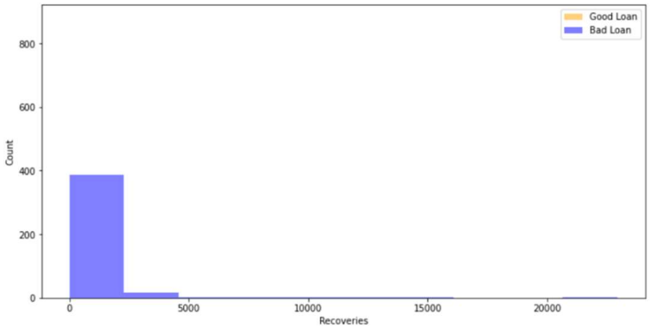


Figure 3.20: Risky Loan by Recoveries

From figure 3.20, it is clear that the recovery amount is between 0 to less than 5000 and the loan not getting repaid is very high.

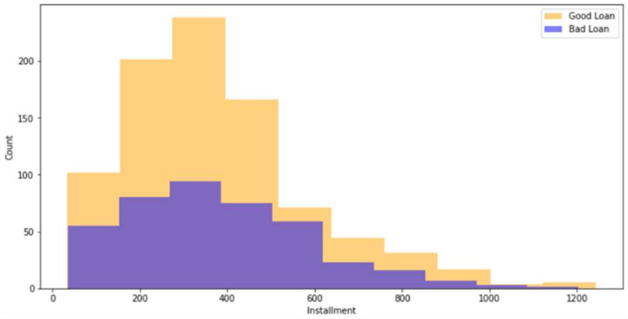


Figure 3.21: Risky Loan by Installment

From figure 3.21, it is clear that people who take loans have a monthly installment amount between 180 – 500 and they seem to repay the loan within time. When the monthly installment amount increases, people tend to not take a loan from a bank, and at 1200 installment amount or more the chances of the loan getting repaid decrease. People like to take loans when the monthly installment is less.

3.5 Proposed Methodology

Research methodology is the specific method utilized to discover, select, process, and analyze data about a topic.

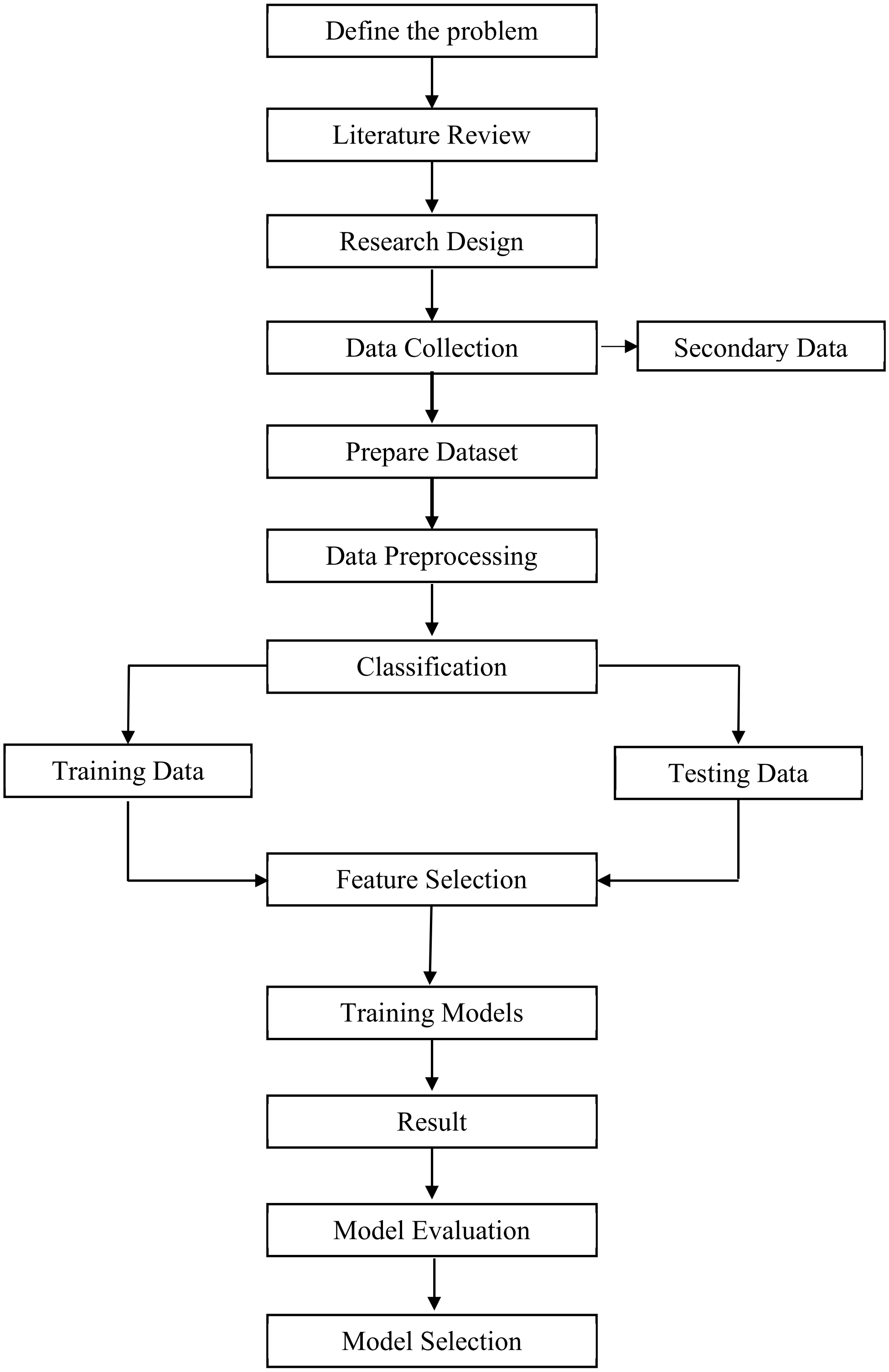


Figure 3.22: Research Methodology

In a research-related paper, the methodology section tells the reader to how to critically evaluate that study's overall validity and dependability.

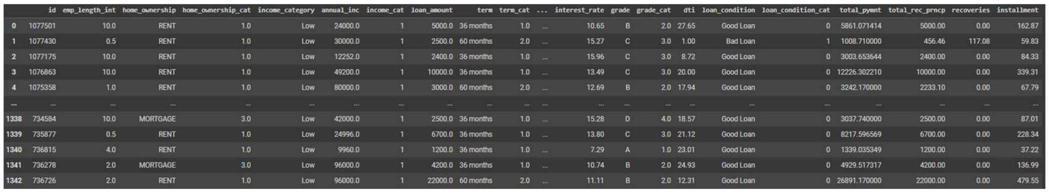
Define the problem: The first step of doing research is to select a topic. Then the existing problem within that research area is needed to be found. In this research, it was detected that financial institution in Bangladesh does not use technology to detect loan risk. Even the available paper on credit risk detection has some flaws which will be answered in this paper.

Literature Review: This step is also very important. Literature Review means studying the existing work in the related research field. Some works have already been done by other researchers. By studying their findings, new research can be conducted. Those results are used to find something new or try to improve what they had found. Generally, some gaps in the existing research are needed to be found. In this research, some works of some researchers were studied. In this paper's case, the accuracy rate of the algorithm models was upgraded by inputting new features to predict whether a borrower is risky or not to lend a loan. The papers that were studied were discussed in the Related Works section of chapter 2.

Research Design: Research refers to the complete plan that a researcher uses while conducting his/her research. The research design stage consists of the steps from making a questionnaire for collecting data to data interpretation, analysis, and discussion of data and finally getting insights from that gathered data. In this case, a questionnaire was not required to be made to gather data. However, some questions were considered while preparing the dataset. During the research design phase, the following: the purpose of the statement, data collection techniques, Methodology, Data Preprocessing, Data Analyzing, Data Interpretation, Training Model, Reviewing Results, and Evaluating the results were considered.

Data Collection: Data collection is the step where information is gathered from the targeted individuals who are the core aspect of the research. First, some questions regarding the problem that is being solved are made. After that, a survey is conducted to collect data from the targeted subjects. Data that is collected this way is called primary data. Generally, data is of two types: primary data and secondary data. For this research, secondary data was collected. It is collected online, more specifically it is collected from Kaggle in the form of a dataset. It was not possible to get primary data because in Bangladesh financial institutes keep their data confidential and don't share it with anyone. So, this research was dependent on the data which was found online [12]. Although the data was collected online, it was good enough to do the research on this topic.

Prepare Dataset: In the data collection step, data is collected by conducting a survey. The questions are based on the attributes that have been selected before making the questionnaire. After collecting data from the subjects, a dataset is formed. The dataset may contain numerical data or categorical data. MS Excel is used to put the data that have been gathered. That is the dataset for this research. After that, the excel file is converted into a CSV file so that it can be read from IDE. IDE doesn't support any normal file format. Since data was not collected by conducting a survey, a full dataset was not needed to be prepared. Some attributes that were not needed for this research were deleted. After the deletion of the unwanted attributes, the dataset was ready. The attributes that were selected were discussed in the data collection procedure section of this chapter. After this step, the data can be processed from IDE. In the dataset, there were 1343 records and 24 features.



	id	emp_length_int	home_ownership	home_ownership_cat	income_category	annual_inc	income_cat	loan_amount	term	term_cat	...	interest_rate	grade	grade_cat	dti	loan_condition	loan_condition_cat	total_pymnt	total_rec_prncp	recoveries	installment
0	1077501	10.0	RENT	1.0	Low	24000.0	1	5000.0	36 months	1.0	...	10.65	B	2.0	27.65	Good Loan	0	5861.071414	5000.00	0.00	162.87
1	1077430	0.5	RENT	1.0	Low	30000.0	1	2500.0	60 months	2.0	...	15.27	C	3.0	1.00	Bad Loan	1	1008.710000	456.46	117.08	59.83
2	1077175	10.0	RENT	1.0	Low	12252.0	1	2400.0	36 months	1.0	...	15.96	C	3.0	8.72	Good Loan	0	3003.653644	2400.00	0.00	84.33
3	1076863	10.0	RENT	1.0	Low	49200.0	1	10000.0	36 months	1.0	...	13.49	C	3.0	20.00	Good Loan	0	12226.302210	10000.00	0.00	339.31
4	1075358	1.0	RENT	1.0	Low	80000.0	1	3000.0	60 months	2.0	...	12.69	B	2.0	17.94	Good Loan	0	3242.170000	2233.10	0.00	67.79
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1338	734584	10.0	MORTGAGE	3.0	Low	42000.0	1	2500.0	36 months	1.0	...	15.28	D	4.0	18.57	Good Loan	0	3037.740000	2500.00	0.00	87.01
1339	735877	0.5	RENT	1.0	Low	24996.0	1	6700.0	36 months	1.0	...	13.80	C	3.0	21.12	Good Loan	0	8217.596569	6700.00	0.00	228.34
1340	736815	4.0	RENT	1.0	Low	9960.0	1	1200.0	36 months	1.0	...	7.29	A	1.0	23.01	Good Loan	0	1339.035349	1200.00	0.00	37.22
1341	736278	2.0	MORTGAGE	3.0	Low	96000.0	1	4200.0	36 months	1.0	...	10.74	B	2.0	24.93	Good Loan	0	4929.517317	4200.00	0.00	136.99
1342	736726	2.0	RENT	1.0	Low	96000.0	1	22000.0	60 months	2.0	...	11.11	B	2.0	12.31	Good Loan	0	26891.170000	22000.00	0.00	479.55

Figure 3.23: Dataset

Data Preprocessing: In this phase, the data is selected to be cleaned. Because in a dataset there may be some missing values, some duplicate values, and some outliers present. Also, the dataset may be imbalanced too. Feature scaling is also done to make the performance of the model better than before. Without preprocessing data, the dataset will not be flawless. Thus, it is more likely that most of the time wrong output will be found because the model will be trained with flawed data. Here, all the mentioned scenarios will be discussed.

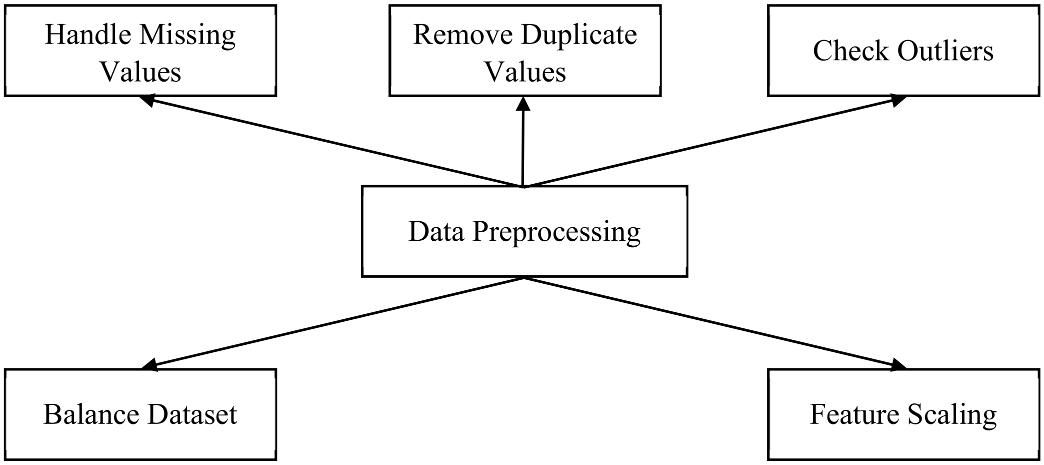


Figure 3.24: Data Preprocessing Steps

Handle Missing Values: When data was collected from the research subjects, some of them don't want to answer all the questions due to privacy. So, some fields of a dataset remain empty. That is also true for the online dataset. So, null fields or empty record fields should be handled. To view missing values, isnull() function and missingno library of python language and a graph are used.

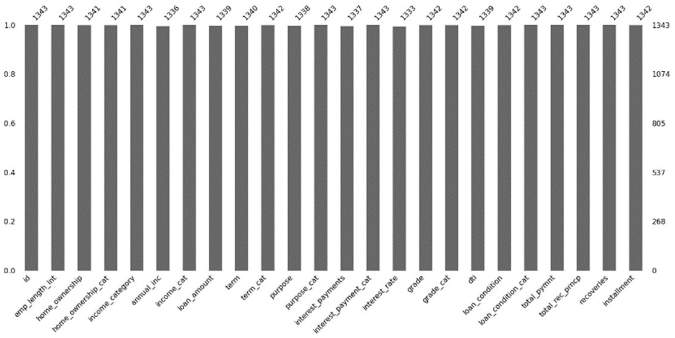


Figure 3.25(a): Missing Value Representation (With missing values)

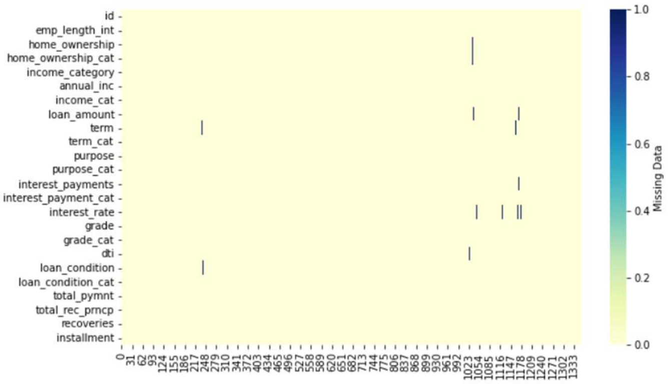


Figure 3.25(b): Missing Value Representation (With missing values)

From figure 3.25(a) and figure 3.25(b), it can be seen that home_ownership (2), home_ownership cat (2), annual_inc (7), loan_amount (4), term (2), term_cat (1), purpose (5), interest_payments (6), interest_rate (10), grade (1), grade_cat (1), dti (4), loan_condition (1), installment (1) missing values.

Missing values can be dealt with in a number of ways. An average of the above and below records can be made and then by putting it in the missing fill or by taking the value of the previous record or below the record of the missing fields and then by putting it in the missing field the missing value problem can be tackled. But, more often than not if this is done to handle missing values then that field will also contain flawed data. So, to deal with missing values properly, the records that contain the missing values will be deleted to make the dataset flawless. This is done for all missing records. The records with missing values are deleted using dropna() function in the data preprocessing steps.

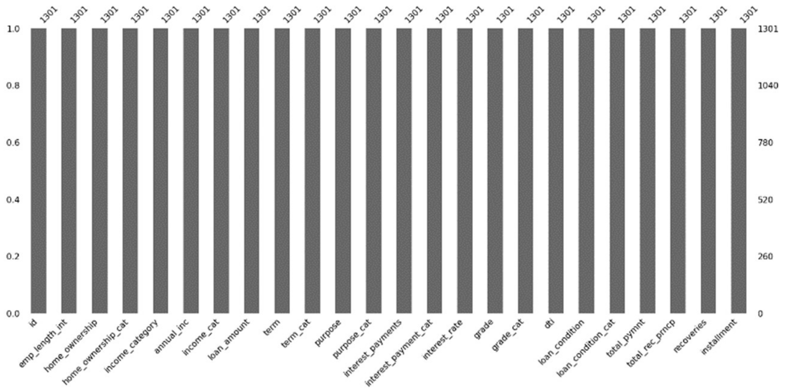


Figure 3.26(a): Missing Value Representation (After dropping missing values)

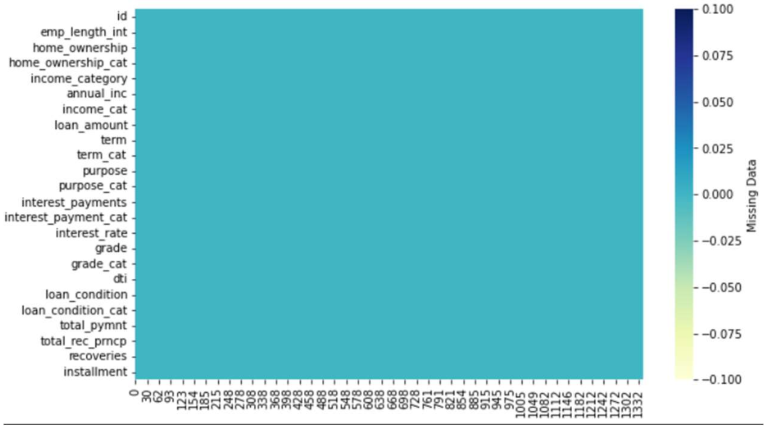
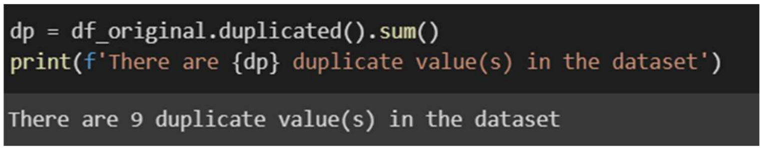


Figure 3.26(b): Missing Value Representation (After dropping missing values)

From figure 3.26(a) and figure 3.26(b), it can be seen that there are no values that are missing in the dataset. After removing the missing values, there were 1301 records in the dataset with 24 features.

Remove Duplicate Values: In a dataset, duplicate values can be present. This mainly happens due to manually inputting data or combining two datasets. Duplicate values need to be removed because unique data is needed to train the models. The more unique the data points are the more accurately the model will perform. It is a popular quote that more data means more perfect results. This is the same for ML models. So, duplicate values will increase data points, but it will not be beneficial to train the ML models. By detecting the number of duplicate values, an idea of how much unique data the dataset has, and if any more data is needed to train the models will be found. Duplicate records were detected using the duplicated() function.

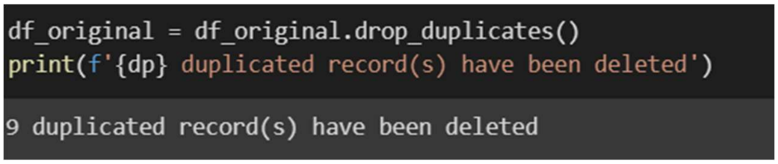


```
dp = df_original.duplicated().sum()
print(f'There are {dp} duplicate value(s) in the dataset')
```

```
There are 9 duplicate value(s) in the dataset
```

Figure 3.27: Duplicate Values Check

The duplicated records were removed using the drop_duplicates() function.



```
df_original = df_original.drop_duplicates()
print(f'{dp} duplicated record(s) have been deleted')
```

```
9 duplicated record(s) have been deleted
```

Figure 3.28: Duplicate Values Deletion

After removing the duplicate values, there were 1292 records left in the dataset with 24 features.

Check Outliers: Outliers are known as data points that do not belong to a certain dataset with other values, because they are different from the other values. They are abnormal observations that get in a dataset due to inconsistent data entry. Outliers cause the ML model to give the wrong prediction. So, the outliers need to be removed. First, the outliers are needed to be detected for every attribute using a boxplot. A boxplot is a standardized technique that is used to display the outliers and it is based on the five-number summary. It tells that the outliers exist in a dataset and it also reveals their values. The boxplot() function is used to view the graph of every attribute and detect outliers.

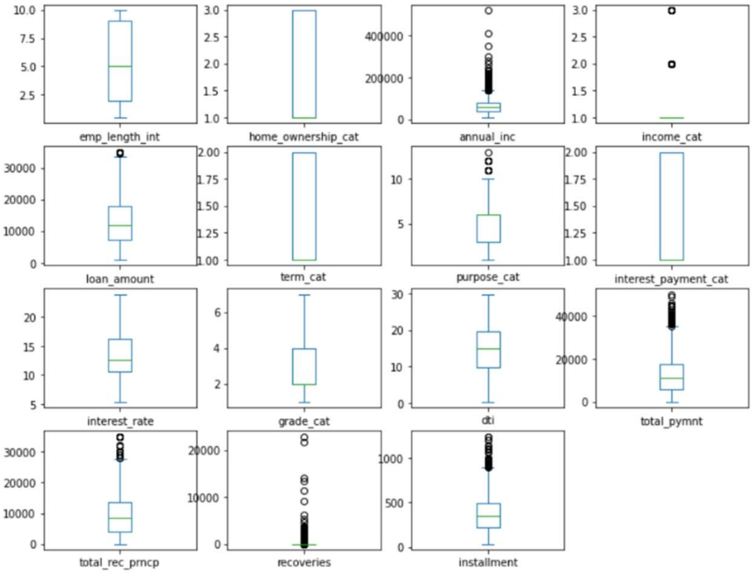


Figure 3.29: Outlier Representation

From figure 3.29, it is seen that there are outliers present in the annual_inc, income_cat, loan_amount, purpose_cat, total_paymnt, total_rec_prncp, recoveries, and installment attributes. However, before removing the outliers, the condition of the features and their values in the dataset will be checked. Firstly, the upper limit and lower limit will be checked using q1, q2, and IQR with the help of the formulas below:

Upper Limit = q3 + (1.5 * IQR) (1)

Lower Limit = q1 – (1.5 * IQR) (2)

For, annual_inc, the upper limit is 137203.75, and the lower limit is -18322.25, However, there is no need to delete any values because there were no values in the negatives and income can be larger than the upper limit. So, the values with outliers will be kept.

For, income_cat, it is not necessary to remove any records because all the records consist of valid values.

For, loan_amount, the upper limit is 34209.375, and the lower limit is -9015.625. An applicant can apply for any amount of loan. After checking the dataset, it can be said that all the records that do not fall between the upper and lower limit are valid.

For, purpose_cat, the records are valid, so it is not necessary to delete any values.

For, total_pymnt, the upper limit is 35243.591, and the lower limit is -11547.843. An applicant can have any amount of total payment based on the amount of loan taken. After checking the dataset, it can be said that all the records that do not fall between the upper and lower limit are valid.

For, total_rec_prncp, the upper limit is 27686.888, and the lower limit is -10111.073. An applicant can have any amount of total recovery principal based on the amount of loan taken and not be able to pay the amount in due time. After checking the dataset, it can be said that all the records that do not fall between the upper and lower limit are valid.

For, recoveries, the upper limit is 0.0, and the lower limit is 0.0. An applicant can have any amount of recovery based on the amount of loan taken and not be able to pay the amount in due time. After checking the dataset, it can be said that all the records that do not fall between the upper and lower limit are valid.

For, recoveries, the upper limit is 0.0, and the lower limit is 0.0. An applicant can have any amount of recovery based on the amount of loan taken and not be able to pay the amount in due time. After checking the dataset, it can be said that all the records that do not fall between the upper and lower limit are valid.

For, installment, the upper limit is 898.833, and the lower limit is -179.148. An applicant can have any amount of monthly installment based on the amount of the loan. After checking the dataset, it can be said that all the records that do not fall between the upper and lower limit are valid.

Balance Dataset: The dataset could be imbalanced. Data imbalance usually tells that the distribution of classes is unequal within a dataset. For example, in the topic of finding bad and good loans, most of the loans are not good, and very few classes are bad loans. In this case, the distribution of one of the classes is way less than the other class(es).

From figure 3.6, it can be seen that 879.0 (68%) of the instances are good loans that will be repaid in due time and 413.0 (32%) of the instances are bad loans that will not be repaid in due time. So, the ratio of good loans and bad loans is not satisfactory. This is also a problem that needs to be considered when training the model. Because, in most cases, the result of the class that has the most distribution will be acquired. Thus, a wrong prediction will be received. So, the dataset needs to be balanced. To balance the dataset, more primary data could be collected to get the distribution of both the classes close to each other which is called oversampling or some of the data points can be removed from the class with the higher instances to get the instances of both the classes closer to each other which is called under-sampling. Data points should not be removed because with more data more accurate results can be acquired. Oversampling can be done by using some techniques that are available in python. First, the categorical data is needed to be transformed into numerical values and then independent attributes should be put in one variable and the dependent variable in another variable. SMOTE technique is used to oversample. SMOTE technique is discussed in the Research Subject and Instrumentation section of this chapter.

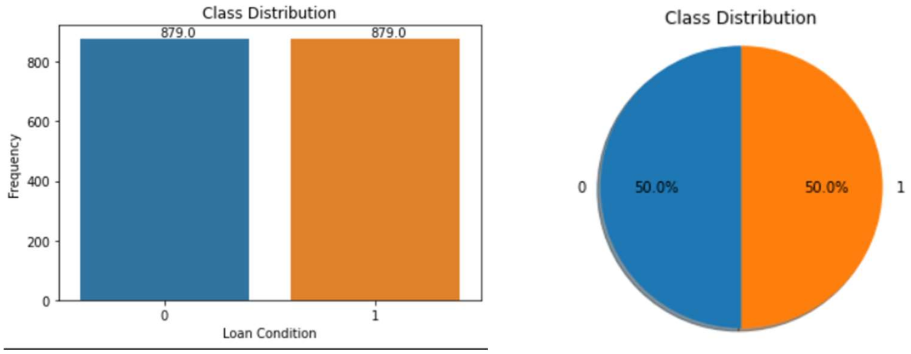


Figure 3.30: Class Distribution (After oversampling)

After oversampling, an equal distribution of classes was present. Because of oversampling the records increased from 1292 to 1758.

Feature Scaling: Feature scaling is a technique utilized to normalize the range of independent attributes of data. Standardization was utilized to scale the attributes. Standardization is a technique of scaling features in which the values of every data are rescaled to fit the original distribution between 1 and 0 it is done by utilizing the mean and the standard deviation calculation and they are the base to find specific values. The distance between data points is then used for plotting similarities and differences. Scaling the features makes it easy for an ML model to learn and understand the problem more fluently.

Classification: This is the step where the data points are separated into different parts. One part consists of the data that are utilized for model training and the other part consists of the data that are utilized for model testing. The model learns from the data within the training variable and tests its prediction with the testing data. Depending on the test, a score that will tell about the prediction accuracy of the model will be acquired. 70% (1230) of the data from the dataset for was taken for training the ML model and 30% (528) of the data from the dataset was taken for testing the trained model.

Training Model: After splitting the data records into training and testing data, the ML model is trained with the records from the training variable. To train the model, various algorithms were used. RF, DT, NB, KNN, LR, and SVM algorithms were utilized to train the model in different cases. These algorithms are some of the most popular and powerful algorithms in the section of ML. Each algorithm is unique in its own way. The algorithms were discussed in the Research Subject and Instrumentation section of this chapter.

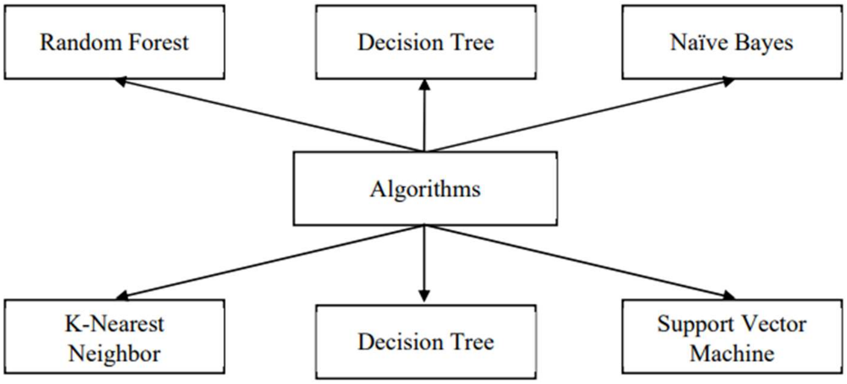


Figure 3.31: Machine Learning Algorithms

Feature Selection: Feature Selection is the technique of reducing the input attributes to the ML model by utilizing only relevant data from the dataset and getting rid of noise from the dataset. It is unnecessary to keep irrelevant attributes. Because they decrease the prediction power of the algorithms. So, feature scaling is important because the prediction power of the algorithms is increased by selecting the most critical attributes and dropping the redundant and irrelevant ones from the dataset. the correlation technique was utilized to check the correlation of independent attributes.

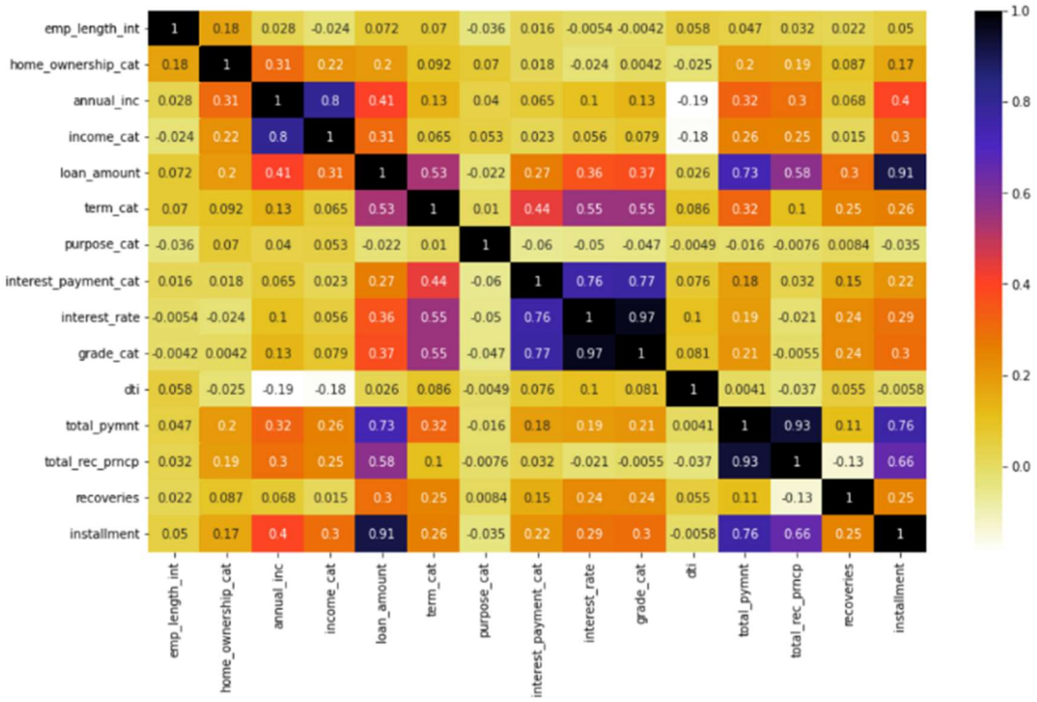


Figure 3.32: Correlation between features (With highly correlated features)

From figure 3.33, it can be seen that there are correlations between multiple attributes. The correlation score between annual_inc and income_cat is 0.8. The correlation score between loan_amount and total_pymnt is 0.73. The correlation between loan_amount and installment is 0.91. The correlation score between interest_payment_cat and interest_rate is 0.76. The correlation score between interest_payment_cat and grade_cat is 0.77. The correlation score between total_pymnt and total_rec_prncp is 0.93. The correlation score between total_pymnt and installment is 0.76. The correlation score between grade_cat and interest_rate is 0.97. Those are the highest correlation score between the two features. The value of 0.70 was considered as the threshold value which will be taken as the highest correlation score that should be accepted between two attributes. One of the attributes from each attribute set that has a correlation value of 0.70 or more was deleted. In this way, 5 highly correlated attributes (income_cat, interest_payment_cat, grade_cat, total_pymnt, installment) were deleted.

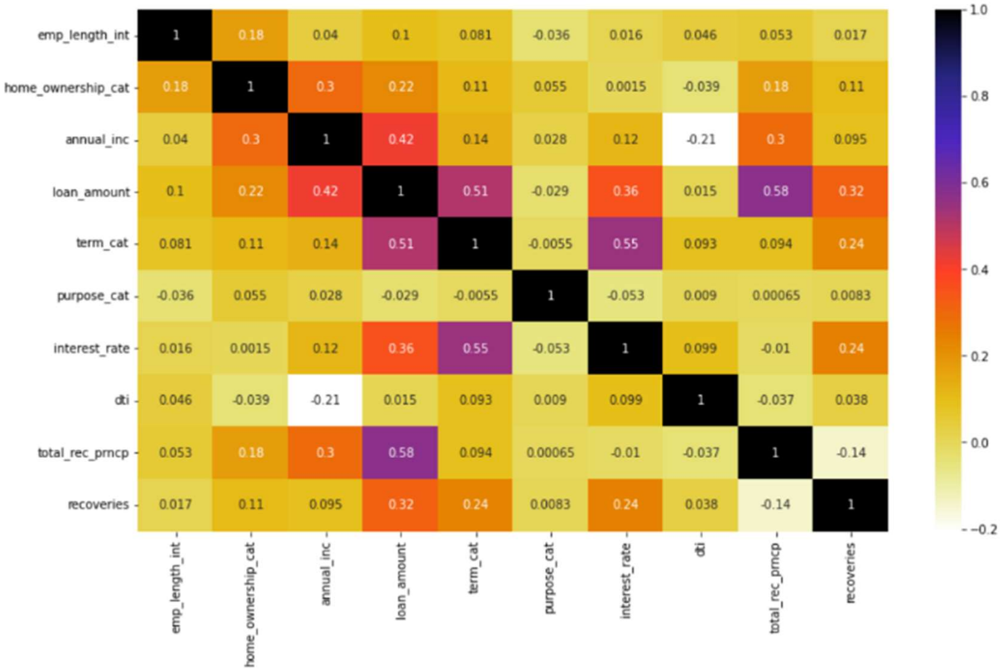


Figure 3.33: Correlation between features (Without highly correlated features)

Now, there is no value that exceeds the threshold value of 0.70.

Result: After training the ML model, the performance score of each model will be checked. The precision, recall, F1-Score, and accuracy of each of the ML algorithms will be checked to find the best algorithm for this research topic. By comparing the performance score of the ML algorithms will be able to find the best algorithm. In this paper’s case, the best algorithm was used to train the model to get the prediction of whether a borrower is risky or not. The result of this research is discussed in chapter 4 in detail.

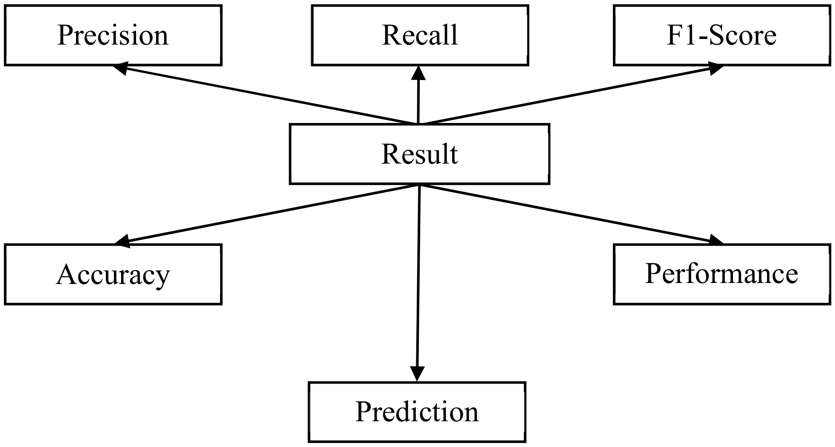


Figure 3.34: Result

Accuracy: Accuracy is the metric for testing the classification of ML models. Accuracy is that fraction of predictions that the ML model got right from the testing dataset. The formula for calculating the score of accuracy is mentioned below:

$$Accuracy = \frac{Numbeer\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \tag{3}$$

Precision: Precision is the ratio of correctly categorized positive samples (TP) to the total number of categorized positive samples (either incorrectly or correctly). Precision assists to show the dependability of the ML model in categorizing the ML model as positive [31] [32]. The formula for calculating precision is given below:

$$precision = \frac{tp}{tp + fp} \tag{4}$$

Where,

tp = True Positive

fp = False Positive

Recall: Recall is the ratio between the number of correctly categorized positive samples as Positive to the total number of positive samples. Recall computes the ability of the ML model to point out positive samples. A higher recall score means the positive samples that are detected are also higher [32]. The recall formula is mentioned in the next page:

$$recall = \frac{tp}{tp + fn} \tag{5}$$

Where,

tp = True Positive

fn = False Negative

F1-Score: F1 Score is that metric that is utilized on a binary classification ML model based on the predictions that are made for the positive class. It is measured with the assistance of a Precision score and Recall score. It is a single score that represents both the Precision score and Recall score. So, the F1-Score can be measured as the harmonic mean of both precision and recall, putting equal weight on each of the scores [32]. The F1-score formula is given below:

$$f1 - score = \frac{2(precision * recall)}{precision + recall} \tag{6}$$

Performance Score: The performance score is a score that is acquired by using the score() function. It gives a score from 0 – 100 that tells about the performance of the trained ML model.

Prediction: After preparing the ML model, values are passed according to the selected features to the trained model. The model gives a result. In this paper's case, the model gives either 0 (Good Loan) or 1 (Bad Loan) as a prediction.

Model Evaluation: Model evaluation is the procedure of utilizing various evaluation metrics to better understand an ML model's performance. K-Fold validation, confusion matrix, and AUC-ROC Curve were utilized to evaluate the trained models. The scores are compared to get the best model from the six trained models. The evaluation metrics were discussed in chapter 4 in detail.

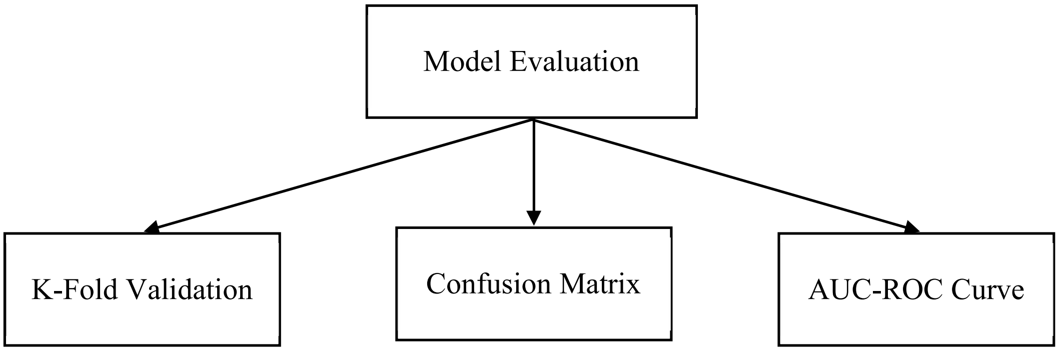


Figure 3.35: Model Evaluation

K-Fold Validation: K-Fold is a cross-validation method that breaks the input dataset into K number of groups of samples of equal sizes. These different samples are called folds individually. For each learning fold, the function for prediction utilizes k-1 folds, and the rest of the folds are utilized for the test dataset [33]. For this research, 3 folds were used to evaluate the trained models.

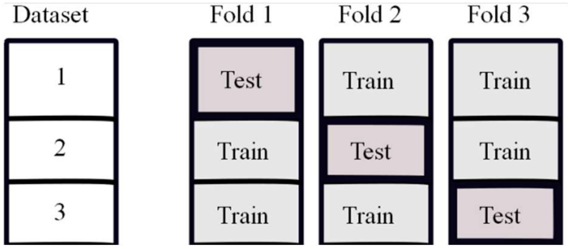


Figure 3.36: 3-Fold Cross-Validation (K-Fold Validation)

Confusion Matrix: The confusion matrix is a matrix utilized to direct the performance of the classification ML models for a test dataset. It can only be formed if the true values for test data are known [34].

TABLE 3.2: CONFUSION MATRIX

Predicted Class/Actual Class	C_1	$\neg C_1$
C_1	True Positives (TP)	False Negatives (FN)
$\neg C_1$	False Positives (FP)	True Negatives (TN)

AUC-ROC Curve: ROC curve is a metric for performance measurement of a classification ML model at various threshold values. ROC curve presents a probability graph to visualize the performance of a classification ML model at various threshold levels [35]. Between two parameters, the curve is plotted between two parameters, which are:

- True Positive Rate or TPR
- False Positive Rate or FPR

FPR and TPR are plotted on the X-axis and the Y-axis respectively in the curve:

TPR: TPR can be calculated by the following formula [35]:

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

FPR: FPR or False Positive Rate can be calculated by the following formula [35]:

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

Here,

TP: True Positive	FP: False Positive
TN: True Negative	FN: False Negative

AUC Curve: AUC computes the 2-D area that is under the whole ROC curve ranging from the points (0,0) to (1,1), as shown in the image which is in the next page:

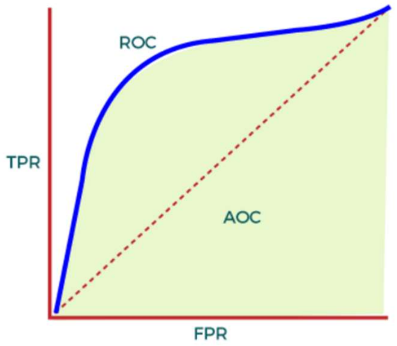


Figure 3.37: AUC-ROC Curve

In the ROC curve, AUC calculates the performance of the binary classifier across various thresholds and gives an aggregate measure. The range of the AUC value is from 0 to 1, which tells that an effective ML model will have an AUC value very near to 1, and thus it will display a good separability measure [35].

Model Selection: Model selection is the final step of selecting the fit ML best model trained with the best algorithm for a certain topic among multiple models that were trained by different algorithms for a training dataset. six different algorithms were utilized to train 6 different models. Their precision, recall, f1-score, accuracy, and performance score were checked and after the result evaluation of the ML models, the best model was selected from the six trained models for this research topic.

3.6 Implementation Requirements

In this research, coding was required to support the findings. Python programming language was used for this topic. Python has an enriched library that helped me in cleaning the dataset, visualizing data, making training and testing data, and using the training data, building the machine model which will give the final prediction of credit risk. Also, the accuracy of the prediction can be tested because python has features for that too. In summary, python fulfilled the implementation requirements for this topic. In summary, the following technologies were used in the implementation of the research topic:

- Language: Python (Version: 3.7.0)
- Open-source web application: Google Colaboratory
- Library: Pandas (Data manipulation and Data analysis)
- Fundamental library for computing: NumPy
- Library: Missingno (Finding missing values)
- Library: Malplotlib (Data visualization)
- Library: Seaborn (Data visualization)
- Library: pydotplus (Data visualization)
- Library: Imbalanced-learn (Handling Imbalance Dataset)
- Library: Scikit-learn (Machine learning)
- Browser: Google Chrome
- Operating System: Windows 11

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

In this chapter, how the models were trained and what performance and accuracy were found from them based on the dataset that was used will be discussed. A total of 6 different ML algorithms like- Random Forest(RF), Decision Tree(DT), Naïve Bayes(NB), KNN, Logistic Regression(LR), and Support Vector Machine(SVM) were used. The results are reported and validated by K-Fold Cross-validation for each of the ML algorithms. The model with the best performance for this topic was elected by the evaluation of the confusion matrix, F1-score, precision, recall, and AUC-ROC Curve.

4.2 Experimental Results

6 different models were trained with 6 different algorithms. Each of the algorithms yields different results. The accuracy, precision, recall, F1-Score, and performance score will be taken as results and a comparison between the results of all ML models will be made to elect the best model for this research topic.

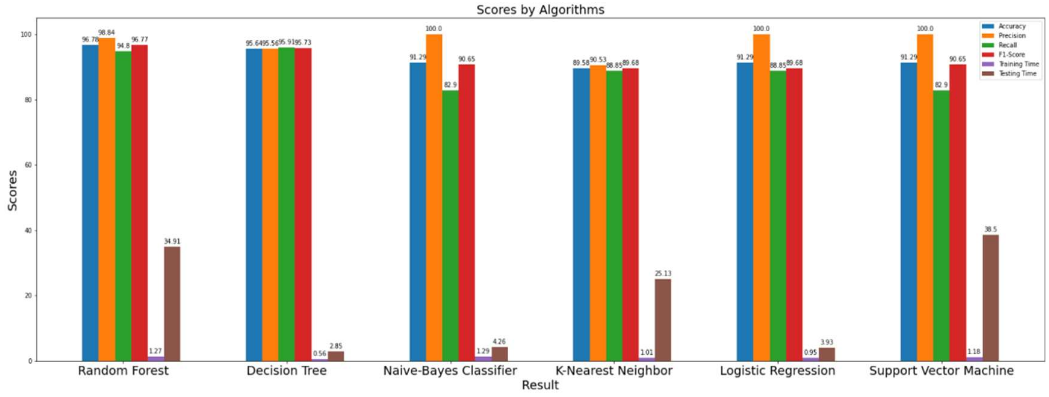


Figure 4.1: Result Comparison (Accuracy, Precision, Recall, F1-Score, Training Time, Testing Time)

From figure 4.1, it is seen that the RF algorithm has the best accuracy score, F1-Score, and Performance Score among all the algorithms. Although it has a lesser precision score compared to NB, LR, and SVM algorithms and a lesser recall score compared to DT, overall, it is the best algorithm for this topic.

If the time to train the model is considered, naïve bayes trained model takes lesser time to complete its training and the RF algorithm takes longer time to be trained because of building subtrees. Testing time is also less for LR and then NB trained model and highest for SVM trained ML model.

TABLE 4.1: RESULT COMPARISON (ACCURACY, PRECISION, RECALL, F1-SCORE, TRAINING TIME, TESTING TIME)

Algorithm (Model)	**Accuracy (Percent)**	**Precision (Percent)**	**Recall (Percent)**	**F1-Score (Percent)**	**Training Time**	**Testing Time**
Random Forest	96.78%	99.84%	94.80%	96.77%	1.27	34.91
Decision Tree	95.64%	95.56%	95.91%	95.73%	0.56	2.85
Naïve Bayes	91.29%	100%	82.90%	90.65%	1.29	4.26
K-Nearest Neighbor	89.58%	90.53%	88.85%	89.68%	1.01	25.13
Logistic Regression	91.29%	100%	88.85%	89.68%	0.95	3.93
Support Vector Machine	91.29%	100%	82.90%	90.65%	1.10	38.50

4.3 Model Evaluation

In this section, the steps of evaluating the trained models will be discussed. K-Fold Cross Validation, Confusion Matrix, and AUC-ROC Curve were utilized to evaluate the ML models.

K-Fold Cross Validation: 3-Fold was utilized to divide the training dataset and testing dataset.

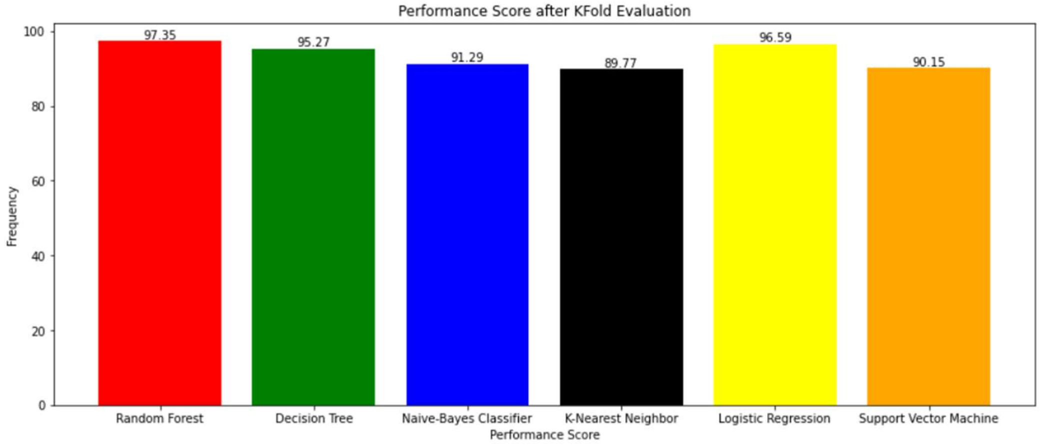


Figure 4.2: Result Evaluation Score (After K-Fold Evaluation)

After evaluating the models using K-Fold Cross-validation, it is seen that RF has the best score among all the algorithms and Logistic Regression is the second-best algorithm in the list. KNN has the lowest performance score among the selected algorithms to train a model.

TABLE 4.2: EVALUATION SCORE

Algorithm (Model)	**Score (In Percent)**
Random Forest	97.35%
Decision Tree	95.27%
Naïve Bayes	91.29%
K-Nearest Neighbor	89.77%
Logistic Regression	96.59%
Support Vector Machine	90.15%

Confusion Matrix

The confusion matrix for all algorithms will be discussed in this section.

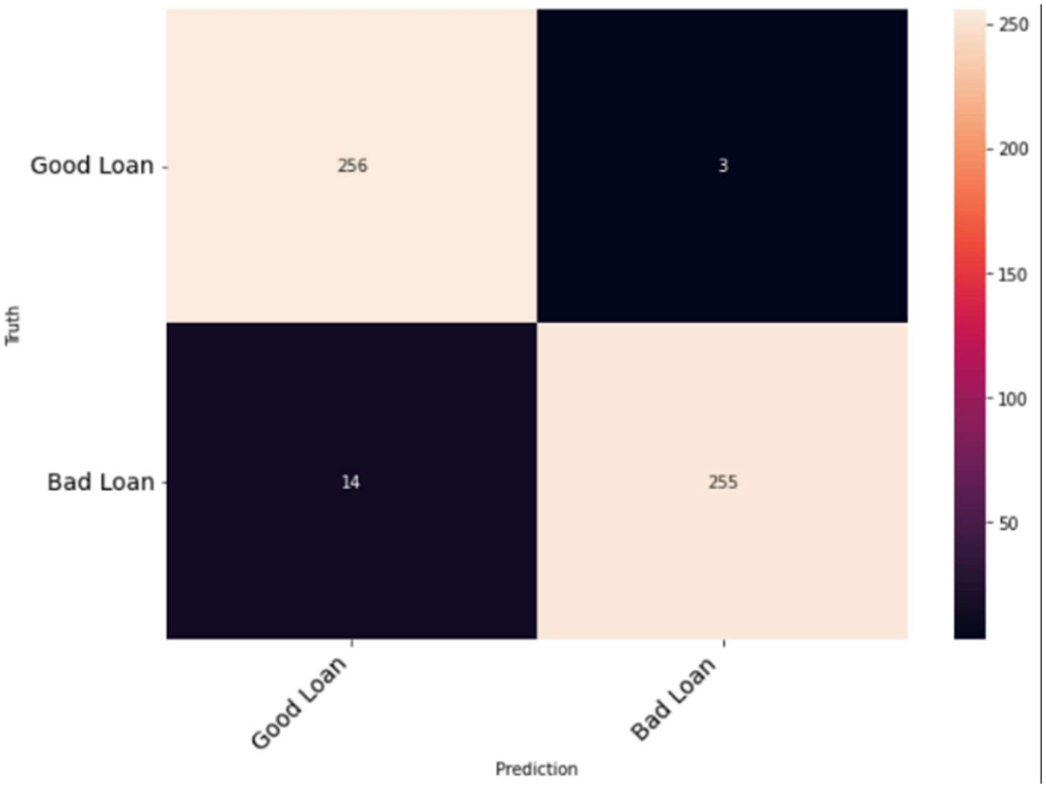


Figure 4.3: Confusion Matrix (Random Forest)

For Random Forest(RF), True Positive(TP) = 256, False Negative(FN) = 3, False Positive(FP) = 14, and True Negative(TN) = 255. This means that there are 256 instances where the loan has been repaid and the ML model's prediction is Good Loan, 3 instances where the loan is repaid but the ML model's prediction is Bad Loan, 14 instances where the loan is not repaid and the ML model's prediction is Good Loan, and 255 instances where the loan is not repaid and the ML model's prediction is Bad Loan.

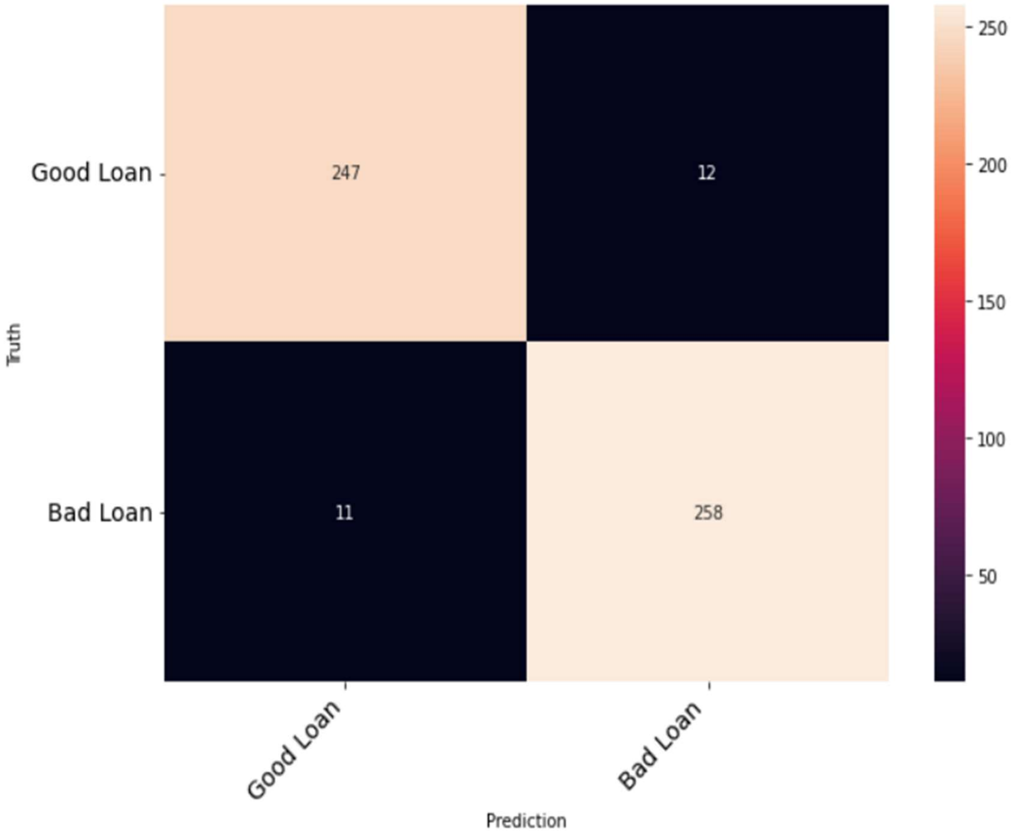


Figure 4.4: Confusion Matrix (Decision Tree)

For Decision Tree(DT), True Positive(TP) = 247, False Negative(FN) = 12, False Positive(FP) = 11, and True Negative(TN) = 258. This means that there are 247 instances where the loan has been repaid and the ML model’s prediction is Good Loan, 12 instances where the loan is repaid but the ML model’s prediction is Bad Loan, 11 instances where the loan is not repaid and the ML model’s prediction is Good Loan, and 258 instances where the loan is not repaid and the ML model’s prediction is Bad Loan.

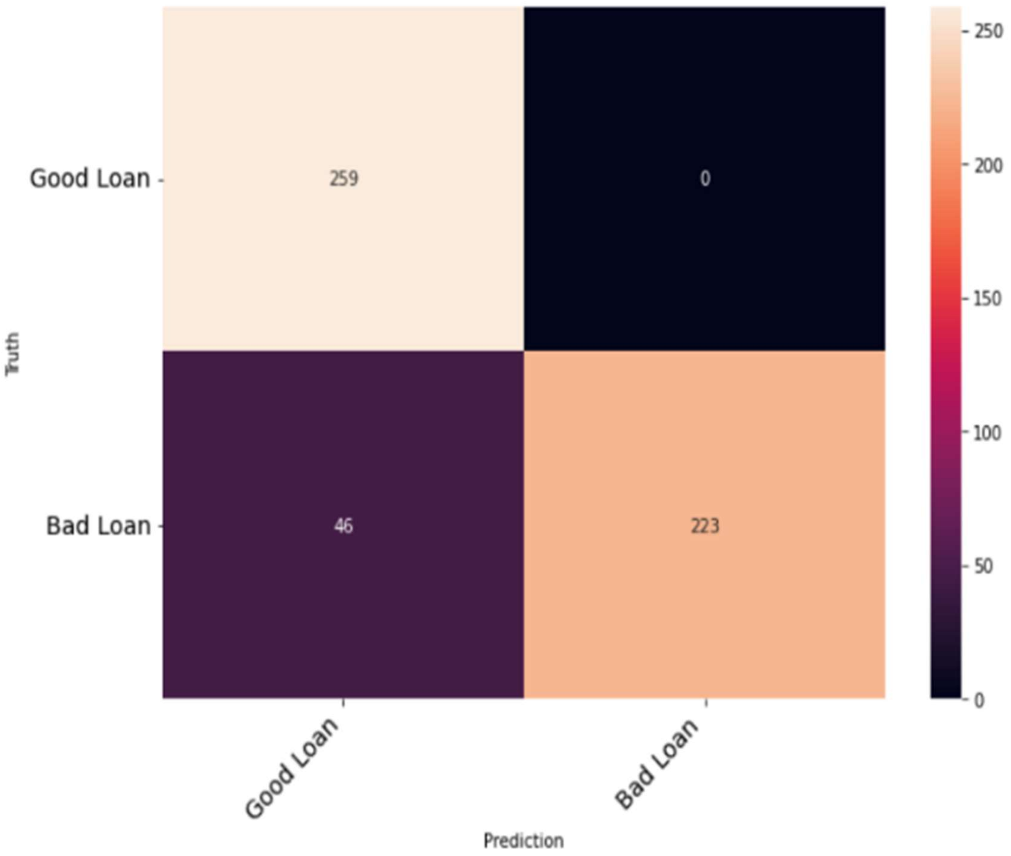


Figure 4.5: Confusion Matrix (Naïve Bayes)

For Naïve Bayes(NB), True Positive(TP) = 259, False Negative(FN) = 0, False Positive(FP) = 46, and True Negative(TN) = 223. This means that there are 259 instances where the loan has been repaid and the ML model's prediction is Good Loan, 0 instances where the loan is repaid but the ML model's prediction is Bad Loan, 46 instances where the loan is not repaid and the ML model's prediction is Good Loan, and 223 instances where the loan is not repaid and the ML model's prediction is Bad Loan.

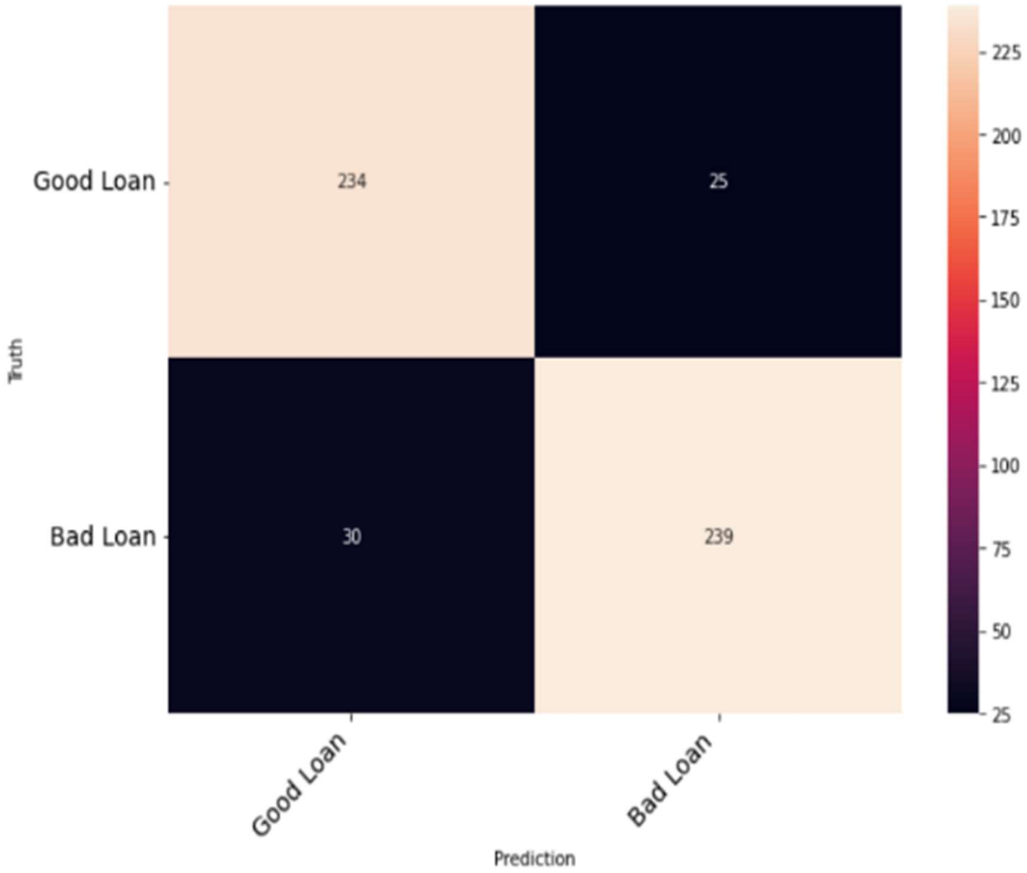


Figure 4.6: Confusion Matrix (K-Nearest Neighbor)

For K-Nearest Neighbor(KNN), True Positive(TP) = 234, False Negative(FN) = 25, False Positive(FP) = 30, and True Negative(TN) = 239. This means that there are 234 instances where the loan has been repaid and the ML model's prediction is Good Loan, 25 instances where the loan is repaid but the ML model's prediction is Bad Loan, 30 instances where the loan is not repaid and the ML model's prediction is Good Loan, and 239 instances where the loan is not repaid and the ML model's prediction is Bad Loan.

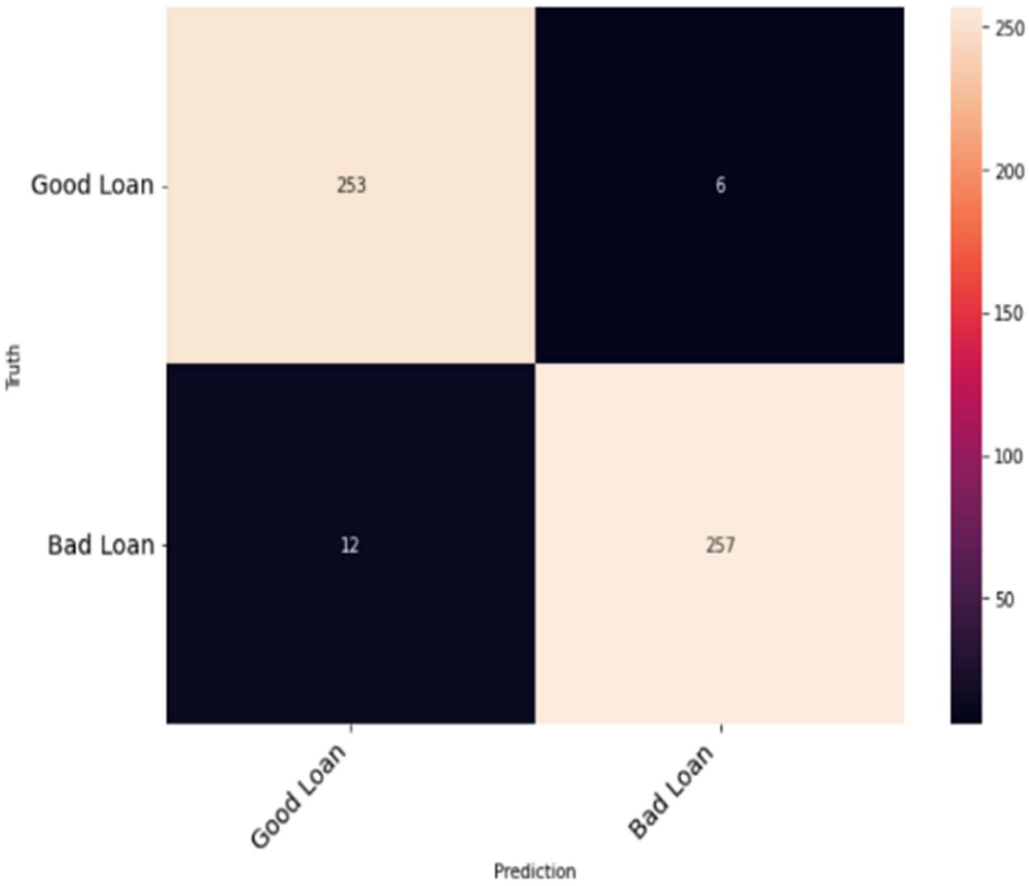


Figure 4.7: Confusion Matrix (Logistic Regression)

For Logistic Regression(LR), True Positive(TP) = 253, False Negative(FN) = 6, False Positive(FP) = 12, and True Negative(TN) = 257. This means that there are 253 instances where the loan has been repaid and the ML model's prediction is Good Loan, 6 instances where the loan is repaid but the ML model's prediction is Bad Loan, 12 instances where the loan is not repaid and the ML model's prediction is Good Loan, and 257 instances where the loan is not repaid and the ML model's prediction is Bad Loan.

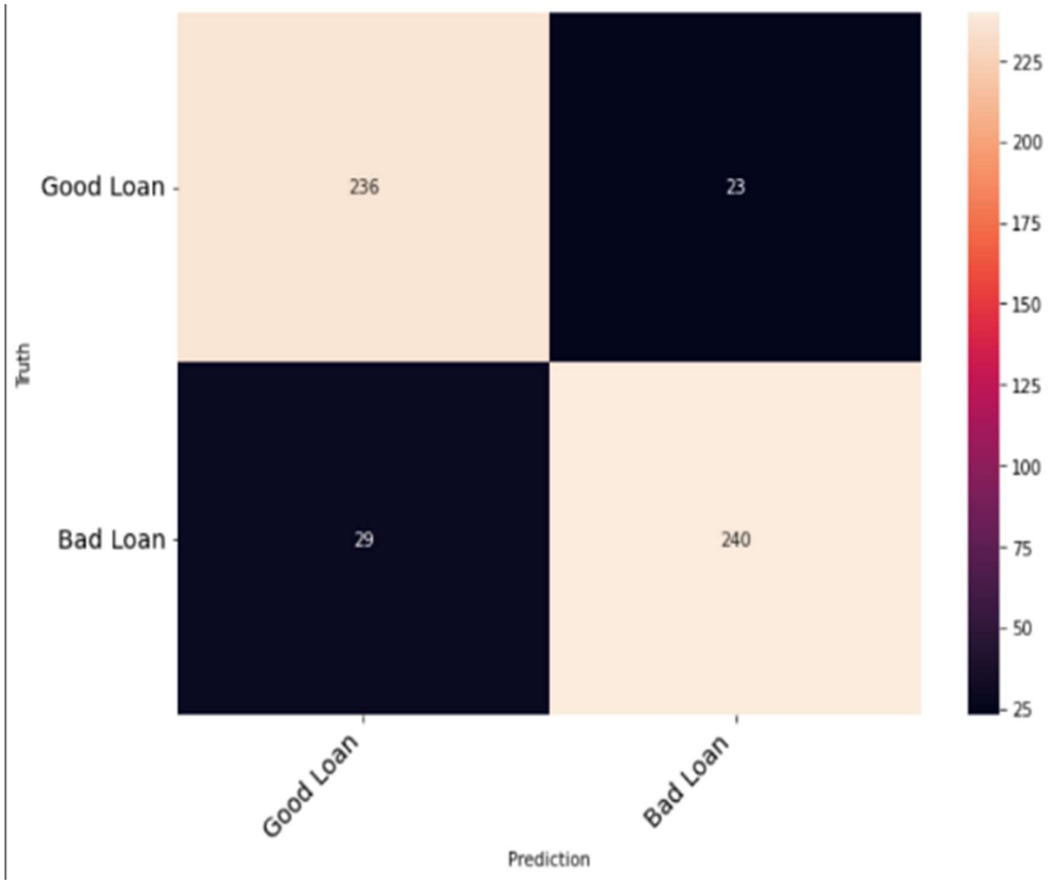


Figure 4.8: Confusion Matrix (Support Vector Machine)

For Support Vector Machine(SVM), True Positive(TP) = 236, False Negative(FN) = 23, False Positive(FP) = 29, and True Negative(TN) = 240. This means that there are 236 instances where the loan has been repaid and the ML model's prediction is Good Loan, 23 instances where the loan is repaid but the ML model's prediction is Bad Loan, 29 instances where the loan is not repaid and the ML model's prediction is Good Loan, and 240 instances where the loan is not repaid and the ML model's prediction is Bad Loan.

After Analyzing the confusion matrix of all the algorithms, it is seen that the true positive instances are better for the Naïve Bayes algorithm, which means it can predict the good loan better than other algorithms. Random Forest is the second-best algorithm based on the same criteria. However, DT is the best algorithm when predicting bad loans. The second-best algorithm is the LR algorithm. However, the performance of the naïve bayes algorithm for predicting bad loans is way lesser and the performances of the DT and LR for predicting good loans are lesser compared to the RF algorithm. So, it can be said that RF is the overall best algorithm for this topic based on the analysis of the confusion matrix.

AUC-ROC Curve

AUC-ROC Curve was utilized to evaluate the ML models as well. There is an AUC value ranging from 0 to 100. A better AUC value indicates a better model. AUC Value over 0.80 is good for any model. Here, the AUC value will be checked for all algorithms, and visualization of the AUC value on the ROC Curve graph will be displayed.

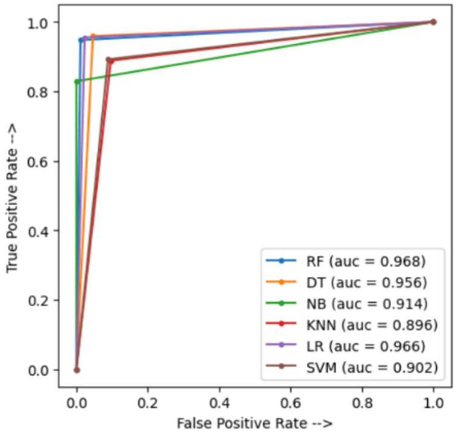


Figure 4.9: AUC-ROC Curve

After Analyzing the AUC-ROC Curve, it is seen that RF has the highest AUC value. Thus, it is the best ML algorithm for credit risk analysis. Support Vector Machine shows the lowest AUC Value among all the algorithms.

TABLE 4.3: AUC VALUE OF TRAINED MODELS

Algorithm (Model)	AUC Value
Random Forest	0.968
Decision Tree	0.956
Naïve Bayes	0.914
K-Nearest Neighbor	0.896
Logistic Regression	0.966
Support Vector Machine	0.902

Feature Importance

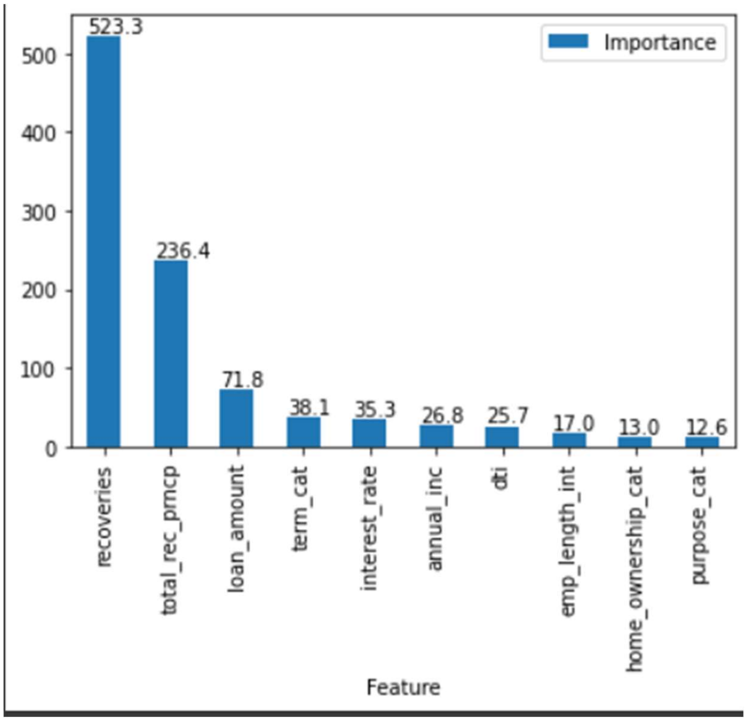


Figure 4.10: Feature Importance

From figure 4.10, it can be seen that recoveries have the highest feature importance with 523.3 followed by total recovery principal, loan amount, term, interest rate, annual income, debt-to-income (dti), employment length, and home ownership. Purpose has the least importance. This feature's importance is for the RF algorithm. Since, RF is the best algorithm to train the ML model for credit risk analysis, the feature importance of the algorithm is described in this section.

Prediction

Sample information was inputted to check what all 6 ML models predicted. Here, 0 = Good Loan and 1 = Bad Loan. The values are inputted in the following order: emp_length_int, home_ownership_cat, annual_inc, loan_amount, term_cat, purpose_cat, interest_rate, dti, total_rec_prncp, recoveries. After using the predict() function, all the trained models have returned 0 for the first set of values and 1 for the second set of values, which means the first client should be issued a loan and the second client fails to receive a loan from a financial institution.

TABLE 4.4: PREDICTION (GOOD LOAN)

Predict([[8, 1, 15000, 2500, 1, 3, 16, 9, 2500, 0]])	
Algorithm	**Prediction**
Random Forest	0
Decision Tree	0
Naïve Bayes	0
K-Nearest Neighbor	0
Logistic Regression	0
Support Vector Machine	0

TABLE 4.5: PREDICTION (BAD LOAN)

Predict([[5, 2, 50000, 5000, 2, 3, 14, 16, 3000, 2000]])	
Algorithm	**Prediction**
Random Forest	1
Decision Tree	1
Naïve Bayes	1
K-Nearest Neighbor	1
Logistic Regression	1
Support Vector Machine	1

4.4 Discussion

6 different algorithms were utilized to train 6 different models. Each of the models has its own accuracy, precision, recall, F1-Score, training time, testing time, cross-validation score, confusion matrix, and AUC value. After analyzing the models, it was found that all of the mentioned scores for all the algorithms are quite close. However, some algorithm performs at high capacity during the testing model (support vector machine) and some other algorithm performs at high capacity during the training model (naïve bayes), some other have a good confusion matrix when it comes to predicting good loan (naïve bayes) and other is good while predicting bad loan (logistic regression). However, to select the best algorithm among them, all the cases had to be considered to do. So, after analyzing all the result scores and validation scores, it was found that Random Forest was the overall best algorithm to train a model for predicting the credit risk of a client with it having an accuracy score of 96.78, which is the best

among all algorithms, a precision score of 99.84 which is only behind NB, LR, and SVM algorithm, recall score of 94.80, which is the second-best score falling only behind DT, F1-Score of 96.77, which the best among all algorithm. The performance score of RF is the highest among the algorithms. The confusion matrix of random forest overall shows the best result and the AUC value is the highest for this algorithm. Therefore, it can be said that RF is the best algorithm to train and test models for credit risk prediction problems.

CHAPTER 5
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Introduction

In this chapter, the impact of this study on society and the environment will be described. The ethical aspects will be mentioned for conducting this research. And at the end of the chapter, a sustainability plan will be given to show the importance of machine learning in this research topic.

5.2 Impact on Society

The financial institution plays an important role in the economic growth of a nation. It generates a huge proportion of cash from providing loans to the borrower. The borrowers are benefitted as well after receiving loans. For instance, the student can pay their tuition fees with it. People can pay their debt with it. Even people take loans to buy property like land, and apartments. Some take loans for starting a business. Some also take loans for doing agricultural work. Those who are students now will lead their nation in the future. People who take debt and cannot repay the debt in time can avoid facing legal action by taking a loan. Starting a business also means increasing employment sources. Agriculture plays a vital role in growing food for a nation. The more food a country can grow the more it can export after meeting its people's needs to bring in cash from other countries. Thus, bank credit has a positive impact on society. On the contrary, because of buying land, the area for cultivating food is reduced. Playing fields for children are disappearing day by day. This can have a severe effect on the mental development of small children. Nowadays, they seem to stay at home all the time. This can also cause autism too. So, the impact of banks can have a negative impact on society as well. However, the purpose of this research is to predict whether one is suitable for a loan or not. From this, the steps for a grating loan will be quicker than before and unworthy applicants won’t get loans. So, people can pay their debt, and tuition fees in time, and they can contribute to society in the future.

5.3 Impact on Environment

Financial institutions can have a negative impact on the environment. The loan can be taken to start a business or industry. Nowadays, there are many industries that are started after taking loans. Some of the industries are the chemical industry, construction, and transport. The chemical industry produces waste that can pollute water and air. Transports are the cause of sound pollution and air pollution. The construction industry requires wood and that comes from cutting trees. Trees produce oxygen and take in carbon dioxide. It also brings rain. So, cutting down trees has an adverse impact on the environment. Also, free spaces are decreasing day by day. And this is all possible because of the startup of most of the companies dependent on bank loans. Since this study it is shown that it is quicker to process loan applications by using machine learning techniques, people can get loans quicker than before to start their industry business. So, the bank has an indirect negative impact on the environment.

On the other hand, green businesses have a positive impact on the environment. Green business refers to that type of business where only sustainable materials are used to make different products. This type of business aims to use raw materials and as little water as possible. It also cuts carbon emissions. In another word, this type of business utilizes used materials in renewable and environment-friendly ways. Even this type of business needs a loan from the banks at the very beginning. So, bank loan also has an indirect positive impact on the environment as well.

5.4 Ethical Aspects

Financial Institutions keep their applicant's information confidential, meaning they don't share that kind of information with the general public. So, conducting this type of research requires one to be extra careful while collecting the data. It is required to protect the data of the debtors and keep it safe. If one gets to have the access to this type of information then the gathered information should be encrypted if it is stored on a computing device. Also, if the information is being stored in a cloud then before storing the information in the cloud, one needs to read the privacy policy of that service provider. One should never disclose this kind of data to anyone, not even to the employee of the bank that the information was collected from. Because not all employees of banks have the access to all kinds of information. The information contains sensitive information about the customer. If the data gets leaked somehow then it could bring trouble not only to the customer but also to the financial institution. The customer's social security number, property details along with other important information could get leaked and the customer will face many problems. On the other hand, the financial institution that provided the data for research purposes will lose customer trust and will lose a customer. So, their cash flow, and revenue will decrease. Thus, they could get bankrupt in the future. The banks should have some kind of policy on what type of information can be provided for research purposes. So, if the provided information is enough for conducting the information then that information should be taken to continue the research work. It is also unethical to take information that is not required for the research. It is also not good practice to disturb a research subject for collecting data if one cannot get the information from banks. Even many people don't want to share simple information about themselves let alone that type of information. If a financial provides the information for research purposes, it is also necessary to provide the outcome of the research to them. The result of the research could be beneficial for them in their day-to-day task. For this study, secondary data was collected due to the banks in Bangladesh don't have any policy to share their customer's information. Therefore, the mentioned situations were not encountered during this research.

5.5 Sustainability Plan

In Bangladesh, financial institutions use a judgmental approach to analyze loan applications. They don't use machine learning to detect the worthiness of a loan applicant. The judgmental approach takes some time to complete all the steps. Therefore, it is always not possible for the loan applicant to pay their dues in time. Sometimes, the right decision doesn't come out of this approach, So, banks see a loss in some cases. Introducing machine learning in this sector not only increases the processing speed but also the accuracy rate. The chances of a loss get lower because of this technology. Since machine learns from data, data in financial institution grows at a high rate every day. So, the accuracy of the system gets higher each day. The features used in this research also show the highest accuracy. Therefore, the banks could introduce these features in their system as well to get the correct prediction. This will bring in huge profits for them. They will also be able to ensure customer satisfaction and reduce the dilemmas of the customers. Another problem with the judgment approach, it requires more trained employees at every step. So, it will take time to train a fresher. In the machine learning approach, it is only required to set up the system and the system will learn automatically from everyday data. So, the requirement for trained employees isn’t a problem anymore. Only a person who will control the system needs to be trained and it will not take that long to train him/her. Thus, by following this research, a financial institution can increase its profit and in the long run, the country will see positive growth in its economy.

CHAPTER 6
SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Introduction

In this chapter, the entire study will be summarized from the beginning to the end so it will be helpful for the readers about the result that was found. A proper conclusion will also be given where the best-performing models will be mentioned. Some recommendations for financial institutions and implications for future studies will be given.

6.2 Summary of the Study

In this research study, various ML algorithms were utilized to train various ML models to get a prediction of whether an applicant is too risky to issue a loan. A dataset from Kaggle was used in this study. The dataset had 24 attributes whereas after data selection there were 10 attributes left. Before that, the dataset had gone through intensive data preprocessing stages like data cleaning, data augmentation, data visualization, and data analysis to fix the dataset, handle the dataset properly, and get to understand the data properly to get a proper insight. A Proper methodology was followed while doing this research. A number of models were built which provide more than 90% of accuracy scores, but only 1 algorithm (KNN) gave lesser than 90% accuracy. The performance score of all ML models was evaluated to check the reliability of each model by using a cross-validation technique, confusion matrix, and AUC-ROC Curve. Finally, the results of all ML models were compared after the evaluation stage to get the best model.

6.3 Conclusions

The main purpose of this study is to build an ML model that can successfully predict whether a loan is classified as good or bad. After training and testing the ML models, it was found that RF trained ML model is the overall best model among the 6 models. It showed an accuracy rate of 96.78%, while the DT tree is the second-best algorithm for this topic with 95.64%, followed by NB, LR, and SVM, all of them having an accuracy score of 91.29%, except KNN had 89.58% of accuracy. It was necessary to go through the result evaluation stage to get further clarification about the accuracy of the ML models. After the result evaluation, the performance scores of all ML algorithms increased. Even after that RF remained the top algorithm with 97.35% of accuracy and LR became the second-best algorithm with a 96.29% of accuracy rate.

6.4 Recommendations

In this study, the most popular and powerful algorithms available to date were utilized to train ML models to get the prediction of whether an applicant is considered risky or not to be granted a loan. 6 different ML algorithms were applied to train 6 different ML models. Among the 6 ML algorithms, RF showed the best accuracy. The result of all the ML models was examined very carefully. Even after that RF remained the best of the lot. Therefore, it is recommended that all financial institutions use this ML algorithm in their day-to-day verification stage to get an early insight of what will be status of the applicants if they were to get a loan from them. It is also recommended that they should use the features that were used in this study because with these selected features, the ML models show better results compared to other researchers' ML models or they can make a custom ML model as per their terms and conditions, and policies.

6.5 Implication for Further Study

Technologies are constantly improving day by day. Even the policies of financial institutions may be updated in the near future. If that happens then there will be a need to update the features that were used for this research. Selecting other features or even changing a few features may result in the accuracy rate of the ML algorithms that were examined during this research. Even, ML algorithms are constantly being updated and new algorithms are being created. However, currently, there is no need to create new ML models with similar features, because ML models keep learning every as millions of data are generated every second. They adjust to the new data very easily and quickly.

For future work, researchers could use different types of ML algorithms to train their ML models with or without changing the features. There are a lot of ML models available in modern times and it is not possible for a single person to examine each of the ML algorithms to train different ML models. Therefore, by examining various ML algorithms the result may change with or without the features that were selected for this study. If an ML model trained by other ML algorithms performs better than this paper's ML models then it would be best for the banking sector to upgrade its process and by doing so, they can reduce the risk of credit loss.

REFERENCES

[1] M. Radhaswami and S. V. Vasudevan, Text Book of Banking, 3rd ed. New Delhi: S. Chand & Company Limited.

[2] A. Bask, et al., “Towards e-banking: the evolution of business models in financial services,” International Journal of Electronic Finance, vol. 5.4, pp. 333-356, 2011.

[3] H. A. Bekhet and B. A. Al-alak, “Measuring e-statement quality impact on customer satisfaction and loyalty,” International Journal of Electronic Finance, vol. 5.4, pp. 299-315, 2011.

[4] K. Curran and J. Orr, “Integrating geolocation into electronic finance applications for additional security,” International Journal of Electronic Finance, vol. 5.3, pp. 272-285, 2011.

[5] T. N. Pandey, et al., “Machine learning–based classifiers ensemble for credit risk assessment,” International Journal of Electronic Finance, vol. 7.3-4, pp. 227-249, 2013.

[6] S. Das, and S. Das. "Credit Risk Management Practices–An Evaluation of Commercial Banks in Bangladesh," ASA University Review, July-December 2007.

[7] T. N. Pandey, S. K. Mohapatra, A. K. Jagadev, S. Dehuri, *“*Credit Risk Analysis using Machine Learning Classifiers,*”* International Conference on Energy, Communication, Data Analytics and Soft Computing, pp. 1-5, 2017.

[8] C.-L. Huang, M.-C. Chen, and C.-J. Wang, “Credit scoring with a data mining approach based on support vector machines,” Expert Systems with Applications, vol. 33.4, pp. 847-856, 2007.

[9] S. Purohit and A. Kulkarni, “Credit evaluation model of loan proposals for Indian Banks,” 2011 World Congress on Information and Communication Technologies, 2011.

[10] A. J. Hamid and T. M. Ahmed, “Developing prediction model of loan risk in banks using data mining,” Machine Learning and Applications: An International Journal (MLAIJ), vol. 3.1, pp. 1-9, 2016.

[11] F. Doko, S. Kalajdziski, and I. Mishkovski, "Credit risk model based on central bank credit registry data," Journal of Risk and Financial Management, vol. 14.3, pp. 138, 2021.

[12] M. N. Ferozi, 2018, "Loan data for Dummy Bank," Kaggle. [Online] Available: https://www.kaggle.com/datasets/mrferozi/loan-data-for-dummy-bank

[13] "Google Colab." Google. https://research.google.com/colaboratory/faq.html/ (Accessed: Nov. 26, 2022).

[14] "What Is Python Used For? A Beginner's Guide | Coursera." Coursera. https://www.coursera.org/articles/what-is-python-used-for-a-beginners-guide-to-using-python/ (Accessed: Nov. 26, 2022).

[15] "Python Pandas | Python Pandas Tutorial – javatpoint." Javatpoint. https://www.javatpoint.com/python-pandas/ (Accessed: Nov. 26, 2022).

[16] "Python Numpy: Tutorial, What It is, Library – Javatpoint." Javatpoint. https://www.javatpoint.com/numpy-tutorial/ (Accessed: Nov. 26, 2022).

[17] "Scikit Learn – Introduction." tutorialspoint. https://www.tutorialspoint.com/ scikit_learn/scikit_learn_introduction.htm/ (Accessed: Nov. 26, 2022).

[18] "What Is Matplotlib In Python? How to use it for plotting? – ActiveState." ActiveState. https://www.activestate.com/resources/quick-reads/what-is-matplotlib-in-python-how-to-use-it-for-plotting/ (Accessed: Nov. 26, 2022).

[19] "pydotplus . PyPI." PyPI. https://pypi.org/project/pydotplus/ (Accessed: Nov. 26, 2022).

[20] "Introduction to Seaborn – Python – GeeksforGeeks." GeeksforGeeks. https://www.geeksforgeeks.org/introduction-to-seaborn-python/ (Accessed: Nov. 26, 2022).

[21] "Using the missingno Python library to Identify and Visualise Missing Data Prior to Machine Learning | by Andy McDonald | Towards Data Science." Towards Data Science. https://towardsdatascience.com/using-the-missingno-python-library-to-identify-and-visualise-missing-data-prior-to-machine-learning-34c8c5b5f009/ (Accessed: Nov. 26, 2022).

[22] "imbalanced-learn documentation — Version 0.10.1." Imbalanced learn. https://imbalanced-learn.org/stable/ (Accessed: Nov. 26, 2022).

[23] "ML | Handling Imbalanced Data with SMOTE and Near Miss Algorithm in Python – GeeksforGeeks." GeeksforGeeks. https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/ (Accessed: Nov. 26, 2022).

[24] "Machine learning, explained | MIT Sloan." MIT MANAGEMENT SLOAN SCHOOL. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained/ (Accessed: Nov. 26, 2022).

[25] "Machine learning random forest algorithm – javatpoint." Javatpoint. https://www.javatpoint.com/machine-learning-random-forest-algorithm/ (Accessed: Nov. 26, 2022).

[26] "Machine learning decision tree classification algorithm – javatpoint." Javatpoint. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm/ (Accessed: Nov. 26, 2022).

[27] "Machine learning naïve bayes classifier – javatpoint." Javatpoint. https://www.javatpoint.com/machine-learning-naive-bayes-classifier/ (Accessed: Nov. 26, 2022).

[28] "K-Nearest Neighbor(KNN) Algorithm for Machine Learning – javatpoint." Javatpoint. https://www.javatpoint.com/k-nearest-nighbor-algorithm-for-machine-learning/ (Accessed: Nov. 26, 2022).

[29] "Logistic Regression in Machine Learning – javatpoint." Javatpoint. https://www.javatpoint.com/logistic-regression-in-machine-learning/ (Accessed: Nov. 26, 2022).

[30] "Support Vector Machine Algorithm – javatpoint." Javatpoint. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm/ (Accessed: Nov. 26, 2022).

[31] "Precision and Recall in Machine Learning – javatpoint." Javatpoint. https://www.javatpoint.com/precision-and-recall-in-machine-learning/ (Accessed: Nov. 26, 2022).

[32] "Performance Metrics in Machine Learning – javatpoint." Javatpoint. https://www.javatpoint.com/performance-metrics-in-machine-learning/ (Accessed: Nov. 26, 2022).

[33] "Cross-Validation in Machine Learning – javatpoint." Javatpoint. https://www.javatpoint.com/cross-validation-in-machine-learning/ (Accessed: Nov. 26, 2022).

[34] "Confusion Matrix in Machine Learning – javatpoint." Javatpoint. https://www.javatpoint.com/confusion-matrix-in-machine-learning/ (Accessed: Nov. 26, 2022).

[35] "AUC-ROC Curve in Machine Learning – javatpoint." Javatpoint. https://www.javatpoint.com/auc-roc-curve-in-machine-learning/ (Accessed: Nov. 26, 2022).

APPENDIX

Reflections of Research

This is my first research on the field of ML. Before this, I had very limited knowledge of ML algorithms and techniques. So, I dealt with various problems during this study. Even finding a suitable dataset was hard to get at one point. I had to rely on a secondary dataset because financial institutions don't want to share data related to their debtors. So, data collection was one of the tougher tasks for me. Since I had limited knowledge of this area, I had to learn about ML from the beginning which was time-consuming, and at the same time, I had to study different papers related to my credit-risk analysis. I have encountered other challenges during coding for data visualization, handling the models, and showing the results that they showed. However, I was able to cope with the challenges. As a result, I have not only completed this study successfully but also acquired a good amount of knowledge as well as enhanced my abilities.

Abbreviations

- RF – Random Forest
- DT – Decision Tree
- NB – Naïve Bayes
- KNN – K-Nearest Neighbor
- LR – Logistic Regression
- SVM – Support Vector Machine
- AI – Artificial Intelligence
- ML – Machine Learning
- SL – Supervised Learning
- ROC = Receiver Operating Characteristic
- AUC – Area Under the ROC Curve

Turnitin Originality Report

Processed on: 13-Jan-2023 16:26 +06
ID: 1992230900
Word Count: 16885
Submitted: 1

CREDIT RISK ANALYSIS USING MACHINE LEARNING ALGORITHMS By Depayan Banerjee

Similarity Index	Similarity by Source	
13%	Internet Sources:	11%
	Publications:	5%
	Student Papers:	7%

2% match (student papers from 11-Feb-2018)
Submitted to Daffodil International University on 2018-02-11

1% match (Internet from 05-Oct-2016)
https://www.researchgate.net/publication/228205841

1% match (Internet from 26-Sep-2022)
https://www.econstor.eu/bitstream/10419/239554/1/1759759708.pdf

1% match (Trilok Nath Pandey, Alok Kumar Jagadev, Suman Kumar Mohapatra, Satchidananda Dehuri. "Credit risk analysis using machine learning classifiers", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017)
Trilok Nath Pandey, Alok Kumar Jagadev, Suman Kumar Mohapatra, Satchidananda Dehuri. "Credit risk analysis using machine learning classifiers", 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017

< 1% match (Internet from 26-Oct-2021)
https://www.researchgate.net/publication/227440487_Measuring_e-statement_quality_impact_on_customer_satisfaction_and_loyalty

< 1% match (Internet from 05-Sep-2022)
https://www.researchgate.net/publication/272237391_Smoothing_decision_boundaries_to_avoid_overfitting_in_neural_network_training

< 1% match (Internet from 26-Jun-2022)
https://www.researchgate.net/publication/356770549_Detection_of_Phising_Websites_using_Machine_Learning_Approaches

< 1% match (Internet from 25-Oct-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5414/192-25-763%20%286%25%29.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 20-Nov-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/3954/P15488%20%2811_%29_.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 21-Nov-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/7225/201-25-868%20%2820_%29.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 03-Jan-2023)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/4842/P15197%20%2819_%29_.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 26-Oct-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/7652/153-15-6334%20%2820_%29.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 20-Nov-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5300/161-15-7068%20%2822_%29.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 19-Nov-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/2549/P11656%20%2819%25%29.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 20-Nov-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/4926/P15171%20%2825_%29_.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 26-Dec-2022)
https://core.ac.uk/download/pdf/11601214.pdf

< 1% match (Internet from 12-Jan-2023)
https://dokumen.pub/internet-of-things-infrastructures-and-mobile-applications-proceedings-of-the-13th-imcl-conference-1st-ed-9783030499310-9783030499327.html

< 1% match (Internet from 02-Nov-2022)
https://dokumen.pub/soft-computing-theories-and-applications-proceedings-of-socta-2020-volume-1-1-9789811617409-9811617406.html

< 1% match (Internet from 28-Jan-2022)