# REPORT ON GERMAN CREDIT DATASET

## INT-247(MACHINE LEARNING FOUNDATION)

**SUBMITTED BY-** DIYALI SHIVHARE

**REGISTRATION NO**- 11801277

**SECTION**- KM001

**ROLL NO**- RKM001A07

# ABSTRACT

Analysis of credit scoring is an effective credit risk assessment technique, which is one of the major research fields in the banking sector. Machine learning has a variety of applications in the banking sector and it has been widely used for data analysis. Modern techniques such as machine learning have provided a self-regulating process to analyze the data using classification techniques. The classification method is a supervised learning process in which the computer learns from the input data provided and makes use of this information to classify the new dataset. This research paper presents a comparison of various machine learning techniques used to evaluate the credit risk. A credit transaction that needs to be accepted or rejected is trained and implemented on the dataset using different machine learning algorithms. The techniques are implemented on the German credit dataset taken from UCI repository which has 1000 instances and 21 attributes, depending on which the transactions are either accepted or rejected. This paper compares algorithms such as Logistic Regression, Support Vector Network, Naive Bayes, k-Nearest Neighbors and Decision trees and the results obtained show that Logistic Regression algorithm was able to predict credit risk with higher accuracy.

## Keywords:

Classification Algorithm, Credit Risk evaluation, Machine learning, supervised learning.

# INTRODUCTION

Since the last decade, the field of banking risk management has bloomed, and the importance of credit risk evaluation has increased in many sectors. The addition of credit scoring and credit risk evaluation was a major advantage to the banking sector. Credit scoring is a statistical study carried out by the financial institutions and the lenders to predict the potential risk, corresponding to a transaction whereas, credit risk evaluation can be defined as identification of the risk levels associated with the credit transaction as to whether the party will meet the commitment towards the agreed terms. Credit risk evaluation can be divided into two categories. In the first category, applicants are classified as "good" and "bad" credit risk. This is called application scoring where the data is categorized into groups based on financial data. In the second category, the payment pattern of the applicant along with payment history.

The objective of this paper is to perform a comparative evaluation of different techniques such as Logistic Regression, k-Nearest Neighbors, Naïve Bayes, Decision tree and Support Vector Machine algorithms in order to find out the most accurate system for determining credit risk. With the purpose to evaluate the techniques, they are tested using a real dataset of German credit data, and the obtained results are used to analyze the better technique, that could be used to accurately evaluate the banking sector's credit risk.

# DATASET

The dataset used for this research purpose is German credit data which categorizes customers based on the set of attributes, as a “good” or “bad” credit risk. This is an open-source dataset which is available on the UCI Machine Learning Repository. The dataset comprises of 1000 instances of 21 attributes including a classification attribute for each instance. The aim of this research is to predict the outcome of each instance as good or bad using the data set which is classified based on features or attributes, utilizing different machine learning classification algorithms, which are applied on the same data set to compare the accuracy of each of them. A total of 21 attributes are used for this analysis, each instance is characterized by the first 20 attributes and the last attribute is used to classify if a credit risk is good or bad. **A good credit risk is denoted by ‘1’ and a bad credit risk is denoted by ‘2’.**

**TABLE-1: Attributes and class of German credit dataset**

| ATTRIBUTE NUMBER | DESCRIPTION | CLASS |
|---|---|---|
| 1 | Status of existing checking account | Categorical |
| 2 | Duration in month | Numerical |
| 3 | Credit history | Categorical |
| 4 | Purpose | Categorical |
| 5 | Credit amount | Numerical |
| 6 | Savings account/bonds | Categorical |
| 7 | Present employment since | Categorical |
| 8 | Installment rate in percentage of disposable income | Numerical |
| 9 | Personal status and sex | Categorical |
| 10 | Other debtors / guarantors | Categorical |
| 11 | Present residence since | Numerical |
| 12 | Property | Categorical |
| 13 | Age in years | Numerical |
| 14 | Other installment plans | Categorical |
| 15 | Housing | Categorical |
| 16 | Number of existing credits at this bank | Numerical |
| 17 | Job | Categorical |
| 18 | Number of people being liable to provide maintenance for | Numerical |
| 19 | Telephone | Categorical |
| 20 | Foreign Worker | Categorical |
| 21 | Credit Risk | Categorical |

The table presents the different attributes and their classes which are either numeric or categorical in nature.

# PROPOSED METHODOLOGY

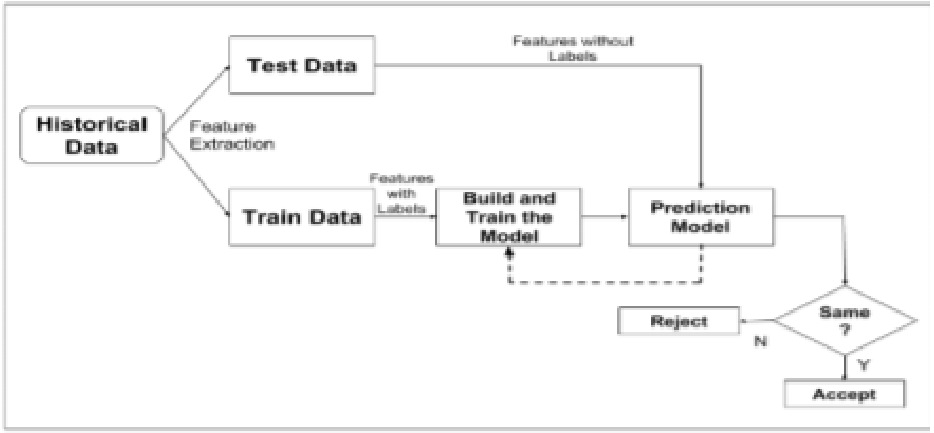
In order to effectively evaluate the credit risk using machine learning classification algorithms, the following architecture is proposed in this paper.

**Step 1**: The data set is split using feature extraction into training data and testing data.

**Step 2**: The various classification algorithms such as Logistic Regression, K Nearest Neighbor, Naive Bayes, Decision Tree and Support Vector Machine are applied to the training data to build a training model.

**Step 3**: A predictive model is built using the test data.

**Step 4**: The predictive model's output is compared to the model built using trained data.



## FLOWCHART OF PROPOSED METHODOLOGY

A brief description regarding the various techniques used to evaluate the model:

### 1. LOGISTIC REGRESSION

Logistic regression can be classified as a statistical technique in which a binomial output with explanatory variables can be modeled. Data is made to fit into a linear regression model; a logistic function is used to predict the categorical dependent variable. A binary logistic regression algorithm can predict only two possible outcomes for a categorical response. A multinomial logistic regression algorithm can predict three or more categories without any order. Ordinal logistic regression function is used to predict three or more categories with ordering. In this technique, generalized linear models (GLM) are used, which is designed to execute the generalized linear model regression on the output of binary data

### 2. K-NEAREST NEIGHBOURS

- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. It can be used for Regression as well as for Classification but mostly it is used for the Classification problems. It is a **non-parametric algorithm**, which means it does not make any assumption on underlying data. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

## 3. NAÏVE BAYES

Naive Bayes classifiers are a group of classification algorithms built on Bayes Theorem. In general, Naïve Bayes is a type of graphical probabilistic model which could be used to create other models based on data or expert opinion. They are used in various fields for prediction, anomaly detection, etc. They consist of a graph with nodes and directed links between them, where each node represents a variable and the arcs represent the relationship among them [15]. The center of Bayesian learning uses the Bayes theorem which states that given a joint probability distribution over events A and B, then the conditional probability is given by

$$P\ (A\ |B) = (P\ (B|\ A) * P(A)\ )\ /\ P(B)$$

In fraud detection, information about the Naïve Bayes is not available, but the set of variables which are the cause of the frauds can be calculated using the same theorem.

## 4. SUPPORT VECTOR MACHINE

Support Vector Machine(SVM) can be described as a supervised machine learning model that can be used to analyze data, for regression as well as classification analysis, using associated learning algorithms. It is commonly used for classification analysis. Each data element in this algorithm is sketched as a point in the m-dimensional area (m is the count of features) where the cost of each feature corresponds to a specific coordinate [9]. A hyper-plane is found, which best suits to classify the two classes appropriately. It is a selective classifier that is formally defined by an independent hyper-plane. Given supervised training data, an optimal hyper-plane is produced as an output that classifies the data.

5. **DECISION TREE**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

## Data-Preprocessing:

The dataset consists of a mixture of categorical and numeric variables. Categorical variable values are limited and based on a finite group whereas numeric variables can take any value from integer to decimal. The pre-processing of the dataset begins with a characterization of the dataset where the lower significance items were removed, and numerical variables were categorized.

## TRAINING and TESTING DATASET:

The dataset is split for training and testing purposes. K-Fold-Cross-Validation technique is commonly used due to the higher accuracy. K-Fold-Cross-Validation technique has been used to split the data that operates on multiple percentage divisions. The data is divided into 3:7 ratio. The training part is given 30% and the testing part is given 70%.

# RESULT AND DISCUSSION

| ALGORITHM | TRAINING ACCURACY | TESTING ACCURACY |
|---|---|---|
| LOGISTIC REGRESSION | 0.76 | 0.75 |
| K NEAREST NEIGHBOURS | 0.76 | 0.68 |
| NAÏVE BAYES | 0.72 | 0.73 |
| DECISION TREE | 0.70 | 0.72 |
| SUPPORT VECTOR MACHINE | 0.72 | 0.72 |

Training as well as Testing Accuracy was highest in Logistic Regression. The training accuracy is 76% and testing accuracy is 75%. The training accuracy of K nearest neighbor algorithm is also same as logistic regression i.e. 76%. The lowest Training Accuracy is given by Decision Tree and the lowest Testing Accuracy was given by K nearest neighbor. Overall, **the best algorithm for classification of German Credit Dataset is Logistic Regression.**

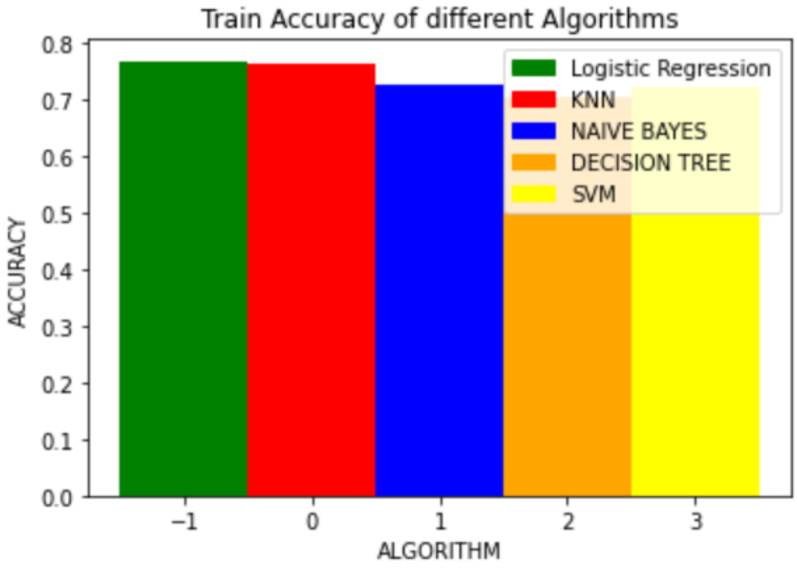


**FIG-1: BAR GRAPH BETWEEN TRAINING ACCURACY OF ALGORITHMS**

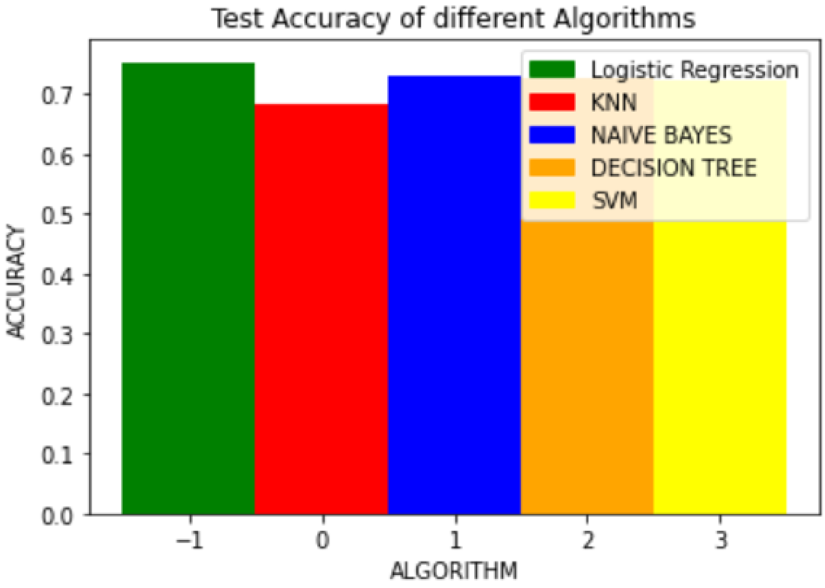


**FIG-2: BAR GRAPH BETWEEN TESTING ACCURACY OF ALGORITHMS**

# CONCLUSION

Global markets are full of risks and many attempts have been made to find quick and efficient ways to predict the future. The introduction of credit scores and credit risk evaluation was a major advantage for the banking sector. In this paper, different machine learning techniques were compared to evaluate the credit risk in the German credit dataset. These have been implemented and tested on various classification algorithms such as Logistic Regression, K Nearest Neighbor, Naïve Bayes, Decision Tree and Support Vector Machine. The techniques are tested by applying them on an existing dataset called German credit with thousand transactions per day. From the above analysis, using the Logistic Regression methodology provides higher accuracy of credit risk evaluation. As future work, various deep learning techniques can be evaluated to see if accuracy increases.

# REFERENCES

1. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
2. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
3. https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html
4. https://machinelearningmastery.com/logistic-regression-for-macine-learning
5. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code

**GITHUB LINK: https://github.com/diyali0811/german_credit_dataset**