# 19EAC381 MACHINE LEARNING LAB USING PYTHON

**An Approach for Prediction of Loan Approval using Machine Learning Algorithm**

A Technical Report

Submitted by

**GROUP 6**

**Group Members:**

B. Shahid (AM.EN.U4EAC21020)

C. Chandu (AM.EN.U4EAC21022)

K. Ruthwik (AM.EN.U4EAC21037)

K. Varsha (AM.EN.U4EAC21038)

L. Ritish Rishi (AM.EN.U4EAC21043)

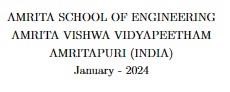
*under the guidance of*

Dr. Manazhy Rashmi, M. Ragesh Rajan *submitted as part of*

19EAC381 PROJECT PHASE I in

ELECTRONICS AND COMMUNICATION ENGINEERING





AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM

AMRITAPURI (INDIA)



# **BONAFIDE CERTIFICATE**

This is to certify that the report entitled **“An Approach for Prediction of Loan Approval using Machine Learning Algorithm”** submitted by **B. Shahid(AM.EN.U4EAC21020)**, **C. Chandu (AM.EN.U4EAC21022)**, **K. Ruthwik (AM.EN.U4EAC21037)**, **K. Varsha (AM.EN.U4EAC21038)**, **L. Ritish Rishi(AM.EN.U4EAC21043)** as part of the 19EAC381 PROJECT PHASE I is a bonafide record of the work carried out by her under my guidance and supervision at Amrita School of Engineering, Amritapuri.

**Signature of Supervisor:**

Dr.Manazhy Rashmi

Assistant professor

Dept.of ECE

**Signature of Co-Supervisor:**

M.Ragesh Rajan

Assistant professor

Dept.of ECE

Date: 22-01-2024

**DEPARTMENT OF ECE**

## **DECLARATION**

**We,**

**B. Shahid (AM.EN.U4EAC21020)**

**C. Chandu (AM.EN.U4EAC21022)**

**K. Ruthwik (AM.EN.U4EAC21037)**

**K. Varsha (AM.EN.U4EAC21038)**

**L. Ritish Rishi (AM.EN.U4EAC21043)**

hereby declare that this technical report entitled **“An Approach for Prediction of Loan Approval using Machine Learning Algorithm”,** is the record of the original work done by us under the guidance of **Dr.Manazhy Rashmi, M.Ragesh Rajan**, Assistant professors, Department of ECE, Amrita School of Engineering, Amritapuri.

**Place: Amritapuri Signature of the Students**

**Date: 22-01-2024**

# **Acknowledgement**

We would like to sincerely thank and appreciate everyone who helped us finish our project. First and foremost, we would want to express our sincere gratitude to **Rashmi mam**, our project supervisor, whose advice and experience were very helpful during the whole process. Their encouraging words and perceptive criticism were really helpful in helping us to improve our work.

We also express our gratitude to **Ragesh sir** for lending us the tools and spaces required to carry out this study. The encouragement of a cooperative and stimulating environment has been greatly aided by the support of our peers and colleagues.

And finally, we would like to thank all the team members for their contribution. To overcome obstacles and complete the project on time, our group's commitment and cooperation were crucial. Without the combined efforts and dedication of all those involved, this project would not have been feasible.

Regards.

Group 6

S5-EAC

# Contents

**Acknowledgement 4**

**List of Figures**

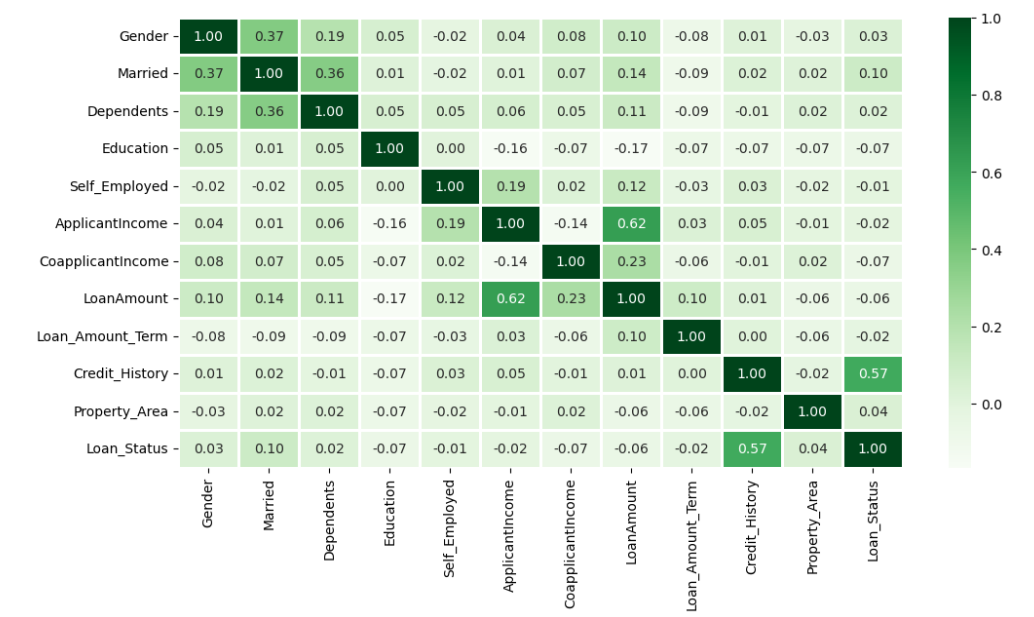
1. **Introduction 1** 
   1. **Introduction** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **1**
   2. **Objective**. . . . . . . . . . . . . . . . . . . . . . . . . . . . .  **1**

|  |  |
| --- | --- |
| **2 Proposed Methodology** | **2** |
| **3 Results and Discussions** | **3** |
| **4 Conclusion** | **4** |

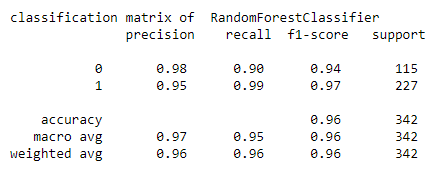
* 1. **Scope** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . **1**

# List Of figures

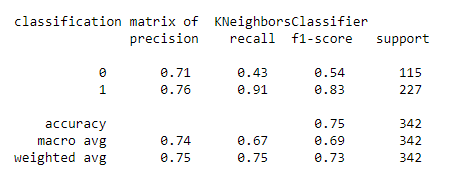
* Correlation heatmap for the given dataset,



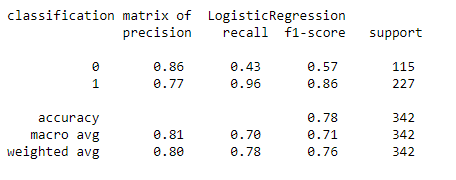
* Classification matrix regarding random forest,



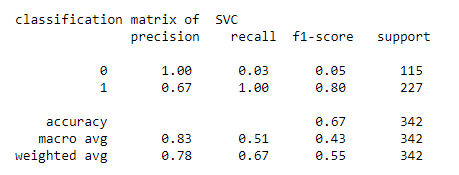
* Classification matrix regarding KNN,



* Classification Matrix for Logistic Regression,



* Classification Matrix for SVM,



Chapter 01

# INTRODUCTION

## 1.1 Background

This project discusses the implementation of a loan eligibility prediction system using Machine Learning models. It explores the application of different machine learning models to analyse and classify loan applicants based on various features. The process involves preprocessing the data, selecting relevant features, and training the different models to make accurate predictions and compare the performances of different models. The article also addresses the importance of model evaluation and performance metrics in assessing the effectiveness of the prediction system. Overall, it serves as a practical guide for utilizing different machine learning models in Python for loan eligibility prediction, catering to individuals interested in applying machine learning techniques to financial decision-making processes.

## 1.2 Objective

The objective of the article is to detect loan approval of a individual using machine learning algorithms and the provided dataset. The study involves the analysis of various features such as **Education, Income, Credit History, Loan Amount term etc.** Additionally, the dataset is pre processed, and different classifiers including SVM, KNN, Logistic Regression, and Random Forest are trained and evaluated for accuracy and precision.

## 1.3 Scope

Loans are a major demand in today's world. Banks earn a significant portion of their profits from this alone. It is advantageous for students to manage their schooling and living expenditures, as well as for people to purchase luxury items such as homes, vehicles, and so on. However, while determining whether the applicant's profile is significant to being given a loan or not. Banks must deal with a wide range of issues. So, we'll be using Machine Learning with Python to make their job easier and forecast if a candidate's profile is relevant or not based on crucial criteria such as marital status, education, applicant income, credit history, and so on.

Chapter 02

# Methodology

## 3.1 Data Collection

The requirements for extending a loan to a specific individual serve as the cornerstone of this project. The dataset includes key attributes such as education, self-employment, income, loan amount, loan amount term, loan status, and so on. The dataset is useful for training and testing machine learning models because it collects data throughout a variety of time periods.

## 3.2 Feature Extraction

Feature extraction is a crucial step in constructing a robust predictive model.

## 3.3 Model Implementation

The project focuses on the implementation of four different machine learning models: Support Vector Machine (SVM) with the Linear Function kernel, Random Forest, K-Nearest Neighbours (KNN), and Logistic Regression. Each model is chosen for its distinct characteristics, resulting in a complete and diversified study.

* SVM with linear Kernel:

Support Vector Machines (SVMs) with a linear kernel are excellent at capturing linear relationships within data, making them useful for applications where the decision boundary is best represented as a straight line or hyperplane. Their simplicity and efficacy apply to high-dimensional datasets such as text and image recognition. Notably, these models lower the risk of overfitting, especially in cases with a large feature-to-sample ratio. Their linear decision function improves interpretability while being memory efficient by using a subset of support vectors. SVMs with linear kernels seek a global optimal solution, making them effective in binary classification and resistant to outliers. These characteristics work together to make them useful in a variety of machine learning applications.

* Random Forest:

Random Forest stands out in machine learning for its excellent accuracy, resistance to outliers, and efficient handling of missing variables. As an ensemble approach, it integrates many decision trees to reduce overfitting while offering robust generalisation. The algorithm's non-parametric character makes it suitable for a variety of jobs that do not require considerable data preprocessing. It excels at classification and regression, provides insights into feature relevance, and manages complex variable interactions. Random Forest's parallelizable architecture provides computational efficiency, giving it a dependable alternative for applications requiring accuracy, interpretability, and robustness to outliers.

* K-Nearest Neighbors (KNN):

The k-Nearest Neighbours (KNN) method is popular because of its simplicity, versatility, and adaptability to different data patterns. With a straightforward methodology and a non-parametric nature, it excels at catching local patterns without assuming data distribution. KNN does not require a formal training step, making it ideal for dynamic datasets, and its resistance to outliers improves performance in noisy data. The computational efficiency of KNN for distance calculations is a significant advantage when working with tiny datasets. Despite these advantages, considerations include susceptibility to irrelevant features and inefficiencies with big datasets. Overall, KNN is a simple and effective solution for situations in which its specific capabilities complement the features of the data.

* Logistic Regression:

Logistic regression is valued for its interpretability, making it an excellent choice when understanding feature impacts is critical. It performs well with tiny datasets, is less prone to overfitting, and thrives in instances involving linearly separable data. The model's probabilistic outputs provide useful probability scores for decision makers. Logistic regression is a feasible solution for binary classification jobs that need transparency and simplicity, as it has minimal sensitivity to outliers and is simple to use.

## 3.4 Evaluation Metrices

Evaluating machine learning models is an important stage in analysing their performance and how effectively they meet the objectives of a certain task. Different evaluation metrics provide insights into various aspects of a model's performance, and the metrics used are determined by the task's unique objectives.

1. Accuracy:

- Accuracy is a regularly used metric for determining the overall validity of model predictions. It is computed as the proportion of accurately anticipated cases to the total number of instances. While accuracy is an important statistic, it may not be enough for assignments with imbalanced classes, in which one class outnumbers the other.

2. Precision:

- Precision is defined as the ratio of true positives to the sum of true and false positives. It assesses the model's ability to anticipate favourable outcomes. Precision is especially crucial when the cost of false positives is significant. For example, in the context of heart disease prediction, precision would reflect how many projected cases of heart disease were correct.

Chapter 03

Machine Learning Models

The design and execution of appropriate machine learning models is critical to the success of predicting loan approvals. In this experiment, four different models—Support Vector Machine, Random Forest, Logistic Regression, and K-Nearest Neighbours (KNN)—were used, each with its own set of skills for predicting heart disease.

4.1 Support Vector Machine (SVM):

Support Vector Machine (SVM) is a machine learning algorithm used for loan approval prediction. In this context, SVM operates by finding an optimal hyperplane that separates the dataset into two classes: approved and rejected loans. The algorithm aims to maximize the margin between these classes, ensuring robust generalization to new data points.

4.2 Random Forest:

A random forest model for loan approval prediction is a machine learning algorithm that leverages an ensemble of decision trees to make accurate and robust predictions regarding whether a loan application should be approved or denied. Each decision tree in the forest is constructed by randomly selecting a subset of features from the dataset and growing the tree based on these features.

4.3 K-Nearest Neighbours (KNN):

A k-Nearest Neighbors (KNN) model for loan approval prediction is a supervised machine learning algorithm used to classify whether a loan application should be approved or denied based on the characteristics of previous loan data. In this context, the algorithm considers the similarity between the features of a new loan application and those of existing approved and denied loans.

4.4 Logistic Regression:

Logistic regression is a statistical technique used in machine learning for binary classification tasks, which means it can predict outcomes with two potential classes. In the context of loan approval prediction, logistic regression can be used to determine whether a loan application will be granted or denied depending on pertinent factors such as income, credit score, employment status, and debt-to-income ratio.

Chapter 04

Results

The loan approval data was analysed using a variety of factors, including education, self-employment, income, loan amount, loan length, and loan status. Exploratory data analysis was performed to better understand the correlations between the characteristics, and various machine learning methods such as SVM, KNN, Logistic Regression, and Random Forest were trained and tested for accuracy and precision. The results showed that Random Forest fared the best in terms of accuracy and precision.

In conclusion, the examination of loan approval data using machine learning algorithms revealed that Random Forest was the most effective classifier in terms of accuracy and precision.

Chapter 05

Conclusion

Finally, our project concentrated on loan approval prediction utilising machine learning methods such as Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Random Forest, and Logistic Regression. After a thorough study, we discovered that Random Forest was the algorithm with the highest accuracy among the tested models. While other algorithms provided useful insights, Random Forest emerged as the most accurate and interpretable approach for predicting loan approval in our dataset. This study emphasises the necessity of considering both accuracy and interpretability when choosing a machine learning model for real-world applications, which are mostly employed in the banking sector.

Chapter 06

References

1. Mohammad Ahmad Sheikh, Amit Kumar Goel , Tapas Kumar, “An Approach for Prediction for loan approval using machine learning”,
2. P. Lyn C. Thomas, David B. Edelman, and Jonathan N. Crook, ”Credit Scoring and Its Applications”, Journal of the Operational Research Society, 2002.