

# **Comparative Analysis of Loan Status Prediction using different Machine Learning Algorithms**

**A MINI PROJECT**

*Submitted by*

**AKSHIT CHAUDHARY (RA2011027010106)**

**SAI SANJANA (RA2011027010109)**

**SIDDHANT SHEKHAR (RA2011027010130)**

*Under the guidance of*

**Dr. E. Sasikala**

**Professor**

**Department of Data Science and Business Systems**

In partial fulfilment for the

Course of

**18CSE392T- Machine Learning-I**

in

**Department of Data Science and Business Systems**



**SCHOOL OF COMPUTING  
COLLEGE OF ENGINEERING AND TECHNOLOGY  
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY  
KATTANKULATHUR – 603203**

**October 2023**



COLLEGE OF ENGINEERING & TECHNOLOGY  
SRM INSTITUTE OF SCIENCE & TECHNOLOGY  
S.R.M. NAGAR, KATTANKULATHUR – 603 203

## BONAFIDE CERTIFICATE

Certified that this mini project report "**Comparative Analysis of Loan Status Prediction using different Machine Learning Algorithms** " is the Bonafide work of **Akshit Chaudhary (RA2011027010106)**, **Sai Sanjana (RA2011027010109)** and **Siddhant Shekhar (RA2011027010130)** who carried out the project work under my supervision.

Dr. E. Sasikala  
Professor  
Department of Data Science and Business Systems  
SRM institute of science and technology

Dr. M Lakshmi  
Professor & HOD  
Department of DSBS  
SRM institute of science and technology

## **ABSTRACT**

Technology has boosted the existence of humankind and the quality of life they live. Every day we are planning to create something new and different. We have a solution for every other problem. We have machines to support our lives and make us somewhat complete in the banking sector candidate gets proofs/ backup before approval of the loan amount. The application approved or not approved depends upon the historical data of the candidate by the system. Every day lots of people apply for the loan in the banking sector but Bank would have limited funds. In this case, the right prediction would be very beneficial using some class-function algorithm. An example the logistic regression, random forest classifier, support vector machine classifier, etc. A Bank's profit and loss depend on the amount of the loans, that is whether the Client or customer is paying back the loan. Recovery of loans is the most important for the banking sector. The improvement process plays an important role in the banking sector. The historical data of candidates was used to build a machine learning model using different classification algorithms. The main objective of this paper is to predict whether a new applicant granted the loan or not using machine learning models trained on the historical data set.

## **TABLE OF CONTENTS**

<b>CHAPTER NO.</b>		<b>TITLE</b>	<b>PAGE NO.</b>
		<b>ABSTRACT</b>	3
		<b>TABLE OF CONTENTS</b>	4
		<b>LIST OF FIGURES</b>	5
		<b>ABBREVIATIONS</b>	6
1.		<b>INTRODUCTION</b>	
	1.1	Aim, Synopsis	7
	1.2	Requirements Specification	8
2.		<b>LITERATURE SURVEY</b>	
	2.1	Literature Review	15
3.		<b>SYSTEM ARCHITECTURE AND DESIGN</b>	
	3.1	Architecture Diagram	16
	3.2	ER Diagram	17
	3.3	Use case Diagram	18
4.		<b>MODULES AND FUNCTIONALITIES</b>	
	4.1	Modules	19
	4.2	Design and Implementation Constraints	20
	4.3	Other Nonfunctional Requirements	21
5.		<b>CODING AND OUTPUT</b>	22-25
6.		<b>RESULTS AND DISCUSSION</b>	26
7.		<b>REFERENCES</b>	27

### **LIST OF FIGURES**

<b>Figure No.</b>	<b>Figure Name</b>	<b>Page No</b>
3.1	Architecture Diagram	<b>10</b>
3.2	Use case Diagram	<b>12</b>
3.3	ER Diagram	<b>13</b>

## **ABBREVIATIONS**

<b>CSS</b>	Cascading Style Sheet
<b>DB</b>	Data Base
<b>ER</b>	Entity Relationship
<b>SQL</b>	Structured Query Language
<b>HTML</b>	Hyper Text Markup Language
<b>UI</b>	User Interface

## **OBJECTIVE**

### **Aim:**

To determine the loan approval system using machine learning algorithms.

### **Synopsis:**

Loan approval is a very important process for banking organizations. The systems approved or rejected the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. In recent years many researchers have worked on loan approval prediction systems. Machine Learning (ML) techniques are very useful in predicting outcomes for large amount of data. In this paper different machine learning algorithms are applied to predict the loan approval of customers. In this paper, various machine learning algorithms that have been used in past are discussed and their accuracy is evaluated. The focus of this paper is to determine whether the loan given to a particular person, or an organization shall be approved or not.

# **REQUIREMENT SPECIFICATIONS**

## **INTRODUCTION**

Prediction of modernized loan approval system based on machine learning approach is a loan approval system from where we can know whether the loan will pass or not. In this system, we take some data from the user like his monthly income, marriage status, loan amount, loan duration, etc. Then the bank will decide according to its parameters whether the client will get the loan or not. So, there is a classification system, in this system, a training set is employed to make the model and the classifier may classify the data items into their appropriate class. A test dataset is created that trains the data and gives the appropriate result that is the client potential and can repay the loan. Predicting a modernized loan approval system is incredibly helpful for banks and the clients. This system checks the candidate on his priority basis. A customer can submit his application directly to the bank so the bank will do the whole process, no third party or stockholder will interfere in it. And finally, the bank will decide that the candidate is deserving or not on its priority basis. The only object of this research paper is that the deserving candidate gets straight forward and quick results.

## **HARDWARE AND SOFTWARE SPECIFICATION**

### **HARDWARE REQUIREMENTS**

- Hard disk : 500 GB and above.
- Processor : i3 and above.
- Ram : 4GB and above.

### **SOFTWARE REQUIREMENTS**

- Operating System : Windows 10
- Software : python
- Tools : Anaconda (Jupyter Notebook IDE)

### **TECHNOLOGIES USED**

- Programming Language : **Python**

## **INTRODUCTION TO PYTHON**

Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is a programming language that lets you work quickly and integrate systems more efficiently.

It is used for:

- web development (server-side),
- software development,
- mathematics,
- System scripting.



### What can Python do?

- Python can be used on a server to create web applications.
- Python can be used alongside software to create workflows.
- Python can connect to database systems. It can also read and modify files.
- Python can be used to handle big data and perform complex mathematics.
- Python can be used for rapid prototyping, or for production-ready software development.

### Why Python?

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc.).
- Python has a simple syntax like the English language.
- Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.
- Python can be treated in a procedural way, an object-orientated way, or a functional way.

### Good to know.

- The most recent major version of Python is Python 3, which we shall be using in this tutorial. However, Python 2, although not being updated with anything other than security updates, is still quite popular.
- Python 2.0 was released in 2000, and the 2.x versions were the prevalent releases until December 2008. At that time, the development team made the decision to release version 3.0, which contained a few relatively small but significant changes that were not backward compatible with the 2.x versions. Python 2 and 3 are very similar, and some features of Python 3 have been backported to Python 2. But in general, they remain not quite compatible.
- Both Python 2 and 3 have continued to be maintained and developed, with periodic release updates for both. As of this writing, the most recent versions available are 2.7.15 and 3.6.5. However, an official End of Life date of January 1, 2020, has been established for Python 2, after which time it will no longer be maintained.
- Python is still maintained by a core development team at the Institute, and Guido is still in charge, having been given the title of BDFL (Benevolent Dictator for Life) by the 12 Python community. The name Python derives not from the snake, but from the British comedy troupe Monty Python's Flying Circus, of which Guido was, and presumably still is, a fan. It is common to find references to Monty Python sketches and movies scattered throughout the Python documentation.
- It is possible to write Python in an Integrated Development Environment, such as Thonny, PyCharm, NetBeans or Eclipse which are particularly useful when managing larger collections of Python files.

### Python Syntax compared to other programming languages.

- Python was designed to for readability and has some similarities to the English language with influence from mathematics.
- Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.
- Python relies on indentation, using whitespace, to define scope, such as the scope of loops, functions, and classes. Other programming languages often use curly brackets for this purpose.

## Python is Interpreted

- Many languages are compiled, meaning the source code you create needs to be translated into machine code, the language of your computer's processor, before it can be run. Programs written in an interpreted language are passed straight to an interpreter that runs them directly.
- This makes for a quicker development cycle because you just type in your code and run it, without the intermediate compilation step.
- One potential downside to interpreted languages is execution speed. Programs that are compiled into the native language of the computer processor tend to run more quickly than interpreted programs. For some applications that are particularly computationally intensive, like graphics processing or intense number crunching, this can be limiting.
- In practice, however, for most programs, the difference in execution speed is measured in milliseconds, or seconds at most, and not appreciably noticeable to a human user. The expediency of coding in an interpreted language is typically worth it for most applications.
- For all its syntactical simplicity, Python supports most constructs that would be expected in a very high-level language, including complex dynamic data types, structured and functional programming, and object-oriented programming.
- Additionally, a very extensive library of classes and functions is available that provides capability well beyond what is built into the language, such as database manipulation or GUI programming.
- Python accomplishes what many programming languages don't: the language itself is simply designed, but it is very versatile in terms of what you can accomplish with it.

## Machine learning

### **Introduction:**

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through learning. In its application across business problems, machine learning is also referred to as predictive analytics.

### **Machine learning tasks:**

Machine learning tasks are classified into several broad categories. In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. For example, if the task were determining whether an image contained a certain object, the training data for a supervised learning algorithm would include images with and without that object (the input), and each image would have a label (the output) designating whether it contained the object. In special cases, the input may be only partially available, or restricted to special feedback. Semi algorithms develop mathematical models from incomplete training data, where a portion of the sample input doesn't have labels. Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set

of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not spam", represented by the Boolean values true and false. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. Examples of a continuous value are the temperature, length, or price of an object. In unsupervised learning, the algorithm builds a mathematical model from a set of data that contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. Unsupervised learning can discover patterns in the data, and can group the inputs into categories, as in feature learning. Dimensionality reduction is the process of reducing the number of "features", or inputs, in a set of data. Active learning algorithms access the desired outputs (training labels) for a limited set of inputs based on a budget and optimize the choice of inputs for which it will acquire training labels. When used interactively, these can be presented to a human user for labeling. Reinforcement learning algorithms are given feedback in the form of positive or negative reinforcement in a dynamic environment and are used in autonomous vehicles or in learning to play a game against a human opponent. Other specialized algorithms in machine learning include topic modeling, where the computer program is given a set of natural language documents and finds other documents that cover similar topics. Machine learning algorithms can be used to find the unobservable probability density function in density estimation problems. Meta learning algorithms learn their own inductive bias based on previous experience. In developmental robotics, robot learning algorithms generate their own sequences of learning experiences, also known as a curriculum, to cumulatively acquire new skills through self-guided exploration and social interaction with humans. These robots use guidance mechanisms such as active learning, maturation, motor synergies, and imitation.

### **Types of learning algorithms:**

The types of machine learning algorithms differ in their approach, the type of data they input and output, and the type of task or problem that they are intended to solve.

#### **Supervised learning:**

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as training data and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an array or vector, sometimes called a feature vector, and the training data is represented by a matrix. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs. An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task. Supervised learning algorithms include classification and regression. Classification algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range. Similarity learning is an area of supervised machine learning closely related to regression and classification, but the goal is to learn from examples using a similarity function that measures how similar or related two objects are. It has applications in ranking, recommendation systems, visual identity tracking, face verification, and speaker verification. In the case of semi-supervised learning algorithms, some of the training examples are missing training labels, but they can nevertheless be used to improve the quality of a model. In weakly supervised learning, the training labels are noisy, limited, or imprecise; however, these labels are often cheaper to obtain, resulting in larger effective training sets.

**Unsupervised learning:**

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labeled, classified, or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. A central application of unsupervised learning is in the field of density estimation in statistics, though unsupervised learning encompasses other domains involving summarizing and explaining data features. Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to one or more predesignated criteria, while observations drawn from different clusters are dissimilar. Different clustering techniques make different assumptions on the structure of the data, often defined by some similarity metric, and evaluated, for example, by internal compactness, or the similarity between members of the same cluster, and separation, the difference between clusters. Other methods are based on estimated density and graph connectivity.

**Semi-supervised learning:**

Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce a considerable improvement in learning accuracy.

**K-Nearest Neighbors**

Introduction In four years of analytics built more than 80% of classification models and just 15- 20% regression models. These ratios can be generalized throughout the industry. The reason for a bias towards classification models is that most analytical problems involve making a decision. For instance, will a customer attrite or not, should we target customer X for digital campaigns, whether customer has a high potential or not etc. This analysis is more insightful and directly links to an implementation roadmap. In this article, we will talk about another widely used classification technique called K-nearest neighbors (KNN). Our focus will be primarily on how does the algorithm work and how does the input parameter effect the output/prediction.

**KNN algorithm**

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique, we generally look at 3 important aspects:

1. Ease to interpret output
2. Calculation time
3. Predictive Power

**Decision tree**

In a decision tree, the algorithm starts with a root node of a tree then compares the value of different attributes and follows the next branch until it reaches the end leaf node. It uses different algorithms to check the split and variable that allow the best homogeneous sets of population. decision trees are widely used in data science. It is a key proven tool for making decisions in complex scenarios. In Machine learning, ensemble methods like decision tree, random forest are widely used. Decision trees are a type of supervised learning algorithm where data will continuously be divided into different categories according to certain parameters. So, in this blog, I will explain the Decision tree algorithm. How is it used? How its functions will cover everything that is related to the decision tree.

What is a Decision Tree?

Decision tree as the name suggests is a flow like a tree structure that works on the principle of conditions. It is efficient and has strong algorithms used for predictive analysis. It has mainly been attributed to internal nodes, branches, and a terminal node. Every internal node holds a “test” on an attribute, branches hold the conclusion of the test, and every leaf node means the class label. This is the most used algorithm when it comes to supervised learning techniques. It is used for both classifications as well as regression. It is often termed as “CART” that means Classification and Regression Tree. Tree algorithms are always preferred due to stability and reliability.

How can an algorithm be used to represent a tree Let us see an example of a basic decision tree where it is to be decided in what conditions to play cricket and in what conditions not to play. You might have got a fair idea about the conditions on which decision trees work with the above example. Let us now see the common terms used in Decision Tree that is stated below:

- Branches - Division of the whole tree is called branches.
- Root Node - Represent the whole sample that is further divided.
- Splitting - Division of nodes is called splitting.
- Terminal Node - Node that does not split further is called a terminal node.
- Decision Node - It is a node that also gets further divided into different sub-nodes being a sub node.
- Pruning - Removal of sub nodes from a decision node.
- Parent and Child Node - When a node gets divided further then that node is termed as parent node whereas the divided nodes or the sub-nodes are termed as a child node of the parent node.

Introduction to Logistics

Logistics refers to the overall process of managing how resources are acquired, stored, and transported to their destination. Logistics management involves identifying prospective distributors and suppliers and determining their effectiveness and accessibility. What are the 3 types of logistics?

Logistics has three types:

inbound,  
outbound, and  
reverse logistics.

What are the 7 R's of logistics?

So, what are the 7 Rs? The Chartered Institute of Logistics & Transport UK (2019) defines them as:

Getting the Right product,  
in the Right quantity,  
in the Right condition,  
at the Right place,  
at the Right time,  
to the Right customer,  
at the Right price.

What is the importance of logistics?

Logistics is an important element of a successful supply chain that helps increase the sales and profits of businesses that deal with the production, shipment, warehousing, and delivery of products. Moreover, a reliable logistics service can boost a business' value and help in maintaining a positive public image. What is logistics in real life?

Logistics is the strategic vision of how you will create and deliver your product or service to your end customer. If you take the city, town, or village that you live in, you can see a very clear example of what the logistical strategy was when they were designing it. What are the 3 main activities of logistics systems? Logistics activities or Functions of Logistics

- Order processing. The logistics activities start from the order processing, which might be the work

of the commercial department in an organization.

- Materials handling.
- Warehousing.
- Inventory control.
- Transportation.
- Packaging.

What are 3PL and 4PL in logistics?

A 3PL (third-party logistics) provider manages all aspects of fulfillment, from warehousing to shipping. A 4PL (fourth-party logistics) provider manages a 3PL on behalf of the customer and other aspects of the supply chain. What are the five major components of logistics?

There are five elements of logistics:

- Storage, warehousing, and materials handling.
- Packaging and unitization.
- Inventory.
- Transport.
- Information and control.

What is logistic cycle?

Logistics management cycle includes key activities such as product selection, quantification and procurement, inventory management, storage, and distribution. Other activities that help drive the logistics cycle and are also at the heart of logistics are organization and staffing, budget, supervision, and evaluation.

Why did you choose logistics?

We chose logistics because it is one of the most important career sectors in the globe and be more excited about it. ... I prefer my profession to work in logistics and it can be a challenging field, and with working in it I want to make up an important level of satisfaction in their jobs.

What is logistics and SCM?

The basic difference between Logistics and Supply Chain Management is that Logistics management is the process of integration and maintenance (flow and storage) of goods in an organization whereas Supply Chain Management is the coordination and management (movement) of supply chains of an organization Here are 6 steps logistics companies should follow to develop a sound logistics marketing plan.

1. Define your service offer. ...
2. Determine your primary and secondary markets. ...
3. Identify your competition. ...
4. Articulate your value proposition. ...
5. Allocate a marketing budget. ...
6. Develop a tactical marketing plan

## **LITERATURE REVIEW**

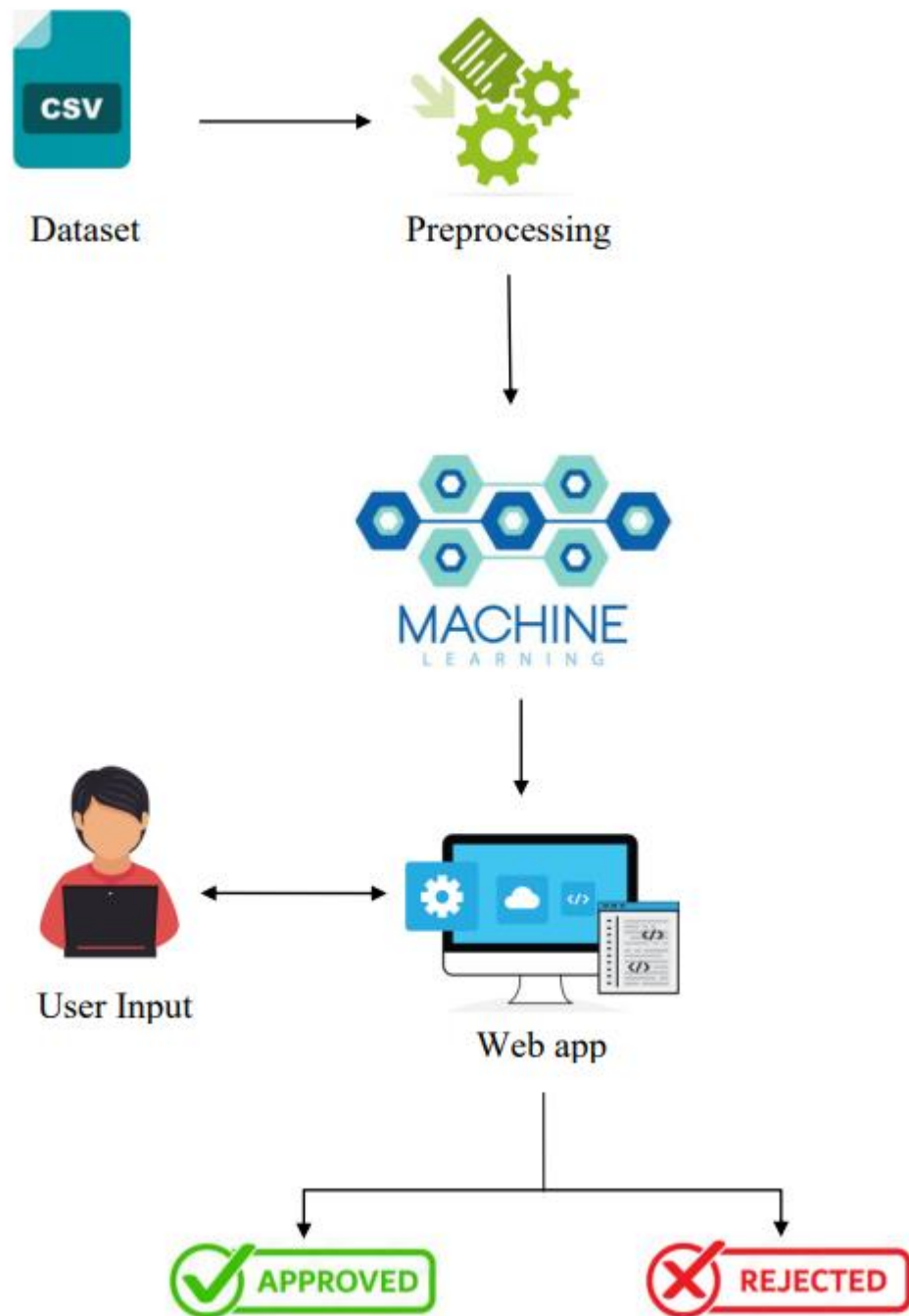
In the domain of loan prediction using machine learning models, a comprehensive literature survey revealed several noteworthy contributions. In the paper titled "Loan Prediction by using Machine Learning Models" (2017) by P Supriya and M Pavani, the authors focused on data collection and preprocessing, leveraging machine learning models, and implementing training and testing modules. During preprocessing, they paid particular attention to outlier detection and removal, as well as imputation processes. Their model, employing gradient boosting techniques, aimed to predict loan approval outcomes, following an 80:20 dataset split. The Decision Tree model stood out with an impressive accuracy of 81.1%.

Another significant study, "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R" (2019) by Sudhamathy G., introduced a risk analysis approach for loan sanctioning. The research encompassed data selection, preprocessing, feature extraction and selection, model construction, prediction, and evaluation. The dataset from the USI repository was employed, and the Logistic Regression classifier was built after meticulous preprocessing. This method achieved a precision of 83.3%.

In the paper "Developing Prediction Model of Loan Risk in Banks using Data Mining" (2020) by Jafar Hamid and Tarig Mohammed Ahmed, three algorithms, namely j48, Bayes Net, and Naive Bayes, were employed to construct predictive models for classifying loan applications as good or bad based on customer behavior and past credit repayment history. The Weka application was used to develop the model, and it was observed that the j48 algorithm outperformed others in terms of accuracy and mean absolute error, making it the preferred choice.

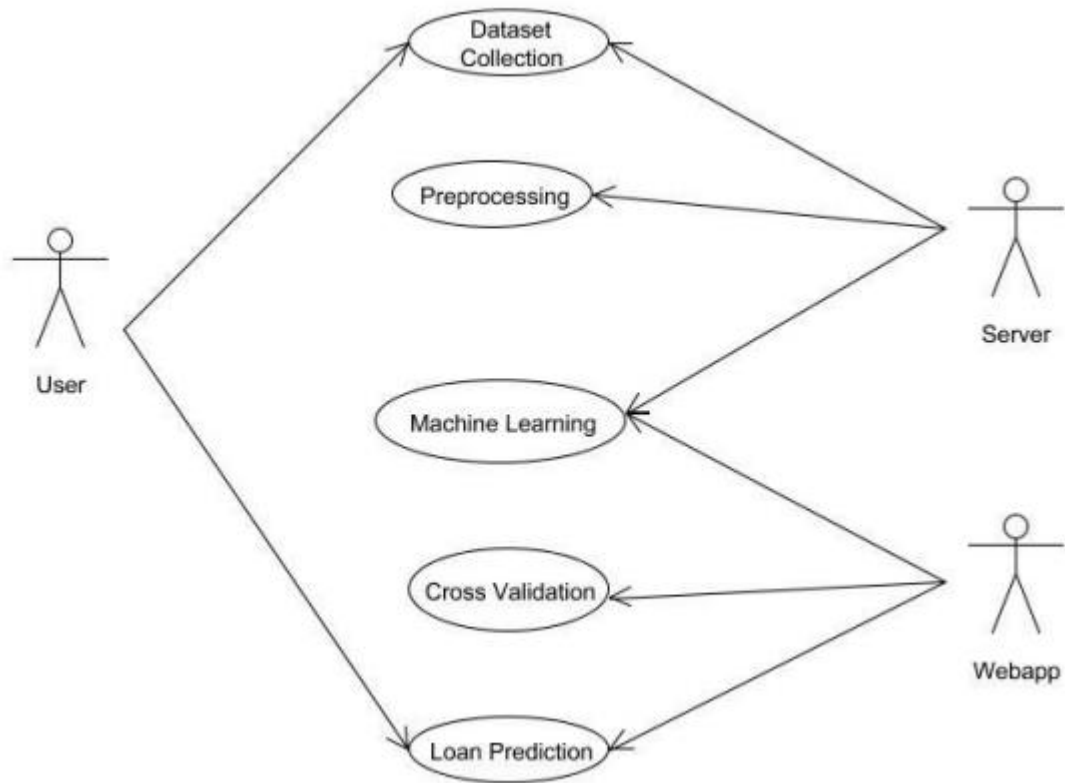
Furthermore, "Loan Prediction Using Ensemble Technique" (2020) by Anchal Goyal and Ranpreet Kaur introduced an ensemble model incorporating eleven machine learning models with nine distinct properties for predicting credit risk in loan applications. This work aimed to evaluate model accuracy, employing parameters like Accuracy, Gini, AUC, Roc, and others while assessing different training algorithms. Real Coded Genetic Algorithms were used to calculate feature importance, facilitating credit risk prediction. The K-fold validation method was employed to ensure the robustness of the predictive model, achieving a maximum accuracy of 81.25% with the Tree model for genetic algorithm. These papers collectively contribute valuable insights into the field of loan prediction and offer diverse methods for improving loan approval and risk assessment processes.

## ARCHITECTURE DIAGRAM

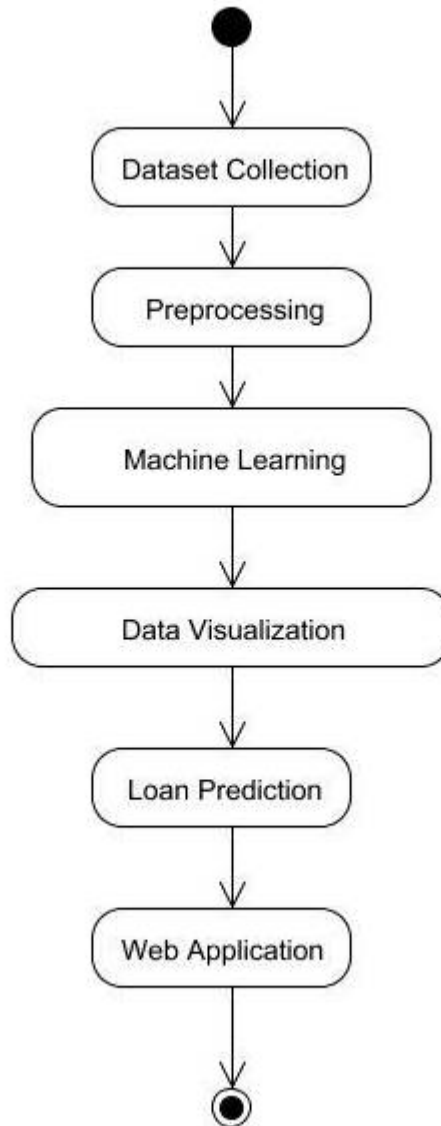




## USE CASE DIAGRAM



## ER DIAGRAM



## **MODULES**

- Dataset collection
- Machine Learning Algorithm
- Prediction

### **MODULE EXPLANATION:**

#### *Dataset collection:*

Dataset is collected from kaggle.com. That dataset has some value like gender, marital status, self-employed or not, monthly income, etc. Dataset has the information, whether the previous loan is approved or not depends on the customer information. That data will be preprocessed and proceed to the next step.

#### *Machine learning Algorithm:*

In this stage, the collected data will be given to the machine algorithm for the training process. We use multiple algorithms to get a high accuracy range of prediction. A preprocessed data set is processed in different machine learning algorithms. Each algorithm gives some accuracy level. Each one is undergoing for the comparison.

- ✓ Logistic Regression
- ✓ Random Forest Classifier
- ✓ Decision Tree Classifier
- ✓ SVM

#### *Prediction:*

Preprocessed data are trained, and input given by the user goes to the trained dataset. The Logistic Regression trained model is used to predict and determine whether the loan given to a particular person shall be approved or not.

## **Design and Implementation Constraints**

### **Constraints in Analysis**

- ◆ Constraints as Informal Text
- ◆ Constraints as Operational Restrictions
- ◆ Constraints Integrated in Existing Model Concepts
- ◆ Constraints as a Separate Concept
- ◆ Constraints Implied by the Model Structure

### **Constraints in Design**

- ◆ Determination of the Involved Classes
- ◆ Determination of the Involved Objects
- ◆ Determination of the Involved Actions
- ◆ Determination of the Require Clauses
- ◆ Global actions and Constraint Realization

### **Constraints in Implementation**

A hierarchical structuring of relations may result in more classes and a more complicated structure to implement. Therefore, it is advisable to transform the hierarchical relation structure to a simpler structure such as a classical flat one. It is rather straightforward to transform the developed hierarchical model into a bipartite, flat model, consisting of classes on the one hand and flat relations on the other. Flat relations are preferred at the design level for reasons of simplicity and implementation ease. There is no identity or functionality associated with a flat relation. A flat relation corresponds with the relation concept of entity-relationship modeling and many object-oriented methods.

## **Other Nonfunctional Requirements**

### **Performance Requirements**

The application at this side controls and communicates with the following two main general components.

- embedded browser in charge of the navigation and accessing to the web service.
- Server Tier: The server side contains the main parts of the functionality of the proposed architecture. The components at this tier are the following.

Web Server, Security Module, Server-Side Capturing Engine, Preprocessing Engine, Database System, Verification Engine, Output Module.

### **Safety Requirements**

1. The software may be safety critical. If so, there are issues associated with its integrity level.
2. The software may not be safety-critical although it forms part of a safety-critical system. For example, software may simply log transactions.
3. If a system must be of a high integrity level and if the software is shown to be of that integrity level, then the hardware must be at least of the same integrity level.
4. There is little point in producing 'perfect' code in some language if hardware and system software (in the widest sense) are not reliable.
5. If a computer system is to run software of a high integrity level, then that system should not at the same time accommodate software of a lower integrity level.
6. Systems with different requirements for safety levels must be separated.
7. Otherwise, the highest level of integrity required must be applied to all systems in the same environment.

# SOURCE CODE

```
comparative_study_loan_status_prediction.ipynb
File Edit View Insert Runtime Tools Help Last edited on October 22
+ Code + Text
Importing the libraries and statistical review

[ ] import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.preprocessing import StandardScaler
import pickle
from sklearn.metrics import recall_score, precision_score, f1_score, accuracy_score

%matplotlib inline

[ ] # loading dataset

[ ] data = pd.read_csv("/content/drive/MyDrive/Datasets/loan_approval_dataset.csv")

[ ] data.head()

loan_id  no_of_dependents  education  self_employed  income_monthly  loan_amount  loan_term  cibil_score  residential_assets_value  commercial_assets_value  luxury_assets_value  bank_asset_value  loan_status
0      1      2      Graduate      No      9600000      29900000      12      778      2400000      17600000      22700000      8000000      Approved
1      2      0      Not Graduate      Yes      4100000      12200000      8      417      2700000      2200000      8600000      3300000      Rejected
2      3      3      Graduate      No      9100000      29700000      20      506      7100000      4000000      33300000      12800000      Rejected
3      4      3      Graduate      No      8200000      30700000      8      467      18200000      3300000      23300000      7900000      Rejected
4      5      5      Not Graduate      Yes      9800000      24200000      20      382      12400000      8200000      29400000      5000000      Rejected

[ ] data.shape

(4289, 13)

[ ] data.describe()

loan_id  no_of_dependents  income_monthly  loan_amount  loan_term  cibil_score  residential_assets_value  commercial_assets_value  luxury_assets_value  bank_asset_value
count  4289.000000      4289.000000  4.269000e+03  4.269000e+03  4289.000000  4289.000000      4.269000e+03      4.269000e+03      4.269000e+03
mean    2135.000000      2498712  5.059124e+06  1.513345e+07  10.900445  599.936051      7.472617e+06      4.973155e+06      1.012631e+07  4.978632e+06
std    1232.498479      1605910  2.806840e+06  9.043363e+06  5.709187  172.430401      6.502637e+06      4.388966e+06      9.103754e+06  3.250185e+06
min      1.000000      0.000000  2.000000e+05  3.000000e+05  2.000000  300.000000      -1.000000e+05  0.000000e+00  0.000000e+00  0.000000e+00
25%    1958.000000      1.000000  2.700000e+06  7.700000e+06  6.000000  453.000000      2.200000e+06  1.500000e+06  7.000000e+06  2.300000e+06
50%    2135.000000      3.000000  5.100000e+06  1.450000e+07  10.000000  600.000000      5.600000e+06  3.700000e+06  1.460000e+07  4.600000e+06
75%    3032.000000      4.000000  7.500000e+06  2.150000e+07  16.000000  748.000000      1.150000e+07  7.600000e+06  2.170000e+07  7.100000e+06
max    4289.000000      5.000000  9.900000e+06  3.950000e+07  20.000000  900.000000      2.910000e+07  1.940000e+07  3.100000e+07  1.470000e+07

[ ] data.isnull().sum()

loan_id      0
no_of_dependents  0
education      0
self_employed  0
income_monthly  0
loan_amount    0
loan_term      0
cibil_score    0
residential_assets_value  0
commercial_assets_value  0
luxury_assets_value  0
bank_asset_value  0
loan_status    0
dtype: int64

[ ] data.columns

Index(['loan_id', 'no_of_dependents', 'education', 'self_employed',
      'income_monthly', 'loan_amount', 'loan_term', 'cibil_score',
      'residential_assets_value', 'commercial_assets_value',
      'luxury_assets_value', 'bank_asset_value', 'loan_status'],
      dtype='object')

[ ] data.columns = data.columns.str.replace(" ", "")

[ ] data.columns

Index(['loan_id', 'no_of_dependents', 'education', 'self_employed',
      'income_monthly', 'loan_amount', 'loan_term', 'cibil_score',
      'residential_assets_value', 'commercial_assets_value',
      'luxury_assets_value', 'bank_asset_value', 'loan_status'],
      dtype='object')

[ ] data['education'].value_counts()

Graduate    2144
Not Graduate  2125
Name: education, dtype: int64

[ ] data['self_employed'].value_counts()

Yes    2150
No     2119
Name: self_employed, dtype: int64

[ ] data['loan_status'].value_counts()

Approved    2656
Rejected    1813
Name: loan_status, dtype: int64

[ ] # we will perform encoding to convert the categorical data into the numerical data

[ ] # first we perform label encoding
# data.replace({'loan_status': ('Approved':1, 'Rejected':0)}, inplace=True)

[ ] data.head()

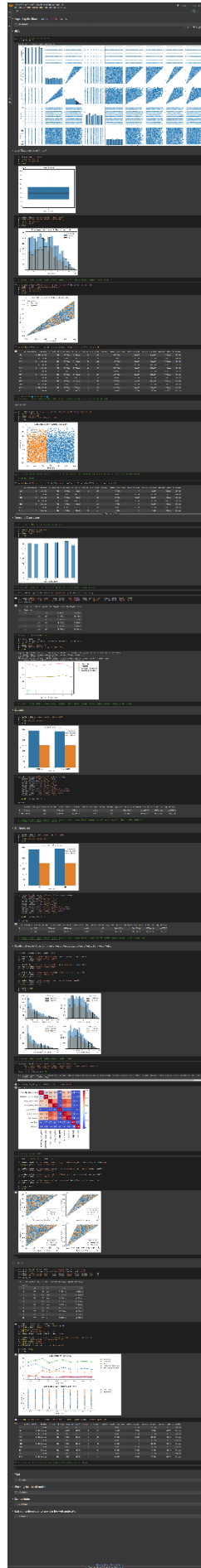
loan_id  no_of_dependents  education  self_employed  income_monthly  loan_amount  loan_term  cibil_score  residential_assets_value  commercial_assets_value  luxury_assets_value  bank_asset_value  loan_status
0      1      2      Graduate      No      9600000      29900000      12      778      2400000      17600000      22700000      8000000      Approved
1      2      0      Not Graduate      Yes      4100000      12200000      8      417      2700000      2200000      8600000      3300000      Rejected
2      3      3      Graduate      No      9100000      29700000      20      506      7100000      4000000      33300000      12800000      Rejected
3      4      3      Graduate      No      8200000      30700000      8      467      18200000      3300000      23300000      7900000      Rejected
4      5      5      Not Graduate      Yes      9800000      24200000      20      382      12400000      8200000      29400000      5000000      Rejected

[ ] # now we will encode the other values
# data.replace({'education': ('Graduate':1, 'Not Graduate':0), 'self_employed': ('Yes':1, 'No':0)}, inplace=True)

[ ] data.head()

loan_id  no_of_dependents  education  self_employed  income_monthly  loan_amount  loan_term  cibil_score  residential_assets_value  commercial_assets_value  luxury_assets_value  bank_asset_value  loan_status
0      1      2      Graduate      No      9600000      29900000      12      778      2400000      17600000      22700000      8000000      Approved
1      2      0      Not Graduate      Yes      4100000      12200000      8      417      2700000      2200000      8600000      3300000      Rejected
2      3      3      Graduate      No      9100000      29700000      20      506      7100000      4000000      33300000      12800000      Rejected
3      4      3      Graduate      No      8200000      30700000      8      467      18200000      3300000      23300000      7900000      Rejected
4      5      5      Not Graduate      Yes      9800000      24200000      20      382      12400000      8200000      29400000      5000000      Rejected



[ ] # we dont need loan_id as it does not signify anything
data = data.drop("loan_id", axis = 1)
```







## Front-End

SRM Bank- Check Loan Eligibility

### Answer the questions

No. of Dependents:

Education:

Self-Employed:

Annual Income:

Loan Amount:

Loan Term:


Cibil Score:

Residential Asset value:



Commercial Asset Value:


Luxury Asset Value:

Bank Asset Value:



[Predict](#)

SRM Bank- Check Loan Eligibility



Loan will be approved 😊

[Check again](#)

## RESULT AND DISCUSSION

- This project is successfully completed with the best model possible.
- The WebApp for the project has been made and hosted online at the Replit platform.
- The result of the comparative study is shown below:

### ▾ Comparision

```
[ ] print("Following are the accuracy scores on test data for the different machine learning algorithm: ")
print(f"Accuracy for Logistic Regression: {log_model_accuracy:.2f}")
print(f"Accuracy for Decision Tree: {dec_tree_model_accuracy:.2f}")
print(f"Accuracy for Random Forest Classifier: {ran_for_model_accuracy:.2f}")
print(f"Accuracy for SVM: {svm_model_accuracy:.2f}")
```

```
Following are the accuracy scores on test data for the different machine learning algorithm:
Accuracy for Logistic Regression: 0.91
Accuracy for Decision Tree: 0.98
Accuracy for Random Forest Classifier: 0.97
Accuracy for SVM: 0.92
```

```
[ ] # since we got the highest accuracy in "decision tree", so we will be using it for our web application.....
```

## **REFERENCES**

1. P. Dutta, “A STUDY ON MACHINE LEARNING LGORITHM FOR ENHANCEMENT OF LOAN PREDICTION”, International Research 3 ITM Web of Conferences 44, 03019 (2022) <https://doi.org/10.1051/itmconf/20224403019ICACC-2022> Journal of Modernization in Engineering Technology and Science, (2021).
2. P. Supriya, M. Pavani, N. Saisushma, N. Kumari and K. Vikas, Loan Prediction by using Machine Learning Models,” International Journal of Engineering and Techniques, (2019).
3. <https://www.irjet.net/archives/V8/i6/IRJET-V8I6582.pdf>
4. <https://ijarsct.co.in/Paper1165.pdf>
5. [https://www.itm-conferences.org/articles/itmconf/pdf/2022/04/itmconf\\_icacc2022\\_03019.pdf](https://www.itm-conferences.org/articles/itmconf/pdf/2022/04/itmconf_icacc2022_03019.pdf)