|  |
| --- |
| **Artificial Intelligence and Machine**  **Learning**  Project Report  Semester-IV (Batch-2022)    **Loan Predictor**                  **Supervised By: Submitted By:**  Mrs. Rishu Taneja Sneha Chaudhary  2210990857  Group -13        **Department of Computer Science and Engineering**  **Chikara University Institute of Engineering & Technology**  **Chitkara University, Punjab**    **1 | AIML – 22CS015** |

# Abstract

The Loan Prediction project aims to predict whether a loan applicant is likely to default on a loan repayment or not. This project is crucial for financial institutions to make informed decisions while approving loan applications, thereby minimizing the risk of default. By utilizing historical data of loan applicants along with their attributes, we aim to develop a predictive model using Artificial Intelligence and Machine Learning algorithms.

The problem of loan default prediction is of significant importance to the banking and finance sector. The ability to accurately predict loan defaulters can help financial institutions manage risks effectively, thereby enhancing their profitability and reducing potential losses.

In this project, we will explore various machine learning techniques to analyze historical loan data, identify patterns, and build a predictive model that can classify loan applicants as either 'high risk' or 'low risk' based on their attributes.

# Table of Contents

1. **Introduction** 
   1. Background ……………………………………………………………..…… Page 4
   2. Objectives .…………………………………………………………...……… Page 5
   3. Significance ……………………………………………………………….… Page 6
2. **Problem Definition and Requirements** 
   1. Problem Definition ………………………………………………………..… Page 7
   2. Software Requirements .…………………………………………….……..… Page 8
   3. Dataset ………………….……………………………………………..….... Page 9
3. **Proposed Design/ Methodology** 
   1. Design and Working ………………………………...…...……………….... Page 10
   2. Folder Structure ……………………………………….………………….... Page 11
   3. Machine Learning Technique ...……………………………...…….…….… Page 14
   4. SVM Model …………..……………………………………………...….... Page 15
4. **Results** 
   1. SVM Model ……...………………………………...…...……………….... Page 21

# Introduction

## Background

Develop a robust predictive model: Utilize Support Vector Machine (SVM) algorithm to develop a predictive model that accurately classifies loan applicants as 'high risk' or 'low risk' based on their attributes.

Improve loan approval process: Enhance the loan approval process by providing a reliable tool for financial institutions to assess the creditworthiness of loan applicants, thus reducing the risk of default.

Minimize financial losses: By accurately identifying high-risk loan applicants, minimize the financial losses incurred by financial institutions due to loan defaults.

Optimize decision-making: Provide financial institutions with a data-driven approach to make informed decisions regarding loan approvals, thereby improving the efficiency and effectiveness of their lending process.

Evaluate model performance: Assess the performance of the SVM model using relevant evaluation metrics such as accuracy, precision, recall, and F1-score to ensure its reliability and effectiveness in loan prediction.

## Objectives

1. Risk Management: Financial institutions face significant risks associated with loan defaults. By accurately predicting the likelihood of loan default using AI and ML models, they can better manage and mitigate these risks, leading to more stable and profitable operations.
2. Efficient Decision Making: AI and ML-powered loan prediction models enable financial institutions to make quicker and more accurate decisions regarding loan approvals. This results in a more efficient lending process, reducing the time taken to approve loans while ensuring that only creditworthy applicants are approved.
3. Cost Reduction: Identifying potential loan defaulters before approving loans helps financial institutions reduce the number of bad loans, thereby minimizing the associated costs such as collection costs, legal fees, and write-offs.
4. Improved Customer Experience: By using AI and ML models, financial institutions can offer a more personalized and seamless experience to loan applicants. Applicants who are likely to be approved can benefit from faster loan processing times, while those deemed high risk can be provided with guidance to improve their creditworthiness.
5. Economic Stability: The efficient functioning of financial institutions is crucial for the overall stability of the economy. By reducing the number of bad loans and minimizing the impact of loan defaults, AI and ML-powered loan prediction models contribute to the overall stability of the financial sector and the economy.

## Significance

**Efficient Decision Making:**

AI and ML-powered loan prediction models enable financial institutions to make quicker and more accurate decisions regarding loan approvals. This results in a more efficient lending process, reducing the time taken to approve loans while ensuring that only creditworthy applicants are approved.

**Cost Reduction:**

Identifying potential loan defaulters before approving loans helps financial institutions reduce the number of bad loans, thereby minimizing the associated costs such as collection costs, legal fees, and write-offs.

**Improved Customer Experience:**

By using AI and ML models, financial institutions can offer a more personalized and seamless experience to loan applicants. Applicants who are likely to be approved can benefit from faster loan processing times, while those deemed high risk can be provided with guidance to improve their creditworthiness.

**Economic Stability:**

The efficient functioning of financial institutions is crucial for the overall stability of the economy. By reducing the number of bad loans and minimizing the impact of loan defaults, AI and ML-powered loan prediction models contribute to the overall stability of the financial sector and the economy.

**Inclusion and Fairness:**

AI and ML models can help reduce human biases in the loan approval process, ensuring that loan decisions are based solely on objective criteria. This can lead to greater financial inclusion by providing fair and equal access to credit for all segments of society.

# Problem Definition and Requirements

## Problem Definition

The loan predictor project aims to develop a machine learning model that can accurately predict whether a loan applicant is likely to default on their loan repayment or not. The problem can be defined as follows:

**Problem Statement**:

Given historical data of loan applicants along with their attributes such as credit score, income, employment status, loan amount, and other relevant features, the task is to develop a predictive model that can classify loan applicants as either 'high risk' or 'low risk' based on their attributes.

Objective:

The main objective of this project is to reduce the risk of loan default for financial institutions by accurately predicting the creditworthiness of loan applicants. This will help financial institutions make informed decisions while approving loan applications, thereby minimizing the risk of default and improving the overall efficiency of the lending process.

.

## Software Requirements

* **Programming Language:** Python - Python is a versatile and widely-used programming language with extensive support for data analysis, machine learning, and web development.

* **Libraries and Frameworks:** 
  1. **NumPy:** NumPy is a powerful library for numerical computing in Python, providing support for large multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.

* 1. **Pandas:** Pandas is a data manipulation library that offers data structures like DataFrame and Series, making it easy to manipulate and analyze structured data.

It provides functionalities for data cleaning, reshaping, and aggregation.

* 1. **Matplotlib:** Matplotlib is a plotting library that enables the creation of various types of static, interactive, and animated visualizations in Python. It offers a wide range of customization options for creating publication-quality plots.

* 1. **Seaborn:** Seaborn is built on top of Matplotlib and provides a high-level interface for creating attractive and informative statistical graphics. It simplifies the process of creating complex visualizations like heatmaps, violin plots, and pair plots.

* 1. **Scikit-learn:** Scikit-learn is a comprehensive machine learning library that provides simple and efficient tools for data mining and data analysis. It offers algorithms for classification, regression, clustering, dimensionality reduction, and model selection.

* **Development Environment:** Jupyter Notebook or any Python IDE (e.g., PyCharm, Visual Studio Code) - Jupyter Notebook provides an interactive computing environment for writing and executing Python code, making it ideal for exploratory data analysis and prototyping machine learning models. Python IDEs offer integrated development environments with features like code editing, debugging, and version control integration.

* **Version Control Git** - Git is a distributed version control system that allows multiple developers to collaborate on a project efficiently. It enables tracking changes to the codebase, managing different versions of the project, and facilitating collaboration through features like branching and merging.

## Dataset

The dataset used in this project is obtained from Yahoo Finance through the yfinance Python library. It comprises historical loan prediction data for the specified loan ticker, covering a defined period.

The dataset includes various fields that are essential for loan price prediction, including but not limited to:

**Input Fields:**

**Loan\_ID**: Unique identifier for each loan application.

**Gender:** Gender of the applicant.

**Married:** Marital status of the applicant (Yes/No).

Dependents: Number of dependents of the applicant.

**Education**: Education level of the applicant (Graduate/Not Graduate).

**Self\_Employed:** Employment status of the applicant (Self-employed/Not self-employed).

**ApplicantIncome:** Income of the applicant.

**CoapplicantIncome:** Income of the co-applicant.

**LoanAmount:** The amount of the loan requested.

**Loan\_Amount\_Term:** Term of the loan in months.

**Credit\_History:** Credit history of the applicant (1 for good credit history, 0 for bad credit history).

**Property\_Area:** Area where the property associated with the loan is located (Urban/Semiurban/Rural).

**Output Field:**

For the loan predictor project, the output field (also known as the target variable) will be **Loan\_Status.**

Loan\_Status is a binary variable indicating whether the loan application was approved or denied.

If Loan\_Status is '1', it means the loan was approved.

If Loan\_Status is '0', it means the loan was denied.

# Proposed Design/ Methodology

## Design and Working

**2. Data Retrieval and Preprocessing:**

* The retrieved data is stored in a pandas DataFrame
* Data preprocessing involves handling missing values and removing duplicate columns from the DataFrame to ensure data cleanliness and consistency.

**3. Data Analysis and Visualization:**

* Descriptive statistics of the loan data are computed
* Various visualizations are generated using Matplotlib, including:

**4. Model Training and Testing:**

* The project employs a Long Short-Term Memory (SVM) model for predicting loan prices.
* The SVM model is trained separately in a Jupyter Notebook (SVM model.ipynb).
* Training data is scaled using scikit learn to ensure numerical stability and convergence during model training.

**5. Prediction and Evaluation:**

* Testing data is prepared by combining dependent and independent variables.
* The trained SVM model is loaded
* Predictions are made using the loaded model on the testing data.
* Predictions are scaled back to their original values for evaluation and comparison with the actual loan prices.
* Predictions vs. original prices are plotted and displayed using Matplotlib for visual evaluation.

## Folder Structure

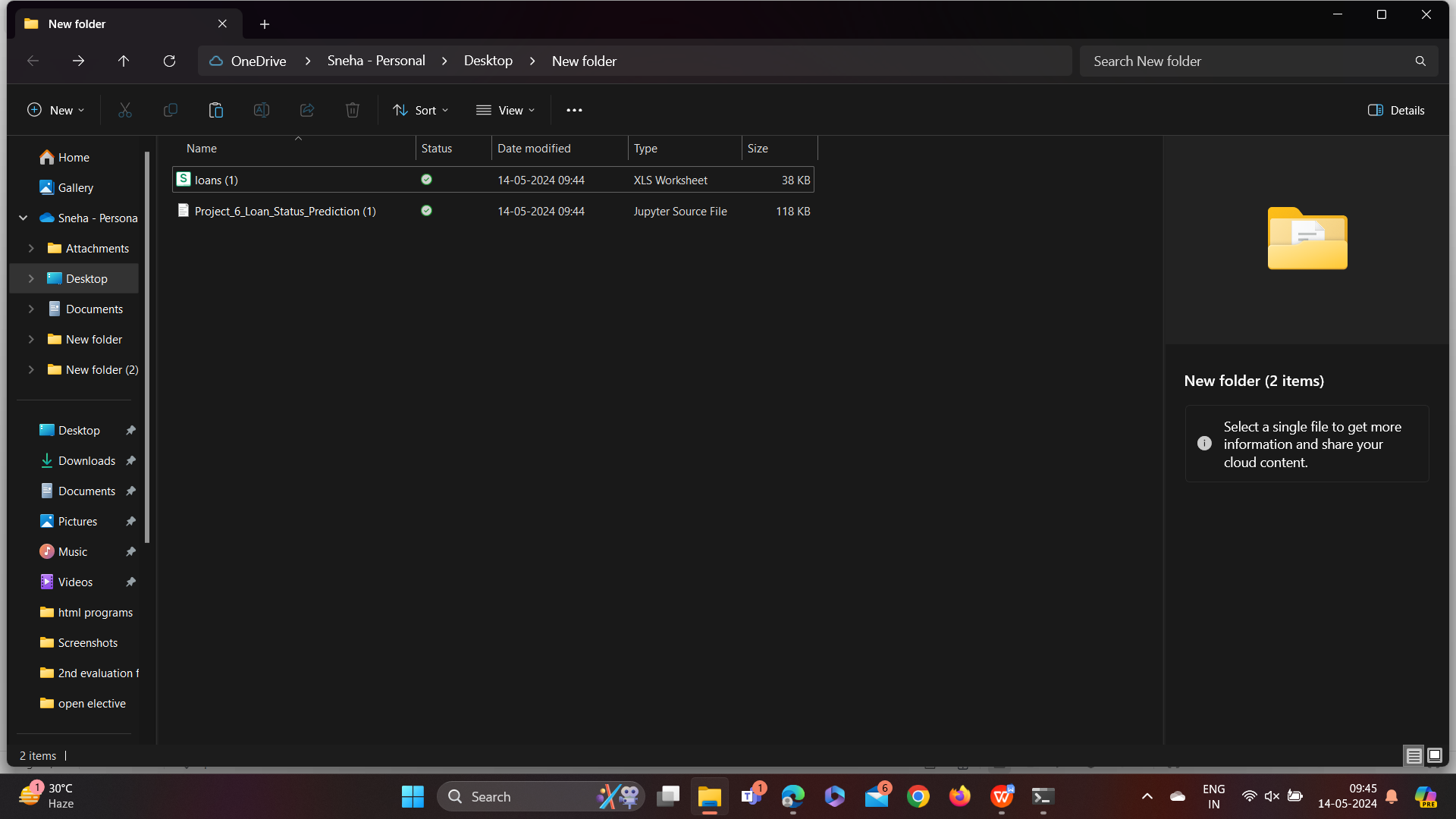


Figure 1 Folder Structure

* **app.py:** The main Python script containing the code for the loan trend prediction application.
* **Loans.csv:** Readme file providing information about the project, its components, and instructions for running the code.

.

## Machine Learning Technique

Loan price prediction typically falls under supervised learning because it involves using historical data (input) with corresponding loan prices (output) to train a model. In this context:

* **Labeled Data:** The historical data consists of various features such as past loan prices, trading volume, financial indicators, and other relevant information. Each data point is associated with the corresponding loan data (output label).

* **Training Process:** During the training phase, the algorithm learns the patterns and relationships between the input features and the corresponding loan prices. It iteratively adjusts its parameters to minimize the discrepancy between its predicted loan prices and the actual observed prices from the historical data.

* **Regression Task:** Loan price prediction is typically formulated as a regression problem where the goal is to predict a continuous value based on the input features. The algorithm aims to approximate the underlying mapping function between the input variables (historical data) and the output variable

* **Supervision:** The historical loan prices act as supervision for the learning process. The algorithm learns from past data, guided by the known outcomes, and aims to generalize this knowledge to make predictions on unseen data.

## SVM Model

Long Short-Term Memory (SVM) is a type of recurrent neural network architecture designed to address the vanishing gradient problem, which occurs when training traditional RNNs on long sequences of data. LSTMs are well-suited for modeling sequential data and have been widely used in various fields, including natural language processing, speech recognition, and time series prediction, such as loan price forecasting.

**Usage for Loan Price Prediction:**

* **Sequence Modeling:** Loan data can be treated as a sequence, where each data point represents the price at a specific time. SVM are well-suited for modeling such sequential data due to their ability to capture temporal dependencies.
* **Feature Extraction:** SVMs can automatically learn relevant features from the input data, such as patterns, trends, and seasonality, without the need for manual feature engineering.
* **Temporal Dynamics:** Loan prices exhibit complex temporal dynamics influenced by various factors such as prediction sentiment, economic indicators, and company performance. SVMs are capable of capturing these dynamics and making predictions based on historical patterns.

**Implementation in the Project:**

* **Model Training:** In the project, an SVM model is trained using historical loan price data. The model learns from the sequential patterns in the data to make predictions about future price movements.
* **Temporal Dependencies:** The SVM model captures temporal dependencies in the loan price data, allowing it to learn from past prices and make predictions based on historical patterns.
* **Long-Term Trends:** SVMs are capable of capturing long-term trends in loan prices, making them suitable for forecasting over extended time horizons.
* **Predictive Performance:** The trained SVM model is evaluated on testing data to assess its predictive performance. Metrics such as Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) may be used to quantify the accuracy of the predictions.

# Results

## SVM Model

