## Artificial Intelligence and

## Machine Learning

Project Report

Semester-IV (Batch-2022)

## LOAN APPROVAL PREDICTION



**Submitted To: Submitted By:**

Ms. Monica Dutta Monu Kumar (2210990583)

Muskan (2210990589)

Muskan Rehni (2210990591)

Nikhil Verma (2210990612)

## Chitkara University Institute of Engineering & Technology,

## Chitkara University, Punjab

# 

# Loan approval prediction plays a crucial role in the banking sector by assisting financial institutions in making informed decisions and mitigating risks associated with lending. This project aims to develop a predictive model that can accurately predict whether a loan application will be approved or not based on various applicant attributes. The project begins with exploratory data analysis (EDA) to gain insights into the dataset and understand the relationship between different features and the target variable. Following EDA, data preprocessing techniques are applied to handle missing values, encode categorical variables, and scale features. Several machine learning algorithms including Gradient Boosting, Random Forest, Decision Tree, K-Nearest Neighbors, and Support Vector Machine are employed to build predictive models. Hyperparameter tuning is performed to optimize the performance of each model. Model evaluation is conducted using metrics such as accuracy, precision, recall, and F1-score. The best-performing model is selected based on its performance and suitability for the problem at hand.Finally, the selected model is deployed for real-world use, enabling financial institutions to automate the loan approval process and make efficient and accurate decisions. This project aims to enhance the efficiency of the loan approval process, minimize default risk, ensure fairness and compliance, and ultimately improve customer satisfaction and financial inclusion.

# ABSTRACT

1. Introduction

2. Problem Definition and Requirements

3. Proposed Design / Methodology

# TABLE OF CONTENTS

4. Results

5. References

# INTRODUCTION:

**1.1 BACKGROUND**

The process of loan approval and risk assessment plays a crucial role in the financial industry, impacting both lenders and borrowers alike. Traditionally, these decisions have been made through manual underwriting processes, often relying on subjective criteria and historical data. However, the limitations of these traditional methods, such as their inability to analyze large volumes of data and identify complex patterns, have led to a growing interest in leveraging artificial intelligence (AI) and machine learning (ML) technologies. The development of AI and ML-based loan prediction systems offers a promising solution to these challenges by automating decision-making processes, improving accuracy, and enhancing transparency in lending practices. In this report, we present a comprehensive analysis of our loan prediction system project, aiming to develop a robust and efficient model for predicting loan approvals and defaults. By leveraging state-of-the-art ML techniques and analyzing extensive datasets, we seek to address the limitations of traditional approaches and contribute to the advancement of fair, efficient, and inclusive lending practices.

* 1. **OBJECTIVE**
* **Develop a Predictive Model**: Build machine learning models to predict whether a loan application will be approved or not based on applicant information.
* **Improve Decision Making**: Provide insights to financial institutions to make informed decisions on loan approvals, thereby reducing the risk of default and increasing profitability.
* **Automate Loan Approval Process**: Create an automated system that can assist loan officers in assessing loan applications, streamlining the approval process, and reducing manual effort.
* **Increase Efficiency**: Enhance the efficiency of the loan approval process by accurately assessing the creditworthiness of applicants, thereby reducing the time taken for loan processing.
* **Minimize Default Risk**: Develop models that can effectively identify high-risk applicants who are likely to default on loans, allowing lenders to take proactive measures to mitigate risks.
* **Ensure Fairness and Compliance**: Ensure that the loan approval process is fair and compliant with regulations by removing biases and making decisions based solely on relevant applicant information.
* **Optimize Resources**: Optimize resource allocation by directing attention to high-potential loan applications while minimizing the time spent on low-probability applications.
* **Improve Customer Satisfaction**: Provide a seamless and transparent loan approval experience for customers by reducing the time and effort required to obtain a loan.
* **Enhance Model Performance**: Continuously refine and improve machine learning models to achieve higher accuracy, precision, and recall in loan approval predictions.
* **Facilitate Financial Inclusion**: Promote financial inclusion by accurately assessing the creditworthiness of underserved populations and offering them access to credit opportunities.

**1.3 SIGNIFICANCE**

Accurate loan prediction is of paramount importance in the financial sector, influencing the stability of lending institutions, the accessibility of credit for individuals and businesses, and the overall health of the economy. By developing a robust loan prediction system, we aim to address key challenges faced by lenders, including the need to assess credit risk efficiently, minimize default rates, and streamline the loan approval process. A well-designed prediction model can help lenders make informed decisions, leading to improved portfolio management, reduced financial losses, and enhanced customer satisfaction. Moreover, a fair and transparent loan prediction system promotes financial inclusion by ensuring equal access to credit opportunities for all individuals, regardless of their demographic or socioeconomic background. Through our project, we seek to harness the power of machine learning to drive positive outcomes for lenders, borrowers, and society as a whole.

1. **PROBLEM DEFINITION AND REQUIREMENTS:**

**2.1 PROBLEM STATEMENT:**

The problem revolves around predicting whether a loan application will be approved or not based on various attributes of the applicant, such as their demographic information, financial history, and loan details.

Challenges:

* Manual assessment of loan applications is time-consuming and prone to errors.
* Risk of approving loans to individuals who may default, leading to financial losses.
* Ensuring fairness and compliance in the loan approval process.
* Need for an automated system to streamline the loan approval process and improve efficiency.

**2.2 REQUIREMENTS**

**2.2.1 SOFTWARE REQUIREMENTS** :

* Python: For data preprocessing, modeling, and evaluation.
* Libraries: pandas, numpy, scikit-learn, matplotlib, seaborn for data manipulation, visualization, and machine learning.
* Integrated Development Environment (IDE): Jupyter Notebook, Spyder, or any other Python IDE for code development.
* Text Editor: Any text editor for writing documentation and reports (e.g., Microsoft Word, Google Docs).

**2.2.2 HARDWARE REQUIREMENTS :**

* Processor: Any modern multi-core processor (e.g., Intel Core i5 or above).
* RAM: Minimum 4GB RAM, recommended 8GB or more for better performance.
* Storage: Sufficient disk space to store datasets and Python libraries.
* Operating System: Windows, macOS, or Linux.

**2.2.3 DATA SETS**

Our loan prediction system project relies on the analysis of comprehensive datasets encompassing various financial, demographic, and credit-related attributes. These datasets are sourced from reputable financial institutions, credit bureaus, and publicly available repositories to ensure data integrity and reliability. Key features included in the datasets encompass borrower demographics, employment history, income levels, credit scores, loan terms, and historical repayment behavior. Additionally, auxiliary data such as economic indicators, market trends, and regulatory factors may be incorporated to enhance the predictive accuracy of the model. Through meticulous data preprocessing and feature engineering techniques, we aim to extract meaningful insights and patterns that can inform the development of robust machine learning models for loan prediction. Furthermore, we adhere to strict data privacy and security protocols to safeguard sensitive information and ensure compliance with regulatory requirements.

1. **PROPOSED DESIGN / METHODOLOGY:**

**3.1 SCHEMATIC DIAGRAM**

+----------------------+

| |

| Data Preprocessing |

| |

+-----------+----------+

|

+-----------v----------+

| |

| Model Building |

| |

+-----------+----------+

|

+-----------v----------+

| |

| Model Evaluation |

| |

+-----------+----------+

|

+-----------v----------+

| |

| Model Deployment |

| |

+———————————+

**3.2 FILE STRUCTURE:**

Loan\_Approval\_Prediction/

│

├── data/

│ ├── train\_u6lujuX\_CVtuZ9i.csv # Raw dataset

│ └── processed\_data.csv # Preprocessed dataset

│

├── notebooks/

│ ├── 01\_Data\_Preprocessing.ipynb # Data preprocessing notebook

│ ├── 02\_Model\_Building.ipynb # Model building notebook

│ ├── 03\_Model\_Evaluation.ipynb # Model evaluation notebook

│ └── 04\_Model\_Deployment.ipynb # Model deployment notebook

│

├── reports/

│ ├── Project\_Report.pdf # Final project report

│ └── EDA\_Visualizations/ # Folder containing EDA visualizations

│ ├── loan\_status\_distribution.png

│ ├── correlation\_heatmap.png

│ └── ...

│

├── src/

│ ├── data\_preprocessing.py # Python script for data preprocessing

│ ├── model\_building.py # Python script for model building

│ ├── model\_evaluation.py # Python script for model evaluation

│ └── model\_deployment.py # Python script for model deployment

│

└── requirements.txt # Python dependencies

**3.3 ALGORITHMS USED**

1. **) SUPPORT VECTOR MACHINE CLASSIFIER**

SVM is a supervised learning algorithm that excels in binary classification tasks by finding the optimal hyperplane that separates the classes in the feature space. By maximizing the margin between the support vectors of different classes, SVM achieves robust and generalizable classification performance, even in high-dimensional feature spaces. Furthermore, SVM can handle non-linear decision boundaries through the use of kernel functions, such as the radial basis function (RBF) kernel, which allows for capturing complex patterns in the data. Through careful tuning of hyperparameters and kernel selection, we aim to harness the full potential of SVM in our loan prediction system, ensuring high accuracy and reliability in determining credit risk and facilitating informed lending decisions.

**2.) DECISION TREE CLASSIFIER**

Decision trees are intuitive and easy to interpret, making them suitable for generating transparent decision rules that lenders can understand and trust. By recursively partitioning the feature space into subsets based on the most discriminative attributes, decision trees identify decision boundaries that effectively separate loan approval and default cases. Furthermore, decision trees can handle both numerical and categorical data, making them suitable for diverse datasets commonly encountered in loan prediction tasks. To mitigate overfitting and improve generalization, we apply techniques such as pruning, limiting tree depth, and implementing ensemble methods like Random Forests. Through extensive experimentation and evaluation, we aim to harness the predictive power of decision trees to enhance the accuracy and transparency of our loan prediction system, ultimately facilitating sound lending decisions and minimizing financial risks.

**3.) RANDOM FOREST CLASSIFIER**

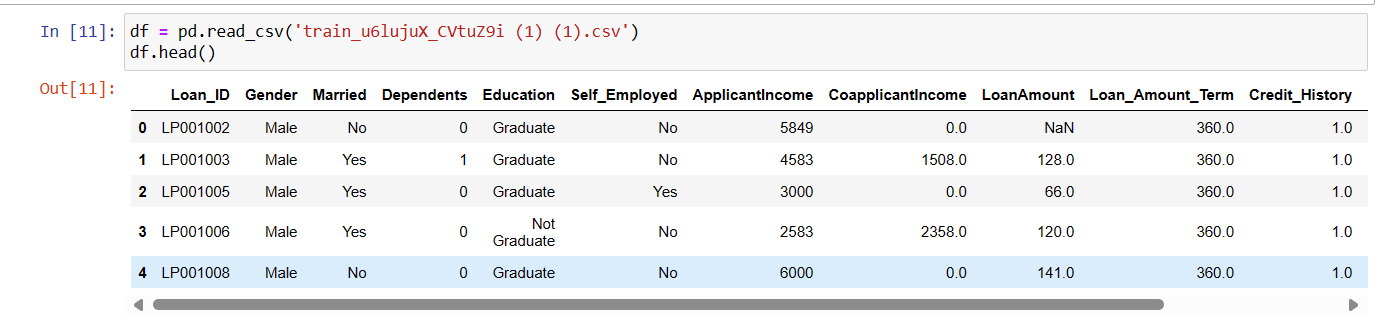
Random Forest is an ensemble of decision trees, where multiple trees are trained on different subsets of the data and their predictions are combined through voting or averaging. This approach helps to mitigate overfitting and improve generalization by reducing the variance of individual trees. Random Forest can handle both numerical and categorical data, making it suitable for heterogeneous datasets commonly encountered in loan prediction tasks. Additionally, it automatically selects informative features and provides estimates of feature importance, aiding in model interpretation and transparency. By fine-tuning hyperparameters such as the number of trees and maximum tree depth, we aim to optimize the performance of our Random Forest model and achieve high accuracy in predicting loan approvals and defaults. Through rigorous evaluation and validation, we validate the effectiveness of Random Forest in our loan prediction system, contributing to more informed lending decisions and improved risk management.

**4.) LOGISTIC REGRESSION**

Logistic Regression models the probability of a binary outcome (e.g., loan approval or default) based on one or more predictor variables. Unlike linear regression, which predicts continuous outcomes, Logistic Regression applies a sigmoid function to the linear combination of input features, producing a probability score between 0 and 1. This probability score is then thresholded to make binary predictions. Logistic Regression is well-suited for scenarios where interpretability is crucial, as it provides easily interpretable coefficients that indicate the strength and direction of the relationships between predictors and the outcome. Moreover, Logistic Regression is computationally efficient and robust to noise, making it suitable for large-scale datasets commonly encountered in loan prediction tasks. By optimizing regularization parameters and feature selection techniques, we aim to develop Logistic Regression models that accurately predict loan approvals and defaults, thereby facilitating sound lending decisions and minimizing financial risks.

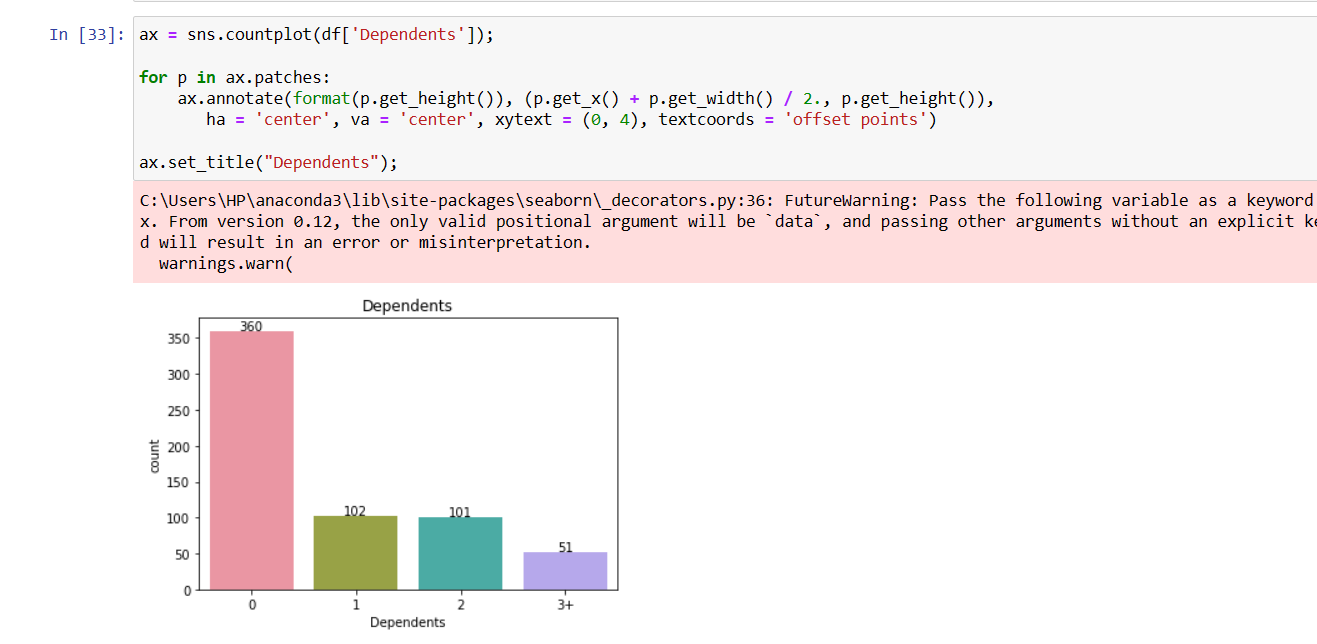
**4. RESULTS:**

****

****

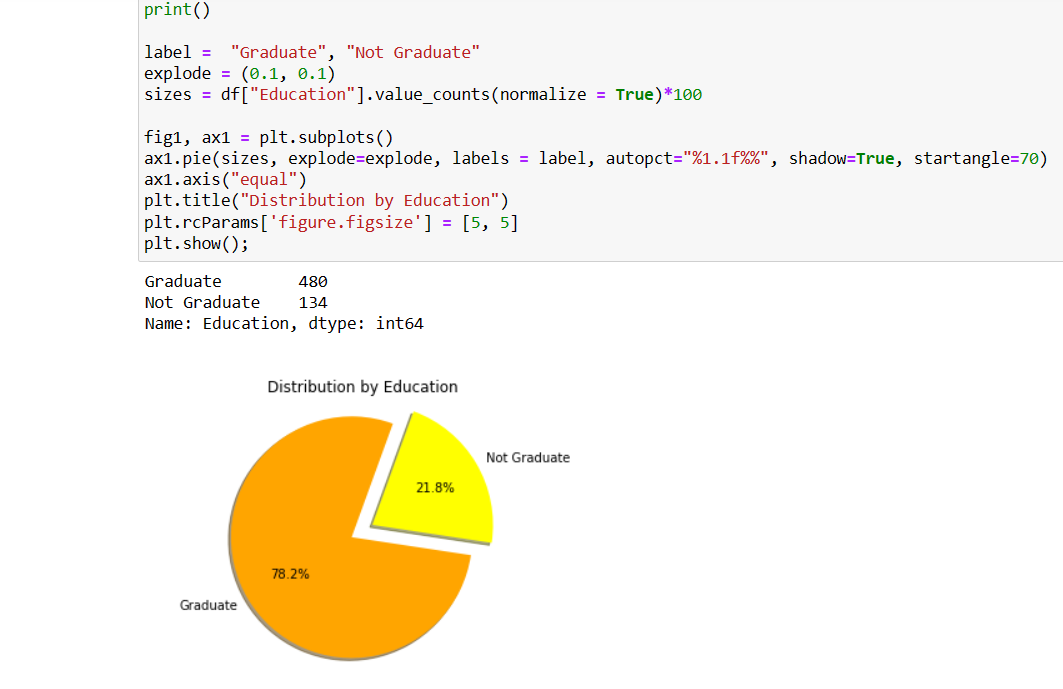
This is the dataset of our project i.e. prediction the loan approved by the bank or not by checking all the details of that person every information.

Our dataset provides all the columns named as: Loan\_id, Gender, Education, Income and many more and on these basis prediction will be made.



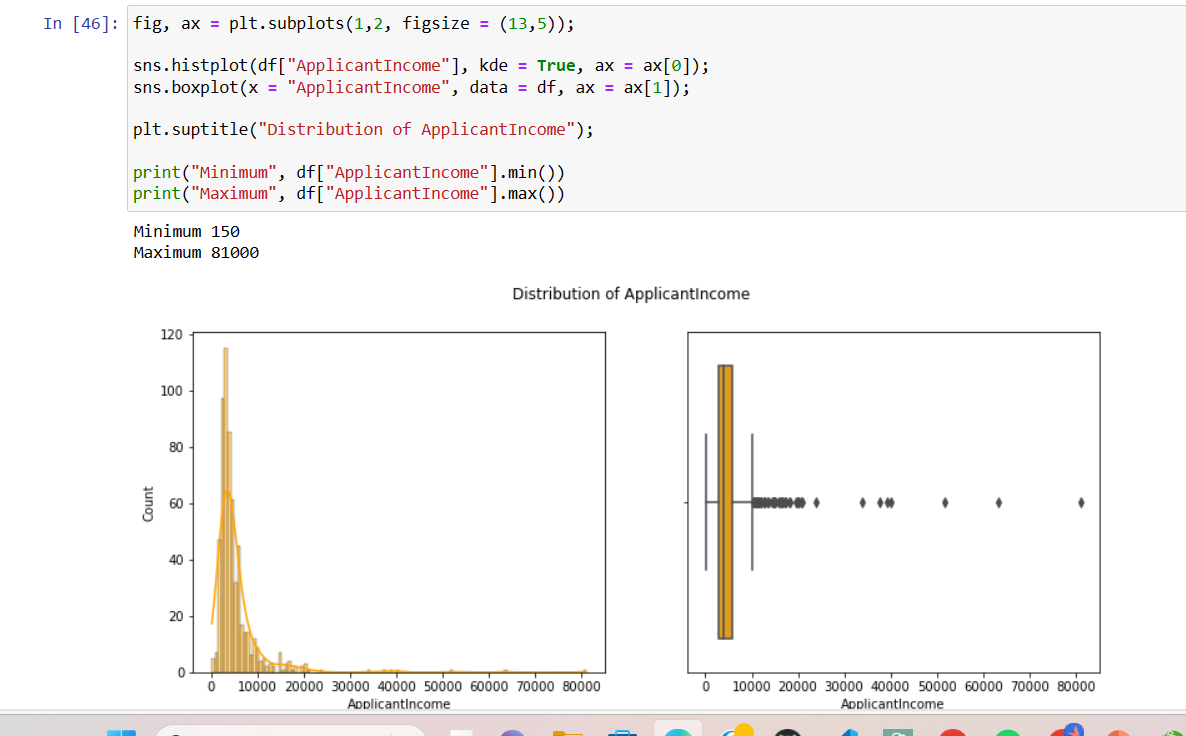
This graph shows the output of of all the number of people in the dependents column in the dataset.

According to this graph there are 360 people who have 0 dependents , 102 people who have 1 dependent, 101 people who have 2 dependents and last 51 people from the data provided has 3 or more number of dependents.



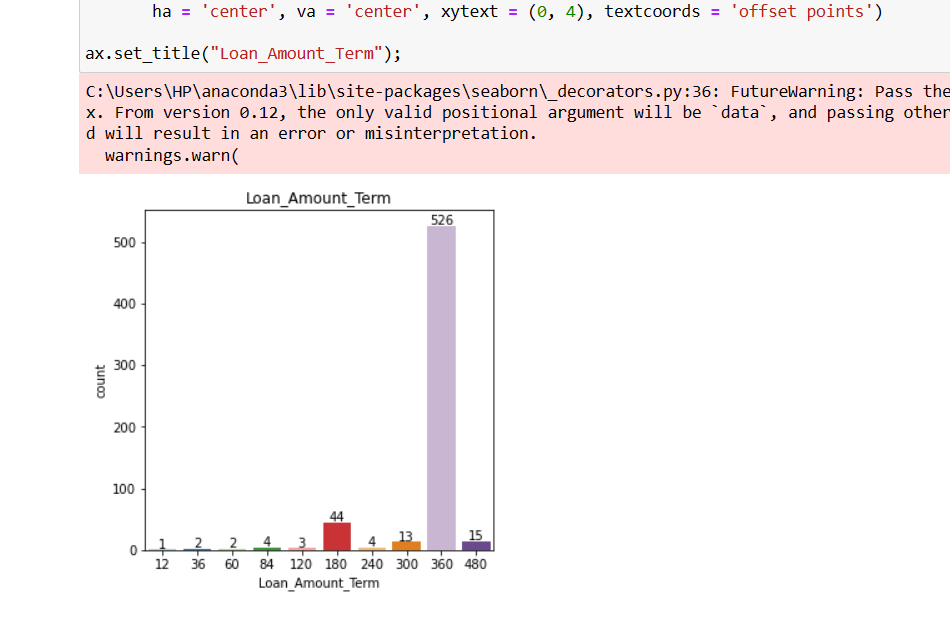
This pie chart shown here explains about the distributions under education column in the dataset.

According to this diagram there are 78.2% people who are graduate and 21.2% people who are not graduate under education field.



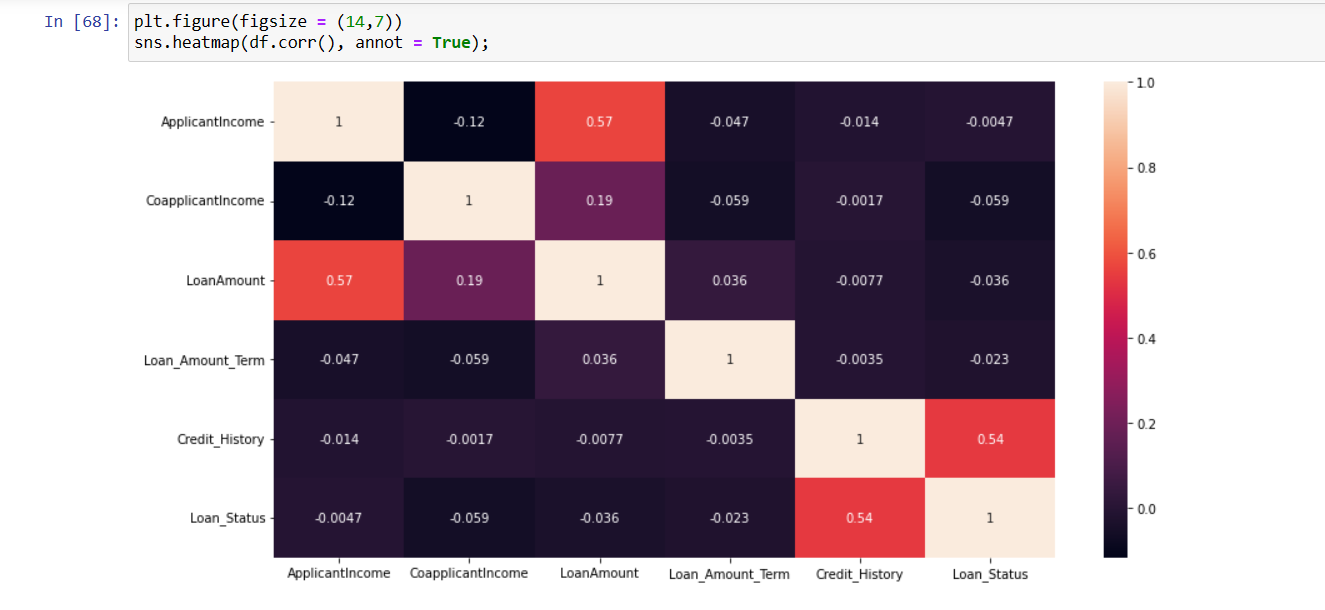
These 2 graphs i.e. histogram and box plot are just the representation of applicantincome column in the dataset which tells the distribution graph and also shows the max/peak point of the graph also the minimum point in the graph.

We used 2 graphs just to compare the graph representation in both ways of our code we have done.



This bar graph tells about number of people and their loan duration for how much time they tool loan from the bank

According to graph above there are 560 people at max who have 360 months duration in the banking record.



A correlation matrix plays a crucial role in machine learning and data analysis. Here are some of its key functions and applications:

1. **Feature Selection**: Identifies and removes redundant or irrelevant features.
2. **Multicollinearity Detection**: Detects high correlations between independent variables to ensure model stability.
3. **Exploratory Data Analysis (EDA)**: Provides an overview of relationships between variables.
4. **Dimensionality Reduction**: Utilized in techniques like PCA to transform variables into uncorrelated components.
5. **Model Interpretation**: Helps in understanding feature importance and model behavior.
6. **Anomaly Detection**: Identifies data quality issues or outliers through unexpected correlations.
7. **Correlation with Target Variable**: Highlights features most relevant for prediction.

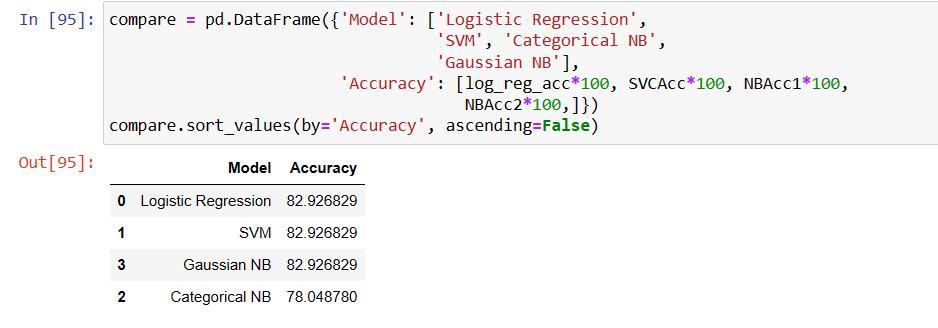




Using cross tabs with seaborn allows to:

* **Examine relationships** between two categorical variables.
* **Summarize and visualize** data in an easy-to-interpret format.
* **Generate heatmaps, bar plots**, and other visualizations to reveal patterns and interactions in categorical data.

Cross tabs simplify the process of analyzing categorical data and making insightful visualizations, which is essential for effective exploratory data analysis.



This is the last output or the final result of what we did in our machine learning project .

At last we have compared all the prediction methods like whose accuracy is more that method will be considered as the best to find out the accuracy of loan approval by the bank system.