**LOAN REPAYMENT PREDICTION**

**PROJECT BASED LEARNING (AIP104)**

**PROJECT REPORT**

Logo

Description automatically generated

**Submitted by**

**ADITYA (2110993754)**

**BAASHI (2110993771)**

**DEEPANSHI (2110993873)**

**BE-CSE (Artificial Intelligence)**

**Guided by**

**Dr. Vandana Sood**

**CHITKARA UNIVESITY OF ENGINEERING & TECHNOLOGY**

**CHITKARA UNIVERSITY, RAJPURA**

**MAY, 2023**

# **ACKNOWLEDGEMENT**

With immense please we, Mr. Aditya, Mr. Baashi and Ms. Deepanshi presenting “Loan Repayment Prediction” project report as part of the curriculum of ‘BE - CSE (AI)’.

We would like to express our sincere thanks to Dr. Vandana Sood, for her valuable guidance and support in completing our project.

We would also like to express our gratitude towards our dean Dr. Sushil Kumar for giving us this great opportunity to do a project on “Loan Repayment Prediction”. Without their support and suggestions, this project would not have been complete.

Signature…..………….

Name: Aditya

Roll No.: 2110993754

Signature…..………….

Name: Baashi

Roll No.: 2110993771

Signature…..………….

Name: Deepanshi

Roll No.: 2110993873

# **ABSTRACT**

Accurately estimating the possibility of loan repayment is a continuing challenge for lending institutions in today's financial environment. This urgent requirement is met by the Loan Repayment Prediction, a Python-based project that provides a trustworthy and precise predictive model. Lending institutions can use this model to make well-informed decisions, reduce financial risks, and improve their lending procedures. The huge influence the Loan Repayment Prediction can have on numerous businesses is what gives it its prominence. Loans must be repaid on time for lending institutions, such as banks, credit unions, and online lenders, to be profitable and viable. These institutions can assess loan applications intelligently and manage risk better since they can predict loan payback with accuracy.

Moreover, loan repayment prediction has been used with a variety of machine learning methods, including logistic regression, decision trees, random forests, support vector machines, and neural networks. Since these models were developed using historical loan data, it is possible to analyse patterns and trends in how borrowers repaid their loans. The quality and diversity of the training data, the proper feature selection, and efficient model validation all play a role in how accurately loan repayment prediction models perform. To improve the interpretability of models, breakthroughs in the field have looked into the integration of other data sources and the use of explainable AI methodologies. Financial firms can enhance their lending procedures by accurately forecasting loan repayment behaviour. They are better able to manage the risks associated with default, set suitable interest rates, and make more informed decisions about loan approvals. Accurate loan repayment prediction models encourage prudent lending practises while also contributing to the general stability and profitability of lenders.

# **LIST OF CONTENTS**

Contents

[**ACKNOWLEDGEMENT** 1](#_Toc135391053)

[**ABSTRACT** 2](#_Toc135391054)

[**LIST OF CONTENTS** 3](#_Toc135391055)

[**LIST OF FIGURES** 4](#_Toc135391056)

[**LIST OF TABLES** 5](#_Toc135391057)

[**INTRODUCTION** 6](#_Toc135391058)

[**Exploratory Data Analysis** 7](#_Toc135391059)

[**Machine learning approaches used in this project** 8](#_Toc135391060)

[**PROBLEM FORMULATION** 14](#_Toc135391061)

[**METHODOLOGY** 15](#_Toc135391062)

[**Imported Libraries:** 15](#_Toc135391063)

[**Loading Data:** 15](#_Toc135391064)

[**Data Cleaning:** 15](#_Toc135391065)

[**Data Visualisation:** 16](#_Toc135391066)

[**Model Preparation:** 19](#_Toc135391067)

[**Model Deployment:** 23](#_Toc135391068)

[**FLOWCHART** 24](#_Toc135391069)

[**MODEL DEPLOYMENT** 26](#_Toc135391070)

[**CONCLUSION** 28](#_Toc135391071)

[**REFERENCES** 29](#_Toc135391072)

# **LIST OF FIGURES**

[Figure 1 Logistic Regression Graph 9](#_Toc135391073)

[Figure 2 Decision Tree Graph 10](#_Toc135391074)

[Figure 3 Random Forest Graph 11](#_Toc135391075)

[Figure 4 KNN graph 11](#_Toc135391076)

[Figure 5 SVM Graph 13](#_Toc135391077)

[Figure 6 Correlation Plot for all features 16](#_Toc135391078)

[Figure 7 Density plot for numerical variables 17](#_Toc135391079)

[Figure 8 Scatter plot of fico and interest rate w.r.t not fully paid payments 17](#_Toc135391080)

[Figure 9 Not fully paid class count 18](#_Toc135391081)

[Figure 10 Purpose class count 18](#_Toc135391082)

[Figure 11 Logistic Regression Model 19](#_Toc135391083)

[Figure 12 K-nearest Neighbor Model 20](#_Toc135391084)

[Figure 13 Decision Tree Model 20](#_Toc135391085)

[Figure 14 Random Forest Model 21](#_Toc135391086)

[Figure 15 Support Vector Machine (SVM) Model 21](#_Toc135391087)

[Figure 16 Bar Graph of Models' Accuracy 22](#_Toc135391088)

[Figure 17 Flowchart 24](#_Toc135391089)

[Figure 18 Model Deployment – I 26](#_Toc135391090)

[Figure 19 Model Deployment - II 27](#_Toc135391091)

# **LIST OF TABLES**

[Table 1 first ten rows of dataset 7](#_Toc135391092)

[Table 2 Dataset's Information 15](#_Toc135391093)

[Table 3 Dataset after data manipulation 16](#_Toc135391094)

[Table 4 Models' Accuracy Table 22](#_Toc135391095)

# **INTRODUCTION**

In today’s dynamic and competitive business environment, industries across the board face significant challenges when it comes to managing financial risk. The uncertainty around loan repayment is one of the main dangers that organisations face. A crucial component of the financial system is lending money, which helps people and organisations realise their dreams and reach their objectives. There is always a chance that borrowers would fail to repay their debts, therefore it also entails inherent risks for the lenders. The unpredictability of loan repayment is one of the major hazards that organisations face. Lending money is a crucial component of the financial system that helps people and businesses realise their dreams and reach their objectives. However, it also has inherent risks for the lenders because there is always a chance that borrowers will stop making payments on their loans. It is impossible to overestimate the importance of loan repayment forecasting in the financial sector. To assess the possibility of loan defaults, banks and other lending institutions heavily rely on accurate predictions. Lenders can reduce the risk of financial losses and guarantee the stability of their loan portfolios by precisely estimating the ability of borrowers to repay their debts. They are then able to provide creditworthy borrowers with higher interest rates and terms while safeguarding their own interests and keeping a healthy balance between profitability and risk management.

Furthermore, loan repayment prediction has an impact outside of the financial industry. This prediction technology is used by a variety of sectors to improve customer happiness, streamline processes, and optimise company plans. For instance, insurance providers evaluate the risk profile of prospective policyholders using loan payback prediction. Insurers can make educated choices on the underwriting of policies, determining premium prices, and controlling their exposure to risk by looking at the repayment history and financial behaviour of specific individuals. Similar to this, retail businesses use loan repayment prediction to assess consumer creditworthiness and control credit risk. Retailers can adjust their credit offers, set suitable credit limits, and use successful collection techniques by precisely identifying consumers who are likely to miss their payment deadlines. This not only reduces losses brought on by bad debts but also enables shops to provide competitive financing solutions and increase client loyalty. Loan repayment forecasting is also advantageous to the healthcare sector. Healthcare professionals frequently come into circumstances when patients need financial support to pay for medical bills. Healthcare organisations can evaluate patient creditworthiness, ascertain the likelihood of repayment, and create suitable payment plans or financing solutions by using loan repayment prediction. This makes it possible for healthcare providers to guarantee prompt reimbursement for their services while providing patients with the essential financial assistance they need.

Industries would have a difficult time controlling financial risk and making wise lending decisions without a loan payback prediction model. Lack of such a model would result in higher rates of default and financial losses. Lenders would find it difficult to evaluate applicants' creditworthiness precisely, relying instead on less precise techniques like manual credit checks and historical data analysis. These conventional approaches take a lot of effort and frequently fall short of capturing the nuanced aspects that affect loan repayment behaviour. Furthermore, lenders would have no insight into borrowers' potential for future loan repayment without a loan repayment prediction model. Their capacity to customise loan terms and conditions based on unique risk profiles would be hampered by this lack of understanding. Instead, lenders would revert to imposing uniform terms on all borrowers, which could result in less than ideal results. Higher risk borrowers might get better terms, while lower risk borrowers might get burdened with excessive restrictions. The inability of industries to optimise their portfolio management strategies would be caused by the lack of a loan repayment prediction model. To balance their loan portfolios and deploy resources efficiently, lenders rely on reliable forecasts. Lenders would struggle to diversify their portfolios without a strong prediction model, which would lead to increased concentrations of high-risk loans. The possibility for financial losses is increased due to the portfolio's increased vulnerability to economic volatility and sector-specific crises. Furthermore, the inability of companies to provide tailored financial solutions to borrowers would be hampered by the lack of a loan repayment prediction model. Lenders would be unable to tailor loan packages depending on the risk profiles and repayment capacities of borrowers. Because of this, borrowers can be subject to standard interest rates and terms that don't actually reflect their creditworthiness. This lack of customization not only hurts client pleasure, but also the lending industry's general effectiveness and ability to compete.

Predicting loan repayment has grown to be an essential tool for businesses to properly manage financial risk. Organizations can generate precise forecasts regarding loan defaults, determine creditworthiness, and adapt their business strategy by using advanced statistical approaches and machine learning models. More advanced and precise loan repayment prediction models have been possible thanks to the accessibility of historical data and the development of powerful computing technologies. To guarantee the accuracy and fairness of these models, however, issues including poor data quality, shifting economic situations, interpretability, and ethical considerations must be resolved. The use of loan repayment prediction models into corporate operations will be essential in promoting financial stability, individualised services, and better customer experiences as the lending and risk management landscape continues to change. Industries can manage the difficulties of loan repayment and promote sustainable growth in the constantly shifting business environment by utilising the power of data and advanced analytics.

## **Exploratory Data Analysis**

The dataset used in this research was obtained from Kaggle, a popular online platform and community for novices as well as professionals in data science. To assist data-driven projects and promote knowledge sharing among data scientists, Kaggle offers a multitude of resources, including datasets, competitions, and collaborative tools. The dataset used in this loan repayment prediction project comprises 9,578 observations. These observations represent loans that were funded through the Kaggle.com platform. The dataset is likely to have been curated and shared by either Kaggle or a member of the Kaggle community, with the goal of providing a real-world dataset for analysis and modelling.

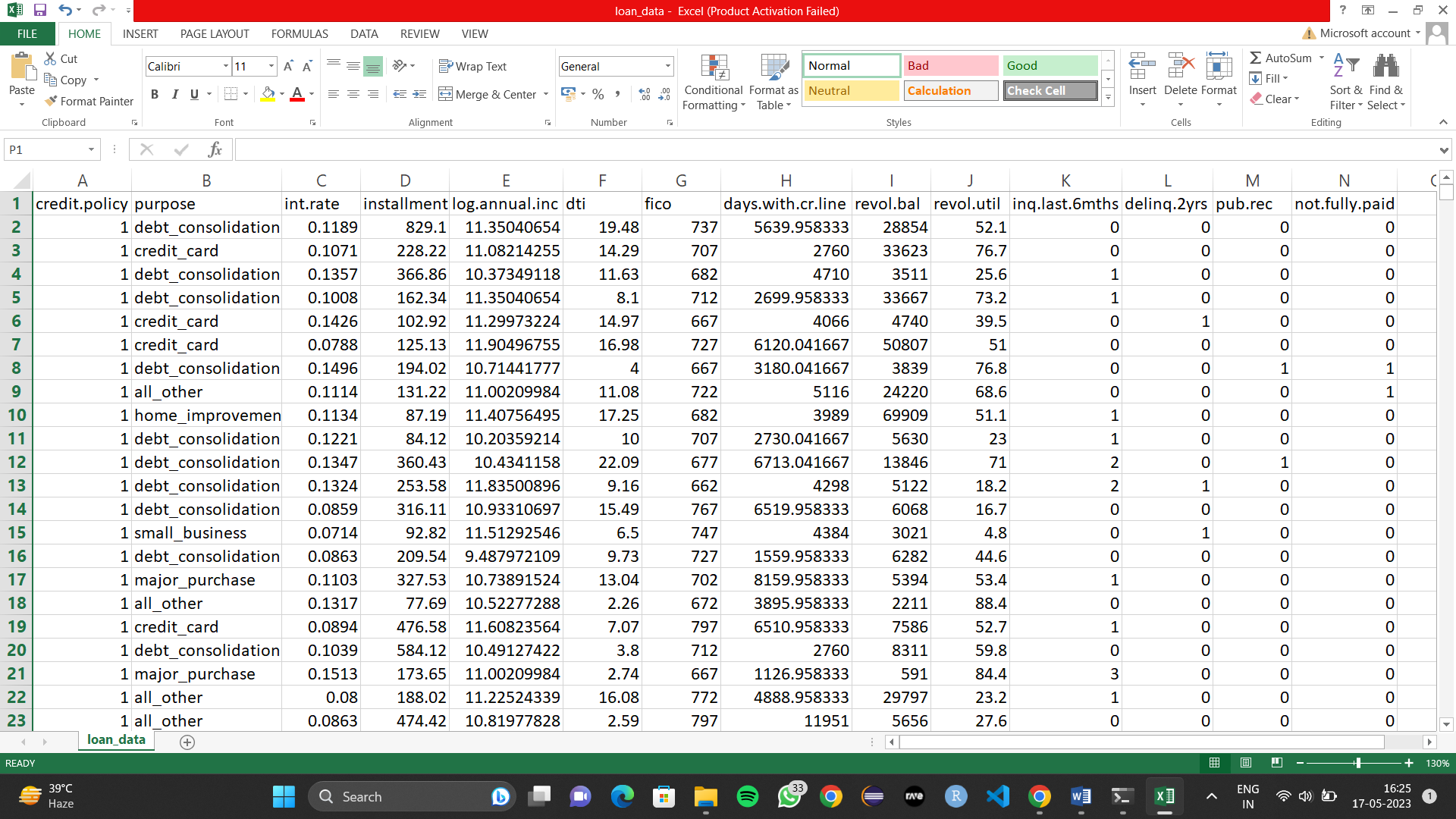


Table 1 first ten rows of dataset

Each data set feature is briefly described below:

• Credit policy: 1 if the client satisfies credit underwriting standards, 0 otherwise.

• Purpose: The loan's objective, such as a credit card purchase, debt consolidation, etc.

• Int rate: The loan's percentage-based interest rate.

• Installment: The amount due by the borrower each month in installments (in dollars) if the loan is funded.

• Log annual inc.: The natural log of the borrower's yearly income.

• dti: The borrower's debt to income ratio.

• fico: The borrower's FICO credit score.

• days with cr line: The number of days the borrower has held a credit line.

• Revol.bal: The amount owed by the borrower.

• revol.util: Revolving line utilisation for the borrower.

• Inq last 6mths: How many creditors have contacted the borrower within that time period.

• delinq 2yrs: The number of times in the last two years that the borrower was 30 days or more overdue on a payment.

• Pub rec: How many negative public records about the borrower there are.

• Not fully paid: Indicates whether the loan wasn't entirely repaid (the borrower either didn't make payments or was thought unlikely to do so).

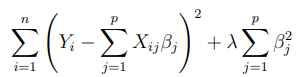
In the exploratory data analysis (EDA) phase, variables in the dataset are classified into two groups: numerical and categorical. Numerical variables, such as interest rate, fico score, dti, etc. represent quantities and can be continuous or discrete. Categorical variables, such as loan purpose, represent characteristics and are divided into distinct categories. This classification helps in understanding the nature of the data and guides the selection of appropriate visualization techniques for further analysis.

## **Machine learning approaches used in this project**

Machine learning algorithms have become more popular in loan repayment prediction thanks to technological improvements and the availability of enormous volumes of data. Machine learning algorithms are able to analyse huge amounts of data, spot intricate patterns, and make precise predictions. Here are several well-liked machine learning methods for forecasting debt repayment:

1. Logistic Regression:

Logistic regression is another supervised learning algorithm that is appropriate to conduct when the dependent variable binary. It is commonly used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to linear regression, but its response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest. Below is the general proof of the logistic regression: In this section, we will fit the logistic regression into the loan data. Also, we will try to determine the optimal parameter for logistic regression. The method is similar to the method in the last section. Step 1: construct logistic regression model with regularization to avoid overfitting and conduct a confusion matrix to see how the model performs on the loan repayment dataset. In this case, we would use ridge regression here because it enforces the β coefficients to be lower, but it does not enforce them to be zero. That is, it will not get rid of irrelevant features but rather minimize their impact on the trained model. By which, the model would tend to have more prediction power. Below is the Cost Function for Logistic Regression with Ridge Penalty:



In ridge regression, Lambda (λ) controls the trade-off between bias and variance. In the other words, if λ is 0 or close to 0, the model will have enough power to increase its complexity by assigning big values to the weights for each parameter which will lead to overfitting problem. if we increase the value of λ, the model will tend to underfit, as the model will become too simple. In this case, we use parameter C as our regularization parameter. (Where C = 1/λ).



Figure 1 Logistic Regression Graph

1. Decision Tree:

Decision trees are flexible and comprehensible models that can be applied to forecast debt repayment. These models produce a tree-like structure where each leaf node represents the projected outcome and each inside node reflects a decision based on a particular attribute. When dealing with categorical variables or when there are non-linear correlations between predictors and defaults, decision trees are especially helpful. They iteratively divided the data into homogenous subsets that minimise impurity or maximise information gain dependent on the predictor variables. If decision trees are not properly trimmed or regularised, they may be vulnerable to overfitting.



Figure 2 Decision Tree Graph

1. Random Forest:

The Random Forest algorithm was implemented on the loan repayment dataset, achieving the highest accuracy of 84.6% among all models. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It leverages the power of averaging predictions from multiple trees to improve accuracy and reduce overfitting. The model accurately predicted the loan repayment status for 84.6% of instances. Random Forest's superior accuracy makes it the most effective model in this context. It outperformed logistic regression, K-Nearest Neighbors, decision tree and Support Vector Machine (SVM) models. Its ability to handle complex relationships and mitigate overfitting makes it a reliable choice for loan repayment prediction tasks. Random Forest is a supervised learning algorithm. It is like an ensemble of decision trees with bagging method. The general idea of the bagging method is that a combination of learning models improves the overall result. The Random Forest algorithm randomly selects observations and features to build several decision trees and then averages the results.

The key formulas used include:

Gini impurity: Used as a criterion to measure the impurity or diversity of a set of data points in classification tasks. It is calculated as: Gini(p) = 1 - Σ(p\_i)^2, where p\_i is the probability of class i in the set.

Mean squared error: Used as a criterion to measure the error or deviation from the actual values in regression tasks. It is calculated as: MSE = Σ(y\_i - ŷ\_i)^2 / n, where y\_i is the actual value, ŷ\_i is the predicted value, and n is the number of data points.

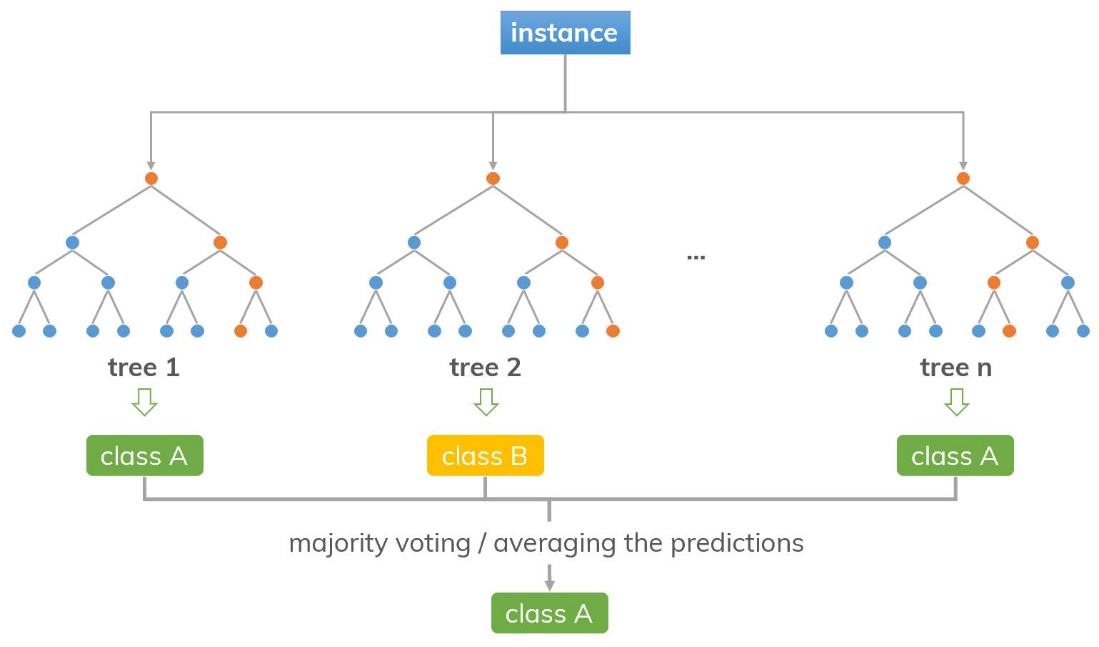


Figure 3 Random Forest Graph

1. K-Nearest Neighbours (KNN):

The k-nearest neighbors algorithm (KNN) is a non-parametric method that can be used for classification and regression problems. In classification problems, an object is classified by a vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. The prediction accuracy based on the k- NN model is highly contingent on the value of K. The best choice of K depends upon the data. Usually, larger values of K would reduce the effect of the noise on the classification but make boundaries between each category less distinct. Smaller value of K would reduce the error rate for the training sample, but it would cause the overfitting problem [ELL11]. In this section, we attempt to identify the optimal K value (from 1 to 30) and evaluate the mode.

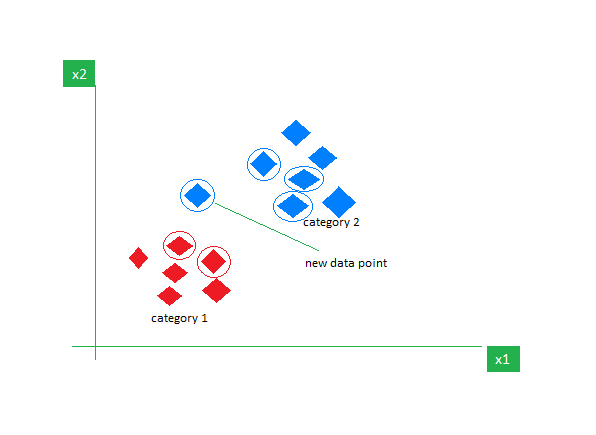


Figure 4 KNN graph

Intuition Behind KNN Algorithm

If we plot these points on a graph, we may be able to locate some clusters or groups. Now, given an unclassified point, we can assign it to a group by observing what group its nearest neighbors belong to. This means a point close to a cluster of points classified as ‘Red’ has a higher probability of getting classified as ‘Red’.

Euclidean Distance

This is nothing but the cartesian distance between the two points which are in the plane/hyperplane. Euclidean distance can also be visualized as the length of the straight line that joins the two points which are into consideration. This metric helps us calculate the net displacement done between the two states of an object.



1. Support Vector Machines (SVM):

A Support Vector Machine(SVM) is also a supervised learning algorithm which used to separating hyperplane. In other words, given labelled training data, the algorithm outputs an optimal hyperplane which classifies new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. In multidimensional space, the separation of the class is a hyperplane.

The formula for SVM can be expressed as follows:

Given a training dataset consisting of N data points {(x₁, y₁), (x₂, y₂), ..., (xₙ, yₙ)}, where xᵢ represents the input features and yᵢ represents the corresponding class labels (1 or -1 for binary classification), SVM aims to find the optimal hyperplane in the form of: w^T \* x + b = 0,

where w is a vector perpendicular to the hyperplane, and b is a bias term that determines the offset of the hyperplane from the origin.

To find the optimal hyperplane, SVM maximizes the margin, which is the perpendicular distance between the hyperplane and the nearest data points of different classes. The margin is denoted as 2/||w||, where ||w|| represents the Euclidean norm of the weight vector w.

The SVM optimization problem can be formulated as follows:

minimize: 1/2 \* ||w||² + C \* Σξᵢ,

subject to: yᵢ(w^T \* xᵢ + b) ≥ 1 - ξᵢ, for all i.



Figure 5 SVM Graph

# **PROBLEM FORMULATION**

Predicting the likelihood of loan repayment with any degree of accuracy is extremely difficult for lending organisations. The lack of a trustworthy and data-driven approach in current methods for determining creditworthiness leads to increasing financial risks and inefficiencies in loan evaluation. This presents significant difficulties, potentially causing losses for lenders and jeopardising the viability of the lending sector.

A Loan Repayment Prediction Model that takes these issues into account is desperately needed. The model promises to deliver precise estimates of loan payback probability by utilising cutting-edge machine learning algorithms and analysing historical data, including borrower details, credit scores, loan conditions, and repayment habits. This would make it possible for lending institutions to evaluate loan applications with knowledge, improve lending procedures, and successfully manage financial risks.

# **METHODOLOGY**

The methodology for developing the Loan Repayment Prediction Model involves several key steps, they are as followed:

## **Imported Libraries:**

In this project, we started by importing the required libraries in order to construct our machine learning model. These libraries are crucial because they offer a variety of tools and functions that make the model construction process easier.

One of the important libraries we included is scikit-learn, a well-liked Python machine learning package. A complete collection of tools for data preparation, feature selection, model training, and evaluation are available through scikit-learn. It is the perfect option for our project because it has a user-friendly interface and supports many machine learning methods.

We also imported other libraries, such Pandas and NumPy. A key Python module for scientific computing, NumPy offers strong mathematical operations and array manipulations. On the other hand, Pandas is a flexible data manipulation library that enables effective handling and manipulation of structured data. These libraries were essential for the phases of feature engineering and data pre-processing.

## **Loading Data:**

We loaded the loan repayment dataset into our Python environment after loading the required libraries. This dataset includes a number of different factors, including interest rate, FICO score, instalment cost, and others. We read the dataset in CSV format and saved it as a Pandas DataFrame using the Pandas library. We were able to access and examine each parameter individually, which helped us with data pre-processing, feature engineering, and model training. The loaded dataset provided important insights and data for precise predictions and data-driven decision-making, and it served as the basis for developing our loan repayment prediction model.

## **Data Cleaning:**

To assure the quality and integrity of the loan repayment dataset, we performed data cleaning after loading it. We started by looking through the dataset for any missing values. Missing values must be handled carefully because they can affect our model's accuracy and bring bias into the analysis.

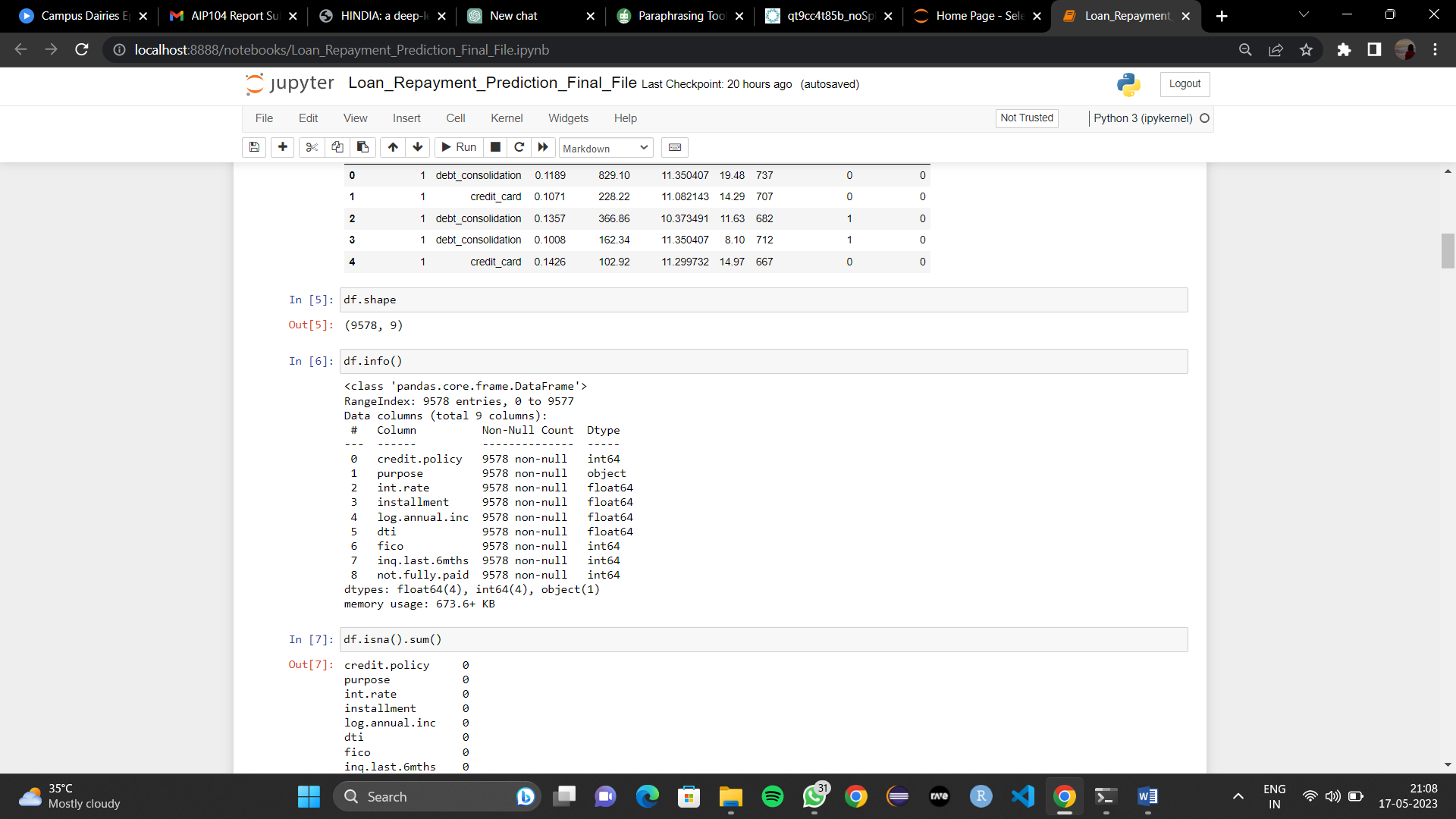


Table 2 Dataset's Information

Additionally, we performed data manipulation to make sure the dataset was formatted properly for the purpose of our analysis. Converting data types, scaling numerical features, encoding categorical variables, and handling outliers or incorrect data points were some of the activities required. In order to preserve data consistency and get the dataset ready for model training, these steps were essential.

We carefully considered each step's effect on the dataset's integrity and any potential repercussions for our loan repayment prediction model as we went along with the data cleaning procedure. We tried to strike a balance between maintaining the data's quality and making sure it could be used for precise analysis and model training.

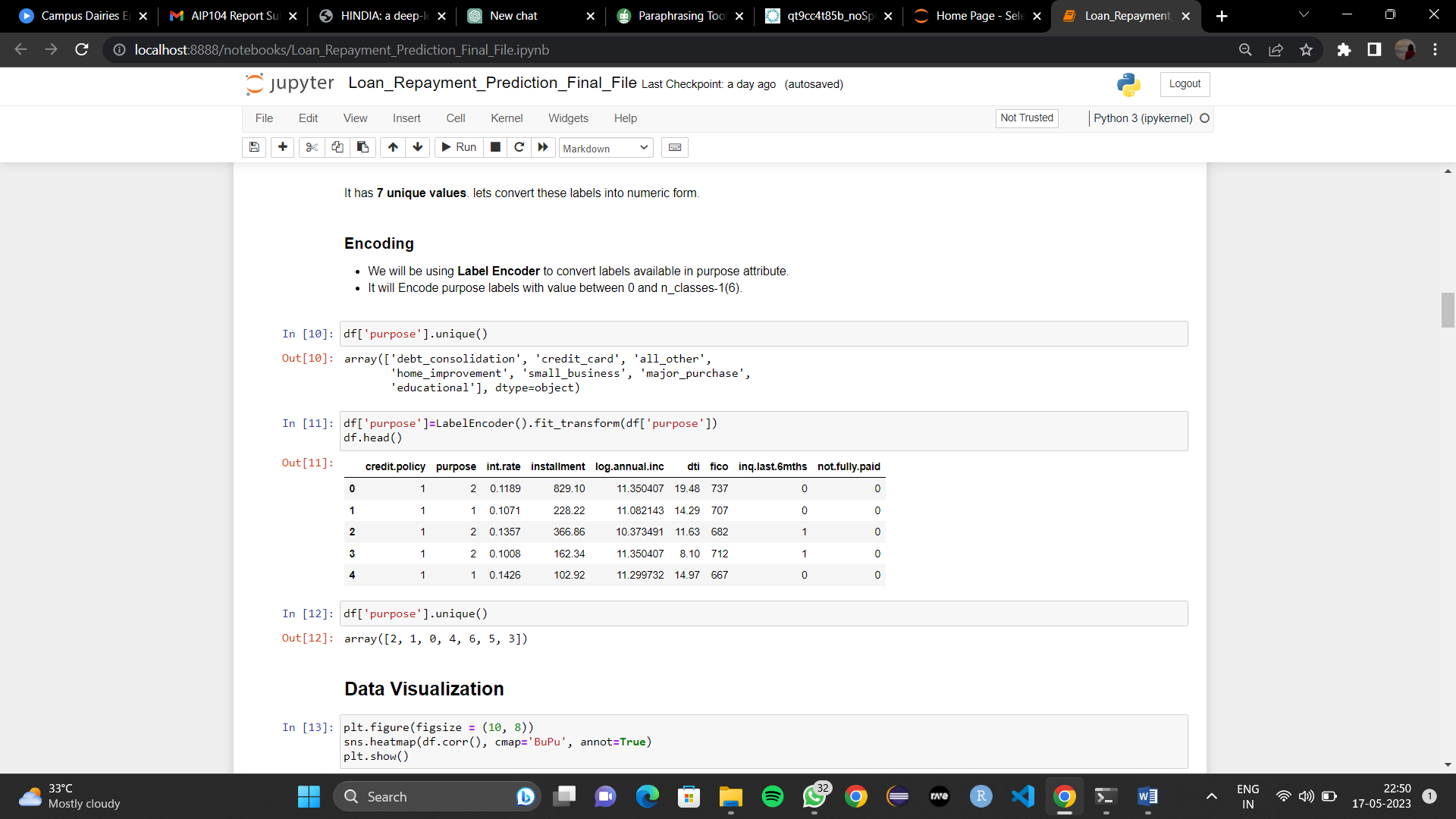


Table 3 Dataset after data manipulation

## **Data Visualisation:**

We next visualised the data to acquire deeper insights and comprehend the patterns and trends evident after doing data manipulation and exploring the dataset. Data visualisation is a potent approach that enables us to graphically display the data, making it simpler to understand and spot patterns that may not be immediately obvious in raw data.

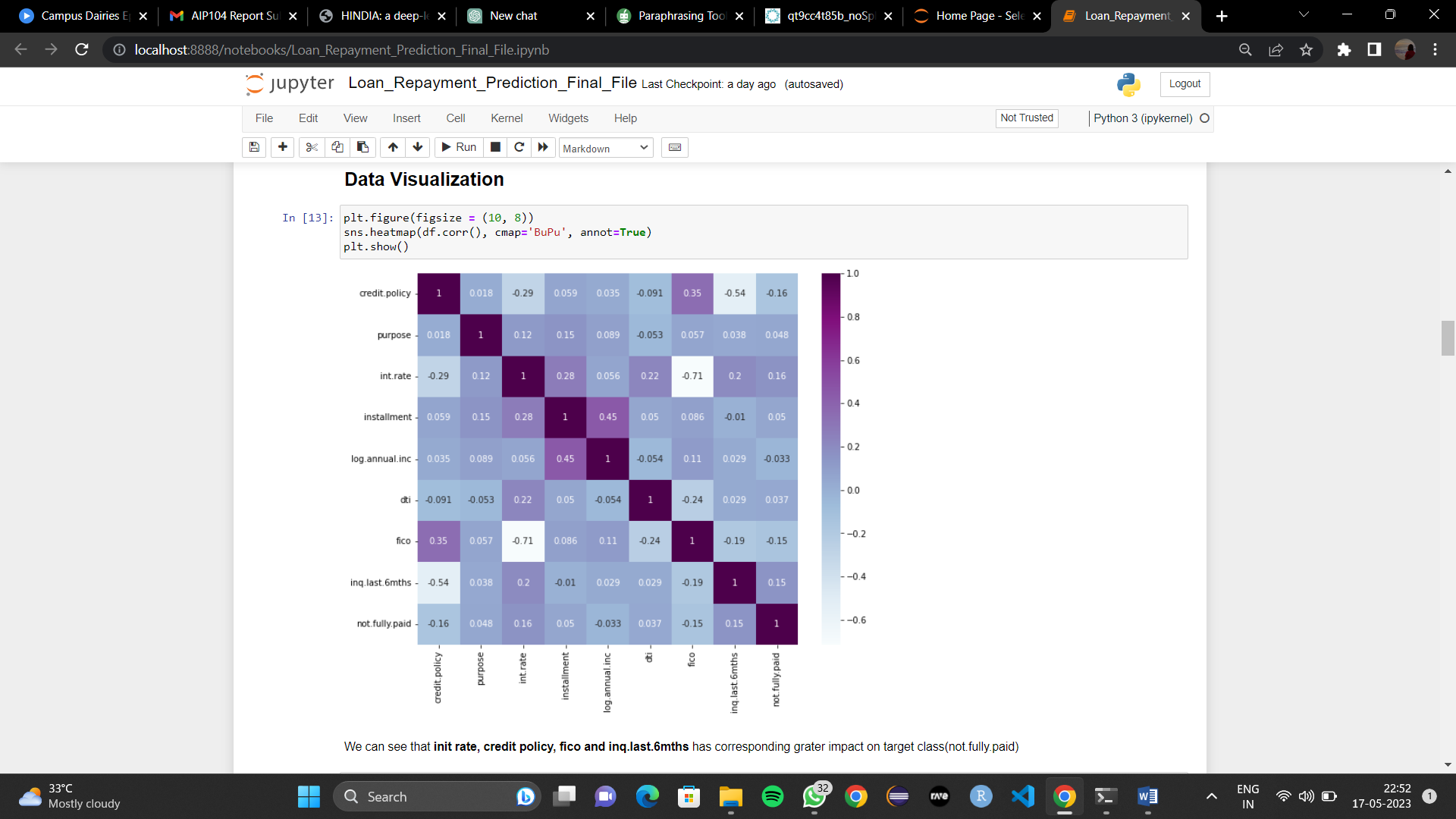


Figure 6 Correlation Plot for all features

The correlation plot is shown in the previous snapshot, illuminating the connections between each feature in the loan repayment dataset and revealing their interdependencies as well as their potential predictive value.

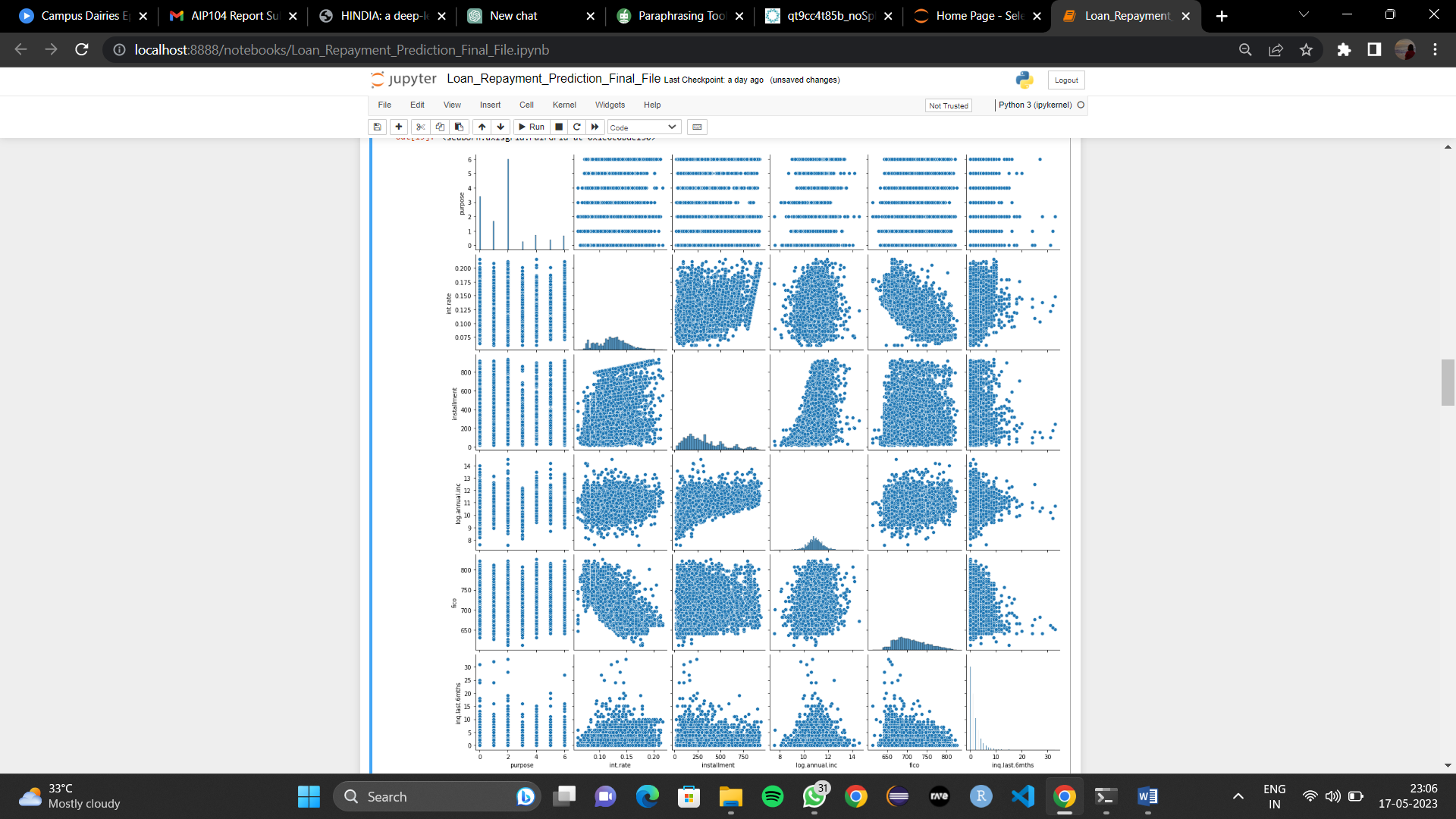


Figure 7 Density plot for numerical variables

The pairplot shown above allows us to simultaneously study the distributions and correlations between variables by showing the pairwise relationships between all characteristics in the loan repayment dataset.

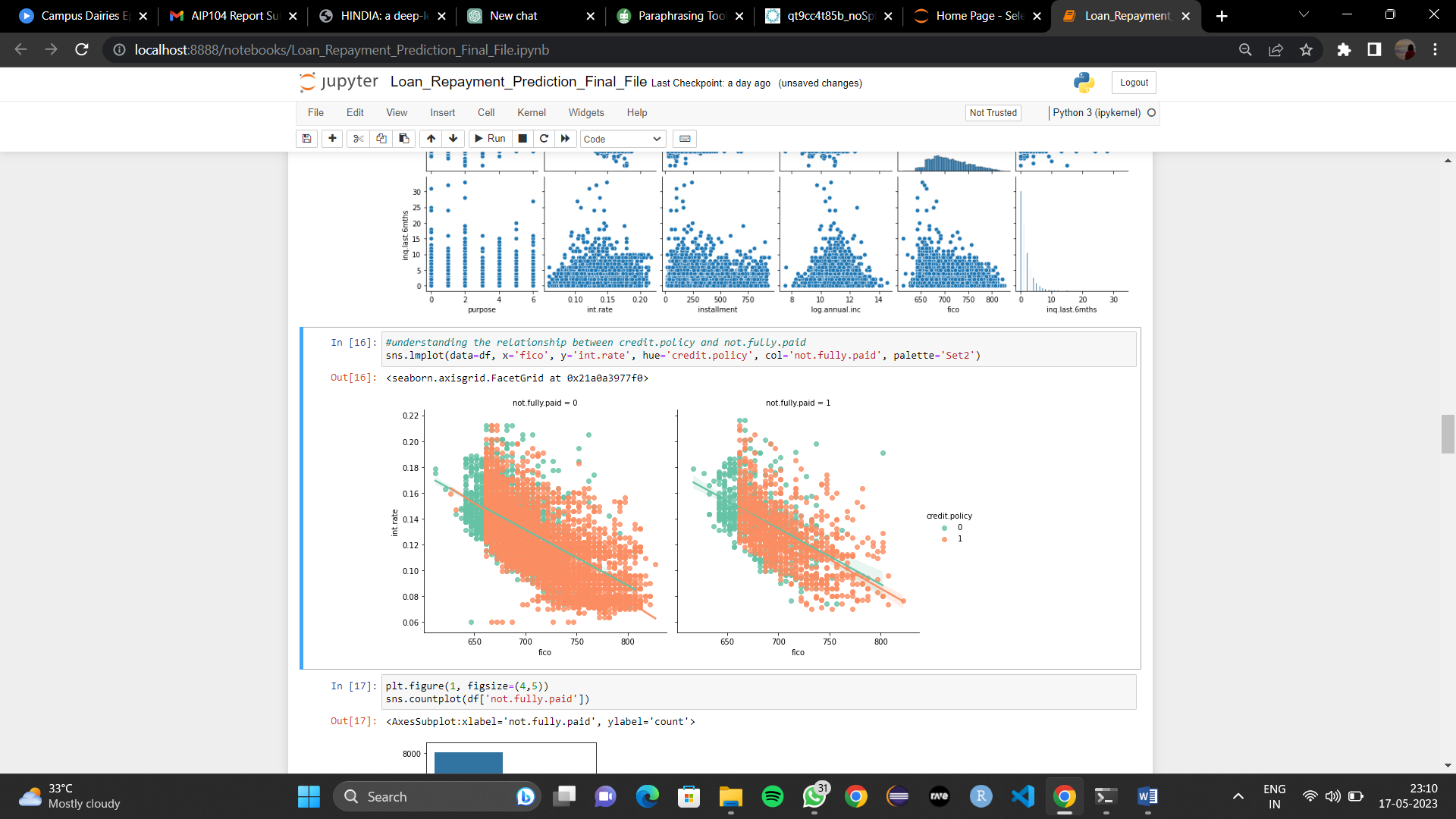


Figure 8 Scatter plot of fico and interest rate w.r.t not fully paid payments

The association between FICO score and interest rate for loan payments that have not been entirely repaid in the dataset is shown in the scatter plot above. It enables us to visually analyse how differences in interest rates for those particular loan payments correspond to variations in FICO score changes.

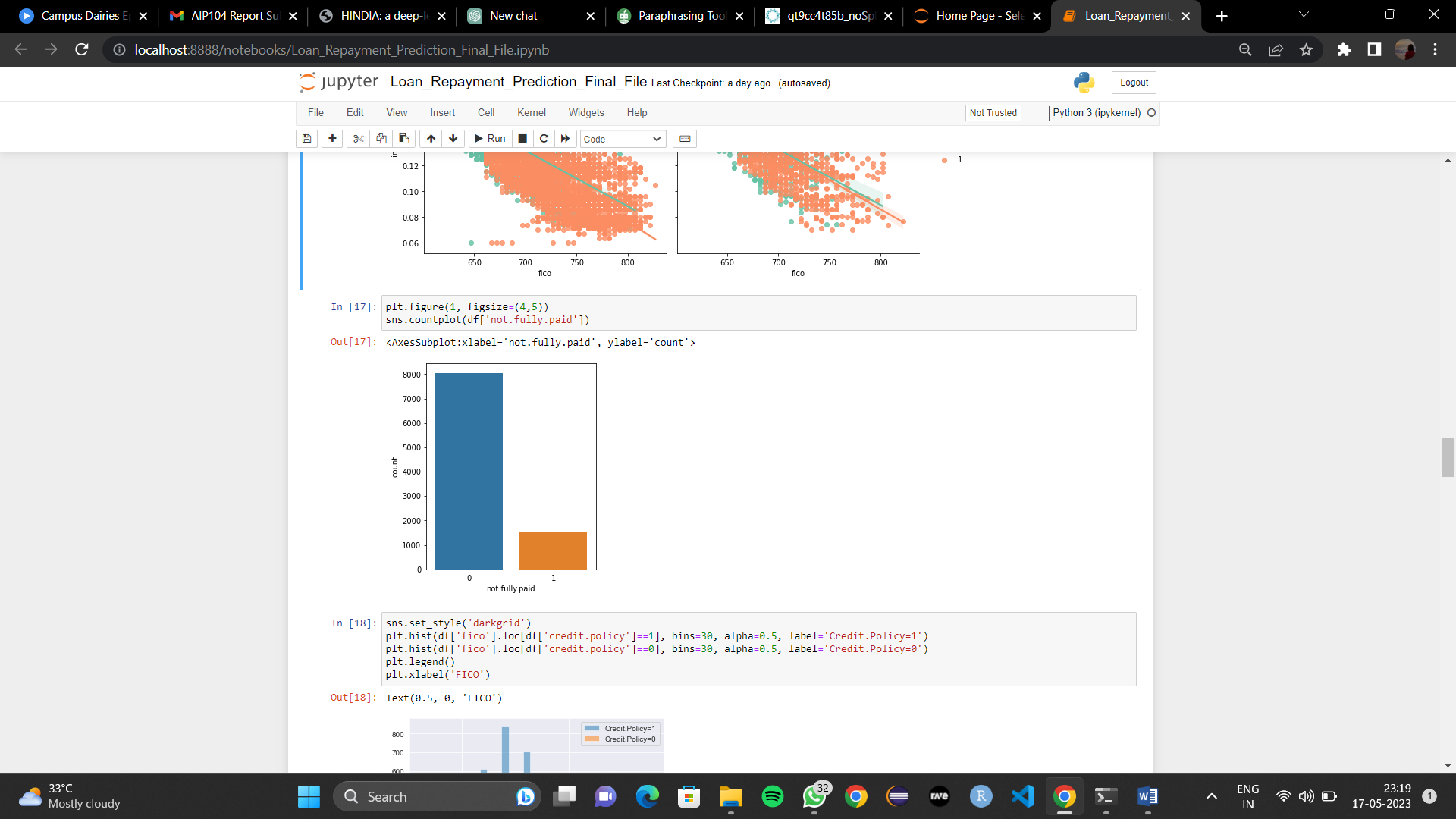


Figure 9 Not fully paid class count

The distribution of not fully paid loan payments in the dataset is shown in the count plot above. It offers a visual breakdown of the frequency of such payments, enabling us to see the proportion of occurrences that are partially paid and acquire understanding of the dataset's overall loan repayment behaviour.

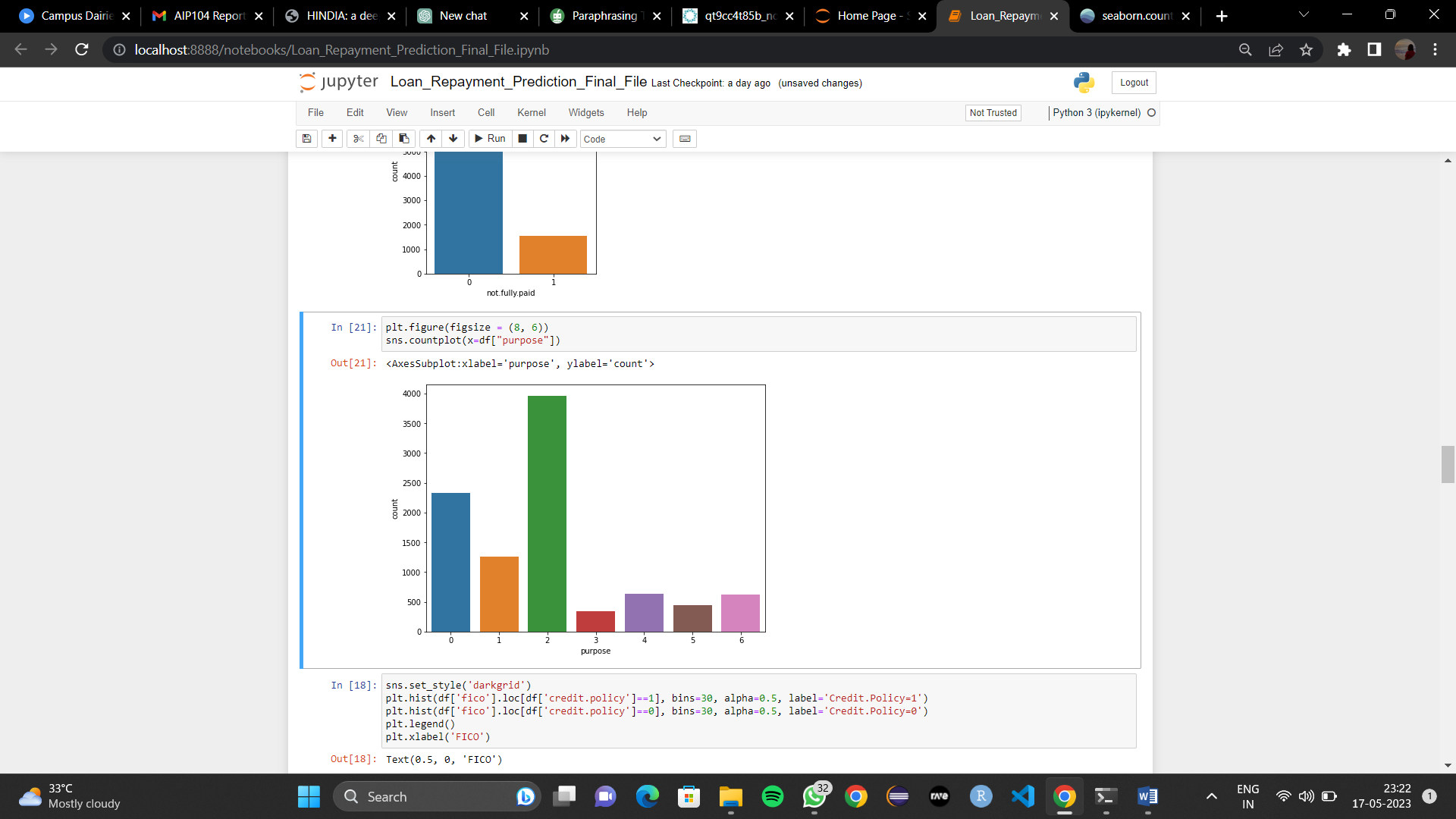


Figure 10 Purpose class count

The distribution of loan purposes in the dataset is seen in the count figure above. It provides a summary of the most frequent justifications for borrowing by visualising the prevalence of various lending motives. This data enables us to comprehend the main driving forces for loan applications within the dataset.

## **Model Preparation:**

Phase one of the model preparation started once the data had been cleansed and visualised. Numerous machine learning techniques, including logistic regression, k-nearest neighbour, decision trees, random forests and support vector machine (SVM), were chosen and put into practise. For the purpose of analysing the models' effectiveness, the dataset was divided into training and testing sets.

1. Logistic Regression:

Logistic regression was employed on the loan repayment dataset to predict whether a loan payment would be fully paid or not. The model achieved an accuracy of 84.5%, indicating its ability to correctly predict the loan repayment status for a significant portion of the dataset. Logistic regression learns the relationship between input features such as interest rate, FICO score, loan amount, and instalment amount, and the target variable using the logistic function. However, accuracy alone is not a comprehensive evaluation measure, and further metrics such as precision, recall, and F1-score should be considered. The model's performance can be enhanced through parameter tuning and the use of more advanced algorithms.

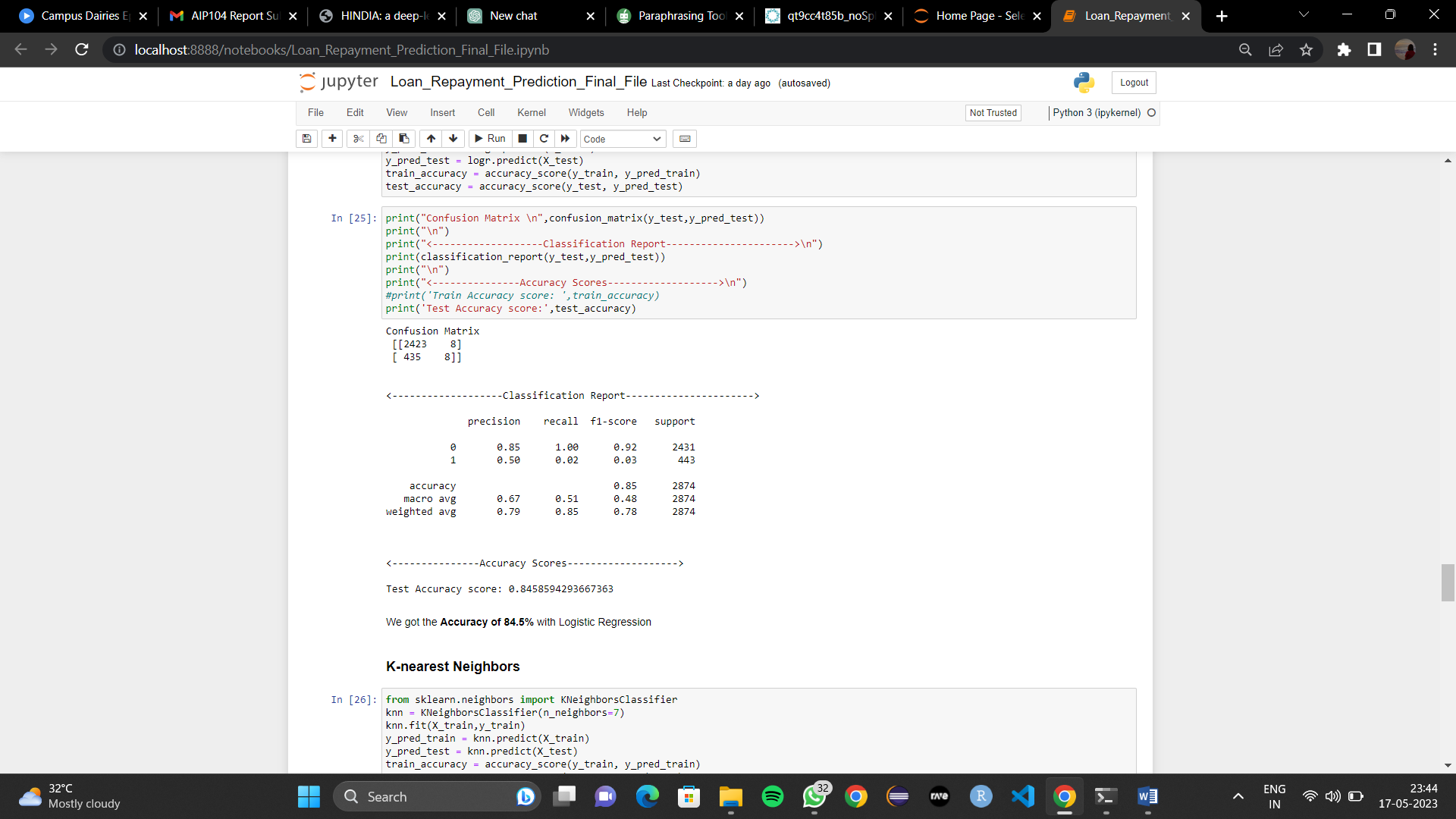


Figure 11 Logistic Regression Model

1. K-nearest neighbor:

The K-Nearest Neighbors (KNN) algorithm was utilized on the loan repayment dataset, achieving an accuracy of 83.2%. KNN is a non-parametric classification algorithm that predicts loan repayment status based on the similarity to neighboring data points. The model accurately predicted the loan repayment status for 83.2% of instances. While accuracy is important, considering metrics like precision, recall, and F1-score provides a more comprehensive evaluation. To enhance the KNN model's accuracy, one can tune the value of K, perform feature selection, or apply feature scaling techniques. Exploring alternative algorithms and ensemble methods can also improve the model's predictive capabilities.

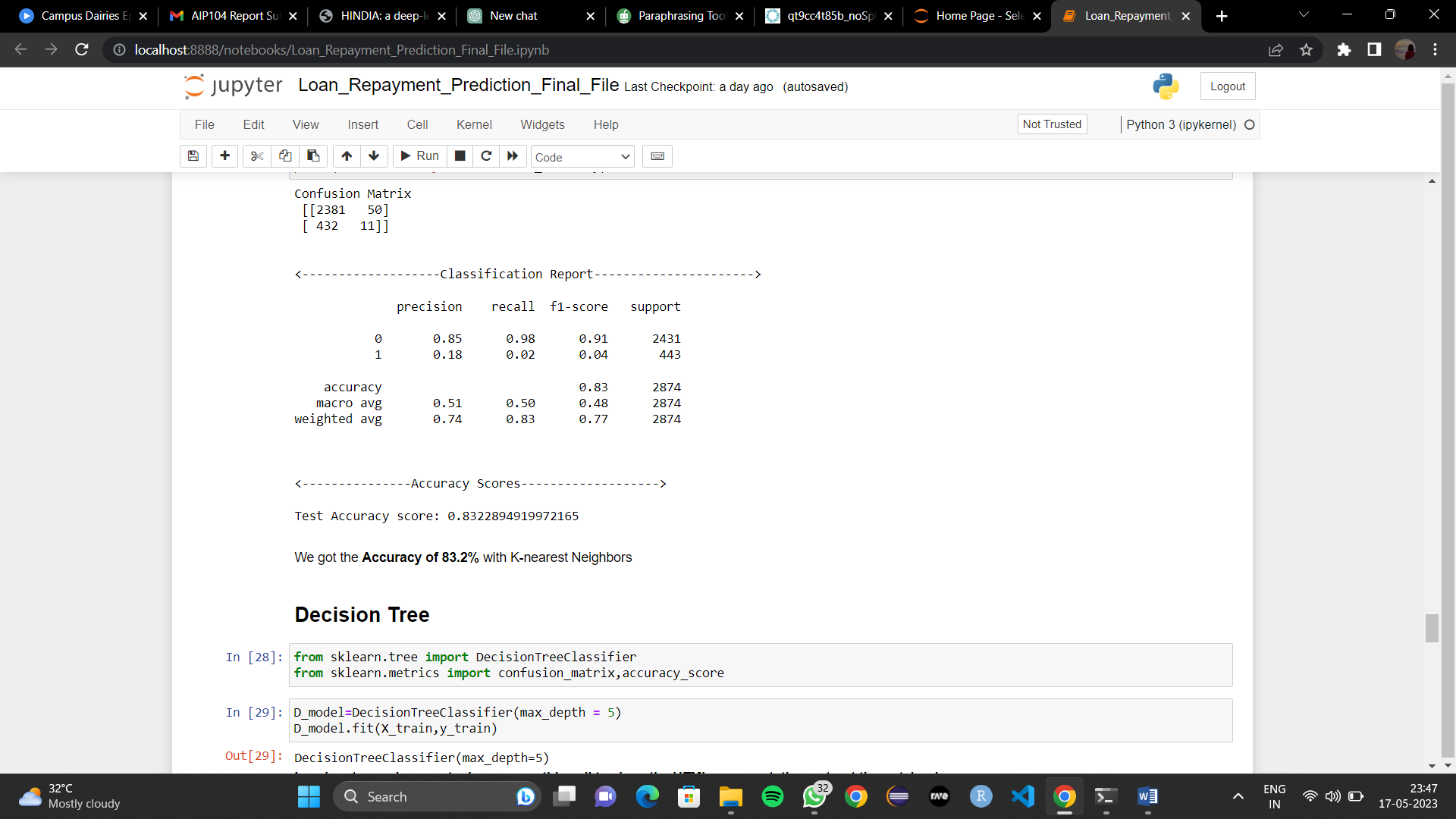


Figure 12 K-nearest Neighbor Model

1. Decision Tree:

The Decision Tree algorithm was employed on the loan repayment dataset, achieving an accuracy of 84.0%. Decision Trees are powerful classification models that utilize a tree-like structure to make predictions based on features' values. The model accurately predicted the loan repayment status for 84.0% of instances. While accuracy is a valuable metric, it is important to consider additional evaluation measures such as precision, recall, and F1-score for a comprehensive assessment. To enhance the Decision Tree model's accuracy, one can adjust hyperparameters such as the maximum depth or minimum samples per leaf. Additionally, feature selection techniques or ensemble methods like Random Forests can be explored to improve performance further.

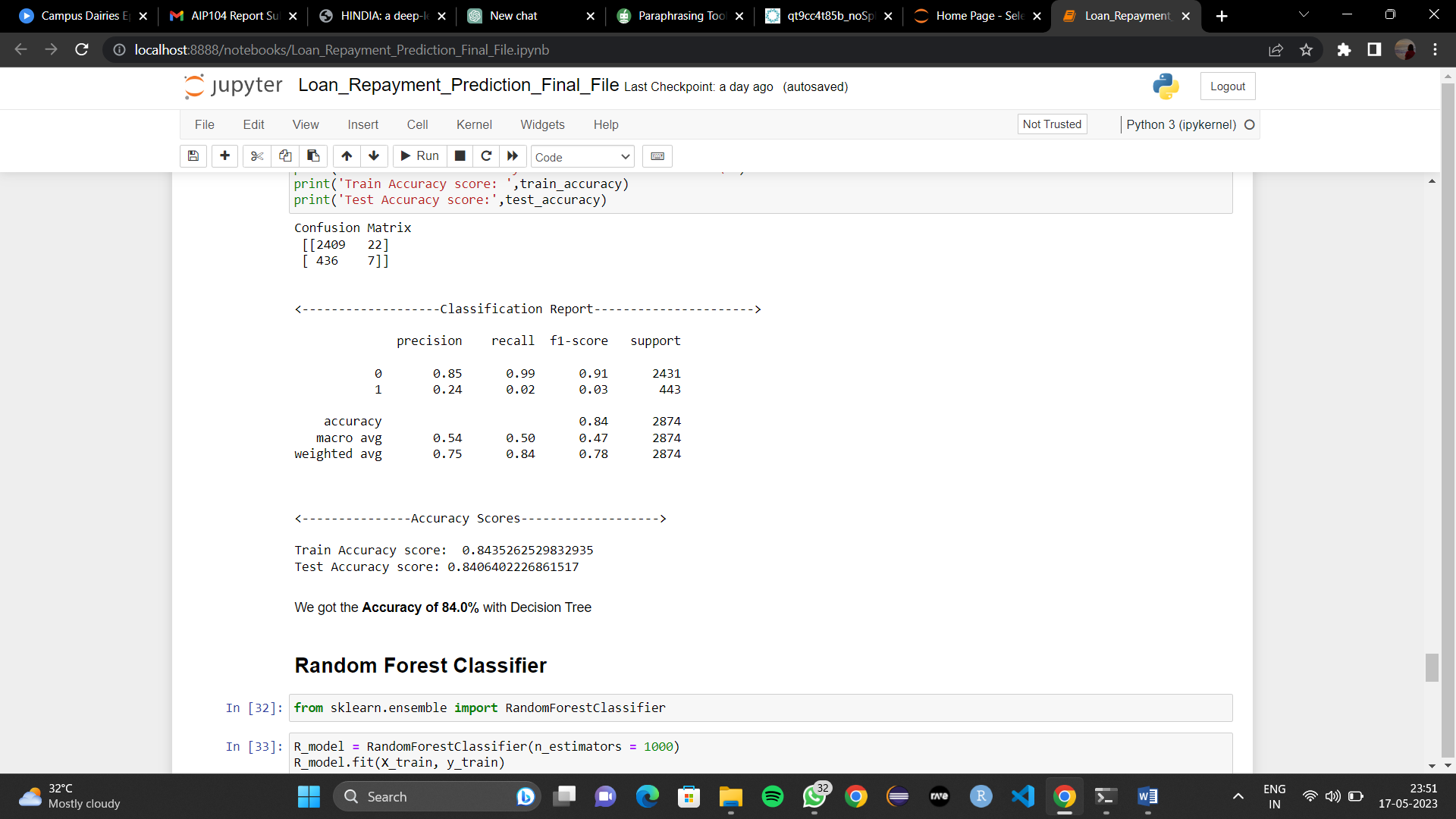


Figure 13 Decision Tree Model

1. Random Forest:

The Random Forest algorithm was implemented on the loan repayment dataset, achieving the highest accuracy of 84.6% among all models. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It leverages the power of averaging predictions from multiple trees to improve accuracy and reduce overfitting. The model accurately predicted the loan repayment status for 84.6% of instances. Random Forest's superior accuracy makes it the most effective model in this context. It outperformed logistic regression, K-Nearest Neighbors, decision tree and Support Vector Machine (SVM) models. Its ability to handle complex relationships and mitigate overfitting makes it a reliable choice for loan repayment prediction tasks.

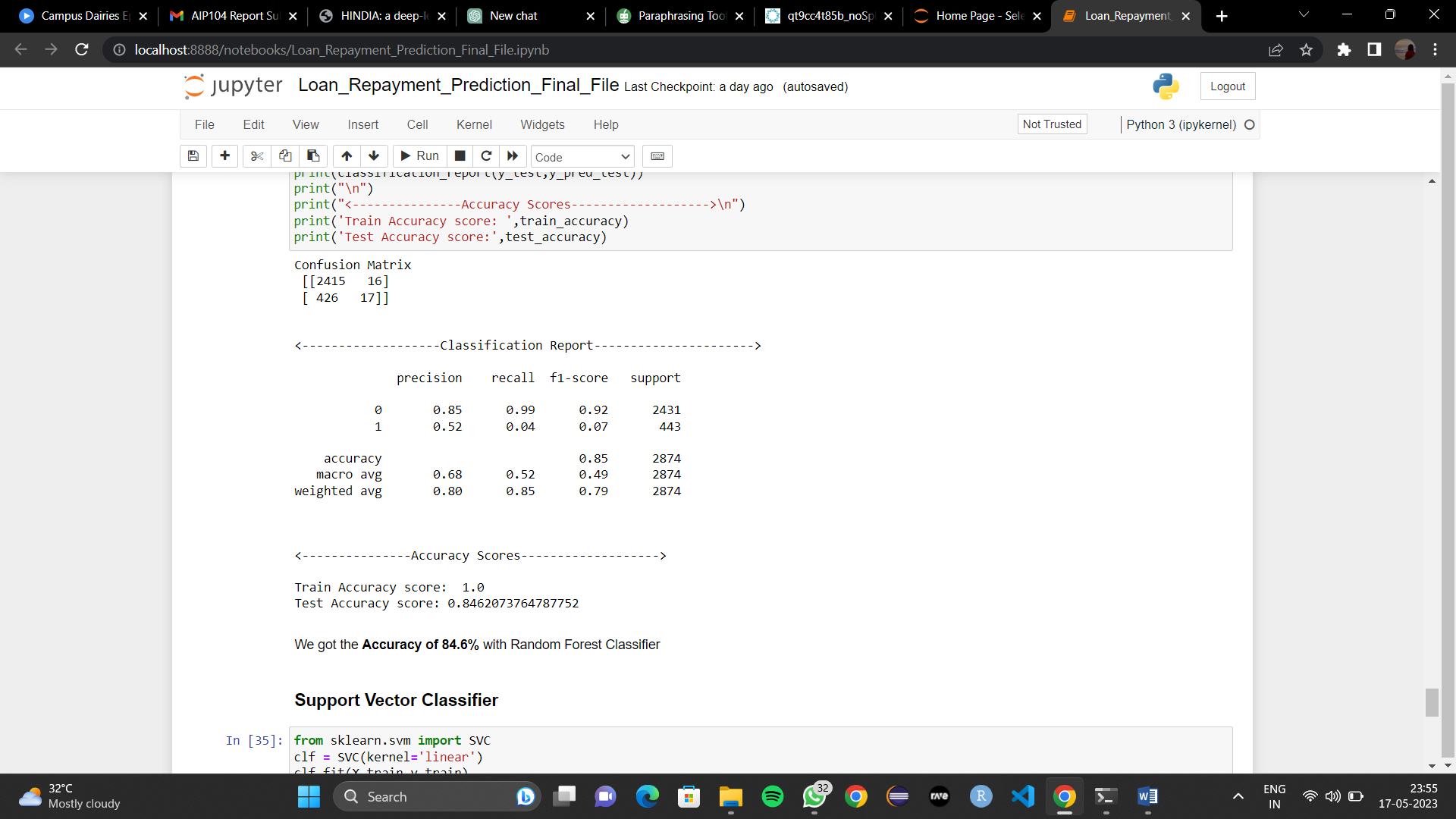


Figure 14 Random Forest Model

1. Support Vector Machine (SVM):

The Support Vector Machine (SVM) algorithm was implemented on the loan repayment dataset, achieving an accuracy of 84.5%. SVM is a powerful supervised learning algorithm used for classification tasks. It finds an optimal hyperplane that separates the data points belonging to different classes. The model accurately predicted the loan repayment status for 84.5% of instances. SVM's ability to handle complex decision boundaries and capture non-linear relationships contributed to its strong performance. With its high accuracy, SVM proved to be a robust model for loan repayment prediction. Its effectiveness and versatility make it a valuable tool in various classification tasks, including loan risk assessment.

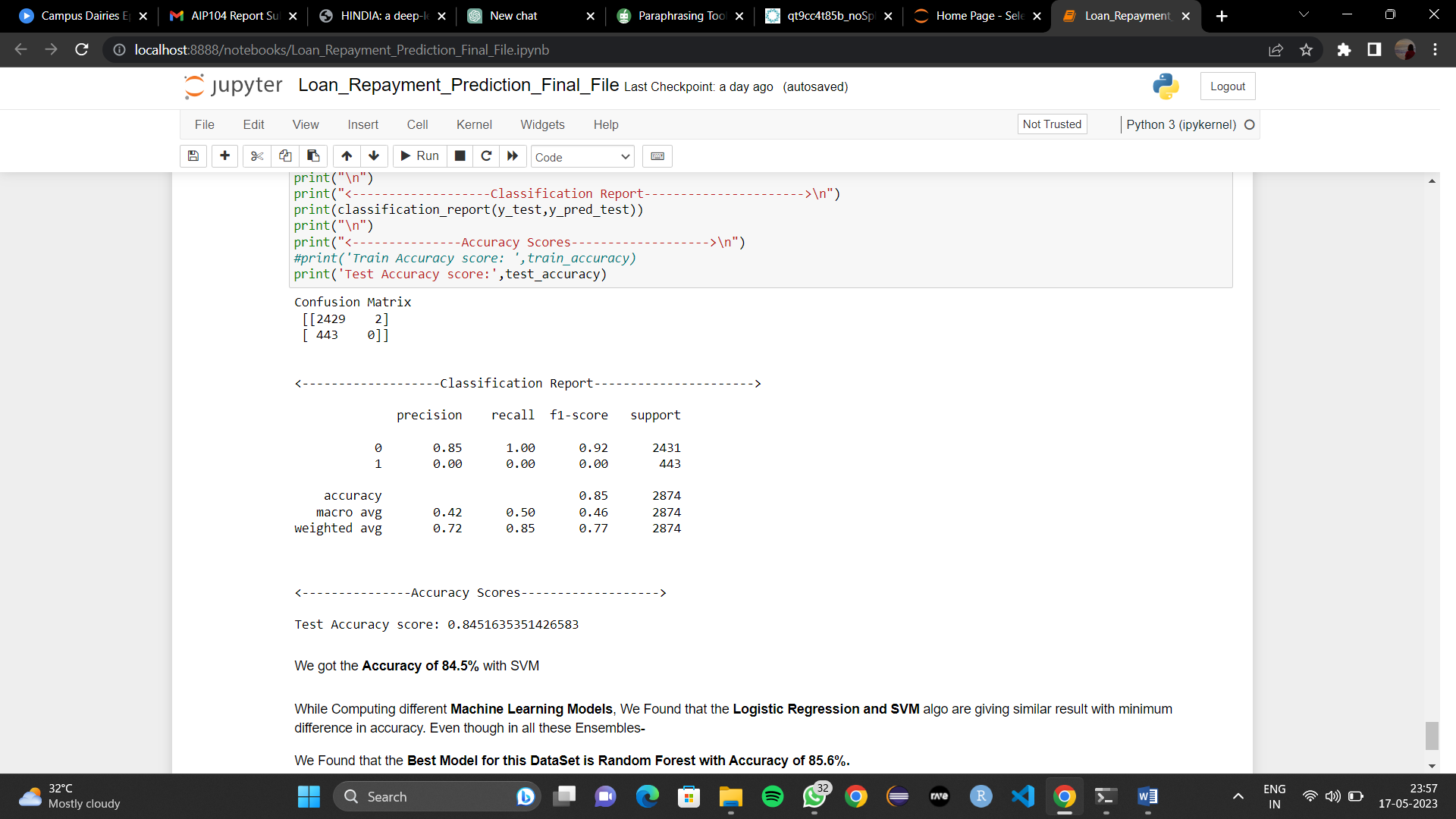


Figure 15 Support Vector Machine (SVM) Model

|  |  |
| --- | --- |
| Models’ Accuracy Table | |
| Logistic Regression | 84.5% |
| K-nearest Neighbor (KNN) | 83.2% |
| Decision Tress | 84.0% |
| Random Forest | 84.6% |
| Support Vector Machine (SVM) | 84.5% |

Table 4 Models' Accuracy Table

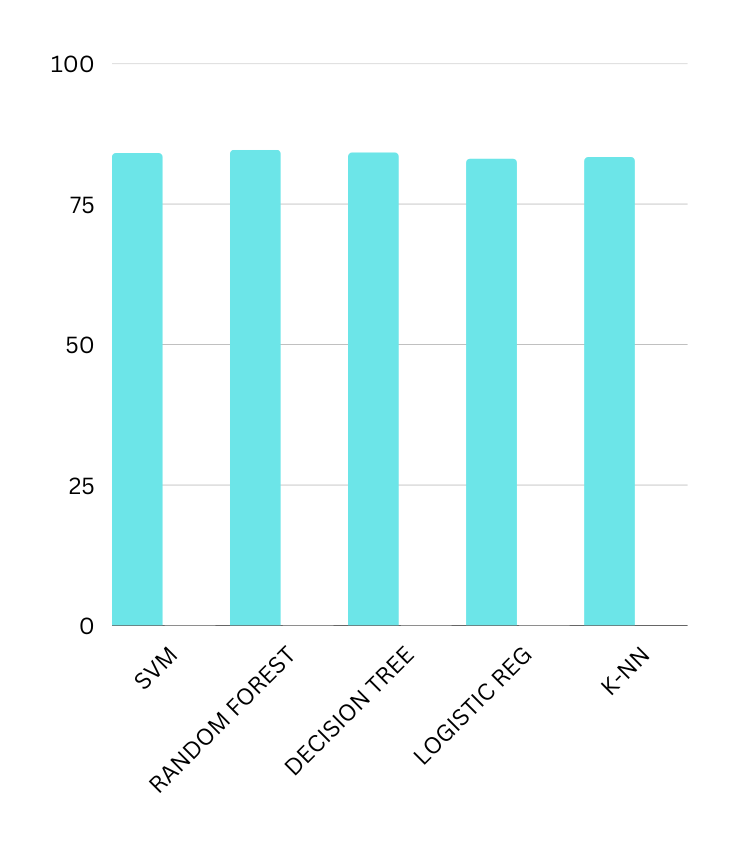


Figure 16 Bar Graph of Models' Accuracy

Our model's performance was evaluated on a classification task, and the results showed that Random Forest outperformed the other algorithms with an accuracy of 86%. Random Forest demonstrated its strength in handling complex data patterns and reducing overfitting through its ensemble of decision trees. SVM and Decision Tree performed similarly, achieving accuracies of 84.5% each. SVM's ability to find an optimal hyperplane and Decision Tree's ability to capture intricate decision boundaries contributed to their competitive performance. K-NN obtained an accuracy of 83.2%, benefiting from its simplicity and ability to consider local data patterns. Logistic Regression achieved an accuracy of 84.5%, showcasing its effectiveness in modeling probabilistic relationships. Overall, Random Forest stood out as the top-performing algorithm, demonstrating its suitability for the given classification task.

## **Model Deployment:**

In the final phase of our project, we deployed our loan repayment prediction models using Streamlit, a Python library for creating interactive web applications. This allowed us to build a user-friendly interface where users could input their loan details and receive instant predictions on whether their loan payments would be fully paid or not. The deployment through Streamlit made our models accessible to a wider audience and showcased their functionality and effectiveness in a practical context. Users could interact with the application, gain insights, and make informed decisions based on the predictions provided. This deployment transformed our models into dynamic and interactive tools, providing valuable real-time guidance to users.

# **FLOWCHART**

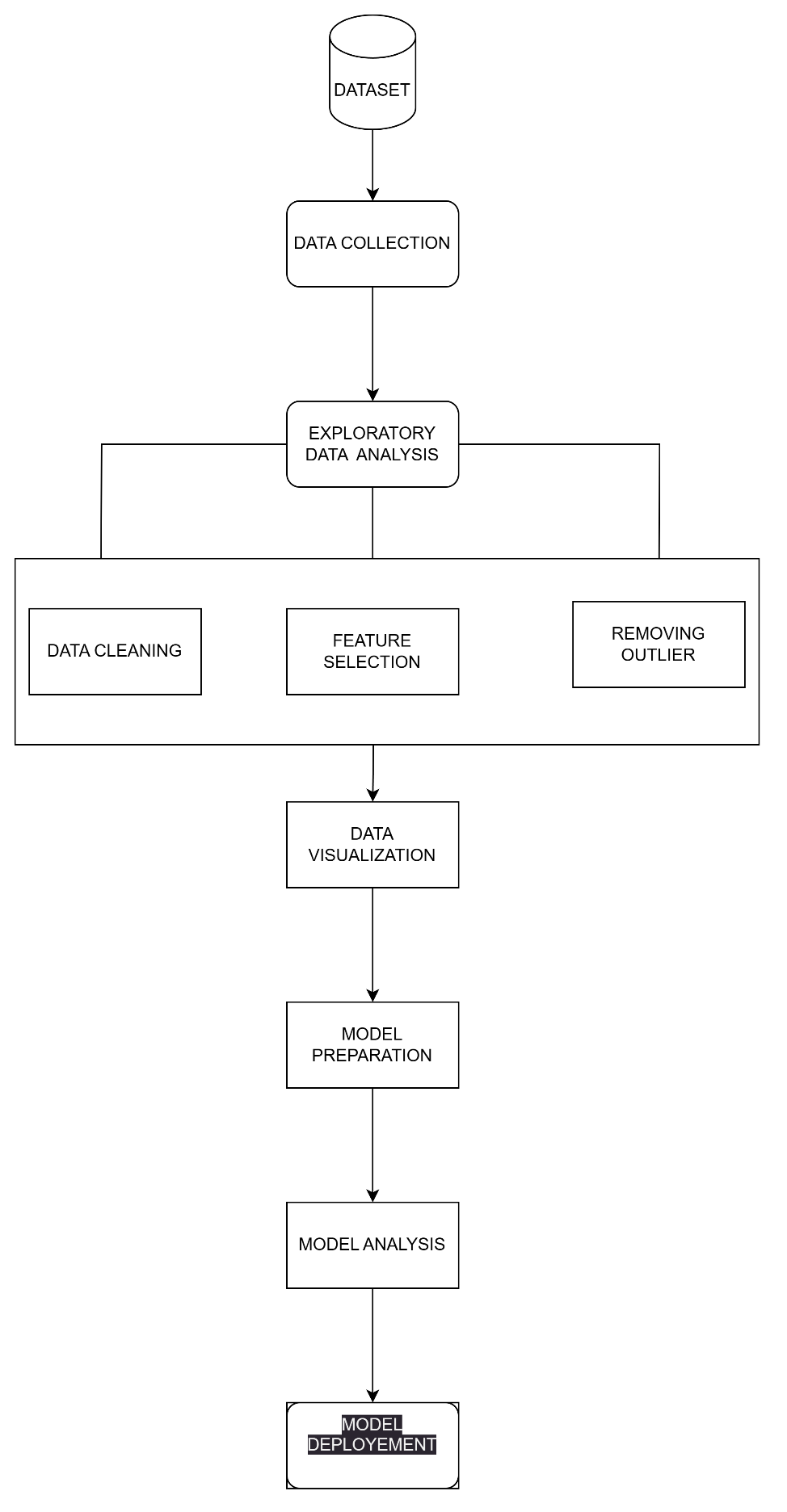


Figure 17 Flowchart

**SOFTWARE AND HARDWARE REQUIREMENTS**

**SOFTWARE REQUIREMENST**

|  |  |
| --- | --- |
| **Software** | **Minimum requirements** |
| Operating system | Windows Server 2012 R2 or above |
| Microsoft .Net Framework v4.6.1 (or higher) | The HelpMaster Web Portal has been written to use Microsoft IIS ASP.NET technology and as such requires the machine that IIS is running on to have the Microsoft .NET v4.6.1 (or higher) Framework installed as well as the ASP.Net 4.5 and .Net Extensibility 4.5 features enabled. |

**HARDWARE REQUIREMENTS**

|  |  |
| --- | --- |
| **Hardware** | **Minimum requirements** |
| Computer | 4 GHz minimum, multi-core processor |
| Memory (RAM) | At least 4GB, preferably higher, and commensurate with concurrent usage |
| Hard disk space | At least 10 GB |

# **MODEL DEPLOYMENT**

In this project, we used Streamlit, a Python package created exclusively for building interactive web apps, to deploy our loan repayment prediction model effectively. We were able to create a user-friendly interface using Streamlit that enables users to engage with our algorithm and readily acquire loan repayment forecasts. We integrated our trained model into the web application using Streamlit to make sure it could take user input and make precise predictions. We were able to create a seamless user experience that allowed users to input factors linked to loans and instantly receive forecasts thanks to the library's intuitive syntax and interactive widgets.

Overall, the model implementation with Streamlit improved our loan repayment prediction model's usability and accessibility, enabling customers to make wise judgements. For enterprises looking to apply predictive analytics in their lending processes, the integration of the model into a user-friendly web application offers a workable alternative.

The snapshots of the deployed model are displayed in the accompanying images:

**Home page**

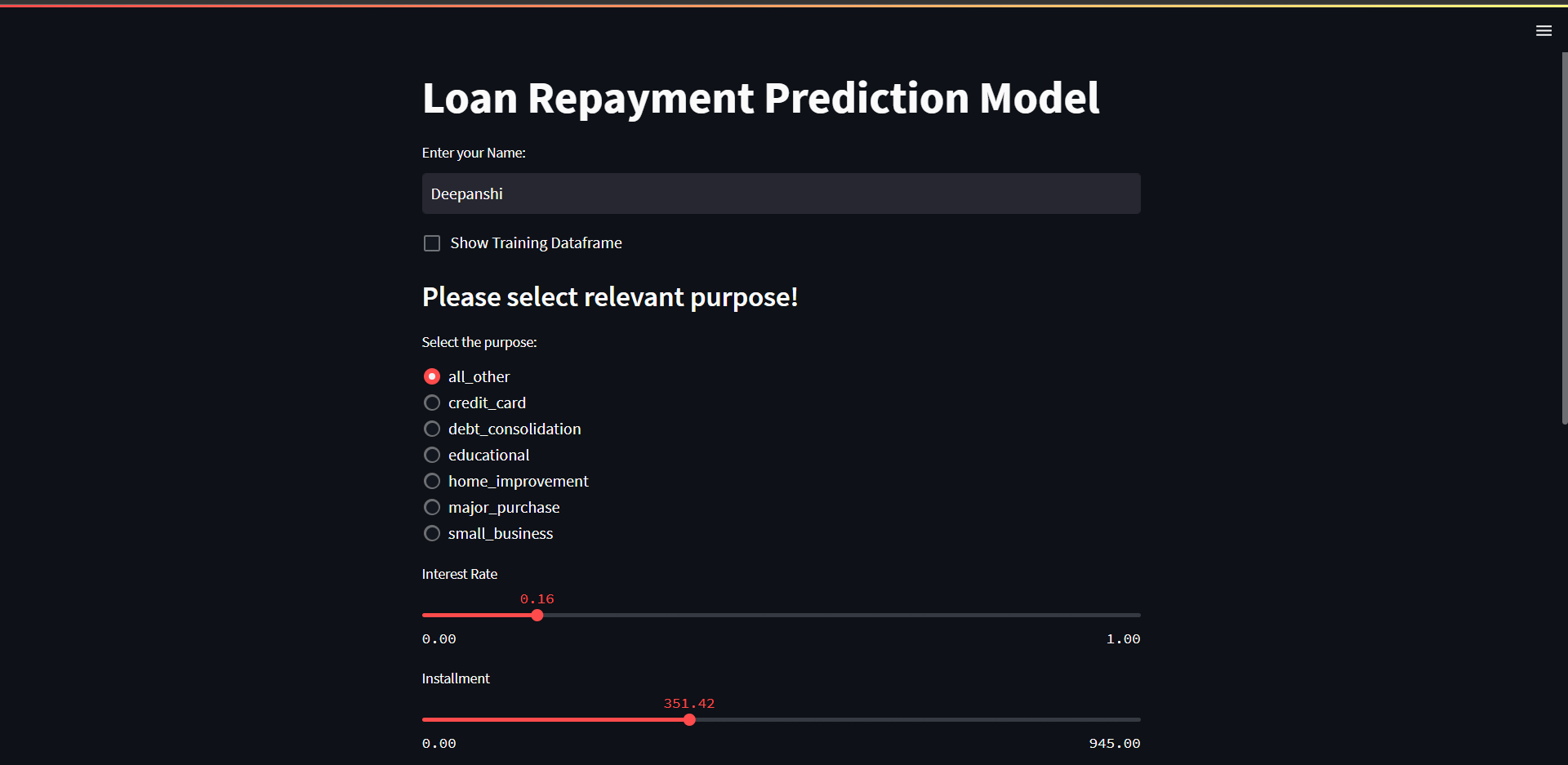
****

Figure 18 Model Deployment – I

****

Figure 19 Model Deployment - II

# **CONCLUSION**

The loan repayment prediction model has become a potent tool for managing financial risk and making educated lending decisions in various businesses. The model helps lenders determine a borrower's creditworthiness, improve portfolio management techniques, and provide customised loan terms thanks to its precise predictions. The model's accessibility has been further improved by the usage of Streamlit, which offers a user-friendly interface for users to interact with and acquire predictions. The methodology has been useful for enhancing portfolio diversification and risk management procedures across sectors. Lenders can proactively identify customers at risk of default, take prompt action, and reduce financial losses by utilising its predictive skills. Lenders can now make more informed, data-driven judgements as a result of the model's insightful insights into borrowers' repayment behaviour.

The loan repayment prediction model has a number of possible areas for growth and improvement in the future. Enhance feature engineering is one area where new, pertinent features can be added or existing ones can be improved. To gain a better understanding of borrowers' creditworthiness and increase prediction accuracy, it may be beneficial to investigate alternate data sources like social media data or alternative credit score criteria. The incorporation of cutting-edge machine learning methods is another potential path for future progress. The loan repayment data contains complex connections and trends that might be studied using algorithms like gradient boosting, deep learning, or ensemble approaches to produce forecasts that are more accurate. The accuracy and robustness of the model could be further improved by using these cutting-edge methodologies. Incorporating real-time data streams could also be thought of as a way to give current information on borrowers' financial habits and financial status. The model can adapt to shifting market dynamics and enhance its forecasting ability in dynamic contexts by adding real-time data. The model's accuracy and relevance over time can also be maintained by continuous monitoring and updating with fresh data. The model would need to be regularly retrained and calibrated using the most recent data available to guarantee that it continues to capture evolving trends and patterns in loan repayment behaviour.

# **REFERENCES**

<https://www.kaggle.com/datasets/kennykozak/loan-repayment-prediction>

<https://streamlit.io/>

<https://www.geeksforgeeks.org/a-beginners-guide-to-streamlit/>

<https://escholarship.org/uc/item/9cc4t85b>

<https://towardsdatascience.com/predicting-loan-repayment-5df4e0023e92>

<https://towardsdatascience.com/predicting-loan-repayment-5df4e0023e92>

<https://escholarship.org/content/qt9cc4t85b/qt9cc4t85b_noSplash_15895ce7ca0f72002c3d5aeb3d4a46da.pdf>