***Table of Content Page***

|  |
| --- |
| Chapter-1 Introduction…………………………………………………………...1 |
| Chapter-2 Technology and Concepts…………………………………………….2 |
| Chapter-3 Methodology …………………………………………………………3 |
| Chapter-4 Result and Discussion………...……………………………………..9 |
| Chapter-5 Future Enhancements………………………………………………..12 |

***List******Of Figures***

|  |
| --- |
| 3.1 Machine learning process……………………………………………..1 |
| 3.1.1.1 Importing Required Library and Dataset…………………….........1 |
| 3.1.1.2 No occurrence of Missing Values for Dataset………………...…..2 |
| 3.1.1.3 Distribution of Dataset Showing the Class imbalance……….…...*.*2 |
| 3.1.1.4 Distribution of Dataset after balancing…………………………....4 |
| 3.1.2 Feature Extraction………………………….......................................4 |
| 3.1.3.1 Importing Required Algorithms*…*………*……*………………….……*…*5 |
| 3.1.3 Testing Algorithms……………………………………….…………6 |
| 3.1.4 Data Visualization…………………………………………………..6 |
| 3.1.6 Streamlit Hosted Website ..…………………………………………7 |
| 4.1.1.1 Localhost Output 1………………………………………………...8 |
| 4.1.1.2 Localhost Output 2………………………………………………...8 |
| 4.1.2.2 Streamlit Output 1……………………………………………..…10 |
| 4.1.2.2 Streamlit Output 2……………………………..………………....10 |

***Chapter 1***

***Introduction***

Embarking on the Credit Card Approval Prediction project, the primary objective is to harness the power of machine learning techniques to create a robust model that can predict with accuracy whether a credit card application is likely to be approved or not. In the contemporary financial landscape, where credit plays a pivotal role in individuals' financial activities, the ability to streamline and enhance the credit approval process is crucial..

The core of this project involves the analysis of historical credit card application data, encompassing a diverse array of variables and features that influence the approval process. Through the application of machine learning algorithms, the goal is to discern meaningful patterns, correlations, and factors that contribute to successful credit card approvals.

The significance of this Credit Card Approval Prediction project lies in its potential to revolutionize and streamline the credit evaluation process. Through the integration of machine learning, the project seeks to contribute to the financial industry's ongoing efforts to make credit decisions more informed, efficient, and fair. By predicting credit card approval outcomes with precision, the project aims to foster a more accessible and responsive credit ecosystem, fostering positive outcomes for both financial institutions and credit applicants.

**1.1 Dataset:**

* The dataset encompasses diverse information about credit applicants, including demographic details, financial status, and employment specifics.
* Essential features include gender, car and property ownership, income, education, marital status, dwelling type, and age.
* Additional attributes such as employment status, job title, family size, account age, and high-risk classification contribute to a comprehensive profile.
* With a unique identifier for each record, this dataset is well-structured for analyzing factors influencing credit risk assessment in financial decision-making.

**1.2 Problem Statement:**

* The core challenge of this project revolves around building a predictive model that can effectively distinguish between genuine and fraudulent transactions.
* The main problem this project involves evaluating multiple machine learning algorithms on credit card dataset and optimizing their hyperparameters to identify the algorithm that achieves the highest predictive accuracy.
* This exploration aims to determine the most suitable approach for solving Credit Card Fraud problem using machine learning techniques.

***Chapter 2***

***Technology and Concepts***

**2.1 Machine Learning:**

* + Machine learning is a method of data analysis that automates analytical model building.
  + It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.
* It automates tasks, enhances decision-making, and enables personalization.
  + Machine Learning algorithm learns from experience E with respect to some type of task T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.
  + It enables computers to learn and improve from experience without explicit programming.
  + ML models make predictions or decisions based on data patterns.

It can be used in two ways:

* + 1. ***Supervised Learning***

Supervised learning is a machine learning paradigm where an algorithm is trained on labeled data, learning to make predictions or classify inputs based on past examples provided in the training set.

* + 1. ***Unsupervised Learning***

Unsupervised learning is a machine learning approach where algorithms analyze

unlabeled data to discover hidden patterns, clusters, or structures without

explicit guidance from predefined outcomes or labels.

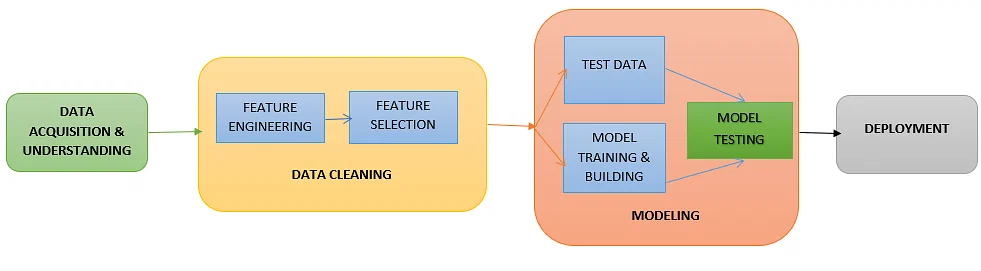
* 1. **Random Forest Algorithm**
* A Random Forest is a powerful ensemble learning technique in machine learning.
* It combines multiple decision trees, each trained on a random subset of the training data, and aggregates their predictions to make more accurate and robust predictions.
* This ensemble approach helps mitigate overfitting, resulting in a more generalized model.
* They also handle missing data and outliers well, making them robust in real-world scenarios.
* Additionally, Random Forests provide a feature importance ranking, helping users identify the most influential features in their data.
* Due to their reliability and effectiveness, Random Forests are widely used in various domains, from finance to healthcare and beyond.

***Chapter 3***

***Methodology***

The project involved developing a Streamlit web application for credit card fraud detection, encompassing all steps of the machine learning pipeline. The application simplifies data preprocessing, model training, and deployment, ensuring user-friendly access to a robust fraud detection system for credit card transactions.

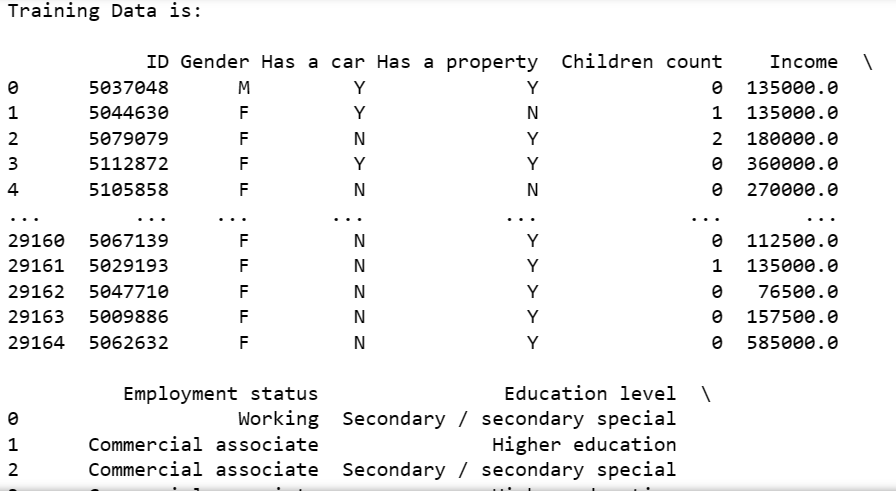
**3.1 The process of Machine learning:**



*Fig 3.1: Machine learning process (*[*source*](https://medium.com/analytics-vidhya/data-science-process-flow-and-machine-learning-fundamentals-6da8c591e8af)*)*

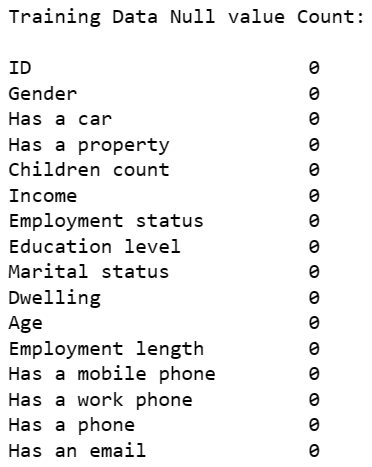
**3.1.1 Data Preprocessing:**

Preparing and preprocessing the dataset to ensure it is suitable for training and testing machine learning models. This involves tasks such as data cleaning, feature engineering, and handling class imbalance.

**3.1.1.1 Importing Dataset:**

*Fig 3.1.1.1: Importing Required Library and Dataset for Data Preprocessing*

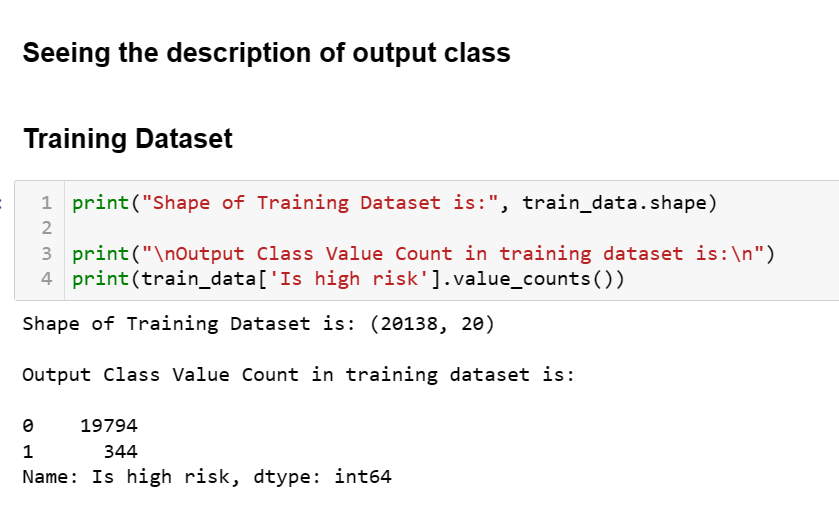
**3.1.1.2 Seeing the Missing Data:**



*Fig 3.1.1.2: No occurrence of Missing Values for Dataset*

We can see that this dataset does not contain missing data.

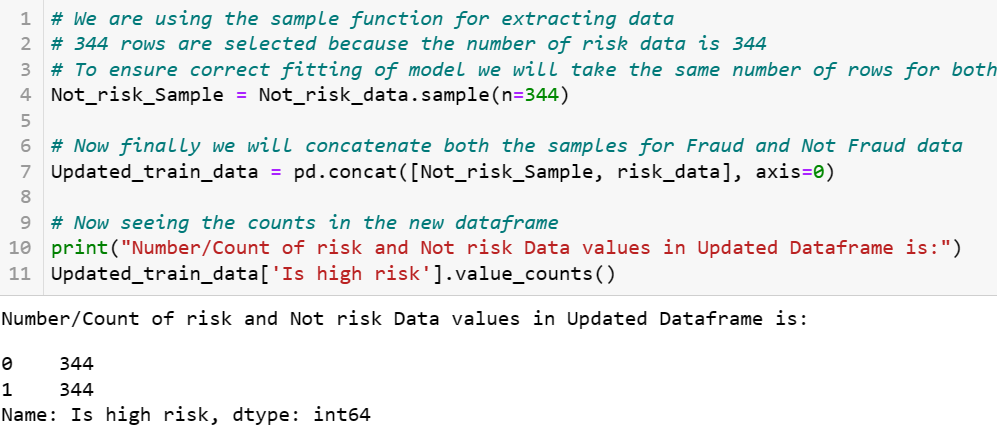
**3.1.1.3 Seeing the Dataset Distribution:**

****

*Fig 3.1.1.3: Distribution of Dataset Showing the Class imbalance*

We can see that there is a huge class imbalance in this dataset so for this we extracted data for both the classes and then we are going to create a new dataframe which will contain equal distribution and then this will serve as the data for our model.

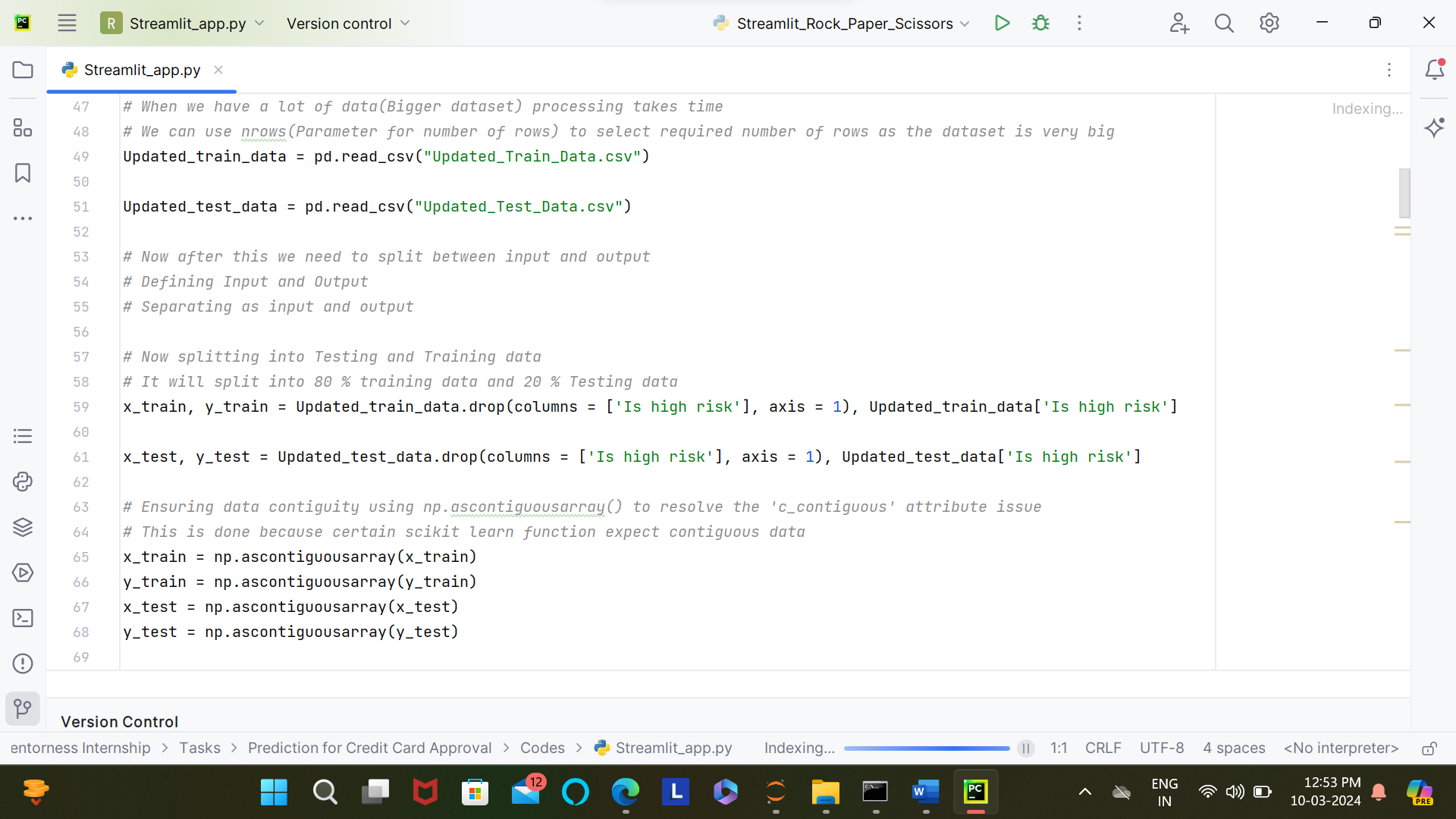
**3.1.1.4 Balancing the Dataset Class Distribution:**

****

*Fig 3.1.1.4: Distribution of Dataset after balancing*

Now in the updated dataframe we have equal number of both Fraud and Not Fraud data and this will help avoid overfitting. This dataframe is saved as csv file which is the dataset for our model.

**3.1.2 Feature Extraction:**

Now our text target is to find the Features for Input and Output in the model.

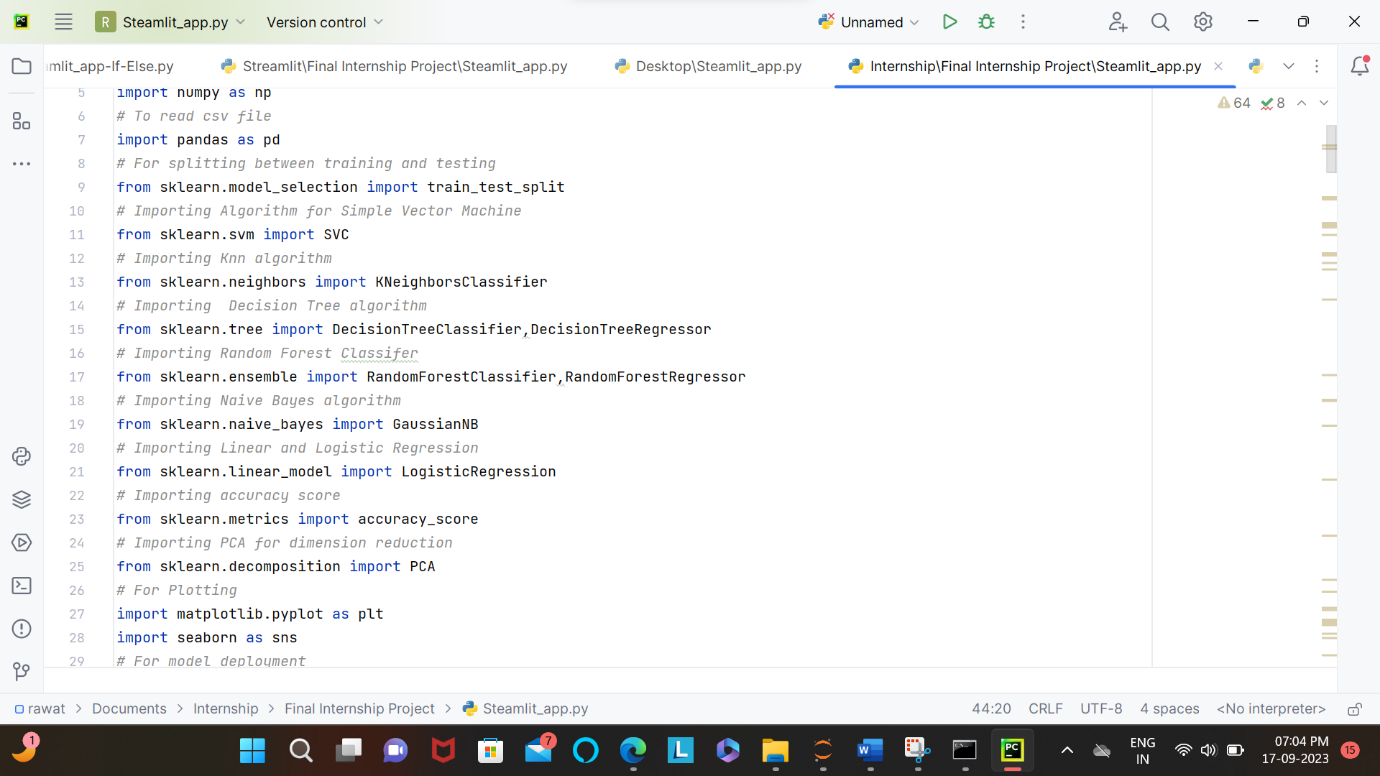
*Fig 3.1.2: Feature Extraction*

Now we have extracted features for input and output and also splitted testing and training data. After Splitting into training and testing data, to ensure compatibility we also convert to numpy contiguous array to ensure stability and compatibility with scikit-learn functions. The next task is model development or choosing algorithm.

**3.1.3 Model/Algorithm Development:**

**3.1.3.1 Exploratory Data Analysis:**

* After feature extraction several supervised learning algorithm are applied in this dataset and the highest accuracy is given by random forest algorithm.
* Thus Random Forest algorithm works best for this dataset.



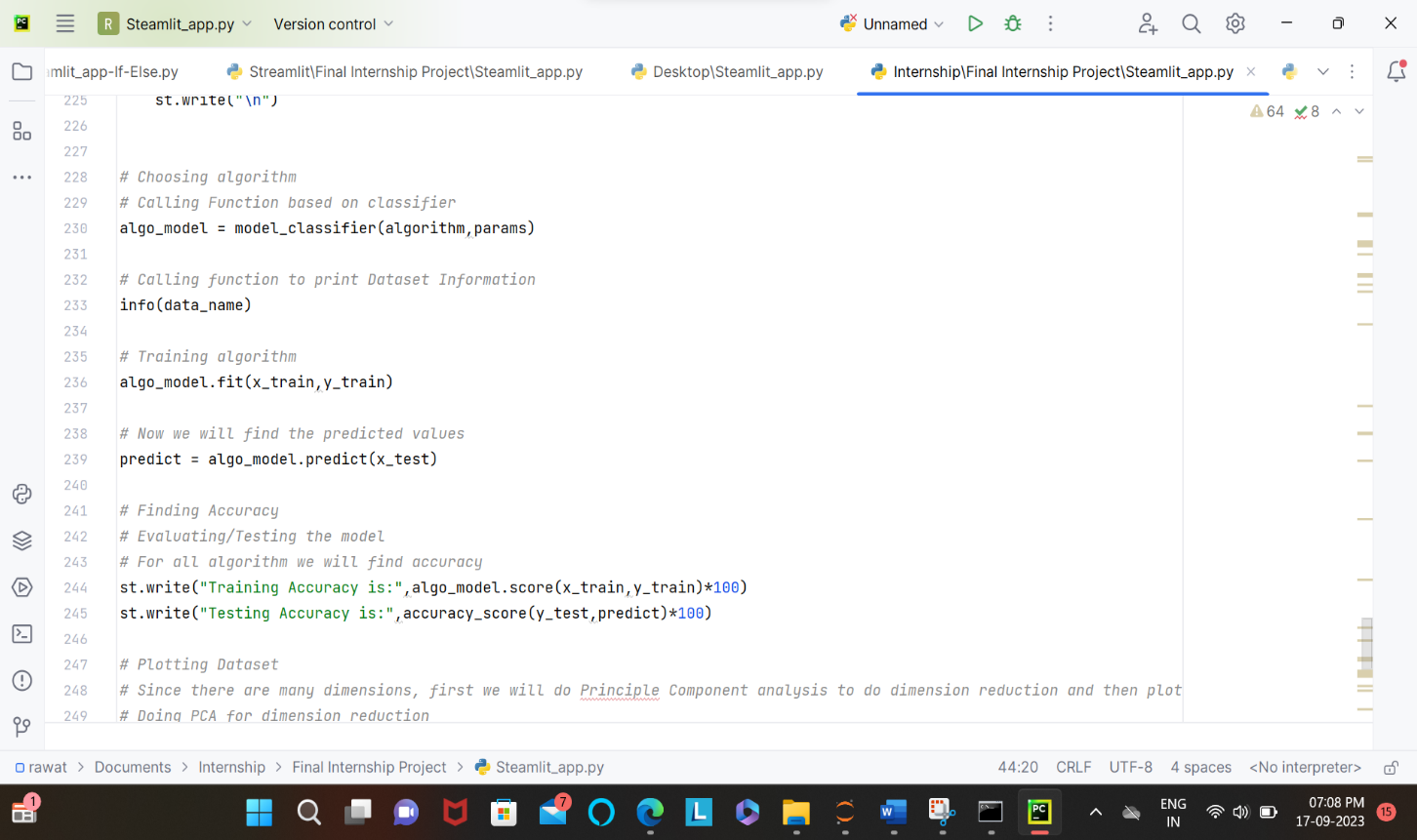
*Fig 3.1.3.1: Importing Required Algorithms*

**3.1.3.2 Hyperparameter Tuning:**

* Hyperparameter tuning is the process of optimizing the settings or configuration of a machine learning model to enhance its performance.
* This involves systematically adjusting hyperparameters, such as learning rates, regularization strengths, and tree depths, through techniques like grid search or random search to find the best combination that yields the highest accuracy or minimizes the loss function.
* Hyperparameter tuning is essential for fine-tuning model behavior and achieving optimal results in various machine learning tasks.
* In this project I have created a streamlit app in which we can see the performance of supervised learning algorithms in this dataset and also tune parameters and see the effect on overall accuracy.

**3.1.3 Model/Algorithm Testing:**

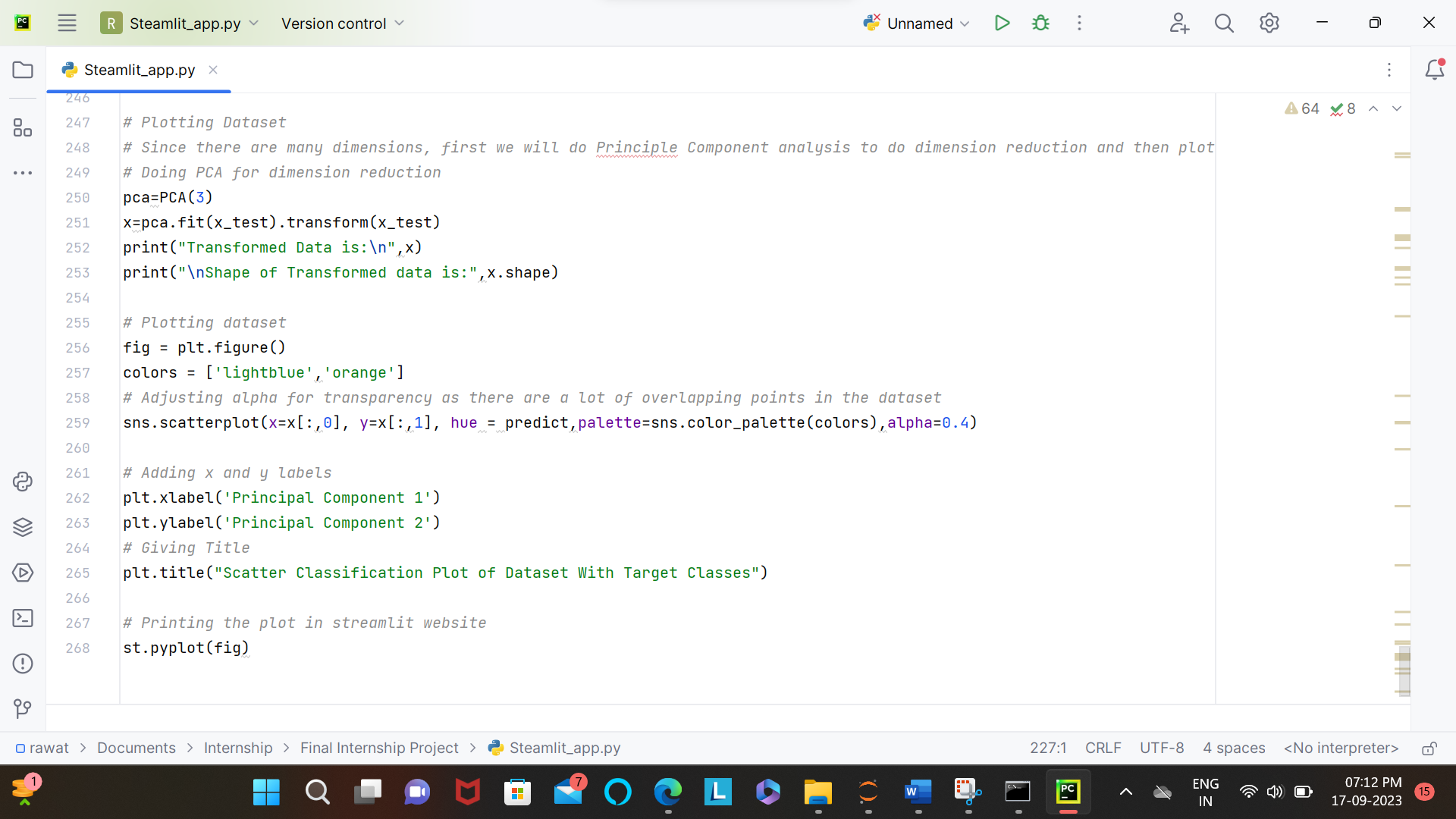
Finally we check the training and testing accuracy for our model for all datasets.



*Fig 3.1.3: Testing Algorithms*

The highest accuracy is given by Random Forest.

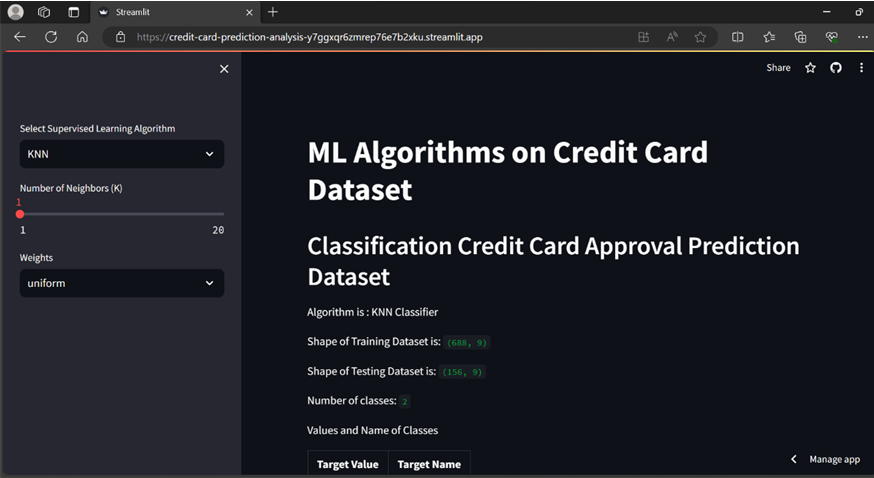
**3.1.4 Data Visualization:**



*Fig 3.1.4: Data Visualization*

First Dimenstionality reduction is done on dataset and then final plotting is done for all algorithms and varying parameters.

* + 1. **Deployment:**
* The final project deployment is done using streamlit.
* In the streamlit app we can see all algorithms and their accuracy on the dataset along with the data visualization according to the parameters and algorithm.
* We can also see the working of various algorithms on this dataset by plotting scatter plot for the algorithm and dataset.
  + 1. **Hosting with GitHub and Streamlit:**
* Streamlit is an open-source Python library that simplifies the creation of web applications for Datascience and ML.
* I have also hosted this app using GitHub and streamlit and it can be seen at: [Streamlit (final-internship-project-nvdx7l2pqe42g2p9ambd8s.streamlit.app)](https://final-internship-project-nvdx7l2pqe42g2p9ambd8s.streamlit.app/)
* Since the repository is private code cannot be seen.
* But anyone with a public URL can access this app.



*Fig 3.1.6: Streamlit Hosted Website*

* For successful deployment we first ensure that it runs successfully in our local machine
* After running it in our local machine we can deploy it using GitHub and Streamlit.
* All the code files are added in the GitHub Repository.
* One requirement file is made with the extention .txt which contains the name of all required libraries or dependencies.
* This file is also uploaded in repository.
* Finally after using all this files, we can host our app in streamlit.

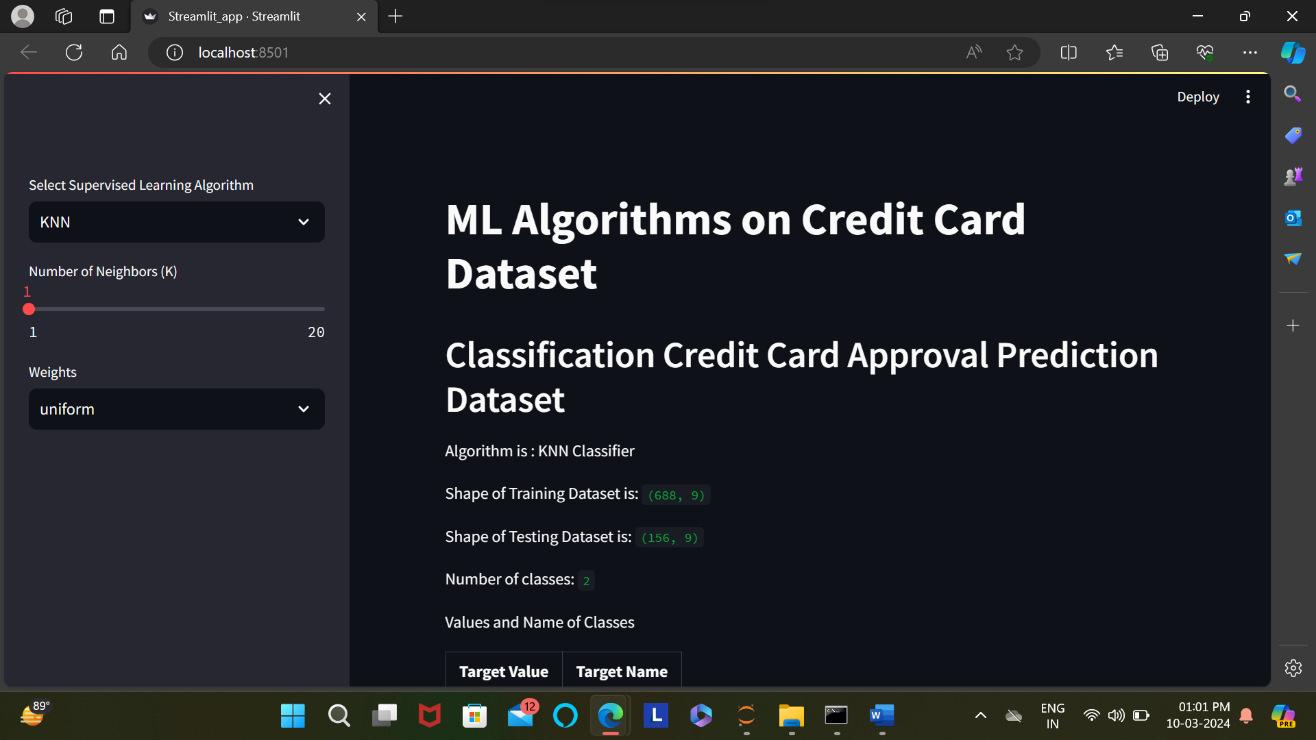
***Chapter-4***

***Result and Discussion***

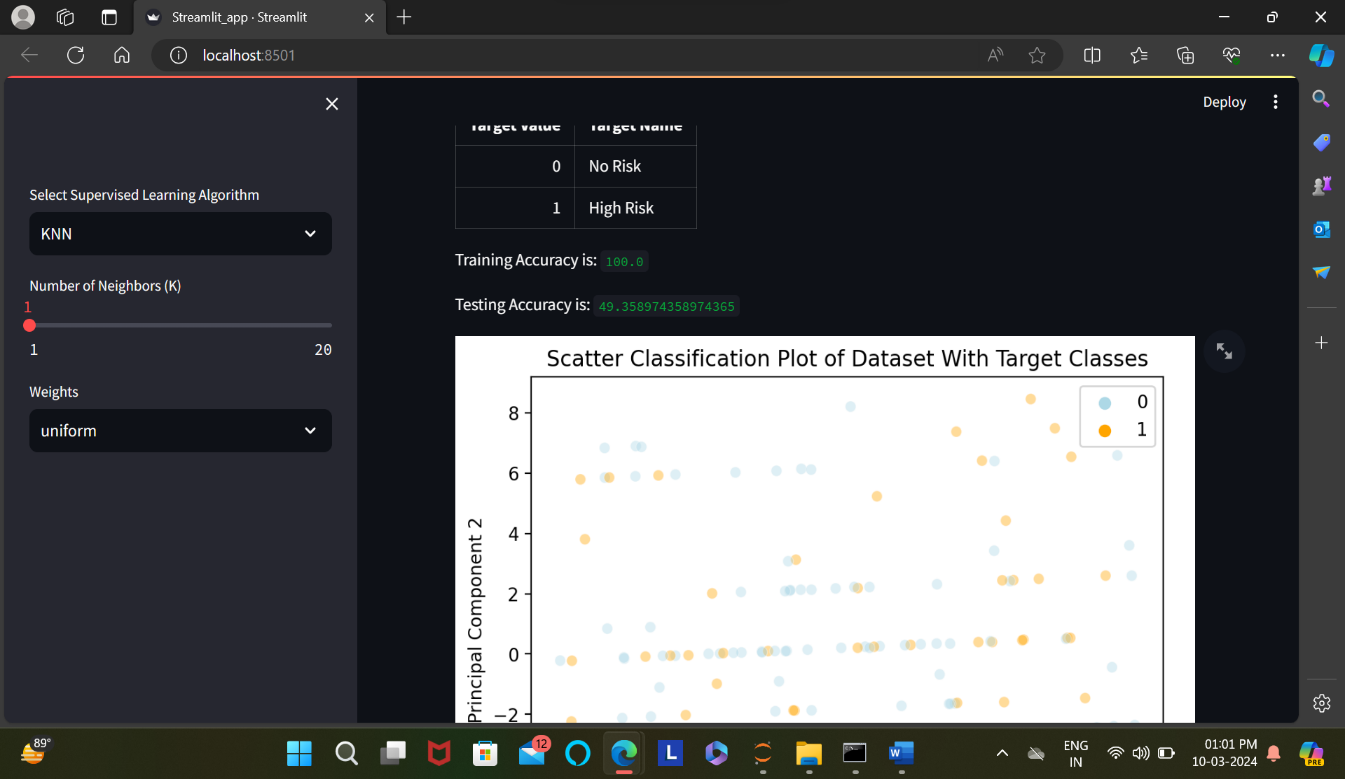
**4.1 Results:**

**4.1.1 Localhost:**

After running the app using Streamlit we can see the output as follows on the localhost:



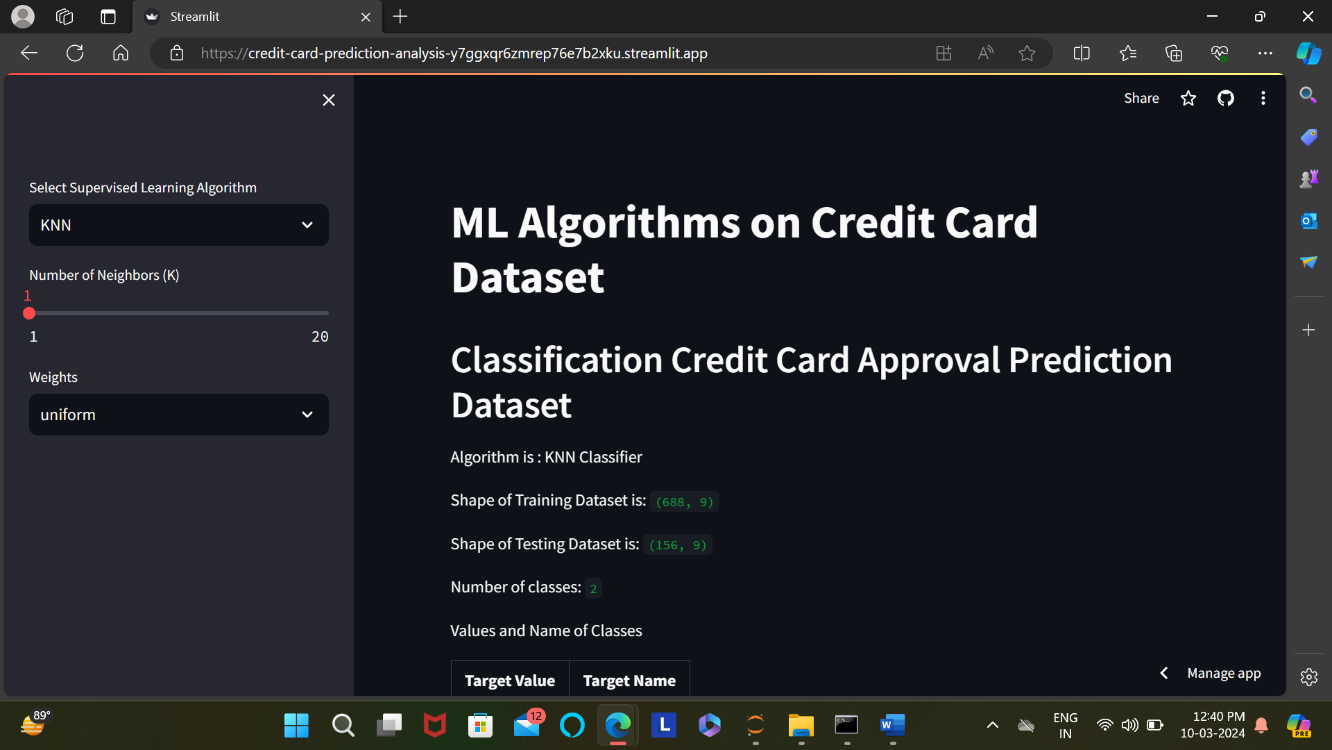
*Fig 4.1.1.1: Localhost Output 1*



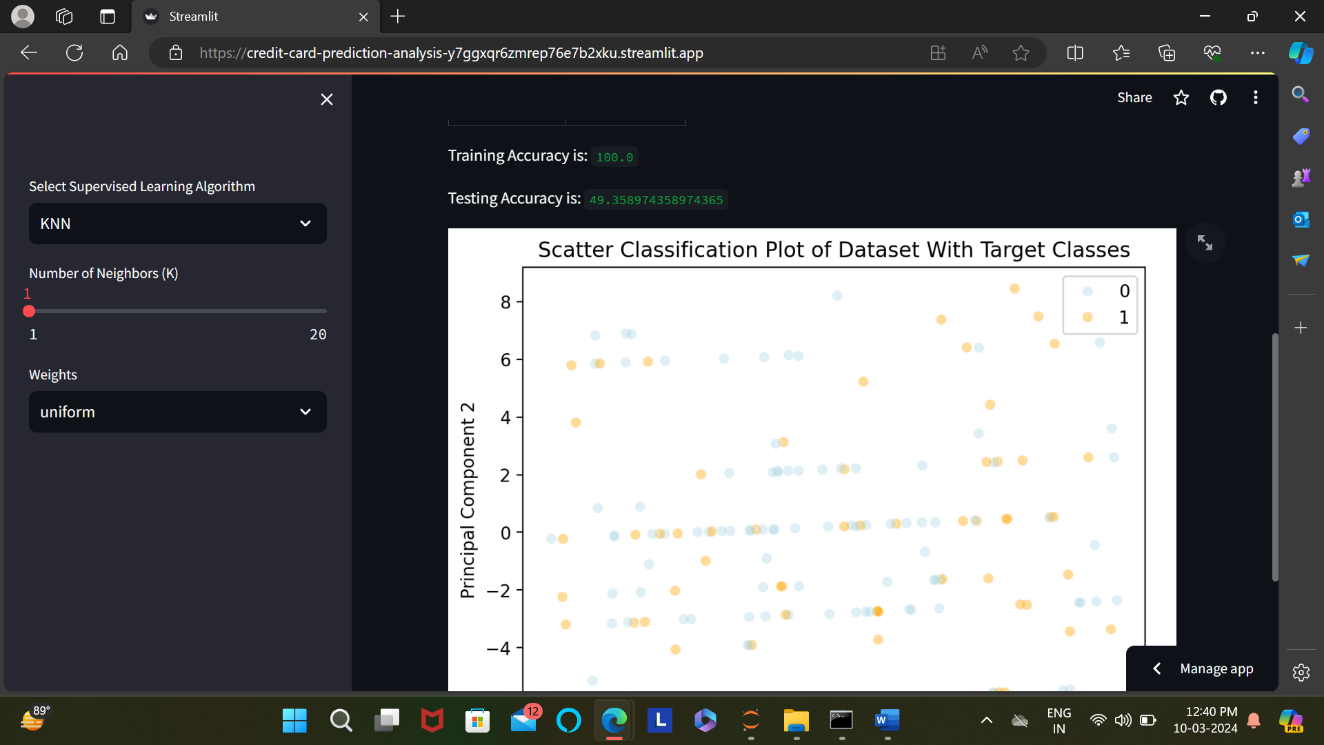
*Fig 4.1.1.2: Localhost Output 2*

**4.1.2 Streamlit Server:**

After visiting the link we can see the output as follows on the browser:



*Fig 4.1.2.1: Streamlit Output 1*



*Fig 4.1.2.2: Streamlit Output 2*

**4.2 Discussion:**

The dataset at hand provides a rich and varied set of attributes crucial for understanding and predicting credit risk in the context of financial decision-making. The inclusion of demographic information such as gender, marital status, and family size, coupled with financial indicators like income and property ownership, offers a holistic view of applicants' profiles.

The binary indicators for mobile phones, work phones, and emails provide insights into applicants' digital connectivity. Moreover, employment-related details, including status, length, and job title, contribute significantly to risk assessment. The dataset's comprehensive nature enables a nuanced exploration of factors that may influence an applicant's creditworthiness.

Analyzing the correlation between attributes like education level, dwelling type, and family size with the 'Is high risk' variable could unveil patterns and potential predictors of credit risk. Employing machine learning algorithms on this dataset has the potential to enhance credit risk prediction models, ultimately aiding financial institutions in making more informed and precise lending decisions.

The dataset's structured format and diverse variables create an excellent foundation for a thorough and insightful analysis of credit risk factors in the given financial context. Furthermore, the dataset's inclusion of temporal factors such as account age and employment length adds a dynamic dimension to the analysis.

These features provide insights into the stability and longevity of an applicant's financial standing, offering a glimpse into their creditworthiness over time. Exploring the relationship between account age and high-risk classification, for instance, may reveal trends indicating whether longer account histories correlate with lower credit risk.

In addition to its practical applications in credit risk assessment, this dataset also holds the potential for sociodemographic studies within the financial sector. By examining the interplay between variables like education level, marital status, and family size, researchers can gain valuable insights into the financial behaviors and preferences of diverse demographic groups. Such insights can contribute to the development of targeted financial products and services that cater to the unique needs and preferences of specific customer segments. Ultimately, this dataset offers a multifaceted lens through which to analyze credit risk and explore broader socioeconomic trends within the realm of financial decision-making.

***Chapter – 5***

***Future Enhancements***

As this project unfolds, several avenues for future enhancements emerge, promising to further elevate the effectiveness of the fraud detection system:

1. **Deep Learning Integration**: The incorporation of deep learning models, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), holds the potential to enhance the model's capacity to detect subtle patterns in transaction data, potentially yielding even higher accuracy.

2. **Anomaly Detection**: The exploration of anomaly detection techniques, including unsupervised learning approaches like Isolation Forests and One-Class SVMs, could prove invaluable in identifying unusual or suspicious transactions that may not adhere to typical patterns.

3. **Ethical Considerations**: With the increasing utilization of AI in finance, ethical concerns such as privacy and fairness need to be thoughtfully addressed. Future enhancements should incorporate robust measures to ensure compliance with privacy regulations and uphold model fairness.

The Credit Card Prediction project constitutes a proactive response to a pressing issue in the financial sector.

By harnessing advanced machine learning techniques and a meticulously curated dataset, this project is poised to make a substantial contribution to bolstering the security and trustworthiness of the financial system.

As it evolves, the integration of deep learning, anomaly detection, continuous monitoring, and ethical considerations will further enhance its capabilities and impact.

***Appendix***

**Objectives:**

The objectives of the Credit Card Fraud Detection model span various phases, from initial design to deployment:

* **Model Design:** Careful selection of machine learning algorithms and the creation of a robust fraud detection model.
* **Hyperparameter Tuning:** Extensive fine-tuning of model hyperparameters to maximize its effectiveness.
* **Real-Time Deployment:** Successful deployment of the model to ensure practical use.

**Technology Stack:**

The technology stack used in this project ensured a seamless user experience and efficient deployment:

* **Streamlit Framework:** Streamlit was chosen for the front-end user interface, guaranteeing a responsive and dynamic user experience.
* **Python:** Python served as the core programming language for model development, preprocessing, and deployment.

**Features:**

The “Credit Card Prediction” model incorporates an array of features to enhance its functionality and accuracy:

* **Algorithm Diversity:** Implementation of multiple machine learning algorithms, including logistic regression, random forests, and gradient boosting, to assess their performance.
* **Hyperparameter Tuning:** Rigorous tuning of hyperparameters to optimize each algorithm's fraud detection capabilities.
* **Parameter Sensitivity Analysis:** Exploration of how varying algorithm parameters impacts prediction detection accuracy and overall performance.

**Preprocessing:**

Preprocessing of the dataset was a pivotal step to ensure its compatibility with various machine learning models:

* **Data Cleaning:** Thorough data cleaning processes were executed to address missing values and outliers.
* **Feature Engineering:** The dataset underwent feature engineering (mainly Feature Extraction) to enhance its predictive power.
* **Scaling and Transformation:** Data scaling and transformation techniques were applied to normalize features and improve model performance.

**Testing:**

Testing played a crucial role in validating the model's reliability and accuracy:

* **Model Validation:** Comprehensive validation to assess the model's ability to detect both legitimate and fraudulent transactions.
* **Performance Metrics:** Utilization of performance metrics such as accuracy score for thorough evaluation.

**Future Enhancements:**

For future iterations of the Credit Card Fraud Detection model, several enhancements and features could be considered:

* **Advanced Algorithms:** Exploration of advanced machine learning and deep learning algorithms to further enhance fraud detection accuracy.
* **Ensemble Methods:** Implementation of ensemble learning techniques to combine the strengths of multiple algorithms.
* **Anomaly Detection**: The exploration of anomaly detection techniques, including unsupervised learning approaches like Isolation Forests and One-Class SVMs, could prove invaluable in identifying unusual or suspicious transactions that may not adhere to typical patterns.
* **Ethical Considerations**: With the increasing utilization of AI in finance, ethical concerns such as privacy and fairness need to be thoughtfully addressed.

The comprehensive appendix offers detailed insights into the Credit Card Fraud Detection project, covering objectives, technology stack, features, preprocessing, testing, and future enhancements. Overall, this project involves precise algorithm selection, hyperparameter tuning, Streamlit-based user interface, data preprocessing, rigorous testing, and ongoing improvements.

***Project Resources***

**Machine learning Process:** <https://www.datacamp.com/tutorial/tutorial-machine-learning-pipelines-mlops-deployment>

**Deployment:**

Credit Card Prediction App: You can access the deployed Credit Card Prediction application by clicking:

<https://credit-card-prediction-analysis-y7ggxqr6zmrep76e7b2xku.streamlit.app/>

**GitHub Repository:**

Codes related to this project can be seen at my GitHub Repository:

GitHub Repository: <https://github.com/madhurimarawat/Credit-Card-Prediction-Analysis>