Online Fraud Prediction Using ML

1. INTRODUCTION:

Fraud detection includes the entire process of institutions to identify fraudulent activities. These activities can be financial as well as other forms such as fraudulent credit card transactions, data theft, or cyberattacks.
Fraud detection is usually done with methods to predict abnormal behaviour, considering predetermined rule flows.
Fraud detection methods are also quite diverse in themselves.
Fraud detection products, which can be customized from the technology used to the workflows defined, to the industry in which they are used, are also developing rapidly to cope with the increasing number of fraud cases because of the rising digitalization.

1.1 Project Overview:

Fraud Detection Using Machine Learning deploys a machine learning (ML) model and an example dataset of credit card transactions to train the model to recognize fraud patterns. The model is self-learning which enables it to adapt to new, unknown fraud patterns.

Use this Guidance to automate the detection of potentially fraudulent activity, and the flagging of that activity for review. Fraud Detection Using Machine Learning is easy to deploy and includes an example dataset but you can modify the code to work with any dataset.

1.2 Purpose:

The main purpose of this project is to predict whether a person effected by fraud or not, based onmultiple factors.

We will be using classification algorithms such as Logistic Regression, KNN, Decision tree, Random forest, AdaBoost and GradientBoost. We will train and test the data with these algorithms. From this the best model is selected and saved in pkl format. We will also be deploying our model locally using Flask.

2. LITERATURE SURVEY

2.1 Existing probems:

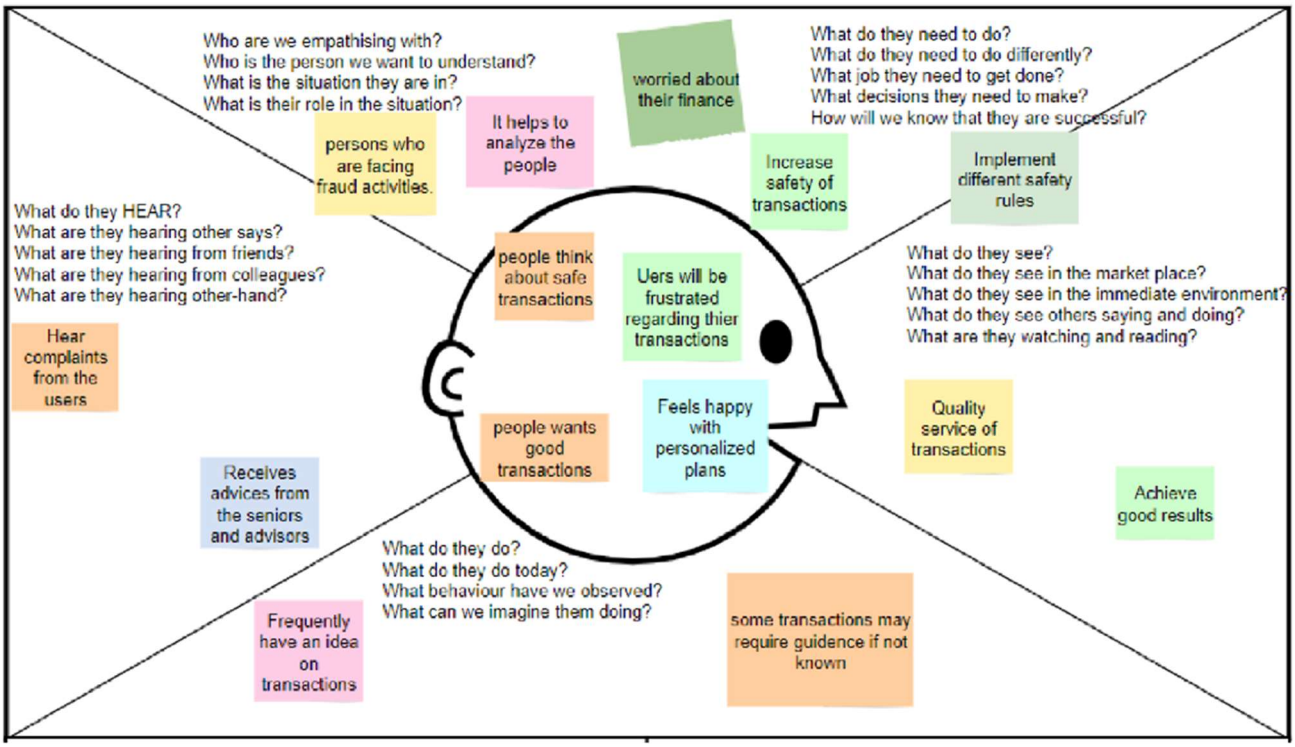
Changing fraud patterns over time is toughest to address since the fraudsters are always in the outlook to find new and innovative ways to get around the systems to commit to the act.

2.2 References:

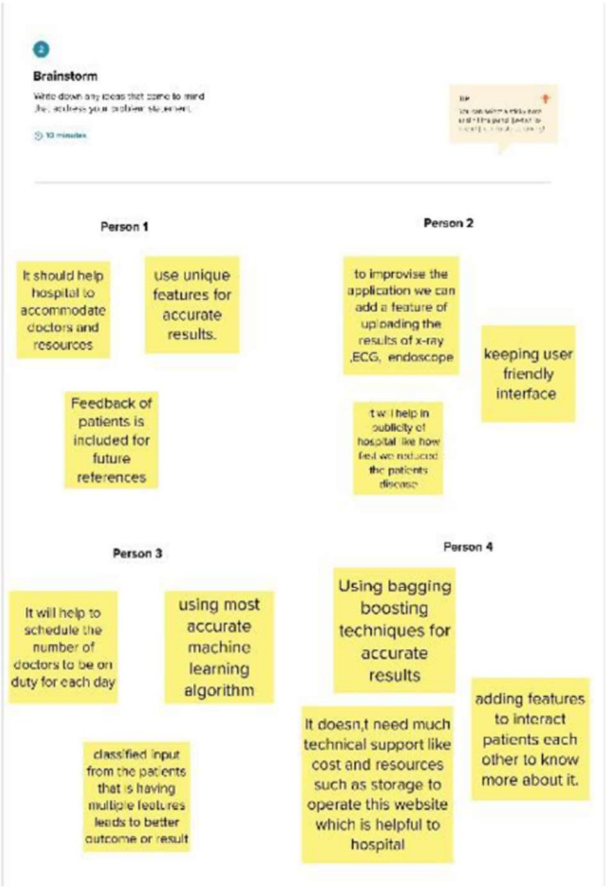
1. Omar,2.Raghavan

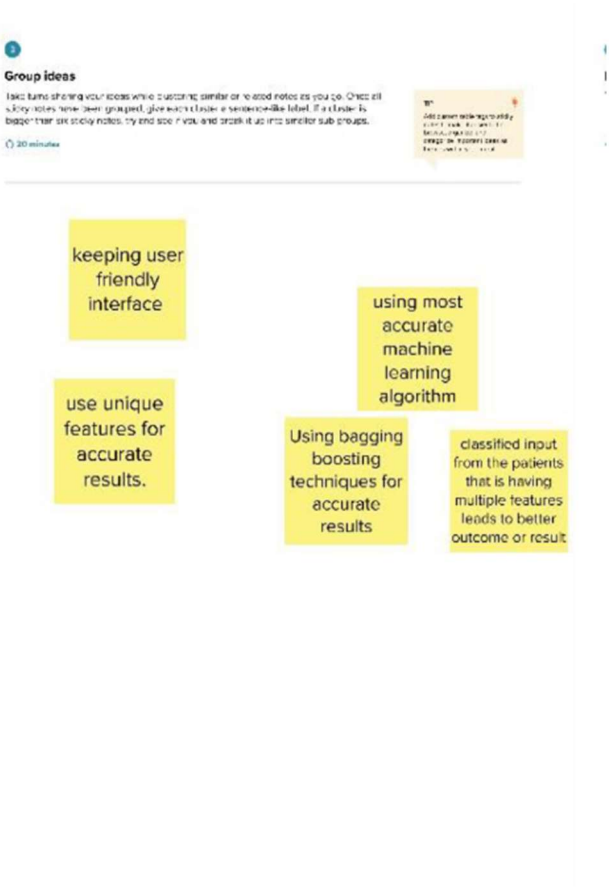
3.IDEATION AND PROPOSED SOLUTION

Empathy Map Canvas



Ideation and Brainstorming:





3. REQUIREMENT ANALYSIS

3.1 Functional Requirement:

- Anaconda navigator:

  Refer to the link below to download anaconda navigator

  Link: https://www.youtube.com/watch?v=1ra4zH2G4o0

- Python packages:
  - Open anaconda prompt as administartor.
  - Type "pip install pandas" and click enter
  - Type "pip install scikit-learn" and click enter
  - Type "pip install matplotlib" and click enter
  - Type "pip install scipy" and click enter
  - Type "pip install pickle-mixin" and click enter
  - Type "pip install seaborn" and click enter
  - Type "pip install Flask" and click enter

3.2 Non-Functional Requirements:

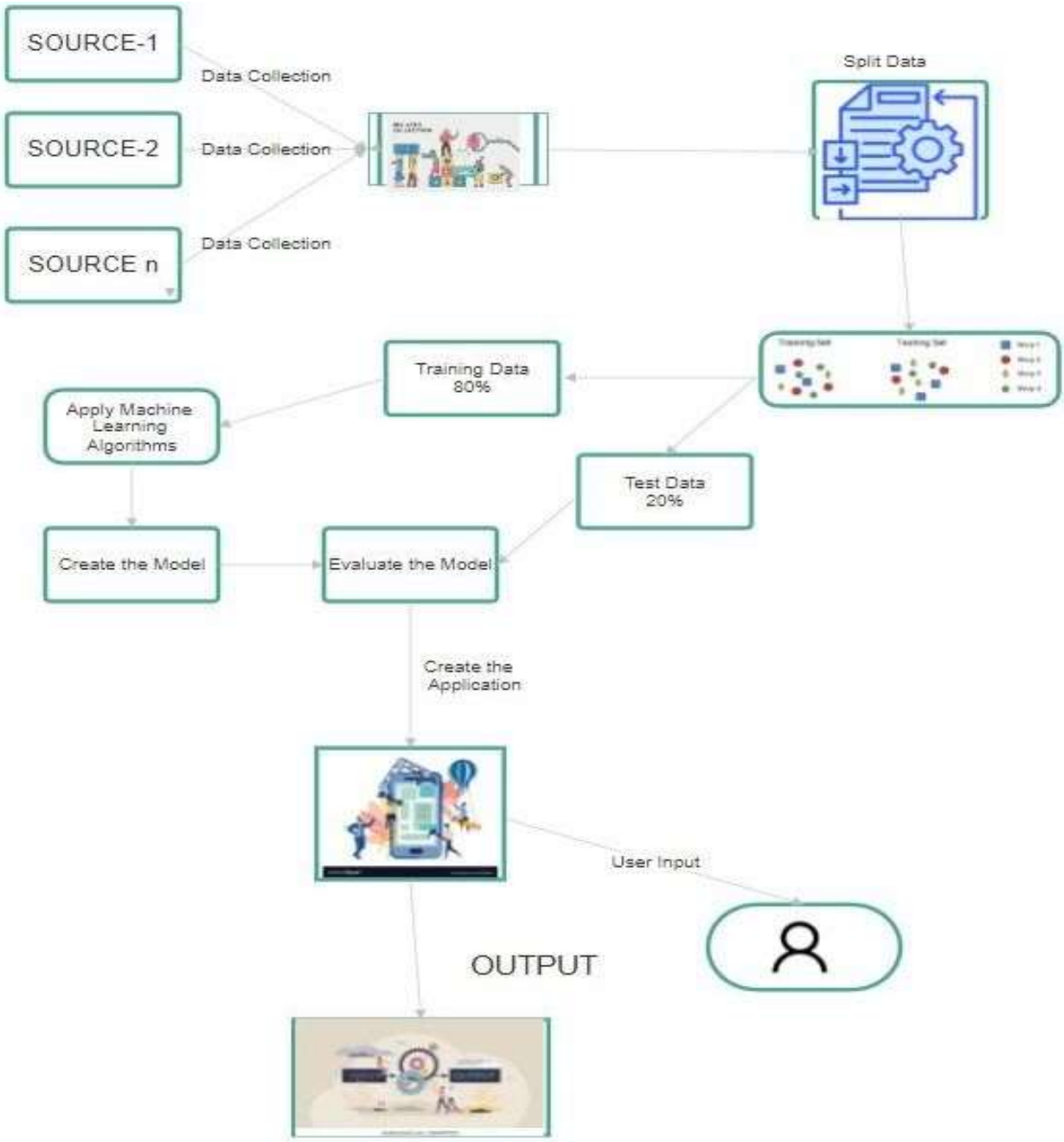
You must have prior knowledge of following topics to complete this project.

- ML Concepts
  - o Supervised learning: https://www.javatpoint.com/supervised-machine-learning
  - o Unsupervised learning:
    https://www.javatpoint.com/unsupervised-machine-learning
  - o Regression and classification
    - ▪ Logistic regression:
      https://www.javatpoint.com/logistic-regression-in-machine-learning
    - ▪ Decision tree:
      https://www.javatpoint.com/machine-learning-decision-tree-classificati
      on-algorithm
    - ▪ Random forest:
      https://www.javatpoint.com/machine-learning-random-forest-algorithm
    - ▪ KNN:
      https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine
      -learning
    - ▪ AdaBoost:
      https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-c
      omplete-guide-for-beginners/
    - ▪ Gradient Boost:
      https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algo
      rithm-a-complete-guide-for-beginners/
    - ▪ Evaluation metrics:
      https://www.analyticsvidhya.com/blog/2019/08/11-important-model-ev
      aluation-error-metrics/

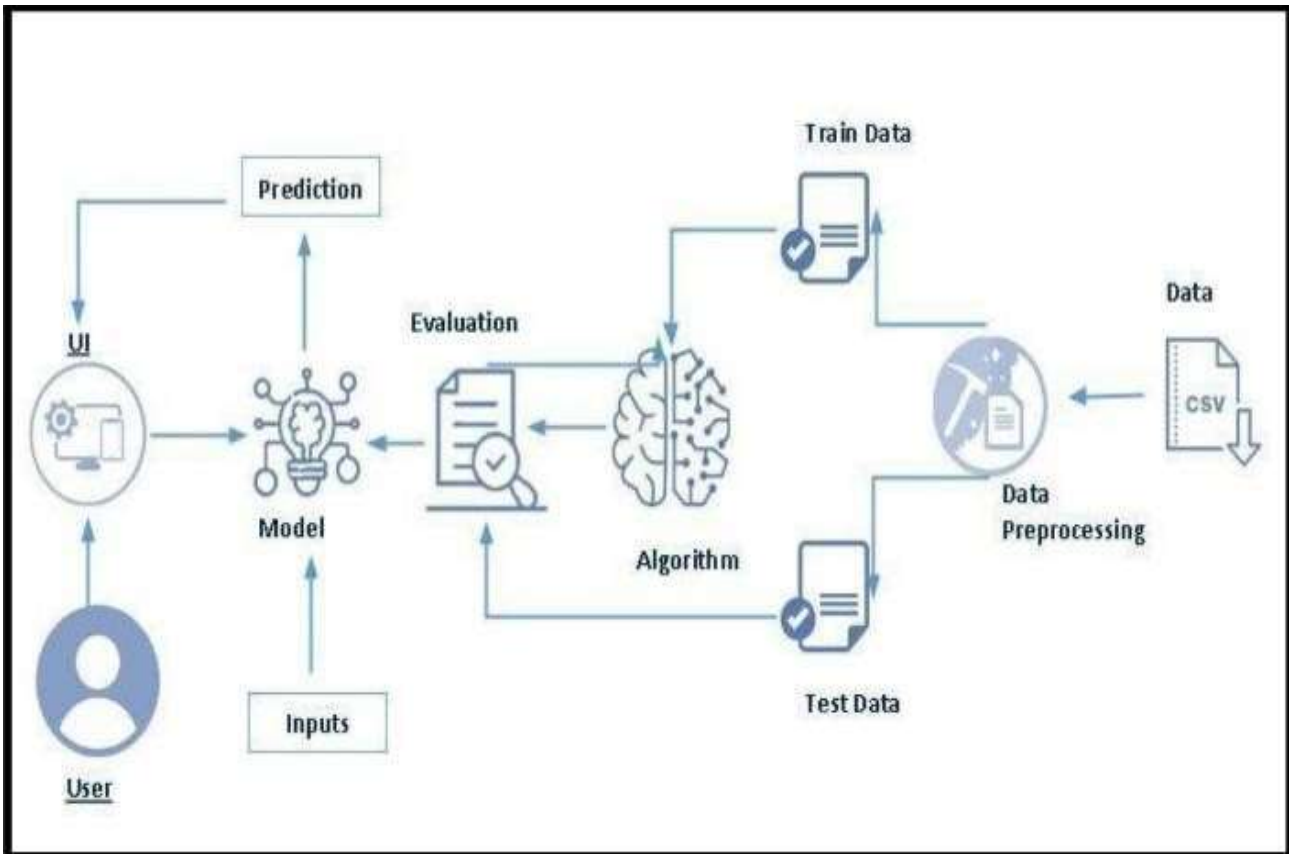
- Flask Basics : https://www.youtube.com/watch?v=lj4I_CvBnt0

4. PROJECT DESIGN

4.1 Data Flow Diagrams and User Stories:

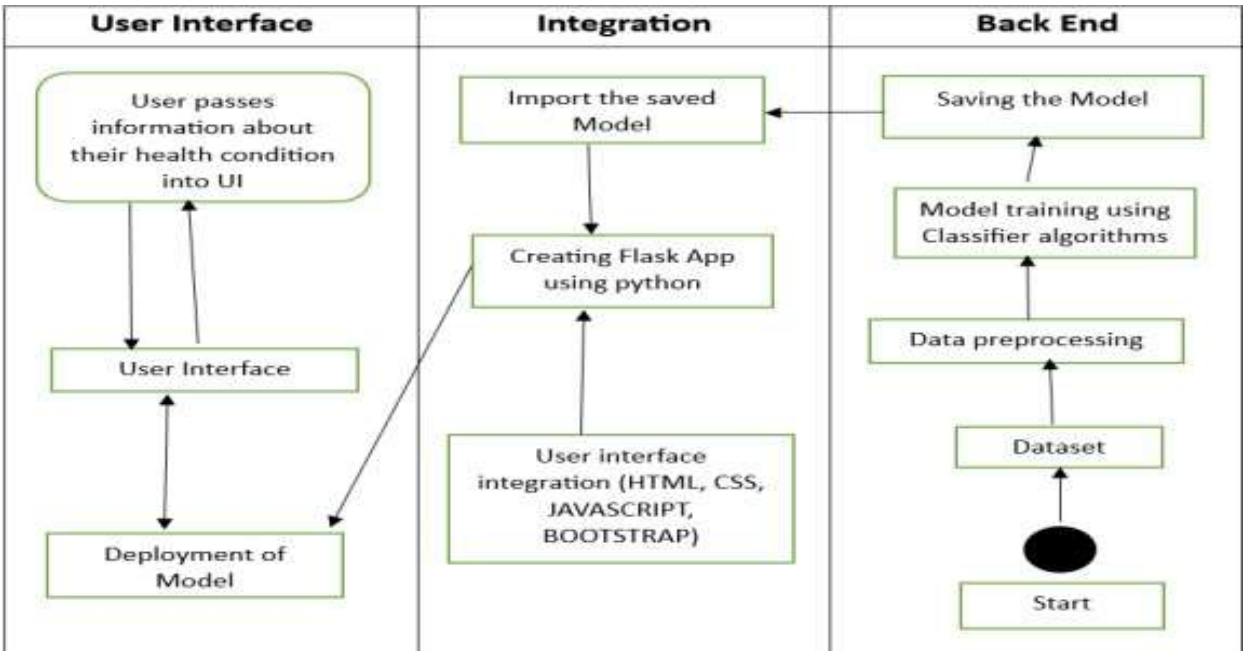


4.2 Solution Architecture:



5. PROJECT PLANNING AND ARCHITECTURE

5.1 Technical Architecture



5.2 Sprint Planning and Estimation:

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Project setup & Infrastructure	USN-1	Set up the environment with the requires tools and frameworks to start the hospital readmission prediction project.	2	High	Naren
Sprint-1	Development environment	USN-2	Make all necessary arrangements to complete the project.	1	Medium	Naren
Sprint-2	Data collection	USN-3	Gather a diverse dataset of readmissions containing different types of features for training the Machine learning model.	2	High	Vamsi
Sprint-3	Data preprocessing	USN-4	Preprocess the collected dataset by handling all types of null values, missing values and selecting correct features for predicting and selecting correct model.	2	High	Naren
Sprint-3	Model development	USN-5	Train the selected machine learning model using pre-processed dataset and monitor its performance on the validation set.	1	Medium	Vamsi
Sprint-4	Training	USN-6	Implement data augmentation techniques to improve the models robustness and accuracy.	2	High	Naren
Sprint-5	Model deployment & Integration	USN-7	Deploy the trained machine learning model as an API or web service to make it accessible for readmission prediction. Integrate the models API into user-friendly web interface for users to give input and predict .	1	Medium	Vamsi
Sprint-5	Testing & quality assurance	USN-8	Conduct thorough testing of the model and web interface to identify and report any issues or bugs. Optimize its performance based on user feedback and testing results	2	High	Naren

5.3 Sprint Delivery and Schedule:

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	3	4 Days	18 October 2023	21 Oct 2023	20	21 Oct 2023
Sprint-2	5	3 Days	22 October 2023	25 Oct 2023	20	25 Oct 2023
Sprint-3	10	7 Days	26 October 2023	2 Nov 2023	20	2 Nov 2023
Sprint-4	1	3 Days	3 November 2023	6 Nov 2023	20	6 Nov 2023
Sprint-5	1	2 Days	7 November 2023	9 Nov 2023	20	9 Nov 2023

6. CODING AND SOLUTIONING

6.1 Feature-1:

We have trained our model with 29 features. But all these features may not be important for

prediction. Hence we will select the features that contribute significantly to the model

performance.

Below is the description of imp_cols:

- discharge_disposition_id : Integer identifier corresponding to 29 distinct values, for

example, discharged to home, expired, and not available

- admission_source_id : Integer identifier corresponding to 21 distinct values, for

example, physician referral, emergency room, and transfer from a hospital

- time_in_hospital : Integer number of days between admission and discharge
- num_medications : Number of distinct generic names administered during the

encounter

● number_emergency : Number of emergency visits of the patient in the year preceding the encounter

● number_inpatient : Number of inpatient visits of the patient in the year preceding the encounter

● diag_1 : The primary diagnosis (coded as first three digits of ICD9); 848 distinct values

● diag_2 : The secondary diagnosis (coded as first three digits of ICD9); 923 distinct values

● max_glu_serum : Indicates the range of the result or if the test was not taken. Values:

">200," ">300," "normal," and "none" if not measured

● glimepiride : glimepiride dosage - Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed

● diabetesMed : Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"

7. Performance Testing

7.1 Performance Metrics

We will compare the confusion matrix, ROC curve and classification report for both models.

In order to obtain these, we will be using the confusion_matrix(),roc_curve() and classification_report() functions from sklearn.metrics.

In [26]:
```
print("Logistic Regression Accuracy Score:", accuracy_score(y_test,y_pred))
print("Decision Tree Accuracy Score: ", accuracy_score(y_test,y_pred_dt))
print("Random Forest accuracy score: ", accuracy_score(y_test,y_pred_rf))
print("Gradient Boosting accuracy score: ", accuracy_score(y_test, y_pred_gb))
```

```
Logistic Regression Accuracy Score: 0.8992214532871973
Decision Tree Accuracy Score:  0.9834775086505191
Random Forest accuracy score:  0.9901384083044983
Gradient Boosting accuracy score:  0.9873702422145328
```

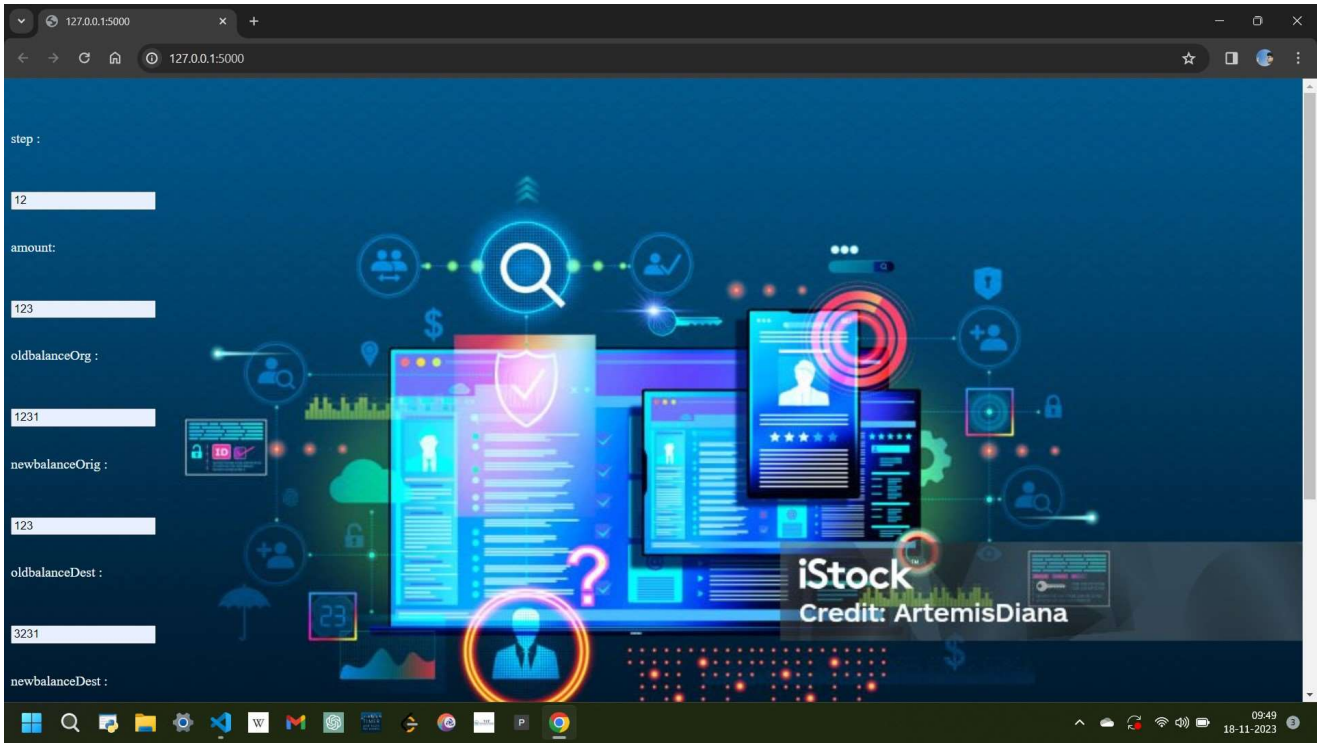
```
print("Random Forest classification report: \n\n"  ,classification_report(y_test,y_pred_rf))
```

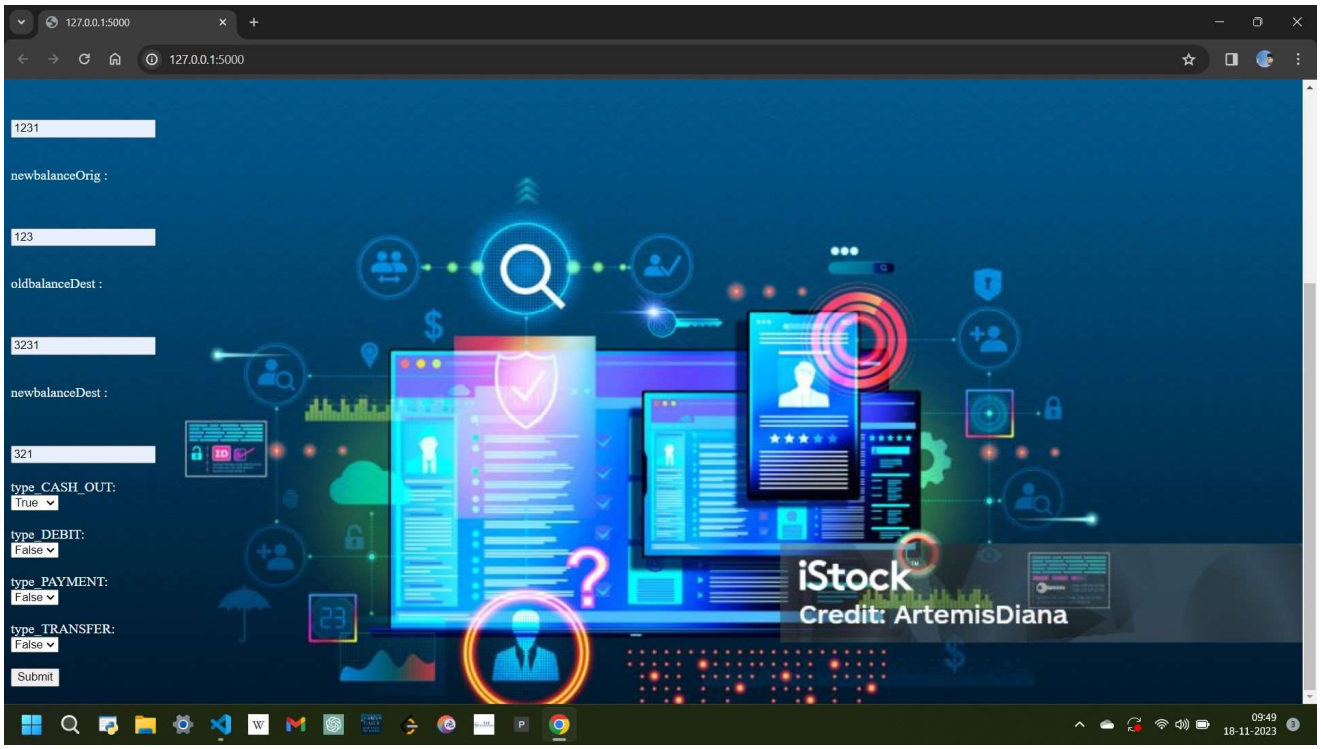
```
Random Forest classification report:

              precision    recall  f1-score   support

           0       0.99      0.99      0.99      5779
           1       0.99      0.99      0.99      5781

    accuracy                           0.99     11560
   macro avg       0.99      0.99      0.99     11560
weighted avg       0.99      0.99      0.99     11560
```

8. Results

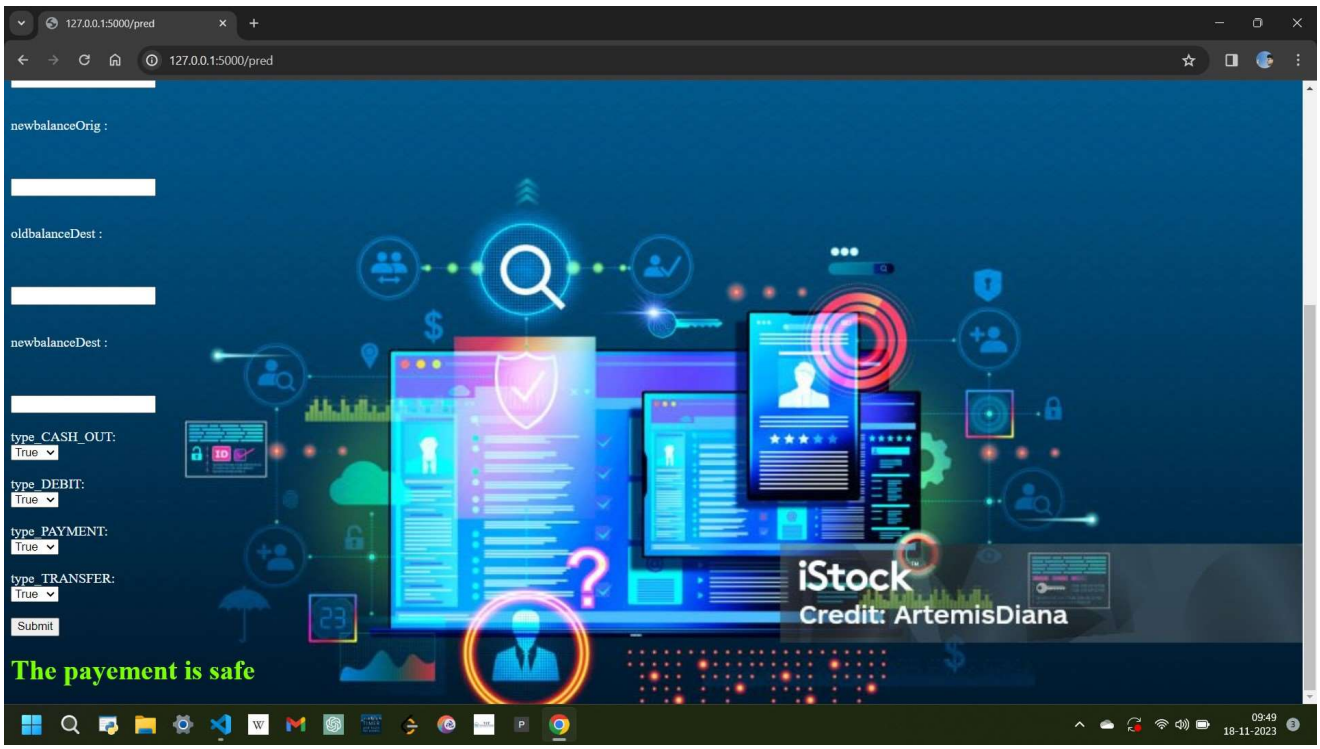
8.1 Output Screenshots

Lets see how our page looks like:


127.0.0.1:5000
127.0.0.1:5000
step :
12
amount:
123
oldbalanceOrg :
1231
newbalanceOrig :
123
oldbalanceDest :
3231
newbalanceDest :
iStock
Credit: ArtemisDiana
09:49
18-11-2023


127.0.0.1:5000
127.0.0.1:5000
1231
newbalanceOrig :
123
oldbalanceDest :
3231
newbalanceDest :
321
type_CASH_OUT:
True
type_DEBIT:
False
type_PAYMENT:
False
type_TRANSFER:
False
Submit
iStock
Credit: ArtemisDiana
09:49
18-11-2023

OUTPUT:


127.0.0.1:5000/pred
127.0.0.1:5000/pred
newbalanceOrig :
oldbalanceDest :
newbalanceDest :
type_CASH_OUT:
True
type_DEBIT:
True
type_PAYMENT:
True
type_TRANSFER:
True
Submit
The payement is safe
iStock
Credit: ArtemisDiana
09:49
18-11-2023

9. Advantages and Disadvantages

Fraud prediction has both advantages and disadvantages. Here aresome of them:

Advantages:

Resource Optimization

Cost Reduction

Predicting fraud can lead to cost savings for user. It allows for proactive interventions and preventive measures to reduce the likelihood of readmission.

Data Quality and Integration:

The accuracy of predictions relies heavily on the quality of the data and its integration from various sources. Incomplete or inaccurate data may compromise the reliability of predictions.

10. Conclusion

Through this project, we created a machine learning model that is able to predict the fraud. The best model was a gradient boosting classifier with optimized hyperparameters. Future scope

Hospital readmission prediction is a field that uses machine learning and data analysis to identify patients who are at high risk of being readmitted to the hospital within a certain time frame after discharge. This can help improve the quality of care, reduce costs, and prevent unnecessary hospitalizations.

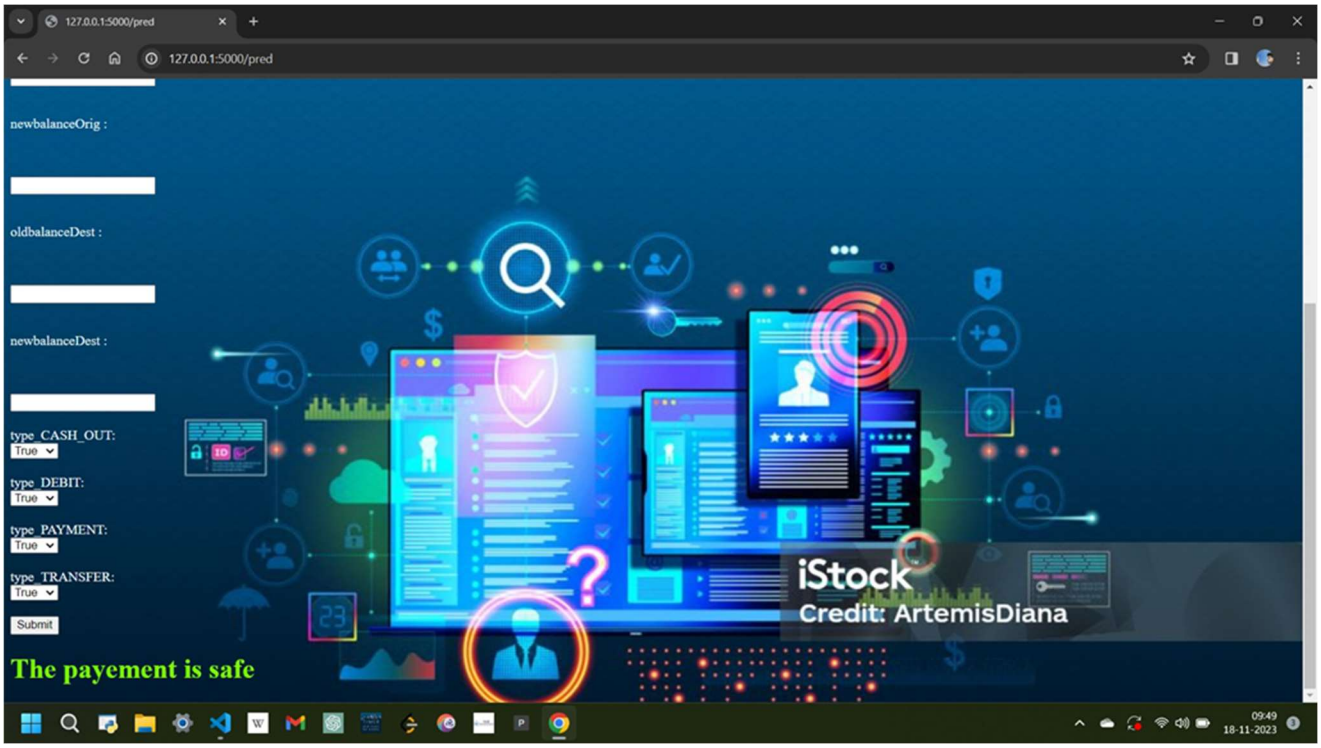
11. Appendix

Source code

All the source code and dataset in kept in the below provided Drive link. Please see the below link.

GitHub & Project Demo link

Git


127.0.0.1:5000/pred
127.0.0.1:5000/pred
newbalanceOrig :
oldbalanceDest :
newbalanceDest :
type_CASH_OUT:
True
type_DEBIT:
True
type_PAYMENT:
True
type_TRANSFER:
True
Submit
The payement is safe
iStock
Credit: ArtemisDiana
09:49
18-11-2023