Hospital Readmission Prediction Using ML

1.INTRODUCTION:

As the healthcare system moves toward value-based care, CMS has created many programs to improve the quality of care of patients. One of these programs is called the Hospital Readmission Reduction Program (HRPP), which reduces reimbursement to hospitals with above average readmissions. For those hospitals which are currently penalized under this program, one solution is to create interventions to provide additional assistance to patients with increased risk of readmission. But how do we identify these patients? We can use predictive modeling from data science to help prioritize patients.

1.1 Project Overview:

Diabetes is a serious metabolic disorder that affects millions of people globally. Early detection and management of diabetes are essential to prevent severe complications. In recent years, machine learning algorithms have become increasingly popular in the medical field to predict the onset of diabetes. This study aims to predict the onset of diabetes using the support vector machine (SVM) and decision tree algorithms.

One patient population that is at increased risk of hospitalisation and readmission is that of diabetes. Diabetes is a medical condition that affects approximately 1 in 10 patients in the United States. So in this project, we will be focusing on hospital readmission prediction for patients who are having diabetes.

This study used the Health Facts database (Cerner Corporation, Kansas City, MO), a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States. The Health Facts data we used was an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States.

1.2 Purpose:

The main purpose of this project is to predict whether a person who is suffering with diabetes and consulting a specific hospital will be readmitted or not, based on multiple 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:

Readmission to the hospital is an undesirable outcome.Thus, there is widespread interest in reducing readmission risk to improve both the patient health and control costs. It has been established that disbetes is an independent risk factor for readmission.

Poor communications, gaps in follow-up care, discharging patients with pending test results and inadequate patient education and discharge instructions have impacted the admission rate.

2.2 References:

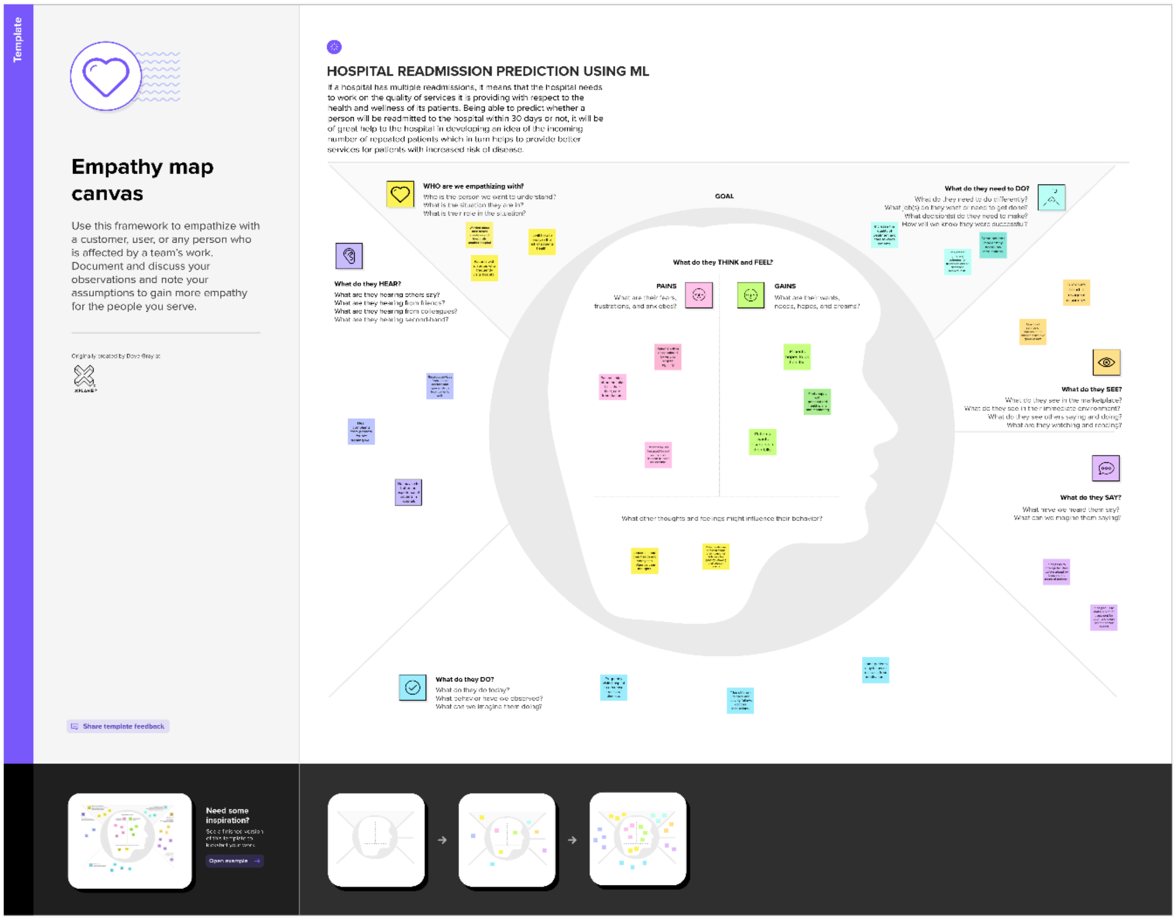
Ostling, Wyckoff, Ciarkowski, Pai, Choe, Bahl, Gianchandani (2017). “The relationship between diabetes mellitus and 30-day readmission rates” in Clinical Diabetes and Endocrinology. 3:1

2.3 Problem Statement Definition:

If a hospital has multiple readmissions, it means that the hospital needs to work on the quality of services it is providing with respect to the health and wellness of its patients. Being able to predict whether a person will be readmitted to the hospital within 30 days or not, it will be of great help to the hospital in developing an idea of the incoming number of repeated patients which in turn helps to provide better services for patients with increased risk of disease.

3.IDEATION AND PROPOSED SOLUTION

3.1 Empathy Map Canvas



Template

Empathy map canvas

Use this framework to empathize with a customer, user, or any person who is affected by a team's work. Document and discuss your observations and note your assumptions to gain more empathy for the people you serve.

Originally created by Dave Gray at

HOSPITAL READMISSION PREDICTION USING ML

If a hospital has multiple readmissions, it means that the hospital needs to work on the quality of services it is providing with respect to the health and wellness of its patients. Being able to predict whether a person will be readmitted to the hospital within 30 days or not, it will be of great help to the hospital in developing an idea of the incoming number of repeated patients which in turn helps to provide better services for patients with increased risk of disease.

WHO are we empathizing with?
Who is the person we want to understand?
What is the situation they are in?
What is their role in the situation?

GOAL

What do they need to DO?
What do they need to do differently?
What job(s) do they want or need to get done?
What decision(s) do they need to make?
How will we know they were successful?

What do they HEAR?
What are they hearing others say?
What are they hearing from friends?
What are they hearing from colleagues?
What are they hearing second-hand?

What do they THINK and FEEL?

PAINS
What are their fears, frustrations, and anxieties?

GAINS
What are their wants, needs, hopes, and dreams?

What do they SEE?
What do they see in the marketplace?
What do they see in their immediate environment?
What do they see others saying and doing?
What are they watching and reading?

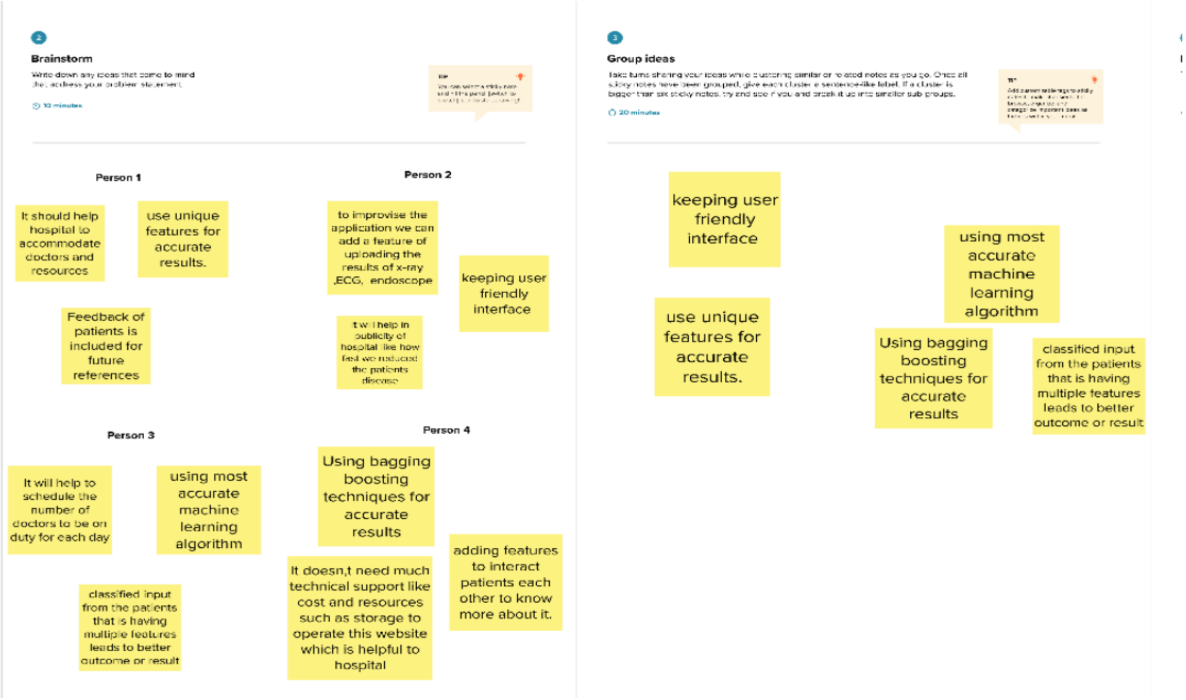
What do they SAY?
What have we heard them say?
What can we imagine them saying?

What other thoughts and feelings might influence their behavior?

What do they DO?
What do they do today?
What behavior have we observed?
What can we imagine them doing?

Need some inspiration?
See a finished version of this template to kickstart your work.

3.2 Ideation and Brainstorming:



2

Brainstorm

Write down any ideas that come to mind that address your problem statement.

10 minutes

Person 1

It should help hospital to accommodate doctors and resources

use unique features for accurate results.

Feedback of patients is included for future references

Person 2

to improvise the application we can add a feature of uploading the results of x-ray ,ECG, endoscope

keeping user friendly interface

it will help in publicity of hospital like how fast we reduced the patients disease

Person 3

It will help to schedule the number of doctors to be on duty for each day

using most accurate machine learning algorithm

classified input from the patients that is having multiple features leads to better outcome or result

Person 4

Using bagging boosting techniques for accurate results

It doesn,t need much technical support like cost and resources such as storage to operate this website which is helpful to hospital

adding features to interact patients each other to know more about it.

3

Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you and break it up into smaller sub groups.

20 minutes

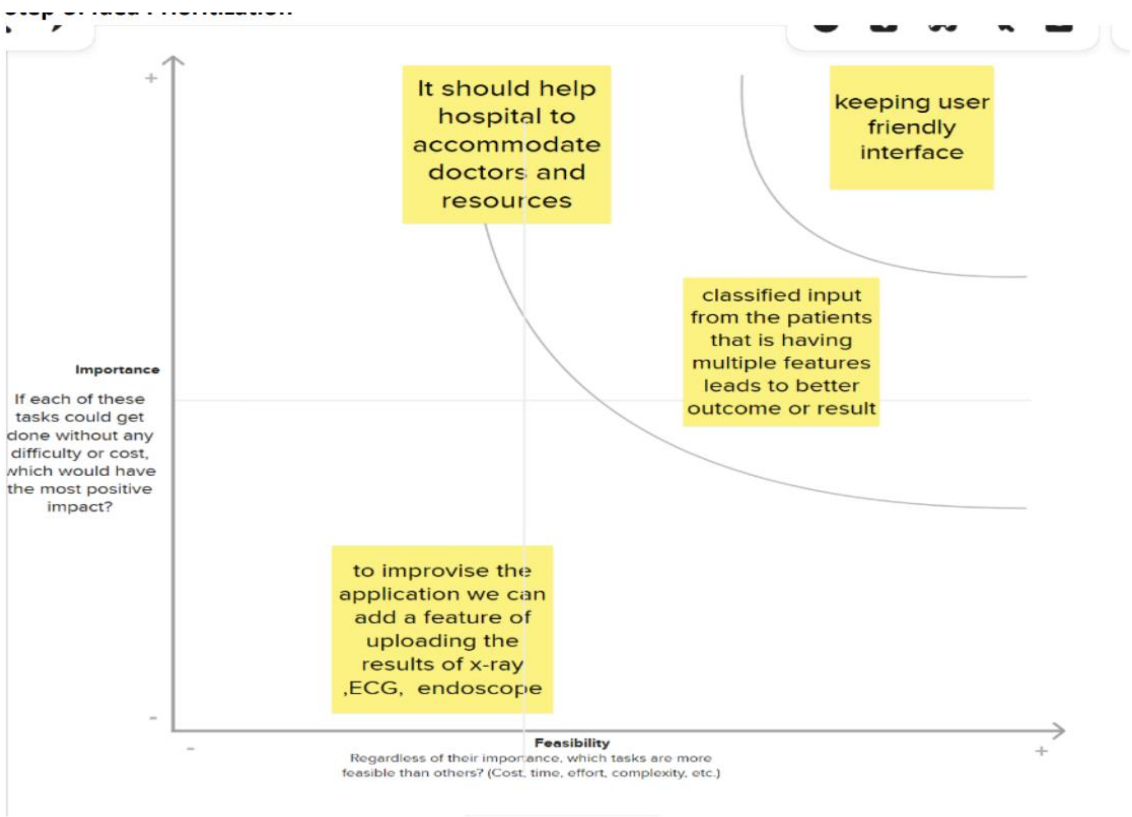
keeping user friendly interface

use unique features for accurate results.

using most accurate machine learning algorithm

Using bagging boosting techniques for accurate results

classified input from the patients that is having multiple features leads to better outcome or result



It should help hospital to accommodate doctors and resources

keeping user friendly interface

classified input from the patients that is having multiple features leads to better outcome or result

to improve the application we can add a feature of uploading the results of x-ray ,ECG, endoscope

Importance

If each of these tasks could get done without any difficulty or cost, which would have the most positive impact?

Feasibility

Regardless of their importance, which tasks are more feasible than others? (Cost, time, effort, complexity, etc.)

4.REQUIREMENT ANALYSIS

4.1Functional 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

4.2 Non-Functional Requirements:

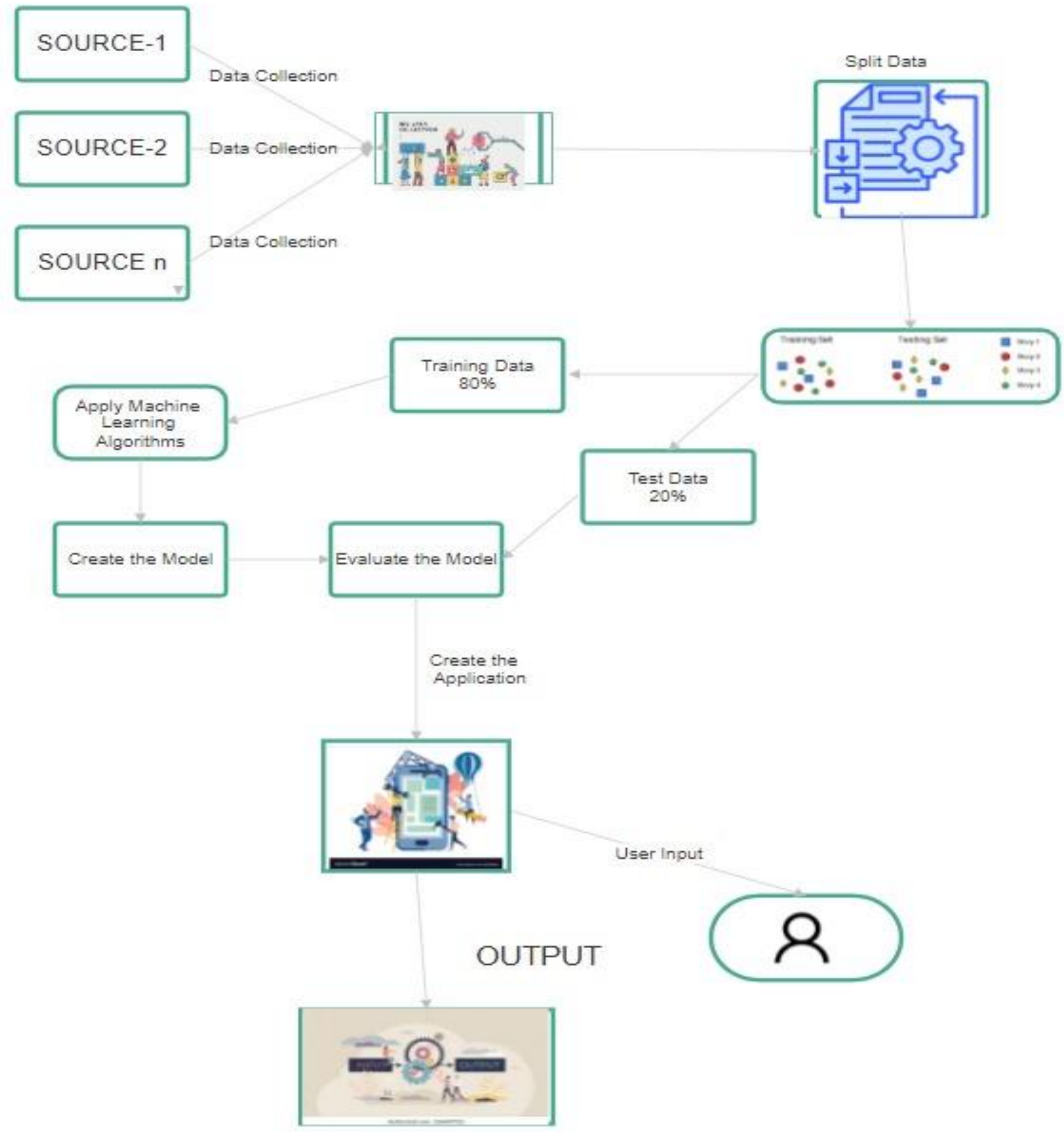
You must have prior knowledge of following topics to complete this project.

- ML Concepts
  - Supervised learning: https://www.javatpoint.com/supervised-machine-learning
  - Unsupervised learning: https://www.javatpoint.com/unsupervised-machine-learning
  - Regression and classification
    - Logistic regression: https://www.javatpoint.com/logistic-regression-in-machine-learning
    - Decision tree: https://www.javatpoint.com/machine-learning-decision-tree-classification-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-complete-guide-for-beginners/
    - Gradient Boost: https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/
    - Evaluation metrics: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

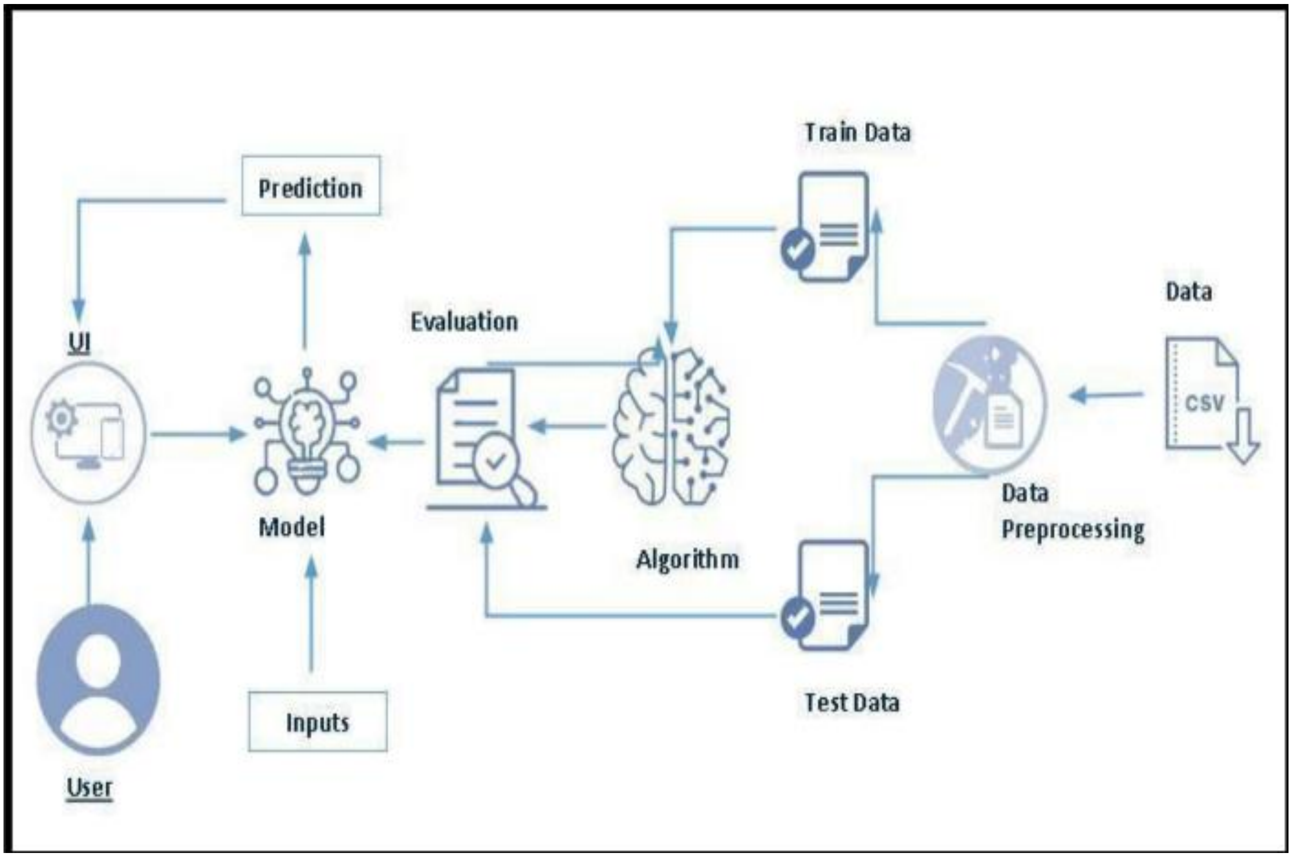
- Flask Basics : https://www.youtube.com/watch?v=lj4I_CvBnt0

5.PROJECT DESIGN

5.1 Data Flow Diagrams and User Stories:

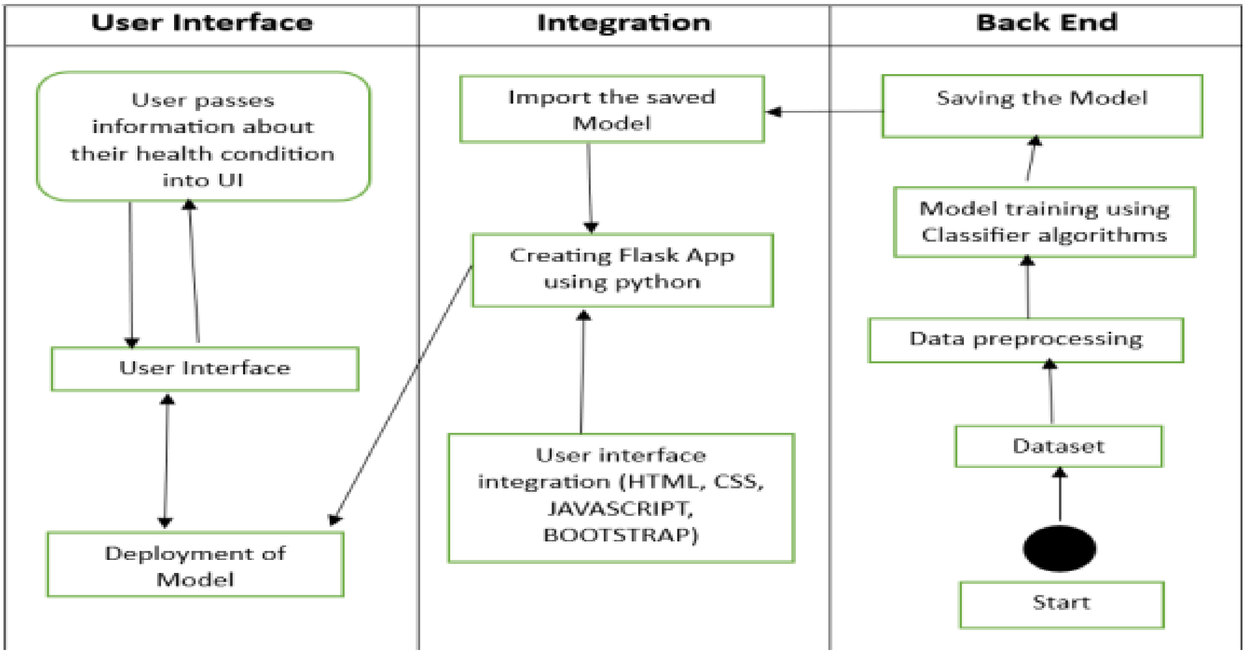


5.2 Solution Architecture:



6.PROJECT PLANNING AND ARCHITECTURE

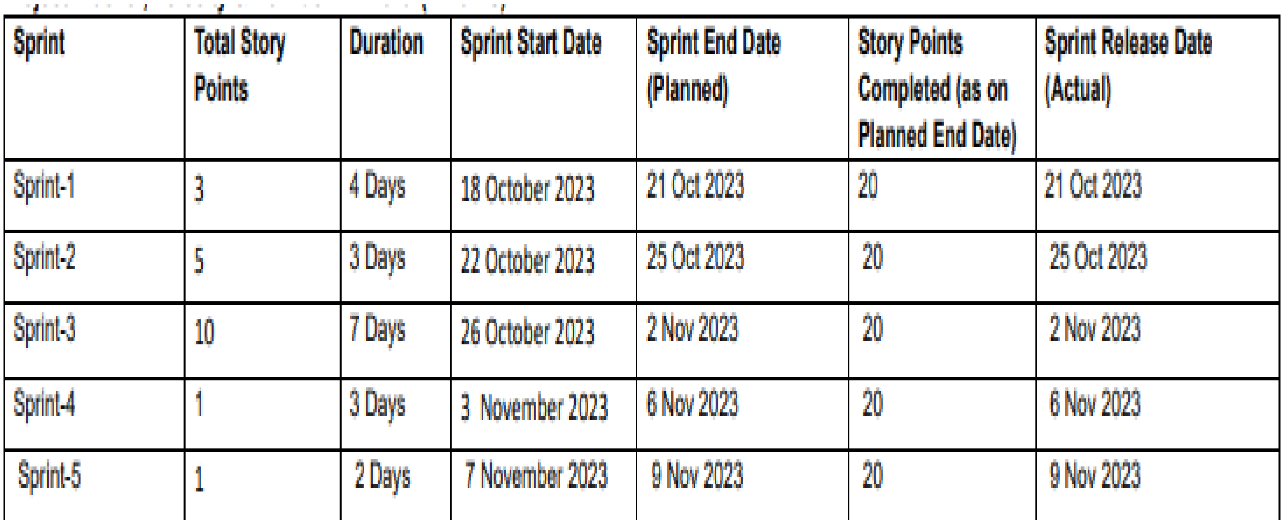
6.1 Technical Architecture


User Interface
Integration
Back End
User passes information about their health condition into UI
User Interface
Deployment of Model
Import the saved Model
Creating Flask App using python
User interface integration (HTML, CSS, JAVASCRIPT, BOOTSTRAP)
Saving the Model
Model training using Classifier algorithms
Data preprocessing
Dataset
Start

6.2 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	V. Sukumar
Sprint-1	Development environment	USN-2	Make all necessary arrangements to complete the project.	1	Medium	V. Sukumar
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	K. Lakshmi Prasanna
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	K. Lakshmi Prasanna, V. Sukumar
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	M. Sumanth
Sprint-4	Training	USN-6	Implement data augmentation techniques to improve the models robustness and accuracy.	2	High	M. Sumanth
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	V. Saikrupa Anjali
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	V. Saikrupa Anjali

6.3 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

7.CODING AND SOLUTIONING

7.1 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:

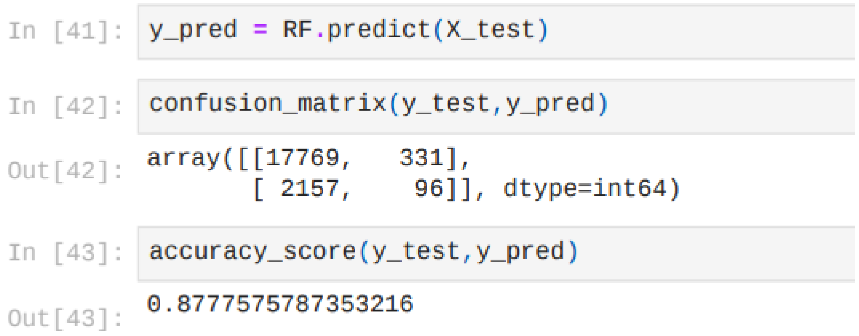
“yes” and “no”

8.Performance Testing

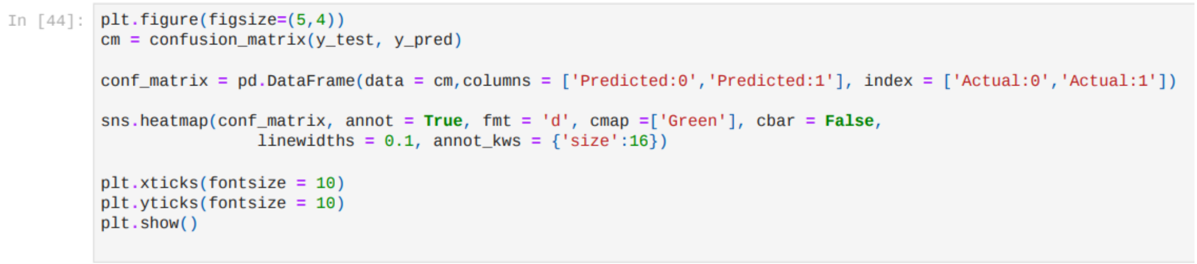
8.1 Performance Metrics

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

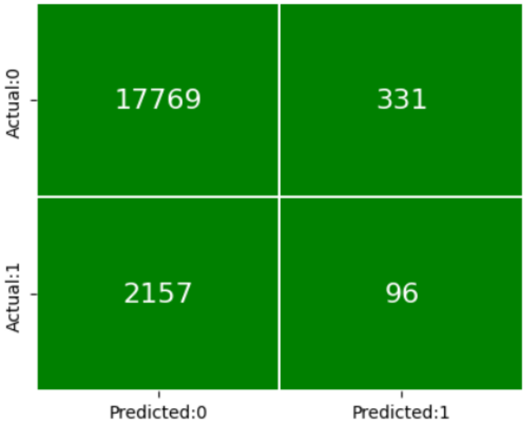
In order to obtain these, we will be using the confusion_matrix(),roc_curve() and classification_report() functions from sklearn.metrics.



```
In [41]: y_pred = RF.predict(X_test)

In [42]: confusion_matrix(y_test,y_pred)

Out[42]: array([[17769,   331],
                [ 2157,    96]], dtype=int64)

In [43]: accuracy_score(y_test,y_pred)

Out[43]: 0.8777575787353216
```



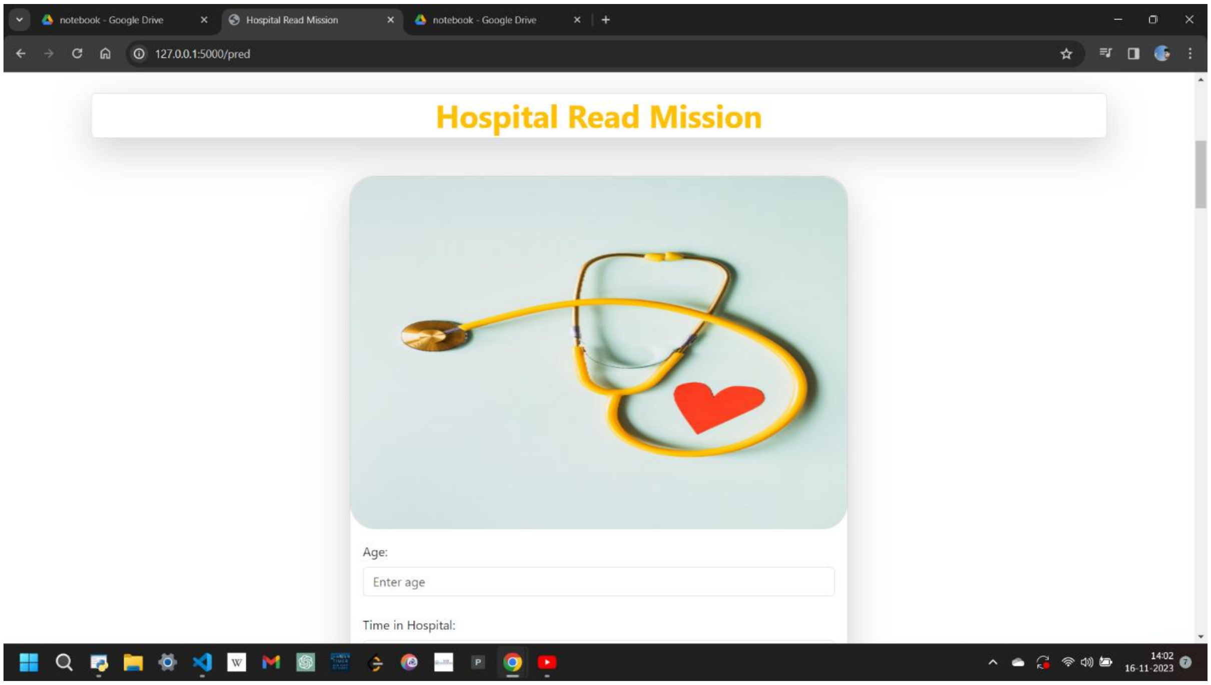
```
In [44]: plt.figure(figsize=(5,4))
         cm = confusion_matrix(y_test, y_pred)

         conf_matrix = pd.DataFrame(data = cm,columns = ['Predicted:0','Predicted:1'], index = ['Actual:0','Actual:1'])

         sns.heatmap(conf_matrix, annot = True, fmt = 'd', cmap =['Green'], cbar = False,
                     linewidths = 0.1, annot_kws = {'size':16})

         plt.xticks(fontsize = 10)
         plt.yticks(fontsize = 10)
         plt.show()
```



9. Results

9.1 Output Screenshots

Lets see how our page looks like:

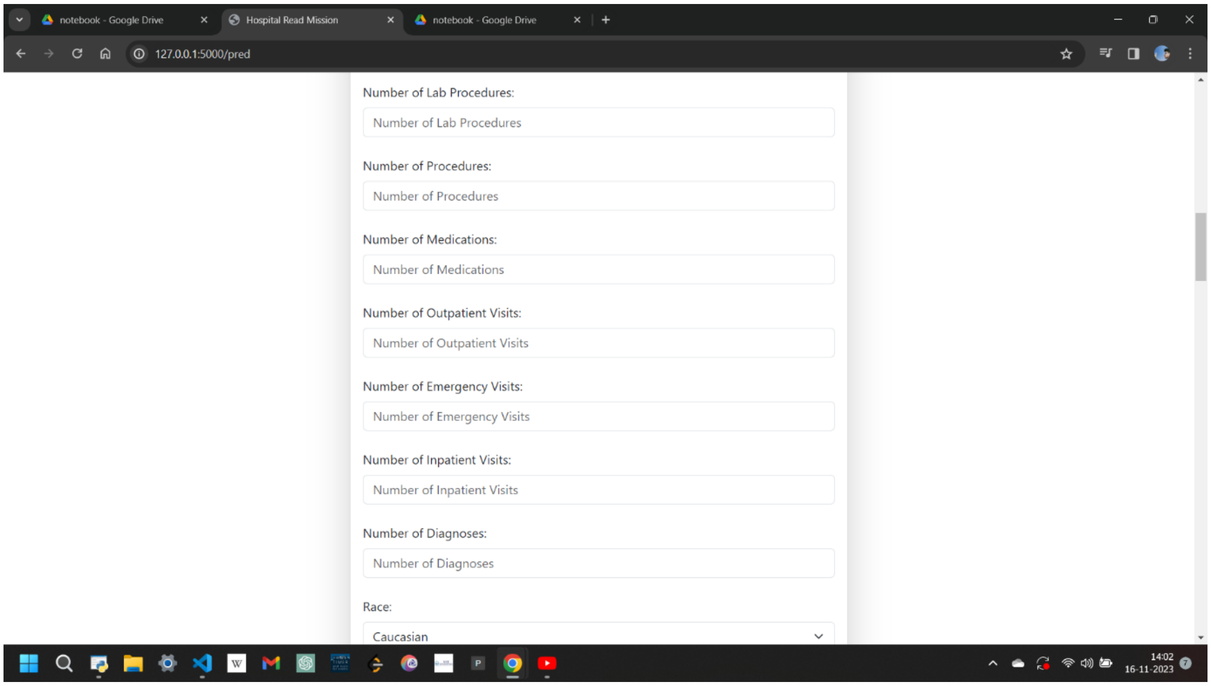


Hospital Read Mission

Age:

Enter age

Time in Hospital:



Number of Lab Procedures:

Number of Lab Procedures

Number of Procedures:

Number of Procedures

Number of Medications:

Number of Medications

Number of Outpatient Visits:

Number of Outpatient Visits

Number of Emergency Visits:

Number of Emergency Visits

Number of Inpatient Visits:

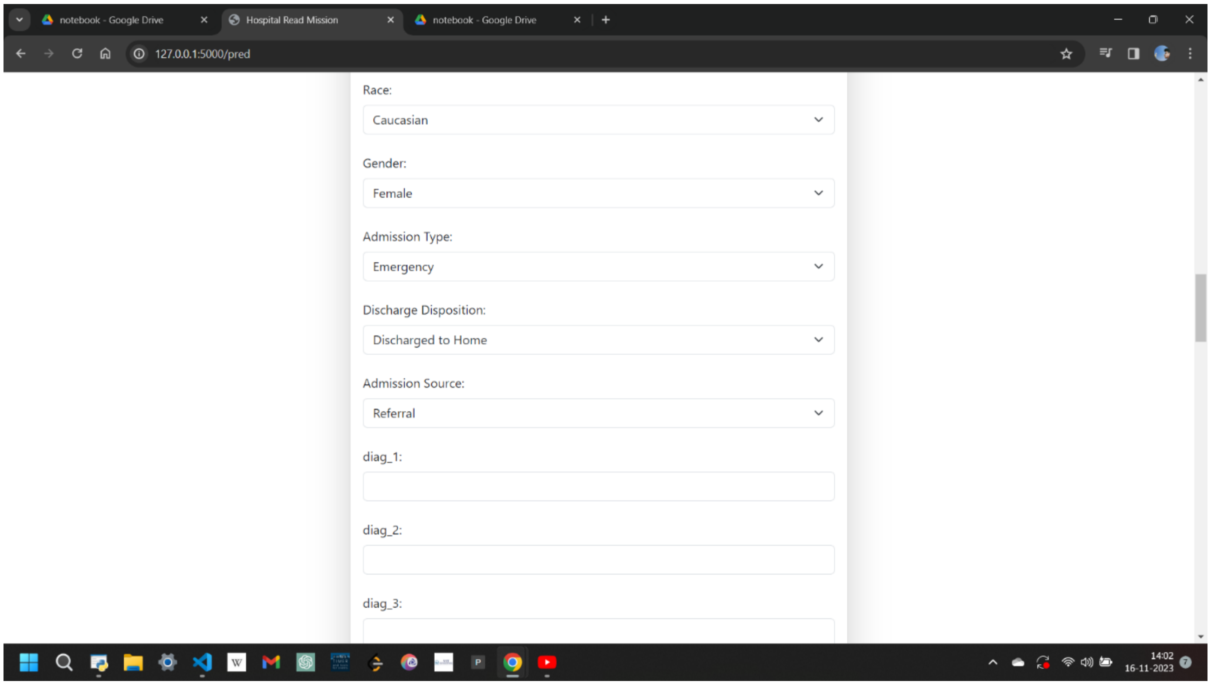
Number of Inpatient Visits

Number of Diagnoses:

Number of Diagnoses

Race:

Caucasian



Race:
Caucasian

Gender:
Female

Admission Type:
Emergency

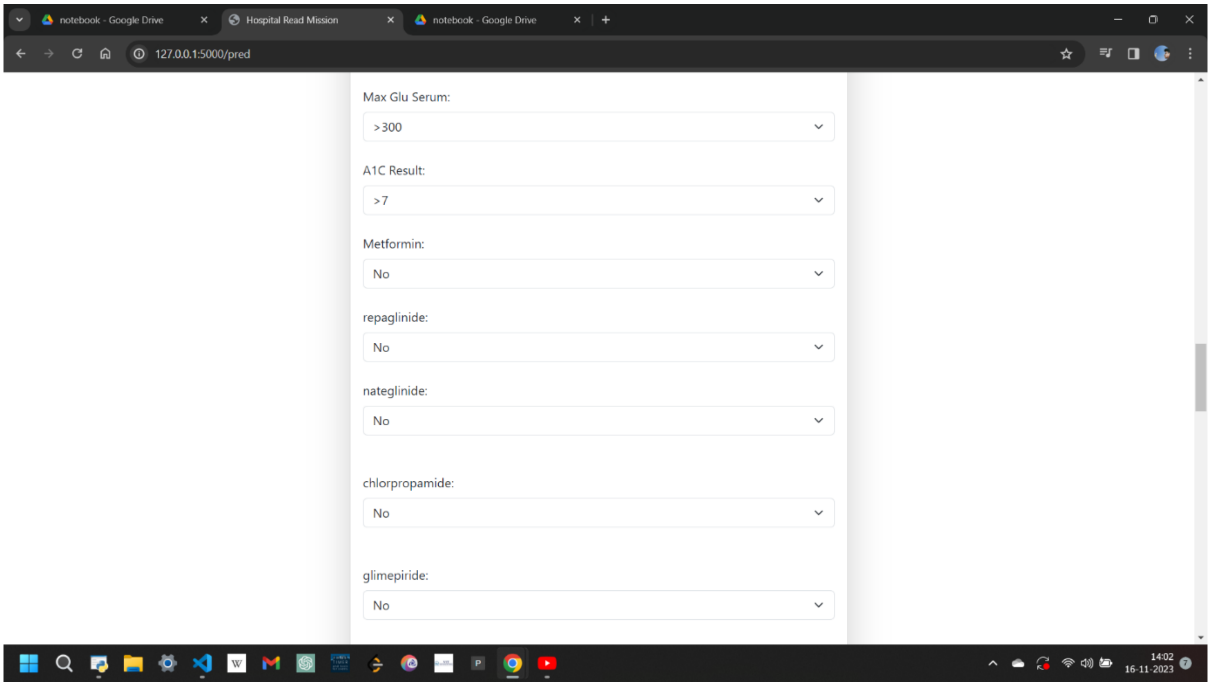
Discharge Disposition:
Discharged to Home

Admission Source:
Referral

diag_1:

diag_2:

diag_3:



Max Glu Serum:
>300

A1C Result:
>7

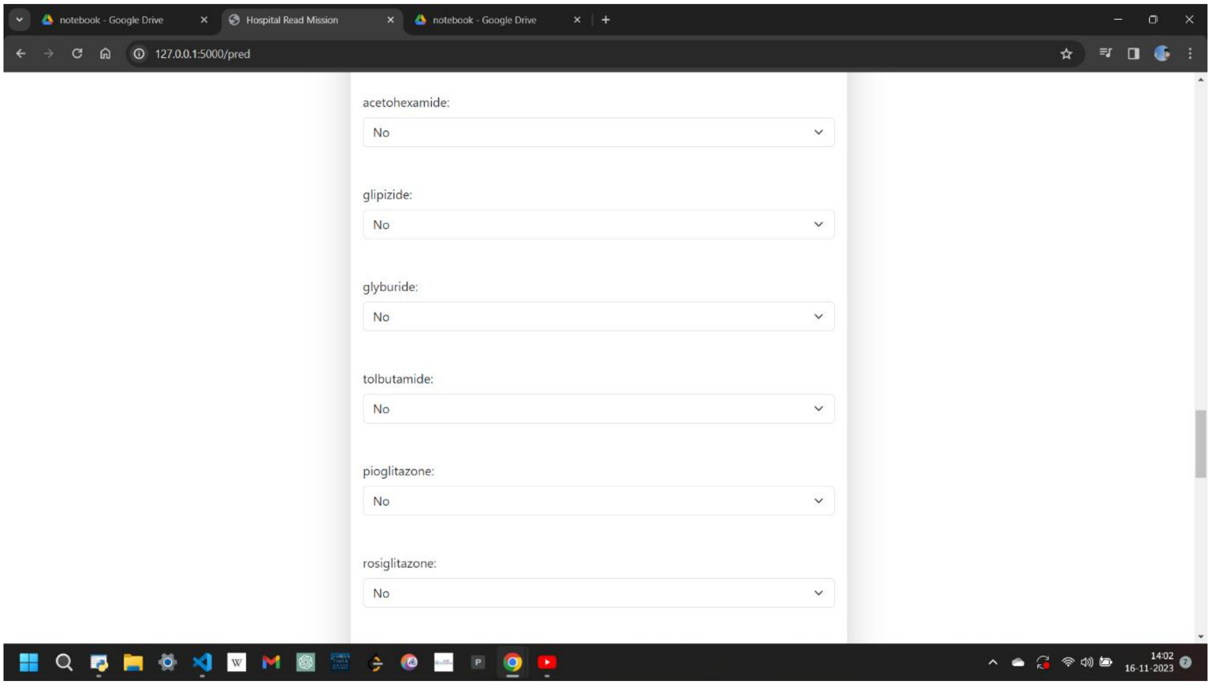
Metformin:
No

repaglinide:
No

nateglinide:
No

chlorpropamide:
No

glimepiride:
No



acetohexamide:
No

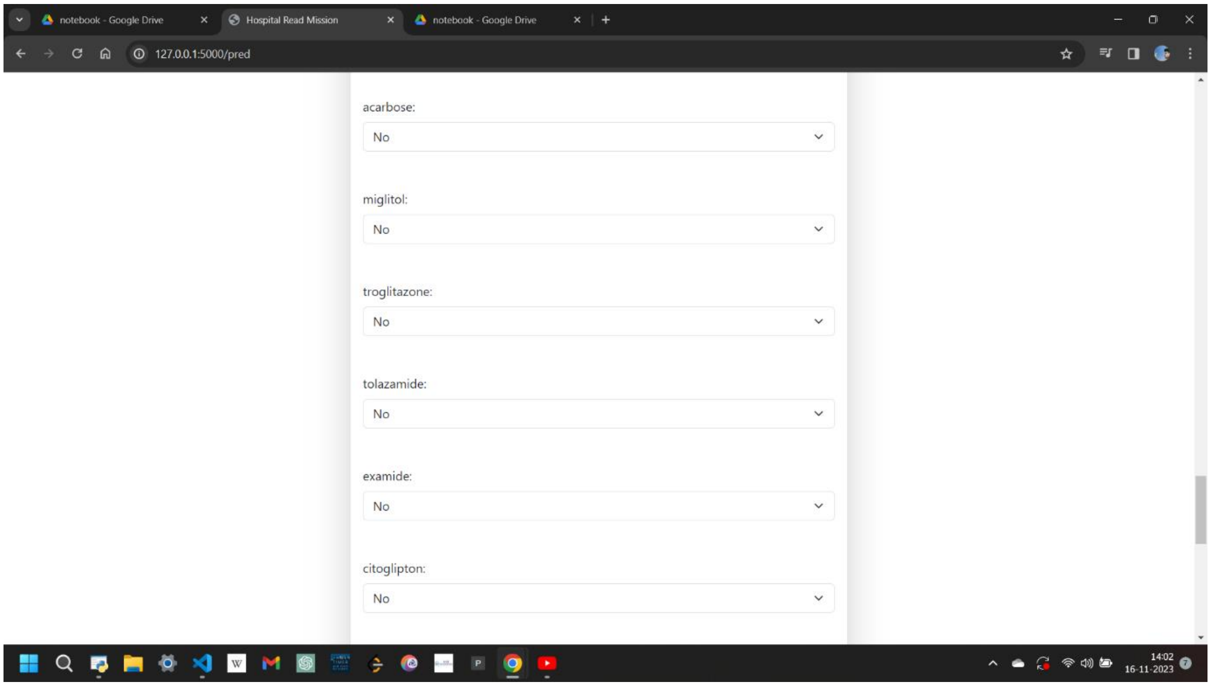
glipizide:
No

glyburide:
No

tolbutamide:
No

pioglitazone:
No

rosiglitazone:
No



acarbose:
No

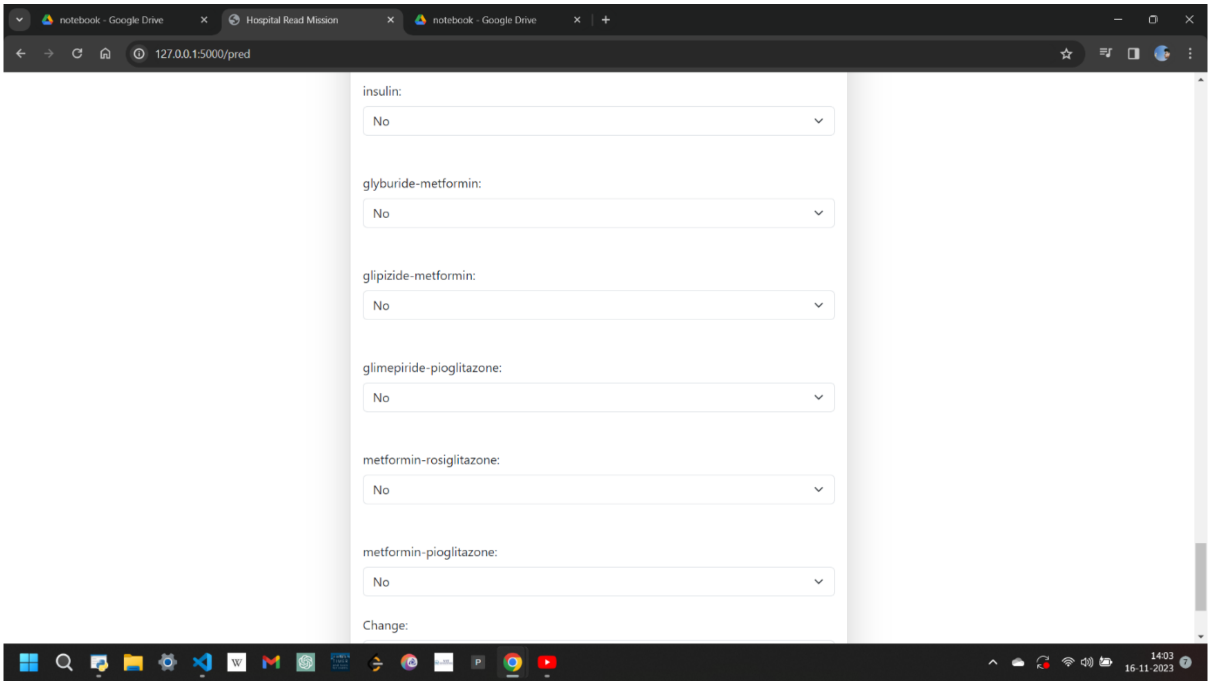
miglitol:
No

troglitazone:
No

tolazamide:
No

examide:
No

citoglipton:
No



insulin: No

glyburide-metformin: No

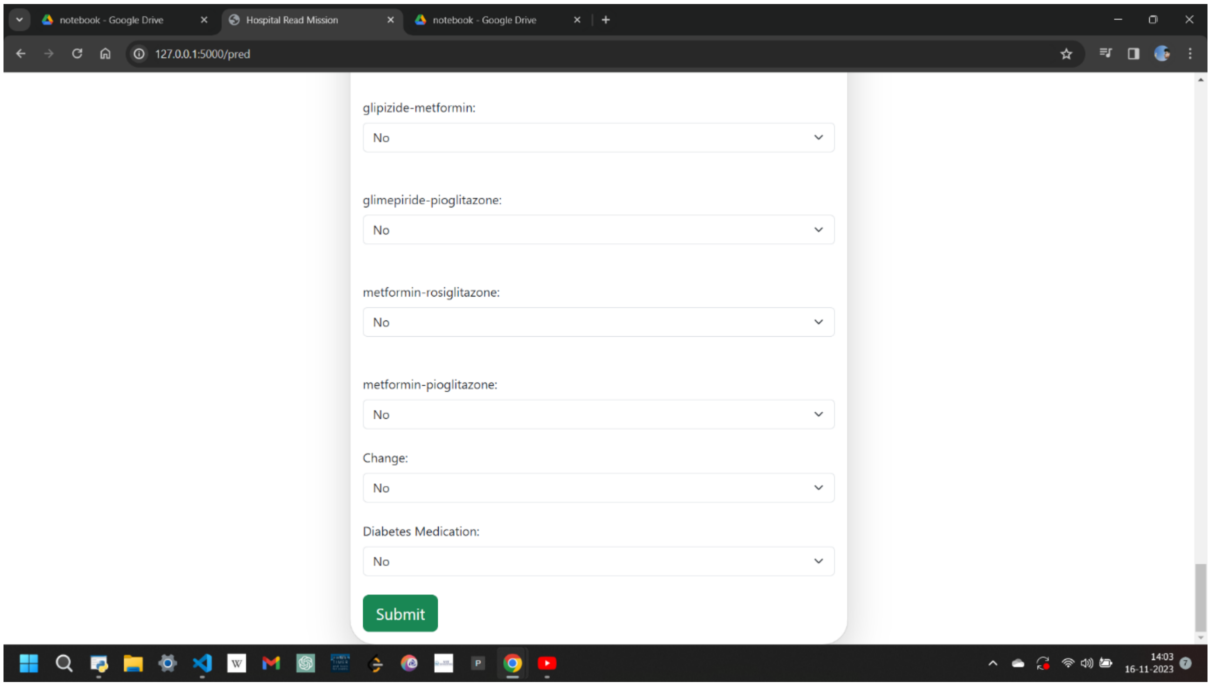
glipizide-metformin: No

glimepiride-pioglitazone: No

metformin-rosiglitazone: No

metformin-pioglitazone: No

Change:



glipizide-metformin: No

glimepiride-pioglitazone: No

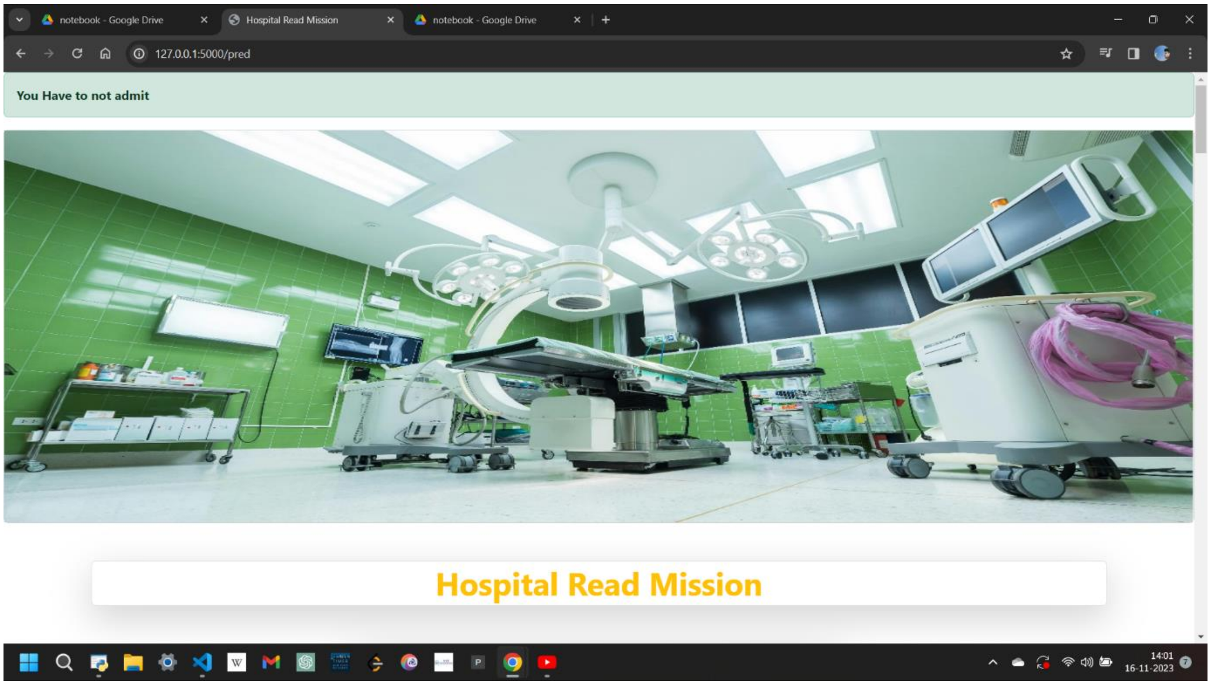
metformin-rosiglitazone: No

metformin-pioglitazone: No

Change: No

Diabetes Medication: No

OUTPUT:



You Have to not admit

Hospital Read Mission

10. Advantages and Disadvantages

Predicting hospital readmissions has both advantages and disadvantages. Here are some of them:

Advantages:

Resource Optimization:

Hospitals can allocate resources more efficiently by predicting which patients are more likely to be readmitted. This includes beds, staff, and medical supplies.

Cost Reduction:

Predicting readmissions can lead to cost savings for both healthcare providers and patients. It allows for proactive interventions and preventive measures to reduce the likelihood of readmission.

Improved Patient Care:

Identifying patients at risk of readmission enables healthcare providers to offer personalized care plans, medication management, and follow-up appointments, leading to better patient outcomes.

Enhanced Patient Satisfaction:

By avoiding unnecessary readmissions, patients experience improved continuity of care and are less likely to be dissatisfied with their healthcare experience.

Quality of Care Monitoring:

Hospitals can use readmission prediction models as a metric for evaluating the effectiveness of their healthcare services and making necessary improvements.

Disadvantages:

Ethical Concerns:

Predicting readmissions may raise ethical concerns related to patient privacy, as it involves the use of sensitive health data.

Potential for Biases:

Models may be biased, leading to disparities in care, especially if they are trained on historical data that reflects existing healthcare disparities.

Overemphasis on Metrics:

There might be a risk of healthcare providers focusing solely on reducing readmissions without considering other important aspects of patient care.

Complexity of Prediction:

Hospital readmission is a complex issue influenced by various factors, making accurate predictions challenging. Overreliance on predictions may result in false positives or false negatives.

Unintended Consequences:

Hospitals might refuse readmission for patients who genuinely need it to avoid penalties or to meet performance metrics, leading to potential negative health outcomes.

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.

11.Conclusion

Through this project, we created a machine learning model that is able to predict the patients with diabetes with highest risk of being readmitted within 30 days. The best model was a gradient boosting classifier with optimized hyperparameters. The model was able to catch 58% of the readmissions and is about 1.5 times better than just randomly picking patients. Overall, I believe many healthcare data scientists are working on predictive models for hospital readmission

12.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.

13. Appendix

Source code

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

https://drive.google.com/drive/folders/1OP9b-QeAtDl7ylgu8y77Gzt-nE9AA5_B?usp=sharing

GitHub & Project Demo link

GitHub link: : https://github.com/smartinternz02/SI-GuidedProject-611650-1699952686

Project Demo link:https://drive.google.com/file/d/1EiN0ci6dtAxdGlpXZejPoZMagLgS-Ica/view?usp=drivesdk