Heart disease detection project

P122 - Group 4

Harshada More

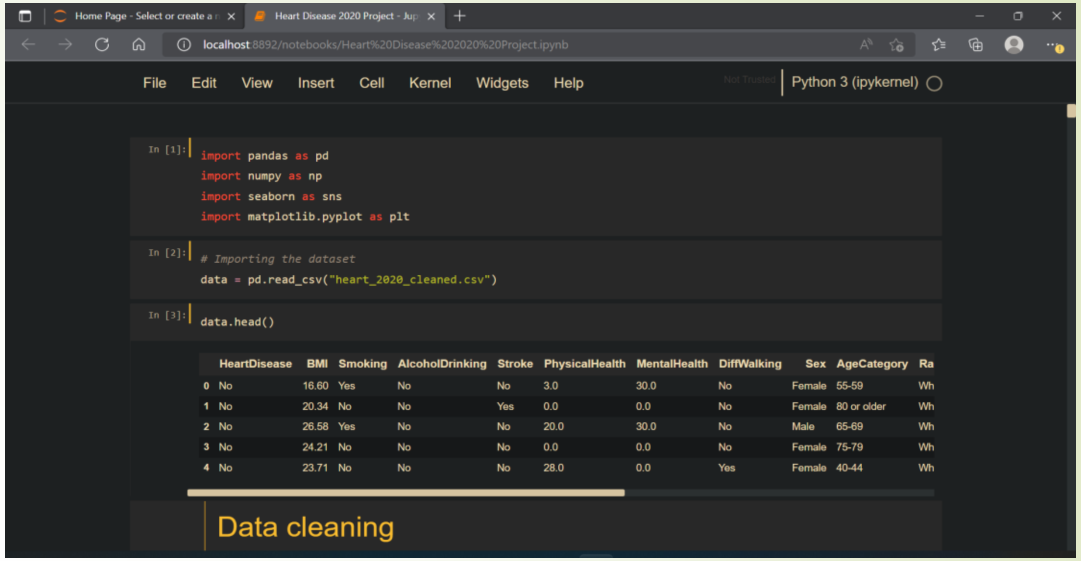
Names of group members:

- Ms. Harshada More
- Ms. Vaibhavi Taide
- Mr. Nagaraja M. R.
- Mr. Ganesh N. R.
- Mr. Parag Wani
- Ms. Mansi Solanki

Importing dataset

Data cleaning

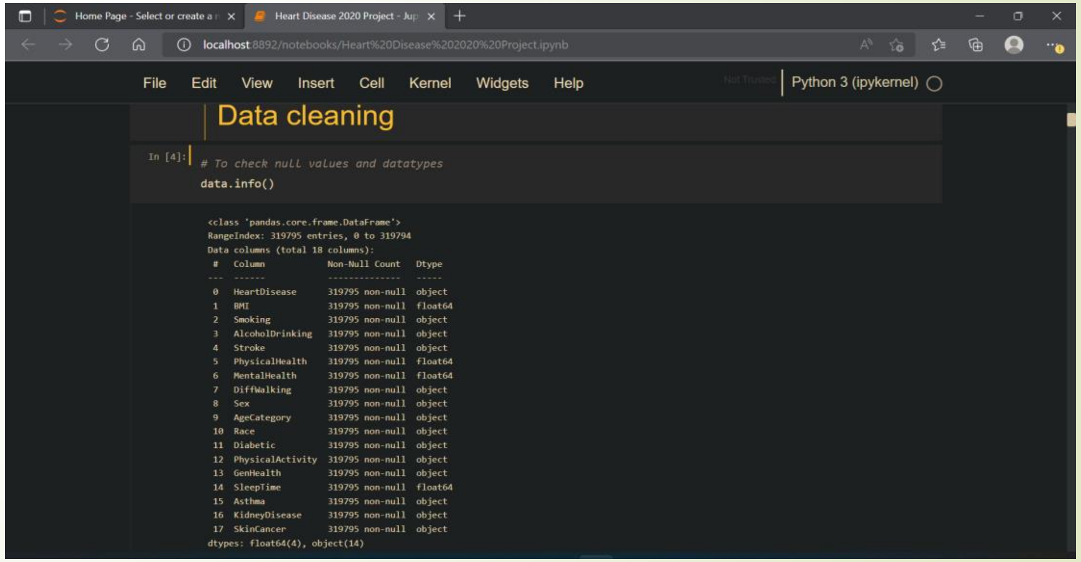
EDA


Home Page - Select or create a n
Heart Disease 2020 Project - Jup
localhost:8892/notebooks/Heart%20Disease%202020%20Project.ipynb
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)

```
In [1]: import pandas as pd
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt

In [2]: # Importing the dataset
        data = pd.read_csv("heart_2020_cleaned.csv")

In [3]: data.head()
```

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Ra
0	No	16.60	Yes	No	No	3.0	30.0	No	Female	55-59	Wh
1	No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	Wh
2	No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	Wh
3	No	24.21	No	No	No	0.0	0.0	No	Female	75-79	Wh
4	No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	Wh

Data cleaning


Home Page - Select or create a n
Heart Disease 2020 Project - Jup
localhost:8892/notebooks/Heart%20Disease%202020%20Project.ipynb
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)

Data cleaning

```
In [4]: # To check null values and datatypes
        data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 319795 entries, 0 to 319794
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   HeartDisease      319795 non-null  object 
 1   BMI               319795 non-null  float64
 2   Smoking           319795 non-null  object 
 3   AlcoholDrinking   319795 non-null  object 
 4   Stroke            319795 non-null  object 
 5   PhysicalHealth    319795 non-null  float64
 6   MentalHealth      319795 non-null  float64
 7   DiffWalking       319795 non-null  object 
 8   Sex               319795 non-null  object 
 9   AgeCategory       319795 non-null  object 
 10  Race              319795 non-null  object 
 11  Diabetic          319795 non-null  object 
 12  PhysicalActivity  319795 non-null  object 
 13  GenHealth         319795 non-null  object 
 14  SleepTime         319795 non-null  float64
 15  Asthma            319795 non-null  object 
 16  KidneyDisease     319795 non-null  object 
 17  SkinCancer        319795 non-null  object 
dtypes: float64(4), object(14)
```

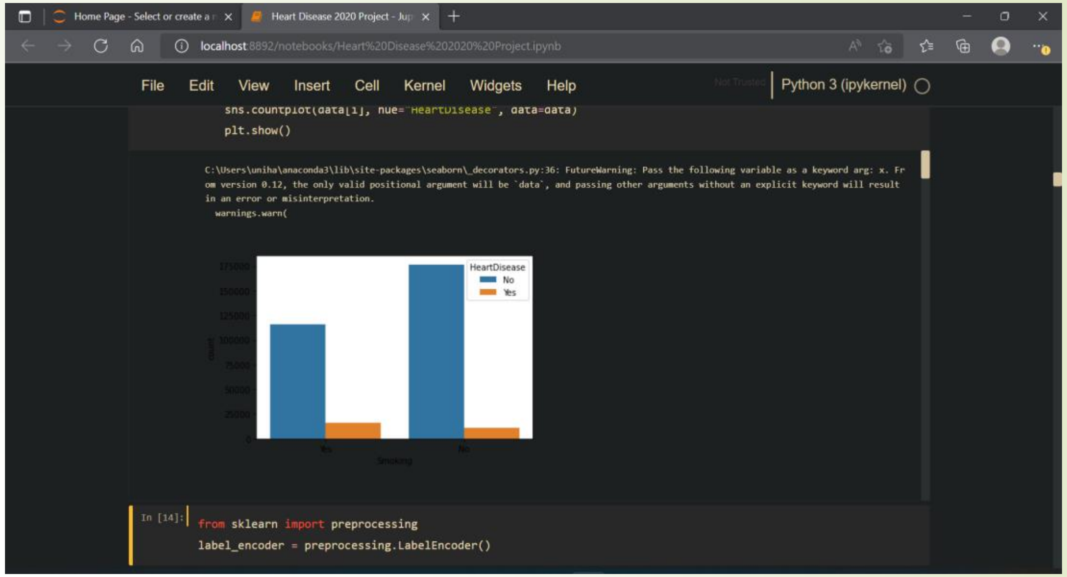
Harshada More

Visualisation

Label encoder

Defining variables

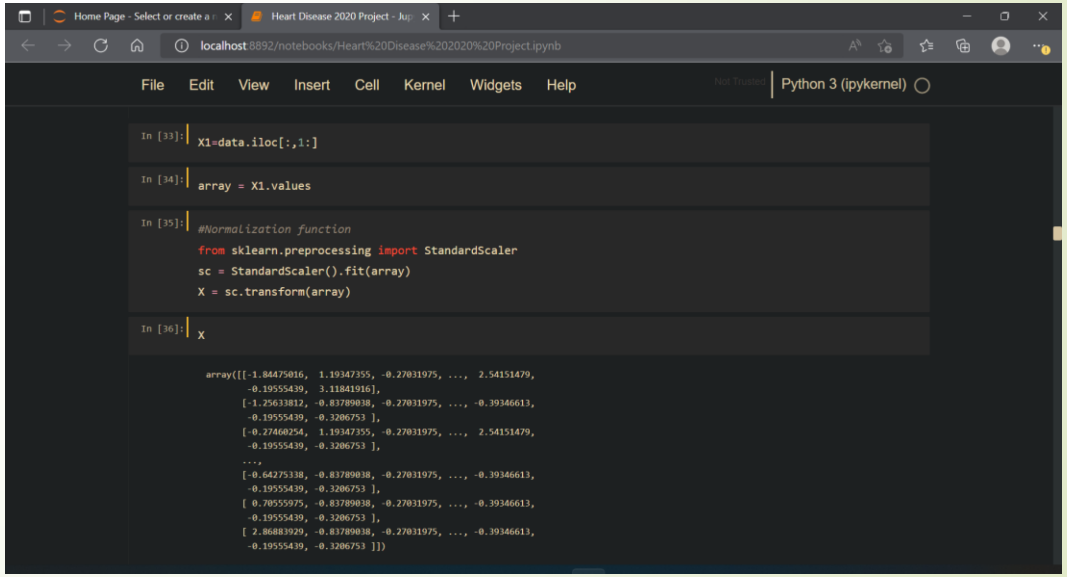
Train-Test split data



```
sns.countplot(data[i], hue="HeartDisease", data=data)
plt.show()
```

```
C:\Users\uniha\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. Fr
om version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result
in an error or misinterpretation.
  warnings.warn(
```

```
In [14]: from sklearn import preprocessing
         label_encoder = preprocessing.LabelEncoder()
```


Home Page - Select or create a
Heart Disease 2020 Project - Jup
localhost:8892/notebooks/Heart%20Disease%202020%20Project.ipynb
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)

```
In [33]: X1=data.iloc[:,1:]

In [34]: array = X1.values

In [35]: #Normalization function
         from sklearn.preprocessing import StandardScaler
         sc = StandardScaler().fit(array)
         X = sc.transform(array)

In [36]: X
```

```
array([[-1.84475016,  1.19347355, -0.27031975, ...,  2.54151479,
        -0.19555439,  3.11841916],
       [-1.25633812, -0.83789038, -0.27031975, ..., -0.39346613,
        -0.19555439, -0.3206753 ],
       [-0.27460254,  1.19347355, -0.27031975, ...,  2.54151479,
        -0.19555439, -0.3206753 ],
       ...,
       [-0.64275338, -0.83789038, -0.27031975, ..., -0.39346613,
        -0.19555439, -0.3206753 ],
       [ 0.70555975, -0.83789038, -0.27031975, ..., -0.39346613,
        -0.19555439, -0.3206753 ],
       [ 2.86883929, -0.83789038, -0.27031975, ..., -0.39346613,
        -0.19555439, -0.3206753 ]])
```

Harshada More

Model analysis performed for the following models:

- Logistic regression
- Decision Tree Classifier
- KNN
- Naïve Bayes
- XGBM
- LGBM
- SVM
- Random forest

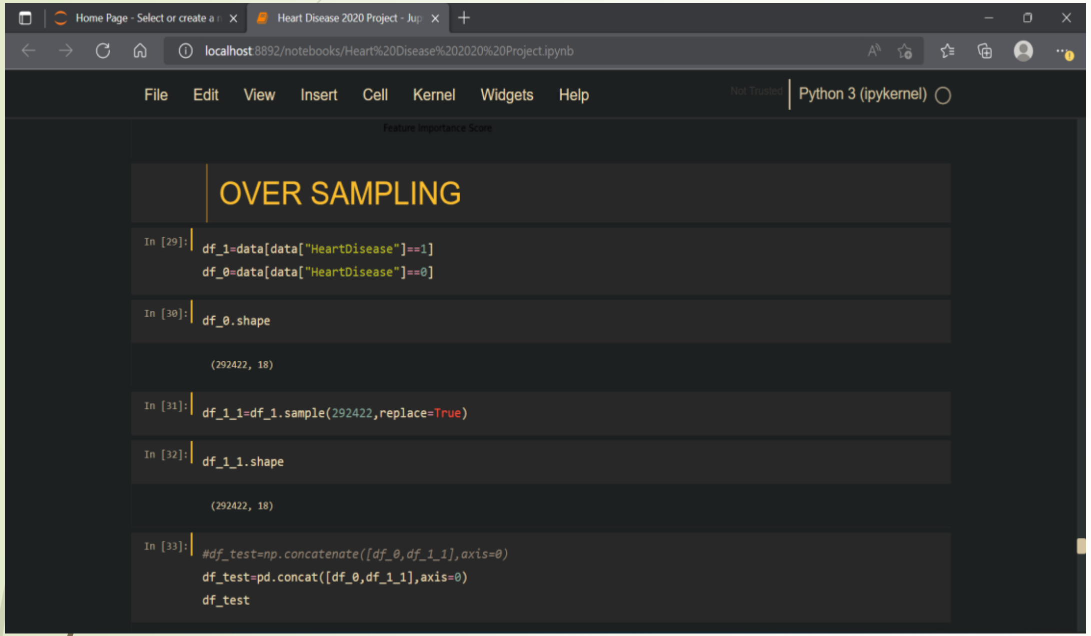
Harshada More

Model analysis performed:

- Confusion matrix
- Model accuracy for training data
- Model accuracy for test data
- Classification report

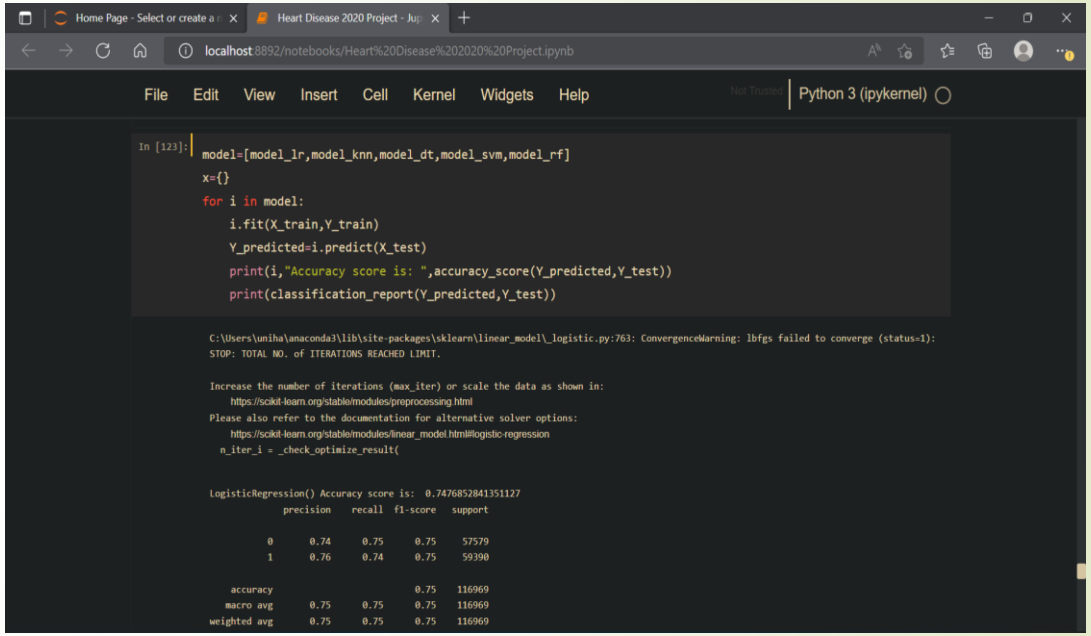
Harshada More

Over sampling model analysis


Home Page - Select or create a
Heart Disease 2020 Project - Jup
localhost:8892/notebooks/Heart%20Disease%202020%20Project.ipynb
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)
Feature Importance Score

OVER SAMPLING

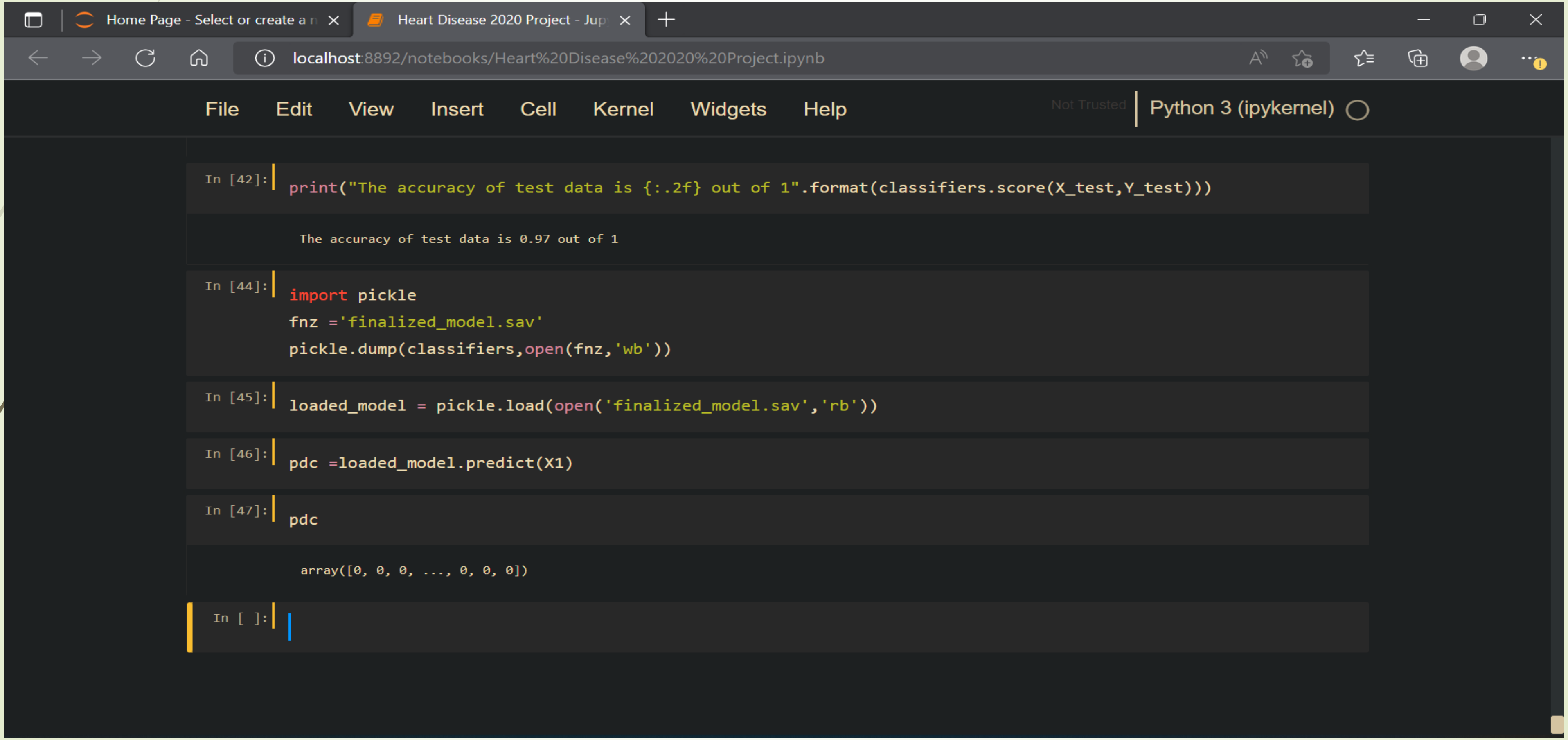
```
In [29]: df_1=data[data["HeartDisease"]==1]
         df_0=data[data["HeartDisease"]==0]

In [30]: df_0.shape

         (292422, 18)

In [31]: df_1_1=df_1.sample(292422,replace=True)

In [32]: df_1_1.shape

         (292422, 18)

In [33]: #df_test=np.concatenate([df_0,df_1_1],axis=0)
         df_test=pd.concat([df_0,df_1_1],axis=0)
         df_test
```


Home Page - Select or create a
Heart Disease 2020 Project - Jup
localhost:8892/notebooks/Heart%20Disease%202020%20Project.ipynb
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)

```
In [123]: model=[model_lr,model_knn,model_dt,model_svm,model_rf]
          x={}
          for i in model:
              i.fit(X_train,Y_train)
              Y_predicted=i.predict(X_test)
              print(i,"Accuracy score is: ",accuracy_score(Y_predicted,Y_test))
              print(classification_report(Y_predicted,Y_test))

C:\Users\uniha\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:763: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(


LogisticRegression() Accuracy score is:  0.7476852841351127
              precision    recall  f1-score   support

           0       0.74      0.75      0.75     57579
           1       0.76      0.74      0.75     59390

    accuracy                           0.75    116969
   macro avg       0.75      0.75      0.75    116969
weighted avg       0.75      0.75      0.75    116969
```

Harshada More

Saving the finalized model – Random forest by using pickle

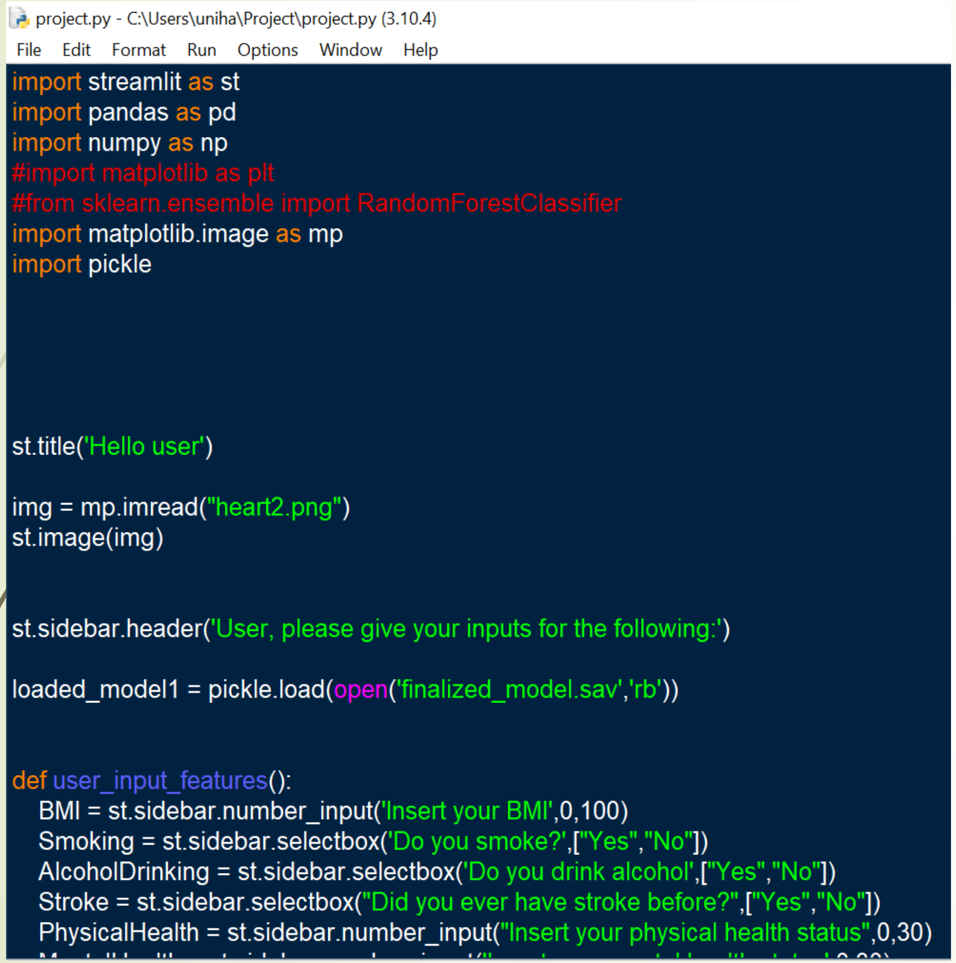


Home Page - Select or create a n
Heart Disease 2020 Project - Jup
localhost:8892/notebooks/Heart%20Disease%202020%20Project.ipynb
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)

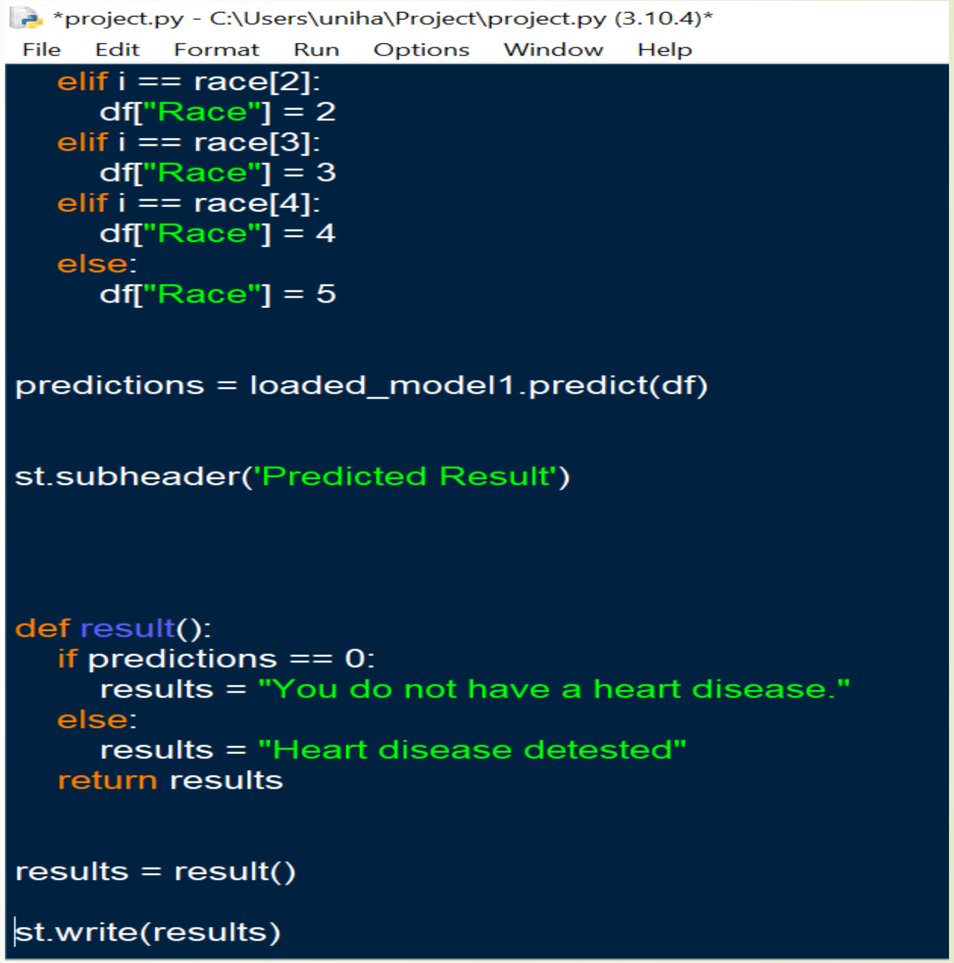
```
In [42]: print("The accuracy of test data is {:.2f} out of 1".format(classifiers.score(X_test,Y_test)))

         The accuracy of test data is 0.97 out of 1

In [44]: import pickle
         fnz ='finalized_model.sav'
         pickle.dump(classifiers,open(fnz,'wb'))

In [45]: loaded_model = pickle.load(open('finalized_model.sav','rb'))

In [46]: pdc =loaded_model.predict(X1)

In [47]: pdc

         array([0, 0, 0, ..., 0, 0, 0])

In [ ]:
```

Harshada More

Streamlit code (Part of the code)


project.py - C:\Users\uniha\Project\project.py (3.10.4)
File Edit Format Run Options Window Help

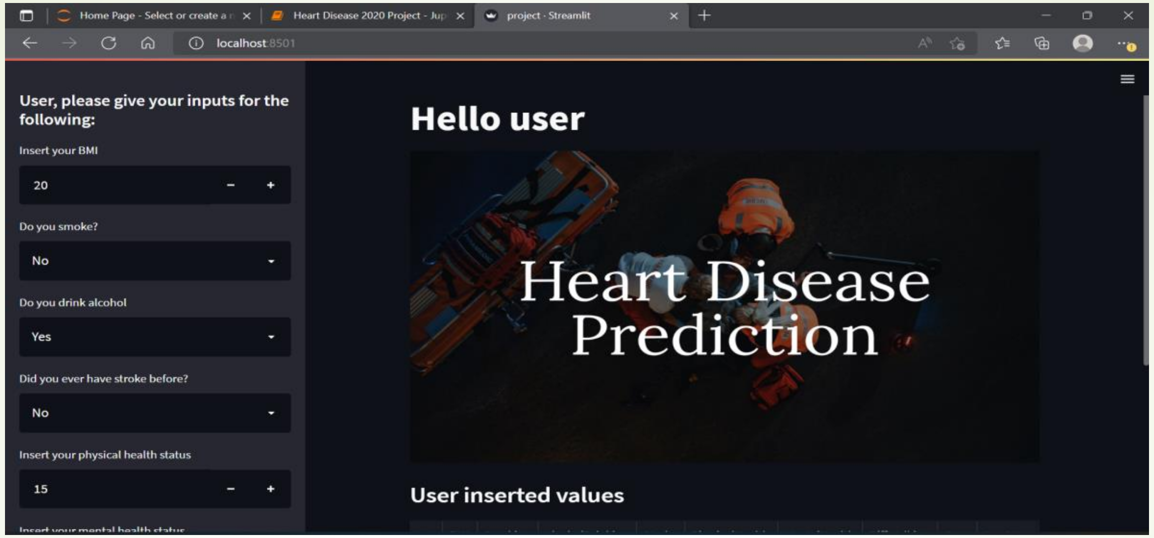
```python
import streamlit as st
import pandas as pd
import numpy as np
#import matplotlib as plt
#from sklearn.ensemble import RandomForestClassifier
import matplotlib.image as mp
import pickle






st.title('Hello user')

img = mp.imread("heart2.png")
st.image(img)


st.sidebar.header('User, please give your inputs for the following:')

loaded_model1 = pickle.load(open('finalized_model.sav','rb'))


def user_input_features():
    BMI = st.sidebar.number_input('Insert your BMI',0,100)
    Smoking = st.sidebar.selectbox('Do you smoke?',["Yes","No"])
    AlcoholDrinking = st.sidebar.selectbox('Do you drink alcohol',["Yes","No"])
    Stroke = st.sidebar.selectbox("Did you ever have stroke before?",["Yes","No"])
    PhysicalHealth = st.sidebar.number_input("Insert your physical health status",0,30)
```


project.py - C:\Users\uniha\Project\project.py (3.10.4)
File Edit Format Run Options Window Help

```python
    elif i == race[2]:
        df["Race"] = 2
    elif i == race[3]:
        df["Race"] = 3
    elif i == race[4]:
        df["Race"] = 4
    else:
        df["Race"] = 5


predictions = loaded_model1.predict(df)


st.subheader('Predicted Result')




def result():
    if predictions == 0:
        results = "You do not have a heart disease."
    else:
        results = "Heart disease detested"
    return results


results = result()

st.write(results)
```

Harshada More

Model final output


Home Page - Select or create a n
Heart Disease 2020 Project - Jup
project · Streamlit
localhost:8501

User, please give your inputs for the following:

Insert your BMI
20

Do you smoke?
No

Do you drink alcohol
Yes

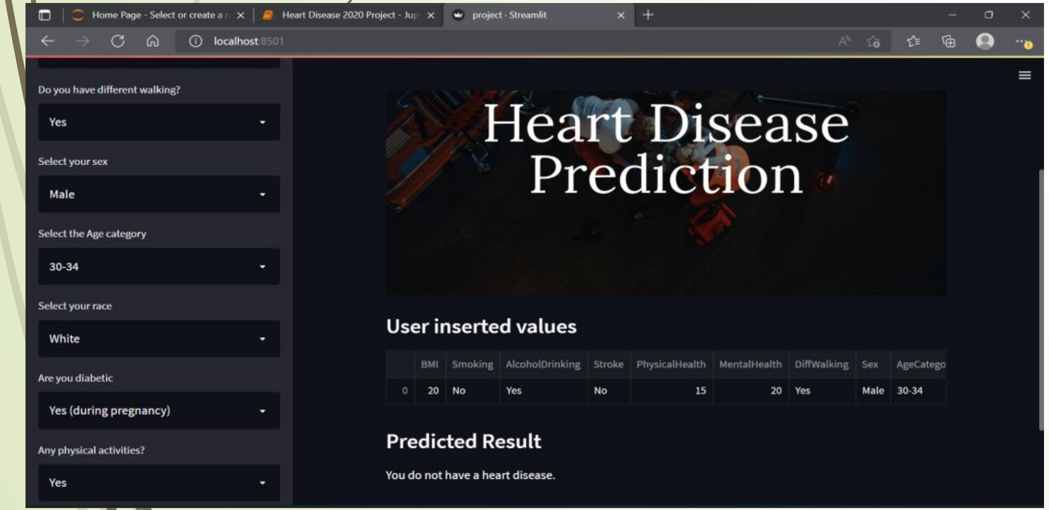
Did you ever have stroke before?
No

Insert your physical health status
15

Insert your mental health status

Hello user

User inserted values


Home Page - Select or create a n
Heart Disease 2020 Project - Jup
project · Streamlit
localhost:8501
Heart Disease Prediction

Do you have different walking?
Yes

Select your sex
Male

Select the Age category
30-34

Select your race
White

Are you diabetic
Yes (during pregnancy)

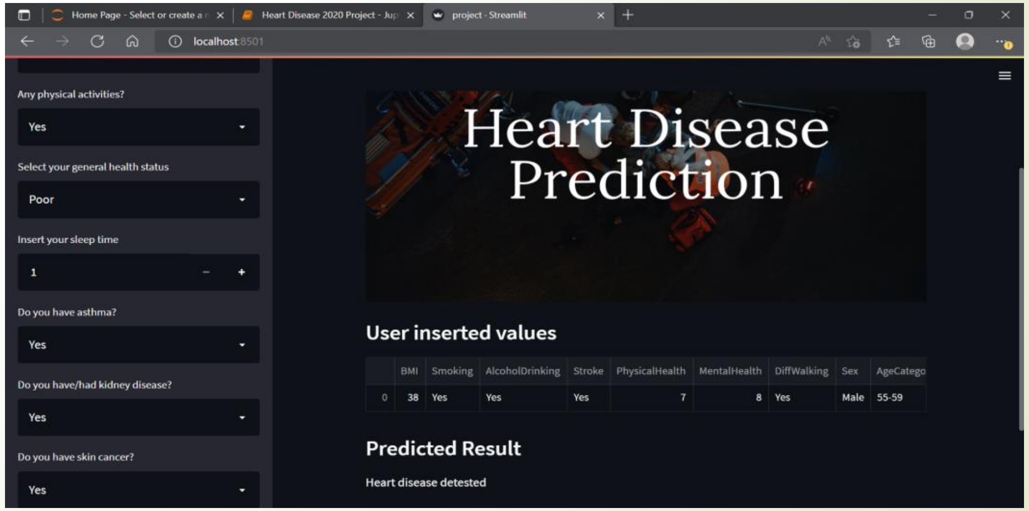
Any physical activities?
Yes

User inserted values

	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCatego
0	20	No	Yes	No	15	20	Yes	Male	30-34

Predicted Result

You do not have a heart disease.


Home Page - Select or create a n
Heart Disease 2020 Project - Jup
project · Streamlit
localhost:8501
Heart Disease Prediction

Any physical activities?
Yes

Select your general health status
Poor

Insert your sleep time
1

Do you have asthma?
Yes

Do you have/had kidney disease?
Yes

Do you have skin cancer?
Yes

User inserted values

	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCatego
0	38	Yes	Yes	Yes	7	8	Yes	Male	55-59

Predicted Result

Heart disease detested

Harshada More

Harshada More