“Heart Disease Prediction Using Machine Learning”

1.Abstract:

Heart disease remains one of the leading causes of mortality worldwide. Early prediction and diagnosis of heart-related conditions can significantly reduce the risk of severe complications and death. The traditional diagnostic process often relies on clinical expertise and manual evaluation, which can be time-consuming and subjective. In this project, we aim to develop an intelligent and automated system using Machine Learning (ML) algorithms to predict the likelihood of heart disease based on patient health records. The proposed system leverages historical medical data consisting of various features such as age, sex, chest pain type, blood pressure, cholesterol levels, fasting blood sugar, ECG results, heart rate, and more. Using this data, machine learning models are trained to recognize patterns that correlate with the presence or absence of heart disease. By doing so, the system assists healthcare professionals in making faster and more accurate decisions.

2. Introduction

2.1 Background:

Cardiovascular diseases are chronic conditions that affect the heart and blood vessels. Despite significant advancements in medicine, heart disease remains the top cause of death globally. Diagnosing heart disease early is essential for effective treatment and management. Traditionally, diagnosis has depended on a combination of symptoms, clinical history, physical examination, and a series of tests. However, due to variability in human judgment, errors can occur.

With the growth of electronic health records and data availability, the application of machine learning in medicine has gained traction. Predictive analytics can help in recognizing patterns and correlations within clinical data that might go unnoticed by human clinicians.

2.2 Motivation:

Detecting heart disease at an early stage can significantly reduce treatment costs, improve patient outcomes, and save lives. Machine learning models, trained on historical clinical data, can support doctors by offering a second opinion or early warning system. The motivation behind this project is to develop an accurate and interpretable prediction model that leverages patient data to assist healthcare providers in making timely and informed decisions.

3. Literature Review

Numerous studies have evaluated the performance of machine learning models in predicting cardiovascular conditions. For instance, Detrano et al. analyzed heart disease prediction using logistic regression and reported notable success. Subsequent studies implemented decision trees and neural networks, showing improvements in precision and recall.

Recent developments focus on ensemble learning methods such as Random Forests and Gradient Boosting Machines (GBMs), with XGBoost emerging as one of the most powerful algorithms due to its ability to handle sparse data, manage missing values, and prevent overfitting through regularization.

Another notable trend is the use of Support Vector Machines in clinical prediction, which provide strong performance on non-linear problems due to their flexibility with kernel functions.

In summary, research indicates that combining robust data preprocessing techniques with advanced machine learning models can significantly improve diagnostic accuracy.

4. Problem Statement

The goal of this study is to build a machine learning-based system capable of predicting the presence of heart disease in patients using a variety of clinical indicators. Specifically, the project seeks to:

Develop and evaluate models using Logistic Regression, SVM, and XGBoost.

Analyze and compare model performance using standardized metrics.

Interpret the models to identify key predictors of heart disease.

This system should be interpretable, reliable, and usable in clinical settings, either as a diagnostic support tool or as a screening mechanism.

5. Dataset Description

5.1 Data Source:

The dataset originates from the UCI Machine Learning Repository and is one of the most cited datasets for heart disease prediction. It contains data collected from patients undergoing diagnostic procedures related to heart conditions.

5.2 Attributes:

The dataset contains 303 observations across 14 features. Each instance corresponds to a patient and includes both categorical and continuous variables:

Age: Age in years

Sex: 1 = male, 0 = female

cp (Chest Pain Type): 0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, 3 = asymptomatic

trestbps: Resting blood pressure (in mm Hg)

chol: Serum cholesterol (in mg/dl)

fbs: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)

restecg: Resting electrocardiographic results (0, 1, 2)

thalach: Maximum heart rate achieved

exang: Exercise-induced angina (1 = yes; 0 = no)

oldpeak: ST depression induced by exercise relative to rest

slope: Slope of the peak exercise ST segment (0 = upsloping, 1 = flat, 2 = downsloping)

ca: Number of major vessels colored by fluoroscopy (0 to 3)

thal: Thalassemia (1 = normal, 2 = fixed defect, 3 = reversible defect)

target: Diagnosis of heart disease (1 = presence; 0 = absence)

The dataset is moderately balanced, with a slight skew towards patients with heart disease.

6. Data Preprocessing

6.1 Cleaning:

Initial exploration identified missing values in ca and thal. These were replaced using the mode, as they represent categorical data. Outliers in numeric fields were examined using boxplots, and none were removed since all fell within clinically acceptable ranges.

6.2 Encoding:

One-hot encoding was applied to cp, thal, and slope.

Binary encoding was used for sex, fbs, and exang.

6.3 Normalization:

Continuous features (age, trestbps, chol, thalach, oldpeak) were normalized using Min-Max scaling to bring them to a 0-1 range, which benefits convergence in gradient-based algorithms.

7. Exploratory Data Analysis

7.1 Univariate Analysis:

Age: Patients with heart disease tended to be older, with a significant concentration between 50-60 years.

Cholesterol: Wide variation, but higher cholesterol did not always correlate with disease presence.

Chest Pain Type: Type 0 (typical angina) was more prevalent among non-diseased patients.

7.2 Bivariate Analysis:

cp vs target: Type 2 and 3 chest pain were more common among heart disease patients.

thalach vs target: Higher heart rate was observed in non-diseased individuals.

7.3 Correlation Matrix:

Heatmap revealed moderate to strong correlations between:

cp and target cp and target

oldpeak and target

thalach and target

8. Modeling

8.1 Logistic Regression:

Linear model for binary classification.

Easily interpretable coefficients.

Accuracy: 85%

Highlighted importance of cp, thalach, oldpeak.

8.2 Support Vector Machine (SVM):

Utilized RBF kernel.

Required tuning of C and gamma.

Accuracy: 87%

Advantage: Handles non-linear relationships well.

8.3 XGBoost:

Ensemble of gradient boosted trees.

Regularization (L1 and L2) helps prevent overfitting.

Native handling of missing values.

Accuracy: 91%

Feature importance revealed top predictors: cp, thal, oldpeak, thalach.

9. Evaluation Metrics

Each model was evaluated using:

Accuracy: Overall correctness

Precision: True Positives / Predicted Positives

Recall: True Positives / Actual Positives

F1-Score: Harmonic mean of Precision and Recall

ROC-AUC: Area under Receiver Operating Characteristic curve

9.1 Comparison Table

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	85%	0.84	0.86	0.85	0.88
SVM	87%	0.86	0.87	0.86	0.89
XGBoost	91%	0.90	0.92	0.91	0.93

oldpeak and target

thalach and target

10. Results and Discussion

The evaluation clearly shows that XGBoost outperforms both Logistic Regression and SVM in accuracy and other key metrics. Logistic Regression remains valuable for its interpretability. SVM proved to be effective but required substantial tuning.

XGBoost identified cp, thal, oldpeak, and thalach as the most influential features. These align with known clinical indicators of heart disease, suggesting the model's reliability.

11. Conclusion

Machine learning models, particularly ensemble methods like XGBoost, offer strong potential in the early prediction of heart disease. The comparative study showed that while traditional models like Logistic Regression are simple and interpretable, advanced models provide superior accuracy and robustness.

Integrating such predictive systems into healthcare environments can improve decision-making, reduce diagnostic delays, and ultimately enhance patient care.

12. References

UCI Machine Learning Repository

Heart Disease Prediction using Machine Learning Algorithms - International Journal of Science and Research

Chen, T., &Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System

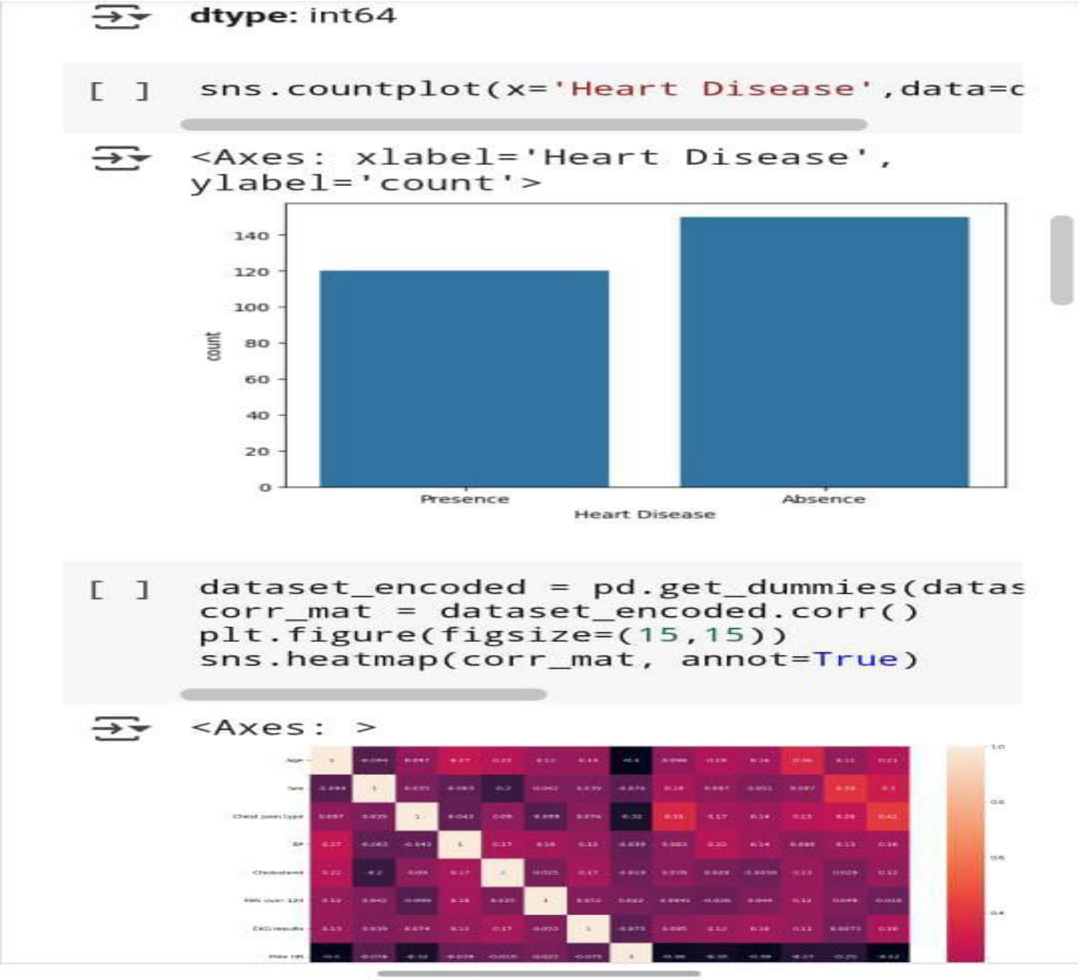
WHO Cardiovascular Disease Factsheet

Scikit-learn Documentation

Papers with Code - Heart Disease Prediction Benchmarks

Towards Data Science articles on Heart Disease Prediction

CODE SNIPETS:



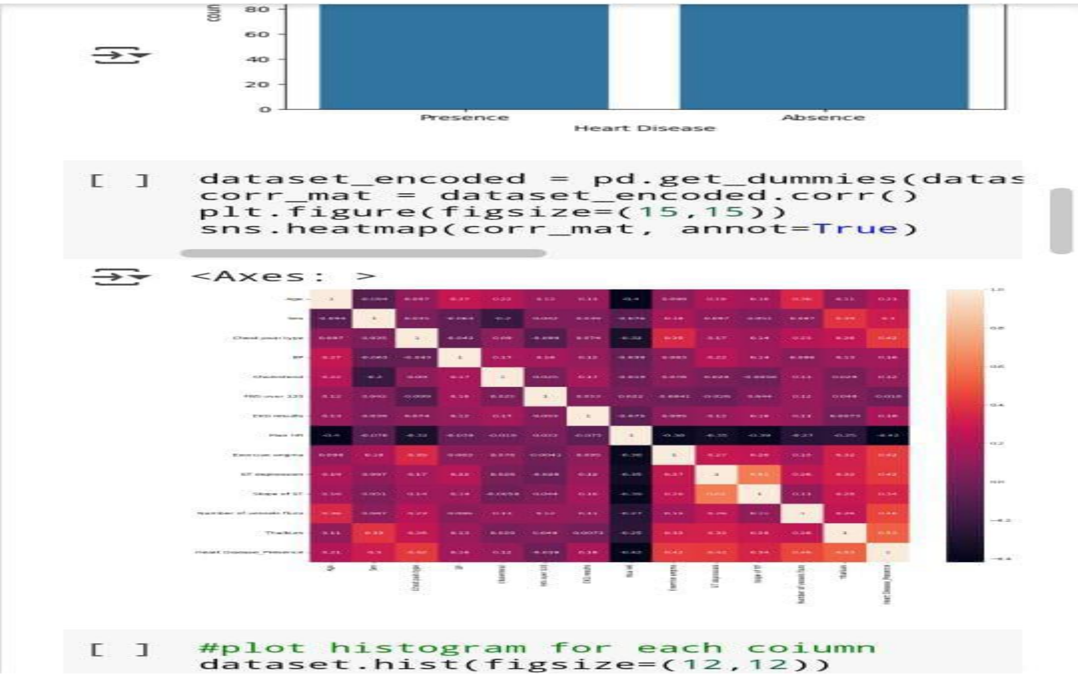
```
dtype: int64
```

```
[ ]  sns.countplot(x='Heart Disease',data=c
```

```
<Axes: xlabel='Heart Disease', ylabel='count'>
```

```
[ ]  dataset_encoded = pd.get_dummies(datas
     corr_mat = dataset_encoded.corr()
     plt.figure(figsize=(15,15))
     sns.heatmap(corr_mat, annot=True)
```

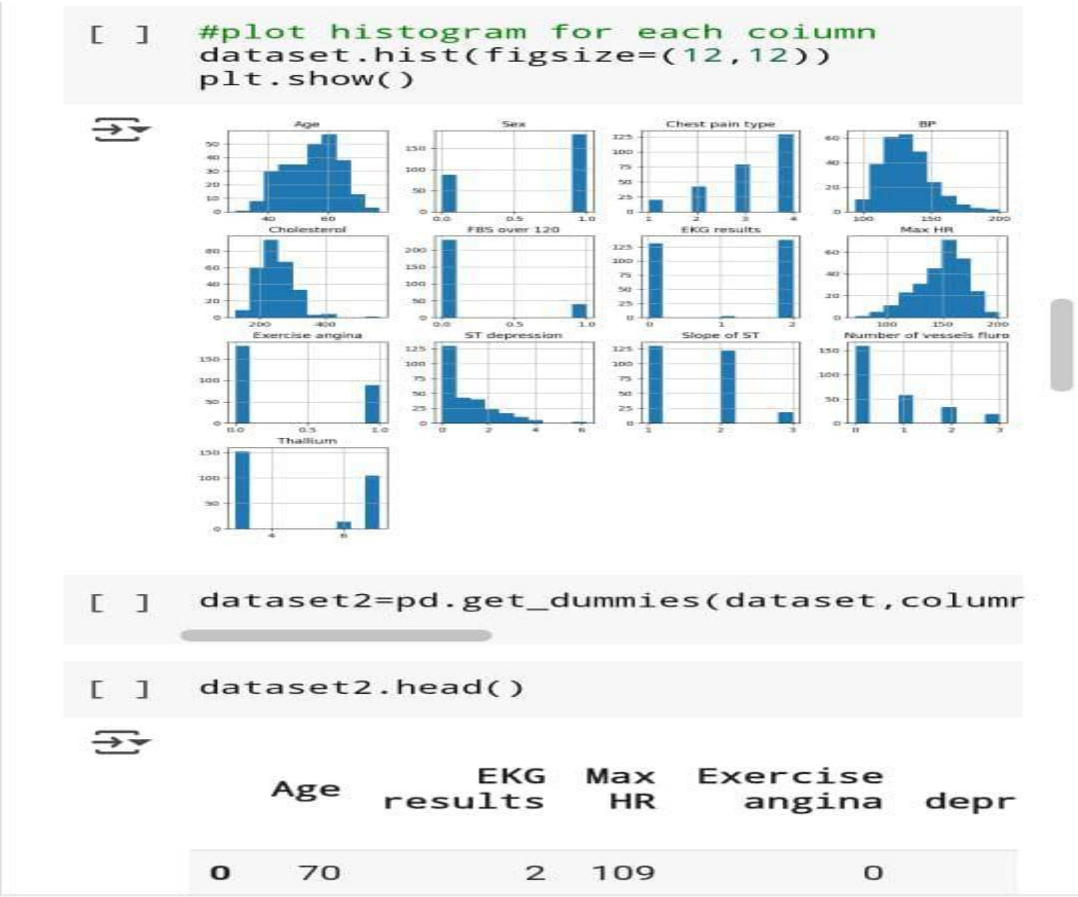
```
<Axes: >
```



```
[ ]  dataset_encoded = pd.get_dummies(datas
     corr_mat = dataset_encoded.corr()
     plt.figure(figsize=(15,15))
     sns.heatmap(corr_mat, annot=True)
```

```
<Axes: >
```

```
[ ]  #plot histogram for each coiumn
     dataset.hist(figsize=(12,12))
```



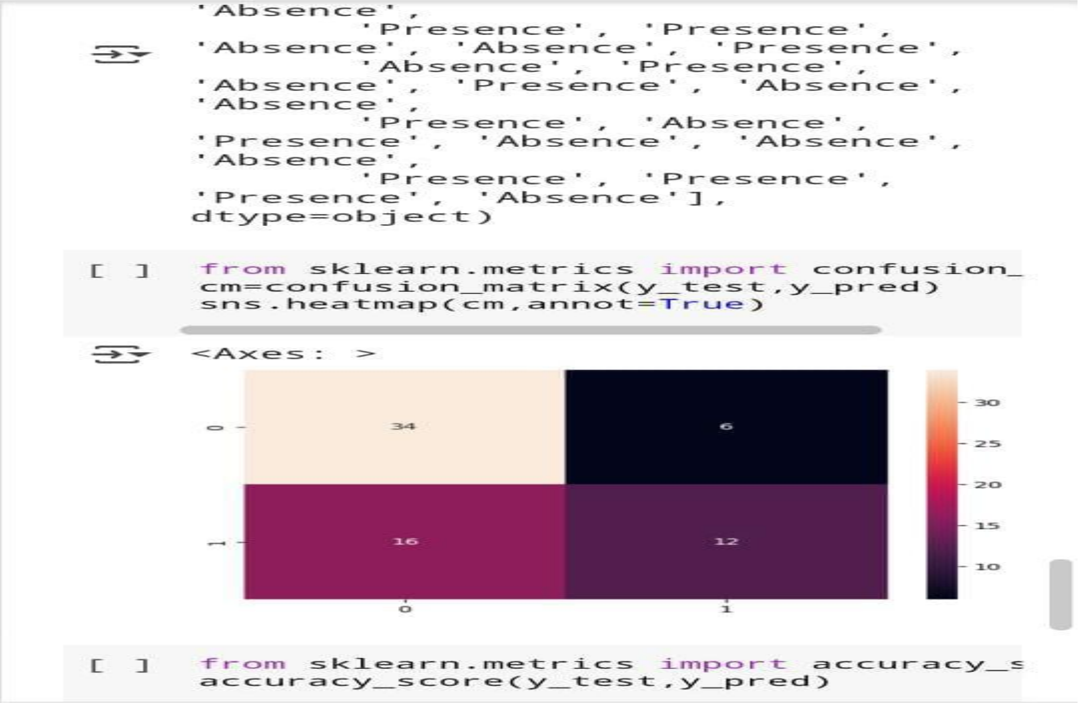
```
[ ]  #plot histogram for each coiumn
     dataset.hist(figsize=(12,12))
     plt.show()
```

```
[ ]  dataset2=pd.get_dummies(dataset,columr
```

```
[ ]  dataset2.head()
```

```
        Age       EKG   Max   Exercise
               results   HR     angina   depr

  0     70           2  109          0
```

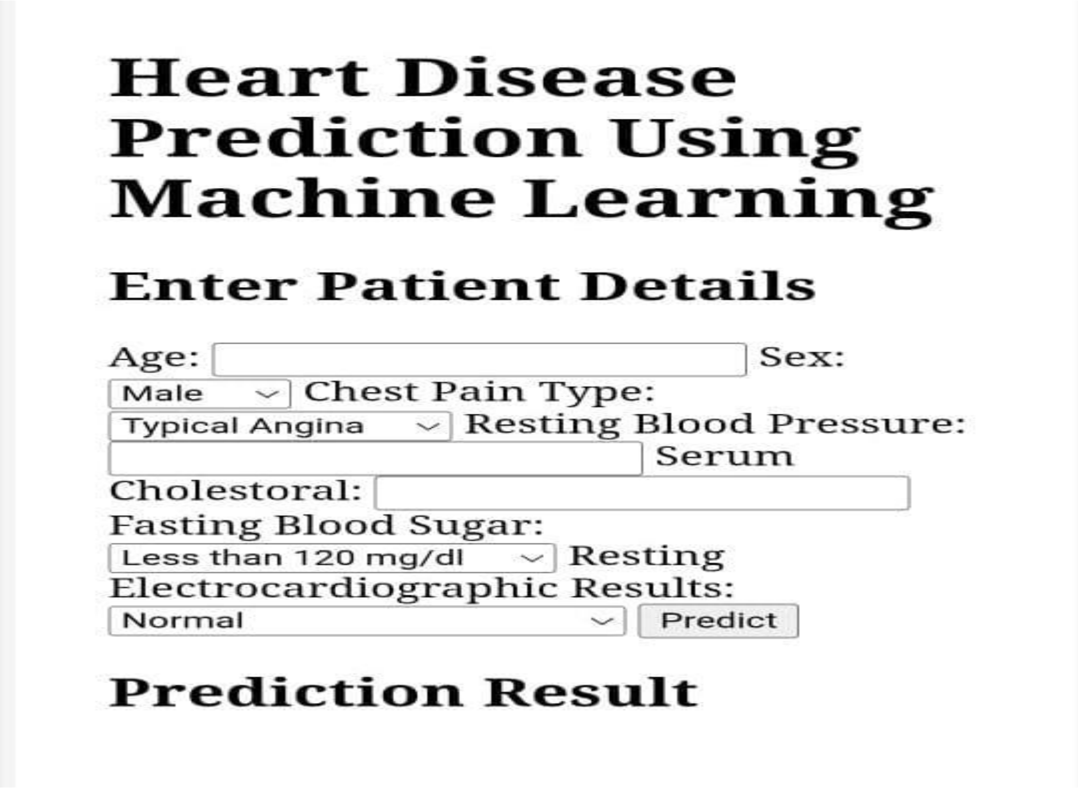


```
'Absence',
       'Presence', 'Presence',
'Absence', 'Absence', 'Presence',
       'Absence', 'Presence',
'Absence', 'Presence', 'Absence',
'Absence',
       'Presence', 'Absence',
'Presence', 'Absence', 'Absence',
'Absence',
       'Presence', 'Presence',
'Presence', 'Absence'],
dtype=object)
```

```
[ ]  from sklearn.metrics import confusion_
     cm=confusion_matrix(y_test,y_pred)
     sns.heatmap(cm,annot=True)
```

```
<Axes: >
```

```
[ ]  from sklearn.metrics import accuracy_s
     accuracy_score(y_test,y_pred)
```

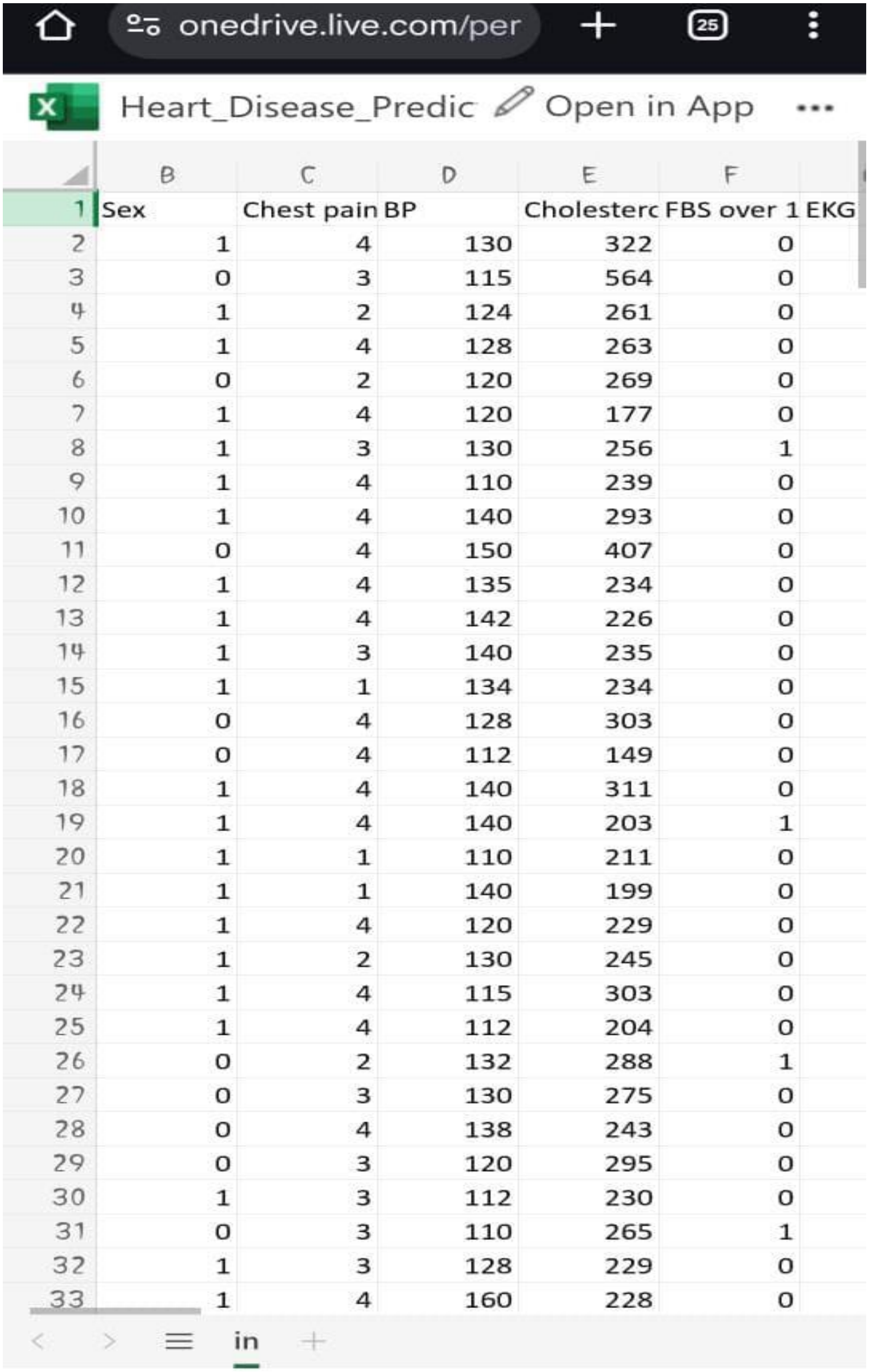


Heart Disease Prediction Using Machine Learning

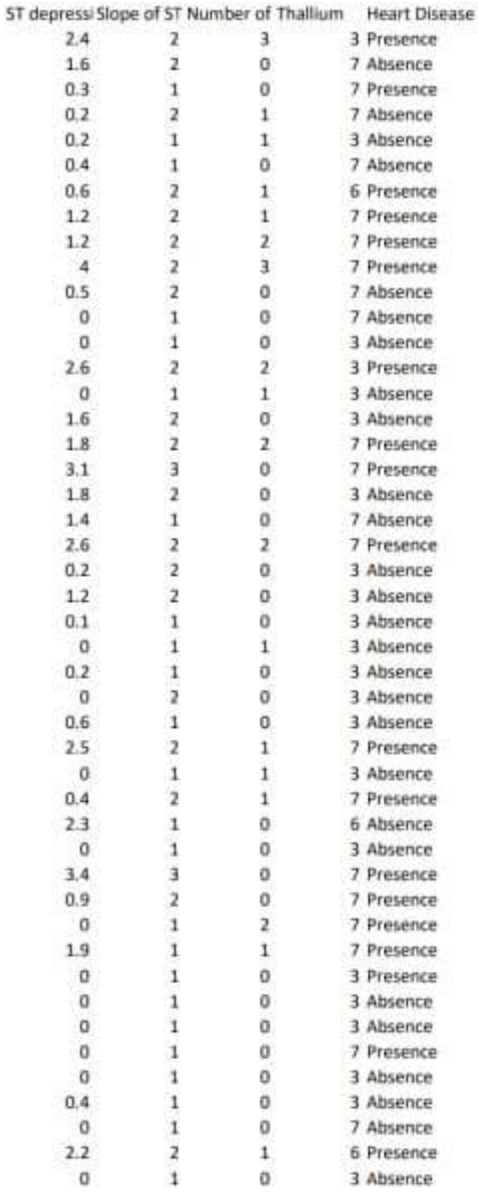
Enter Patient Details

Age: ▭ Sex: Male ⌄ Chest Pain Type: Typical Angina ⌄ Resting Blood Pressure: ▭ Serum Cholestoral: ▭ Fasting Blood Sugar: Less than 120 mg/dl ⌄ Resting Electrocardiographic Results: Normal ⌄ Predict

Prediction Result



	B	C	D	E	F	
1	Sex	Chest pain	BP	Cholesterc	FBS over 1	EKG
2	1	4	130	322	0	
3	0	3	115	564	0	
4	1	2	124	261	0	
5	1	4	128	263	0	
6	0	2	120	269	0	
7	1	4	120	177	0	
8	1	3	130	256	1	
9	1	4	110	239	0	
10	1	4	140	293	0	
11	0	4	150	407	0	
12	1	4	135	234	0	
13	1	4	142	226	0	
14	1	3	140	235	0	
15	1	1	134	234	0	
16	0	4	128	303	0	
17	0	4	112	149	0	
18	1	4	140	311	0	
19	1	4	140	203	1	
20	1	1	110	211	0	
21	1	1	140	199	0	
22	1	4	120	229	0	
23	1	2	130	245	0	
24	1	4	115	303	0	
25	1	4	112	204	0	
26	0	2	132	288	1	
27	0	3	130	275	0	
28	0	4	138	243	0	
29	0	3	120	295	0	
30	1	3	112	230	0	
31	0	3	110	265	1	
32	1	3	128	229	0	
33	1	4	160	228	0	



ST depressi	Slope of ST	Number of	Thallium	Heart Disease
2.4	2	3	3	Presence
1.6	2	0	7	Absence
0.3	1	0	7	Presence
0.2	2	1	7	Absence
0.2	1	1	3	Absence
0.4	1	0	7	Absence
0.6	2	1	6	Presence
1.2	2	1	7	Presence
1.2	2	2	7	Presence
4	2	3	7	Presence
0.5	2	0	7	Absence
0	1	0	7	Absence
0	1	0	3	Absence
2.6	2	2	3	Presence
0	1	1	3	Absence
1.6	2	0	3	Absence
1.8	2	2	7	Presence
3.1	3	0	7	Presence
1.8	2	0	3	Absence
1.4	1	0	7	Absence
2.6	2	2	7	Presence
0.2	2	0	3	Absence
1.2	2	0	3	Absence
0.1	1	0	3	Absence
0	1	1	3	Absence
0.2	1	0	3	Absence
0	2	0	3	Absence
0.6	1	0	3	Absence
2.5	2	1	7	Presence
0	1	1	3	Absence
0.4	2	1	7	Presence
2.3	1	0	6	Absence
0	1	0	3	Absence
3.4	3	0	7	Presence
0.9	2	0	7	Presence
0	1	2	7	Presence
1.9	1	1	7	Presence
0	1	0	3	Presence
0	1	0	3	Absence
0	1	0	3	Absence
0	1	0	7	Presence
0	1	0	3	Absence
0.4	1	0	3	Absence
0	1	0	7	Absence
2.2	2	1	6	Presence
0	1	0	3	Absence