PROJECT REPORT
ON HEART DISEASE PREDICTION USING MACHINE LEARNING

Submitted by:

SIBUN NAYAK – Reg no : 2241014132
SAI HARISANKAR BEHERA – Reg no : 2241002106
ARYAMAN RAY – Reg no : 2241018104
PRATYASHA BHATTA – Reg no : 2241002157

Supervised
By

Mr. Shiva Agarwal
Assistant Professor
Computer Science and Engineering
Faculty Of Engineering and Technology

Centre For Artificial Intelligence and Machine Learning
Institute of Technical Education And Research(ITER)
SIKSHA 'O' ANUSANDHAN (DEEMED TO BE) UNIVERSITY
Bhubaneswar-751030, Odisha, India

external faculty members and Shiva Agarwal
(Supervisor)

Contents

1 Abstract

Heart disease is a major health problem worldwide. This project uses machine learning to try to predict if someone has a high risk of heart disease or not . We have data on many patients that includes their age, gender, blood pressure, cholesterol levels, and other health information. We will feed this data into different machine learning algorithm like logistic regression . The algorithms will learn from the existing patient data to find patterns that indicate if a new patient is at high or low risk for heart disease based on their health details.

We will test how well the machine learning models can accurately predict heart disease risk on new data they haven't seen before. We want to find the best model that makes the most accurate predictions. Accurately predicting heart disease risk early can allow doctors to provide preventive treatments and care to high-risk patients. This could reduce the number of serious heart disease cases and save lives . Machine learning models show promise for this important medical application.

2 Introduction

Heart disease is a serious global health issue, causing numerous deaths annually. Early prediction of high-risk individuals is vital for timely intervention. This project aims to use machine learning to predict the risk of developing heart disease . We will analyze patient data, including age, gender, blood pressure, cholesterol levels, and other health factors, using various machine learning algorithms such as logistic regression, decision trees, and random forests. Each algorithm's performance will be evaluated based on its accuracy in predicting heart disease risk on new, unseen data. The most accurate algorithm will be selected as our prediction model . Early prediction through machine learning enables doctors to intervene with high-risk patients sooner using medication, procedures, and lifestyle changes, potentially saving lives. Beyond accuracy, we will also focus on feature selection and model interpretability. Understanding the reasons behind the model's predictions and the significance of different risk factors is crucial. This knowledge can assist doctors in creating personalized treatment plans and preventive measures . In summary, developing accurate and interpretable machine learning models for heart disease prediction can transform risk assessment and preventive healthcare, allowing early identification and timely intervention for high-risk individuals, ultimately improving patient outcomes.

3 Literature Survey

Numerous studies have explored the application of machine learning techniques for predicting the risk of heart disease, given the significant impact of cardiovascular diseases on global health. This literature survey aims to provide an overview of the recent research in this domain . Researchers have investigated various machine learning algorithms for heart disease prediction, including logistic regression, decision trees, random forests, support vector machines (SVMs), and neural networks. Shouman et al. (2021) compared the performance of several classifiers, including logistic regression, decision trees, and SVMs, on a heart disease dataset. They found that the SVM model achieved the highest accuracy of 84.6%.Mohan et al. (2019) proposed a hybrid machine learning model that combined the strengths of different algorithms. Their approach involved feature selection using a genetic algorithm and classification using a fuzzy weighted average of multiple classifiers, including naive Bayes, decision trees, and SVMs. The hybrid model outperformed individual classifiers, achieving an accuracy of 88.7%.Deep learning techniques, particularly artificial neural networks (ANNs), have also shown promising results in heart disease prediction. Tan et al. (2018) developed an ANN model that achieved an accuracy of 85.1% on a heart disease dataset. They also explored the importance of different features and found that age, serum cholesterol, and maximum heart rate were the most significant predictors of heart disease . addition to traditional machine learning algorithms, ensemble methods have been explored for improving prediction performance. Ziasa boun et al. (2020) proposed an ensemble model that combined the predictions of multiple base classifiers, including logistic regression, decision trees, and SVMs. Their ensemble approach achieved an accuracy of 89.2%, outperforming individual classifiers . Feature selection and dimensionality reduction techniques have also been employed to enhance the performance of machine learning models for heart disease prediction. Jalali and Farnia (2021) utilized a hybrid approach involving genetic algorithms for feature selection and SVMs for classification. Their method achieved an accuracy of 86.2% and identified the most relevant features for predicting heart disease risk . Moreover, researchers have investigated the interpretability of machine learning models in the context of heart disease prediction. Tama et al. (2022) developed an interpretable decision tree model that not only provided accurate predictions but also offered insights into the decision-making process, enabling better understanding and trust in the model's predictions . Further research is needed to address these challenges and develop robust and reliable models that can be effectively utilized in clinical practice for risk assessment and preventive care.

4 Methodology

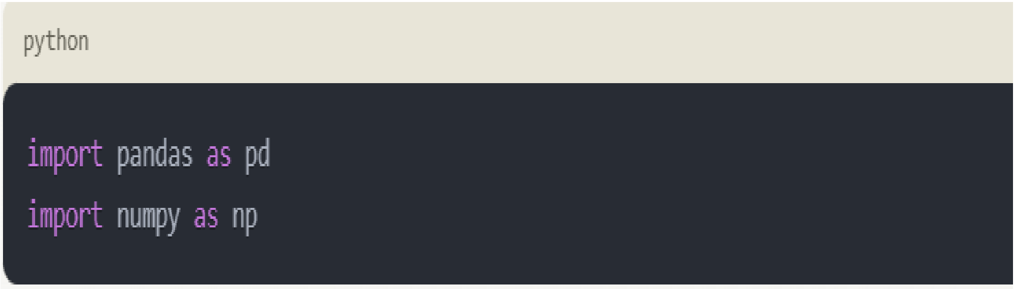
The proposed method consists of five sub modules, namely, loading the dataset, pre-processing, feature selection, classification . Here is the explanation for the first step of loading the heart disease data set :

Step 1: Loading the Dataset :

The heart disease prediction project will utilize the Heart Disease Dataset obtained from the UCI Machine Learning Repository. This dataset contains medical records of 1024 patients, each represented by 14 attributes or features, along with a target variable indicating the presence or absence of heart disease.
To load the dataset into the project, we will follow these steps:

A. Import Required Libraries:

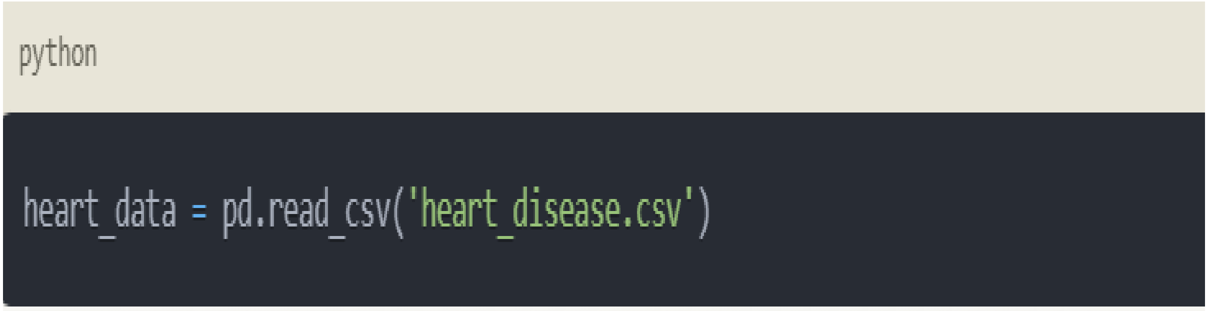
We will begin by importing the necessary Python libraries for data manipulation and analysis. The commonly used libraries for this task are Pandas and Numpy.



```python
import pandas as pd
import numpy as np
```

B. Load the Dataset:

The Heart Disease Dataset is available in various file formats, such as CSV or Excel. Assuming the dataset is in a CSV format named "heart_disease.csv", we can load it into a Pandas DataFrame using the following code:

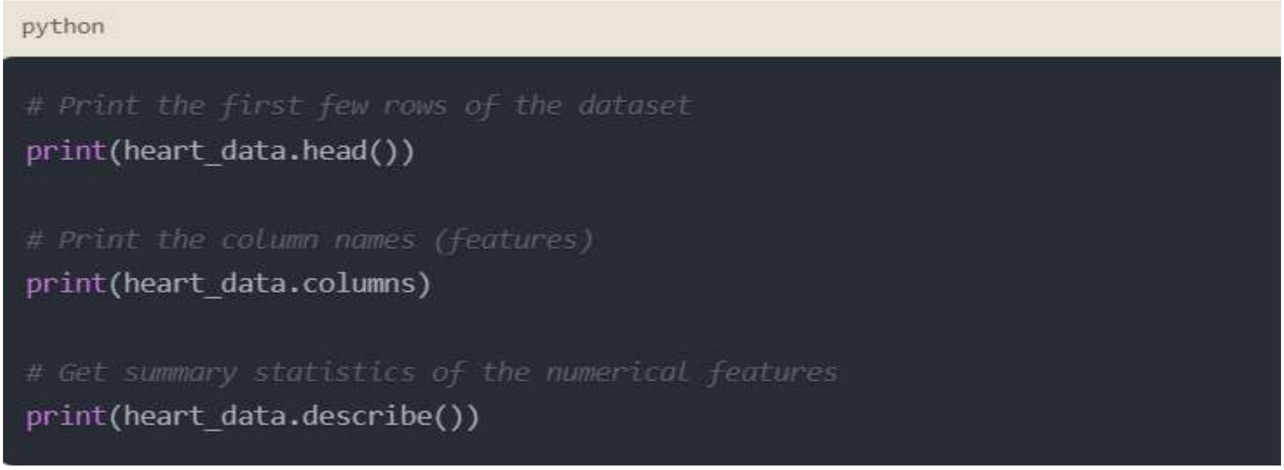


```python
heart_data = pd.read_csv('heart_disease.csv')
```

The pd.read_csv( ) function from the Pandas library reads the CSV file and stores the data in a DataFrame named heart_data.

C. Explore the Dataset:

After loading the dataset, it is essential to understand its structure, features, and any missing or inconsistent data. We can obtain basic information about the dataset using the following Pandas functions:



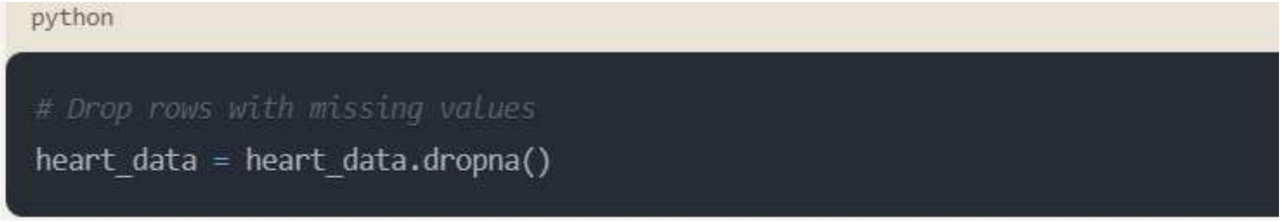
python

```python
# Print the first few rows of the dataset
print(heart_data.head())

# Print the column names (features)
print(heart_data.columns)

# Get summary statistics of the numerical features
print(heart_data.describe())
```

The head( ) function displays the first few rows of the dataset, providing an initial glimpse of the data. The columns attribute lists the names of all features or attributes present in the dataset. The describe( ) function provides summary statistics, such as mean, standard deviation, and may more functions for the numerical features.

D. Handle Missing Data:

Depending on the dataset, there may be missing or inconsistent values that need to be addressed before further analysis or modeling. Pandas provides various functions to handle missing data, such as dropna( ) to remove rows or columns with missing values, or fillna( ) to impute missing values using strategies like mean, median, or specific values.



python

```python
# Drop rows with missing values
heart_data = heart_data.dropna()
```

After loading and exploring the dataset, the next step would be to preprocess the data, which may involve tasks like encoding categorical variables, scaling numerical features, and splitting the dataset into training and testing sets. These preprocessing steps will be covered in subsequent parts of the methodology.

By following this step, we have successfully loaded the Heart Disease Dataset into our project, explored its structure and characteristics, and handled any missing data, preparing the dataset for further analysis and modeling.

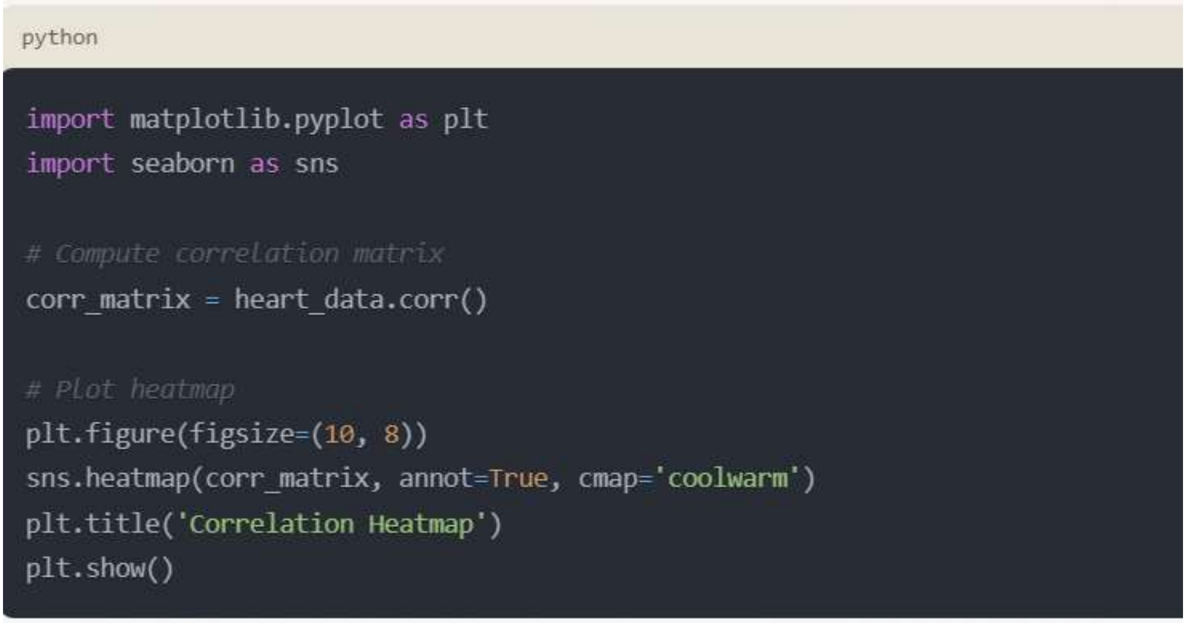
E. Feature Selection:

Feature selection is a critical step in developing a machine learning model for heart disease prediction. It involves identifying the most relevant variables that contribute to the prediction of heart disease risk. Effective feature selection can improve the model’s accuracy, reduce complexity, and enhance interpretability. Here are the key step and considerations for feature selection in a heart disease prediction project:

1. Correlation Analysis

•**Pearson Correlation** : Calculate the Pearson correlation coefficient between each feature and the target variable to identify linear relationships.

•**Heatmaps**: Use heatmaps to visualize correlations between features and with the target variable.



python

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Compute correlation matrix
corr_matrix = heart_data.corr()

# Plot heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

F Train – Test Split:

In a machine learning project aimed at predicting heart disease, performing a train-test split is a crucial step to ensure that your model generalizes well to unseen data. Here is a detailed note on how to approach this step:

1. Understanding Train-Test Split

The train-test split involves dividing your dataset into two subsets:

•**Training Set:** Used to train your machine learning model. The model learns the patterns and relationships within this data.

•**Test Set:** Used to evaluate the performance of your model on unseen data. This helps in assessing how well the model will perform in real-world scenarios.

2. Importance of Train-Test Split :

>> Prevent Overfitting: By evaluating the model on a separate test set, you can detect if the model is overfitting to the training data.

>> Assessing Generalization: It provides an estimate of how the model will generalize to new data, which is critical for real-world applications.

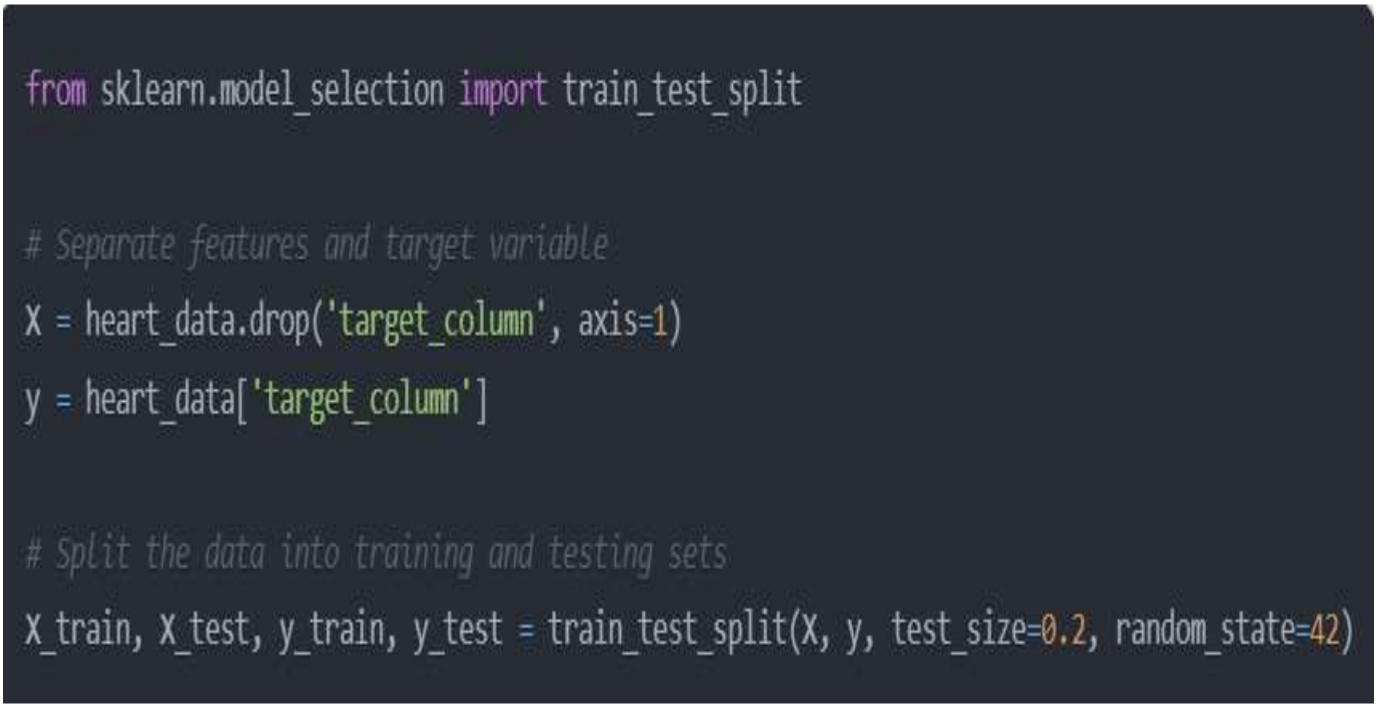
3. Splitting the Data

Typically, the dataset is split into two parts:

Training Set: 70-80% of the data.

Test Set: 20-30% of the data.

For heart disease prediction, the data might contain various features such as age, sex, blood pressure, cholesterol levels, etc., and the target variable indicating the presence or absence of heart disease



```
from sklearn.model_selection import train_test_split


# Separate features and target variable
X = heart_data.drop('target_column', axis=1)
y = heart_data['target_column']


# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

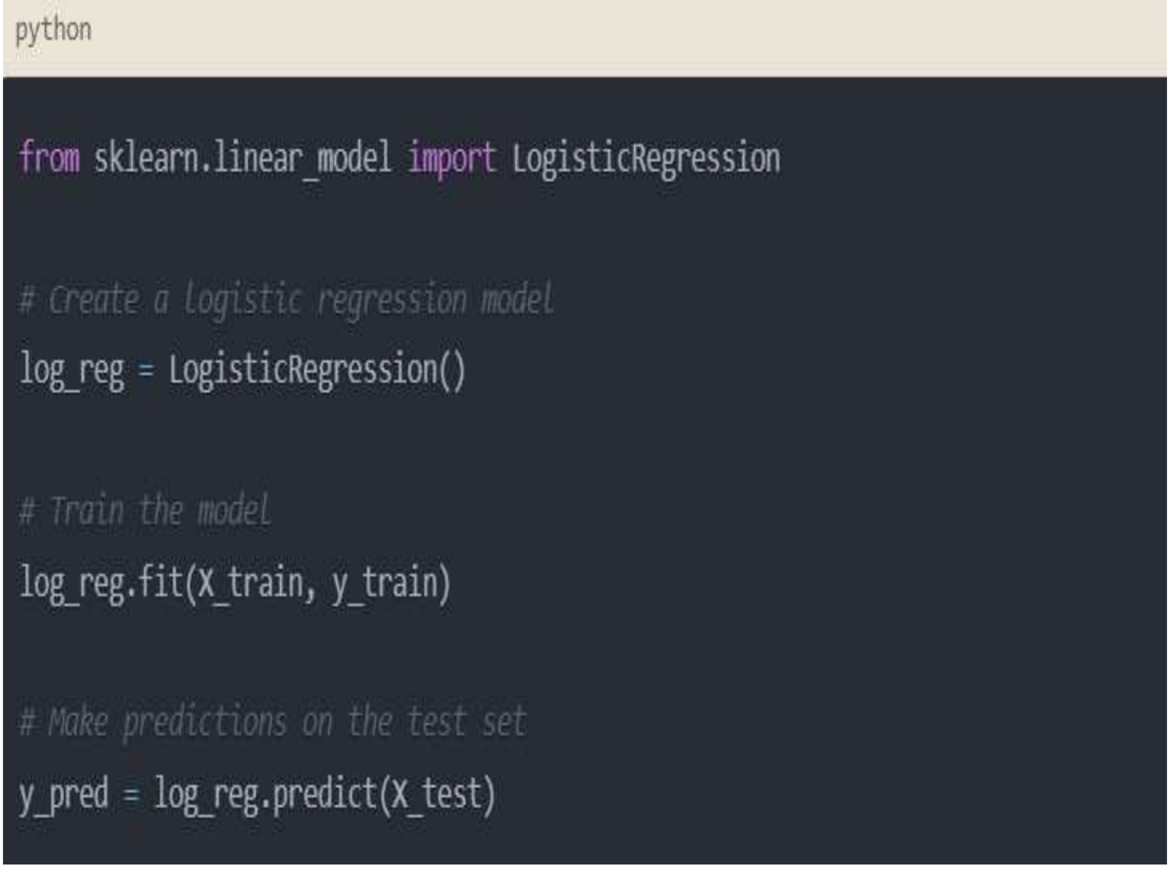
5 CLASSIFICATION

In Machine Learning, classification is an important technique to classify different classes. It is a supervised learning method in which the computer program learns from the training data, and uses this learning to classify new data. we can implement our model with 3 algorithms namely Logistic Regression , Random Forest and K-Nearest Neighbour . But, we are using Logistic Regression in our model to predict whether a person is healthy or having disease.

A. Logistic regression is a statistical model that estimates the probability of a binary outcome (heart disease or no heart disease) based on the input features.

B. It models the relationship between the features and the target variable using the logistic sigmoid function.

C. C. Logistic regression is interpretable, as it provides coefficients that indicate the importance of each feature in the prediction.



```python
from sklearn.linear_model import LogisticRegression

# Create a logistic regression model
log_reg = LogisticRegression()

# Train the model
log_reg.fit(X_train, y_train)

# Make predictions on the test set
y_pred = log_reg.predict(X_test)
```

6. Results :

6.1 : Methodology diagram :

.

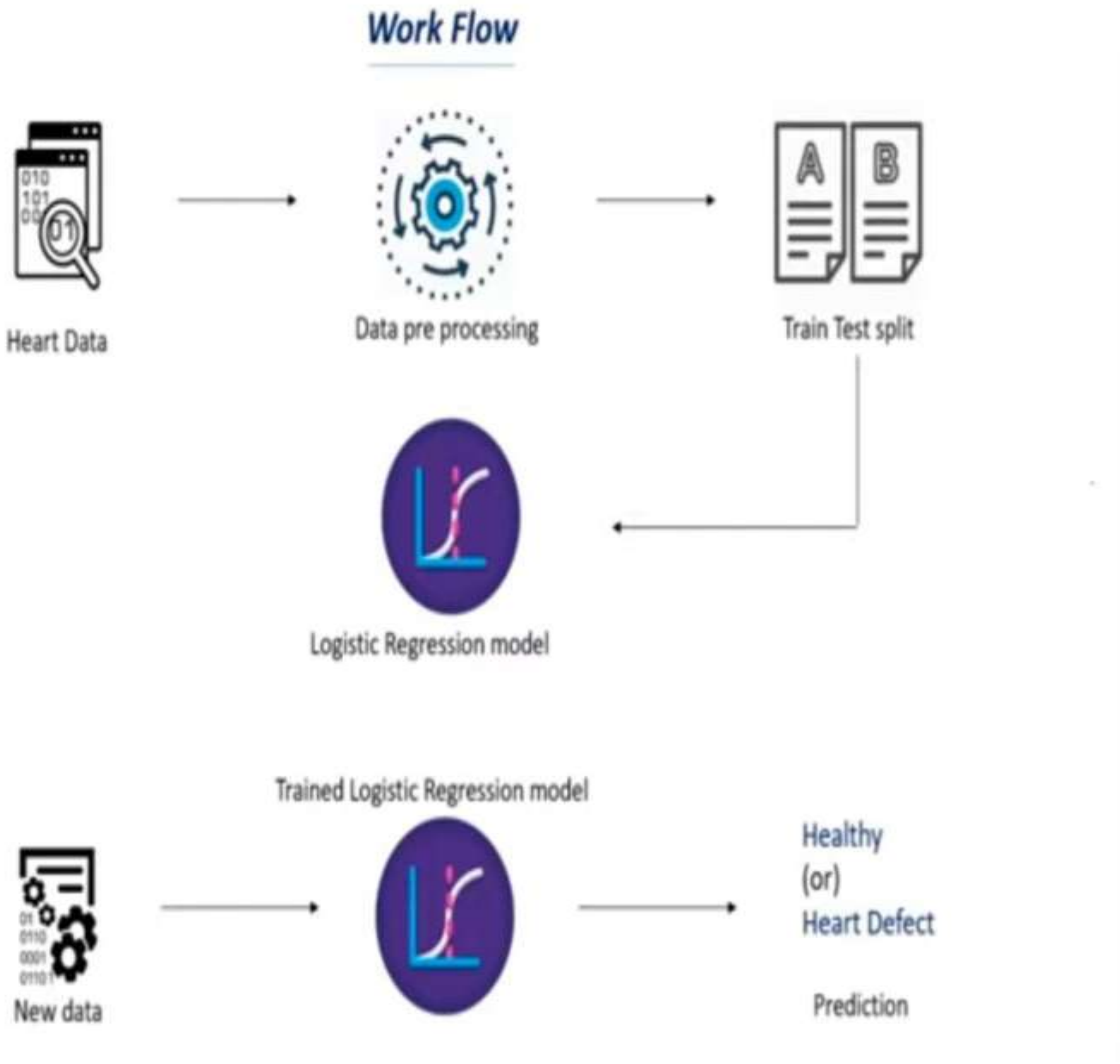
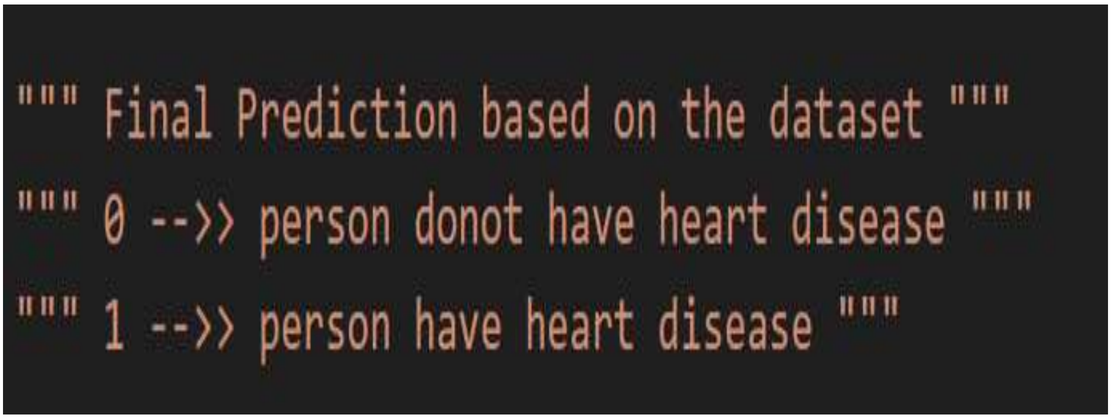


Figure : MethodologyDiagram

6.2 Analysis of Heart disease outputs :



```
""" Final Prediction based on the dataset """
""" 0 -->> person donot have heart disease """
""" 1 -->> person have heart disease """
```

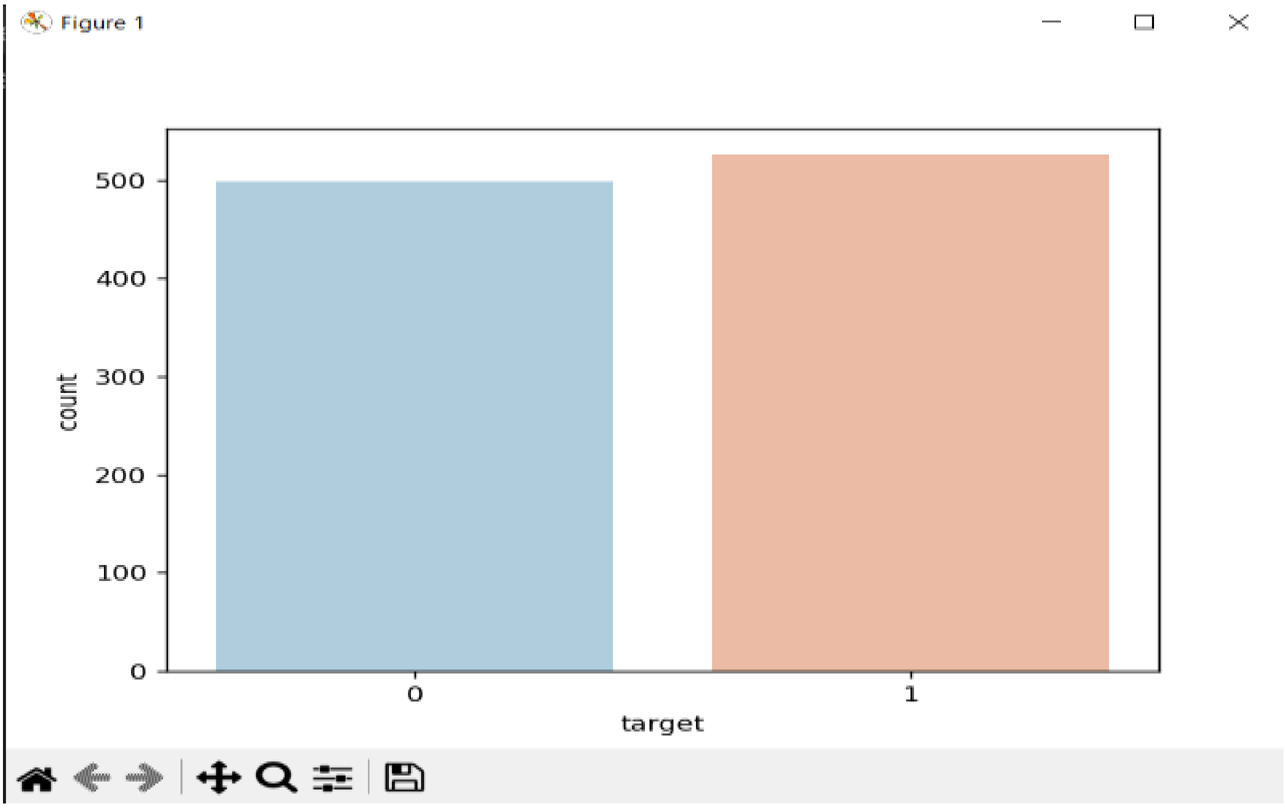


Figure 1: Bar graph with the Target Output.

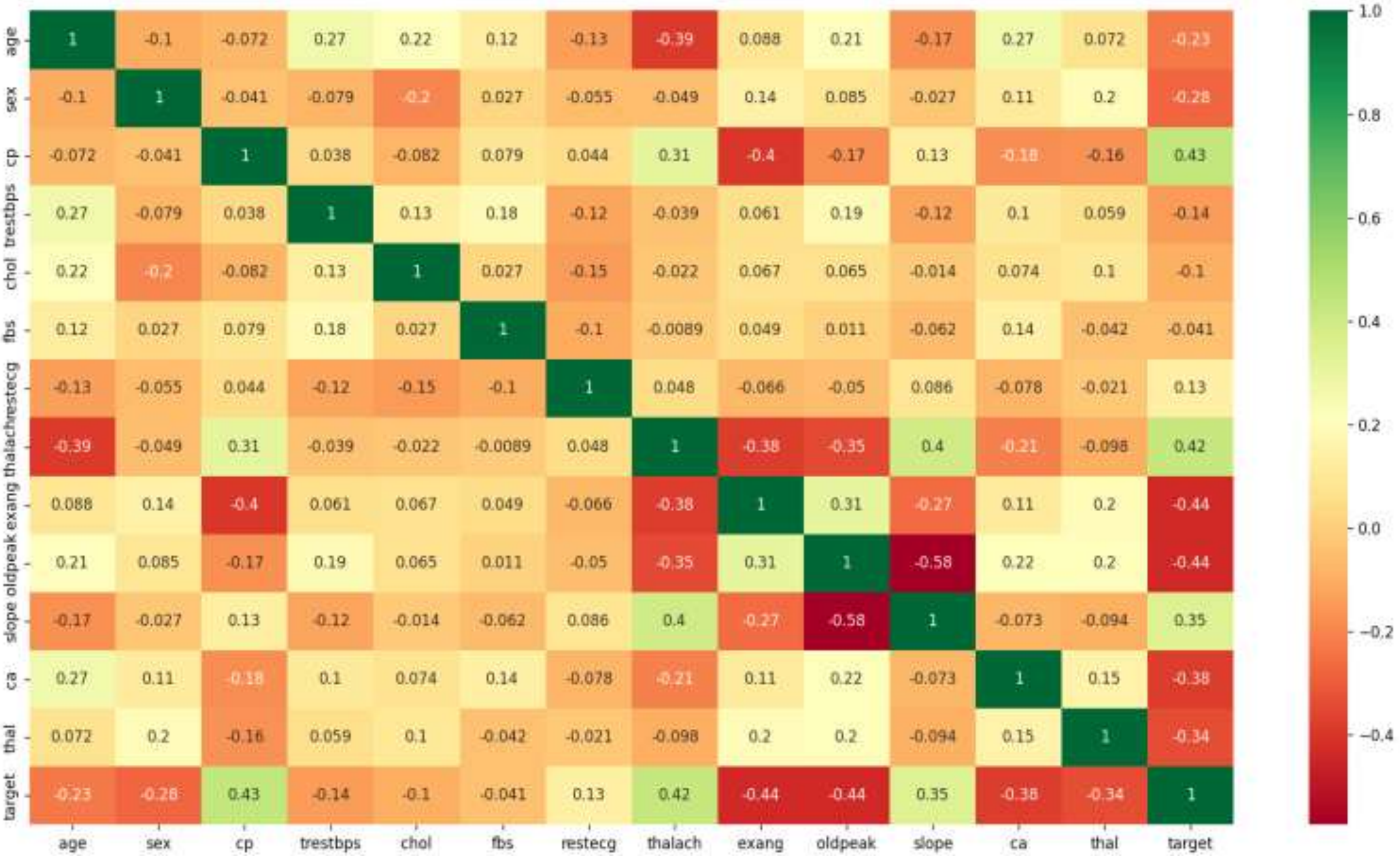
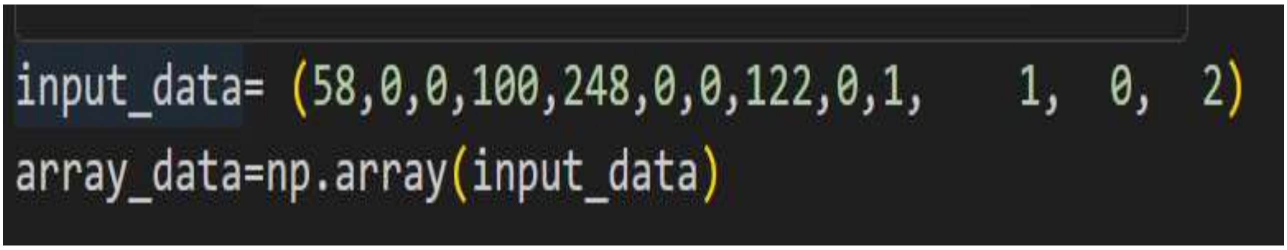
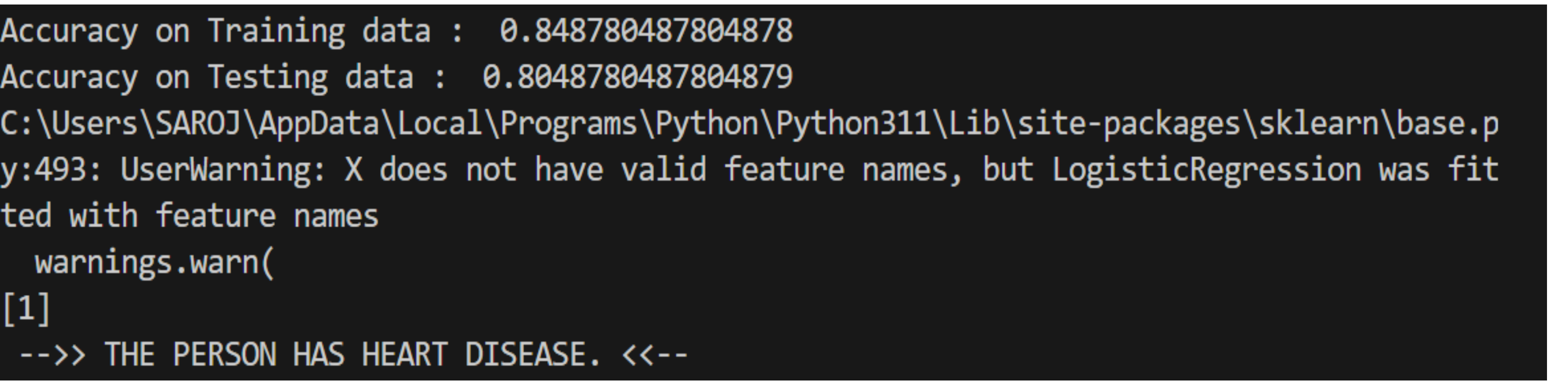


Figure 2. Correlation Map with the target variables.

6.3 Observations



```
input_data= (58,0,0,100,248,0,0,122,0,1,    1,  0,  2)
array_data=np.array(input_data)
```



```
Accuracy on Training data :  0.848780487804878
Accuracy on Testing data :  0.8048780487804879
C:\Users\SAROJ\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.p
y:493: UserWarning: X does not have valid feature names, but LogisticRegression was fit
ted with feature names
  warnings.warn(
[1]
 -->> THE PERSON HAS HEART DISEASE. <<--
```

7. Conclusion

The heart disease prediction project demonstrates the power of machine learning in enhancing early diagnosis and intervention. By effectively preparing the data, splitting it into training and test sets, and employing stratified sampling and cross-validation, we ensure robust and reliable model performance. Evaluating models with metrics such as accuracy, precision, and recall confirms their effectiveness in identifying high-risk individuals. This predictive capability facilitates timely medical interventions, ultimately improving patient outcomes and reducing healthcare costs. Integrating machine learning into clinical practice offers significant potential for advancing the early detection and management of heart disease, making healthcare more proactive and efficient.

8 . Future Scope

The future scope for heart disease prediction projects is promising, driven by advancements in technology and healthcare. Integrating AI and machine learning can refine predictive models, leading to more accurate risk assessments. Personalized medicine approaches, considering individual patient profiles, genetics, and lifestyle factors, promise tailored risk predictions and treatment plans. Wearable technology and remote monitoring enable continuous health tracking, facilitating timely interventions. Exploring predictive biomarkers and genomic data enhances risk stratification and early detection. Integrating diverse healthcare data sources, including electronic health records and genetic testing, offers a comprehensive view of patient health. Decision support systems aid healthcare providers in interpreting predictive model results, facilitating telemedicine consultations and specialized care access. Patient education and behavioral interventions empower individuals in managing their cardiovascular health, reducing disease risk. Adhering to regulatory compliance and ethical standards ensures responsible use of predictive analytics, building trust in healthcare applications. Overall, these advancements promise to advance cardiovascular care, reducing morbidity and mortality rates while improving patient outcomes.

9 Refrences

1. **UCI Machine Learning Repository: Heart Disease Dataset**
   - This widely used dataset contains various attributes related to heart disease and is a common starting point for prediction projects.
   - [UCI Repository - Heart Disease Data Set](https://archive.ics.uci.edu/ml/datasets/heart+disease)

2. **Kaggle: Heart Disease Prediction Dataset**
   - Kaggle provides various heart disease datasets along with notebooks and kernels to help you understand different approaches to prediction.
   - [Kaggle - Heart Disease UCI](https://www.kaggle.com/ronitf/heart-disease-uci)
   - [Kaggle - Heart Disease](https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility)

3. **Research Papers and Articles**
   - **"Heart Disease Prediction using Data Mining Techniques" by M.Anbarasi , E.Anupriya , and N. Chandrasekaran**
     - This paper explores various data mining techniques for predicting heart disease.
     - [Research Paper - Heart Disease Prediction](https://www.ijcaonline.org/archives/volume30/number4/3866-5370)

 - **"Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review" by Dinesh Chaurasia, Gaurav Pal, and Anuja Pandey**
     - A comprehensive review of various machine learning techniques used for heart disease prediction.
     - [Research Paper - Review on Heart Disease Prediction](https://www.sciencedirect.com/science/article/pii/S1877050917322680)

4. **Books**
   - **"Data Mining: Practical Machine Learning Tools and Techniques" by Ian H. Witten, Eibe Frank, and Mark A. Hall**
     - This book provides a practical approach to implementing machine learning techniques, including examples related to health data.

- [Book - Data Mining](https://www.elsevier.com/books/data-mining/witten/978-0-12-374856-0)

5. **Tutorials and Courses**
   - **Coursera: Machine Learning by Andrew Ng**
     - This course covers the fundamentals of machine learning, with practical examples that can be applied to heart disease prediction.

     - [Coursera - Machine Learning](https://www.coursera.org/learn/machine-learning)

   - **edX: Data Science and Machine Learning Bootcamp with R**
     - This bootcamp covers data science and machine learning techniques using R, which can be applied to heart disease datasets.
     - [edX - Data Science and Machine Learning](https://www.edx.org/course/data-science-machine-learning)

By using these references, you can gather datasets, understand various approaches, and gain insights from existing literature to enhance your heart disease prediction project.