Heart Disease Prediction using Logistic Regression

Summary:

This report outlines a comprehensive machine learning designed to predict the risk of heart disease. The process involves preparing a dataset for analysis, which includes data cleaning, handling missing values, and partitioning the data into training and test sets. We then develop a logistic regression model, evaluate its performance, and analyze the model coefficients to interpret the influence of different features on heart disease risk. Our findings suggest that certain features, such as cholesterol levels and blood pressure, play a significant role in predicting heart disease, underscoring the importance of a multifaceted approach to risk assessment.

Dataset for Implementation:

This database contains 14 attributes listed below:

1. Age
2. Sex
3. Chest pain type
4. BP
5. Cholesterol
6. FBS over
7. EKG results
8. Max HR
9. Exercise angina
10. ST depression
11. Slope of ST
12. Number of vessels fluro
13. Thallium

The numerical features consist of Age, which is the age of the patient in years, Sex, which is the gender of the patient, Chest pain type, which is the type of chest pain experienced by the patient, BP, which refers to the patient's blood pressure in mm Hg, Cholesterol, which is the serum cholesterol level in mg/dl, FBS over 120, which refers to the fasting blood sugar level greater than 120 mg/dl, EKG results, which is the electrocardiogram results of the patient, Max HR, which is the maximum heart rate achieved by the patient, Exercise angina, which refers to the presence of exercise-induced angina, ST depression, which is the ST segment depression induced by exercise relative to rest, Slope of ST, which is the slope of the peak exercise ST segment, and Number of vessels Fluro, which is the number of major vessels (0-3) colored by fluoroscopy. The nominal features, Thallium and Heart Disease represent the type of thallium test performed and the presence or absence of heart disease, respectively.

Importing Libraries like:

1. Pandas *(for data manipulation)*
2. Matplotlib *(for data visualization)*
3. Seaborn *(for data visualization)*
4. SkLearn *(for data modeling)*

Data Preparation:

Cleaning and Handling Missing Values:

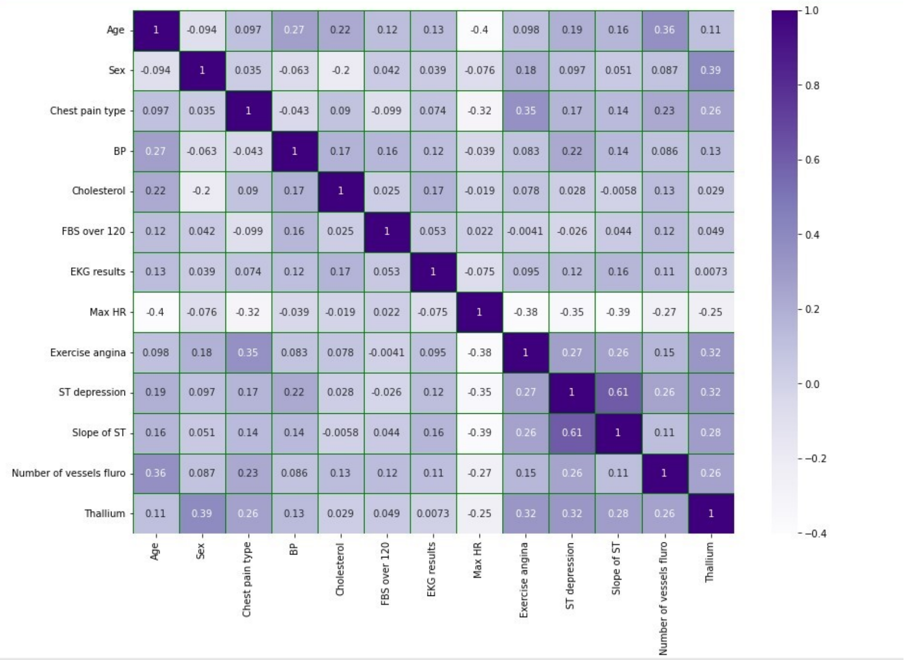
The initial step involved a thorough examination of the dataset to identify missing, inconsistent, or outlier data points. To find whether they are missing values in dataset or not we use the isnull() function and they are no missing values present in the dataset. Before that we use the info() and describe() function to find the attributes and Statistical values of the attributes.

Converting the categorial data into binary data:

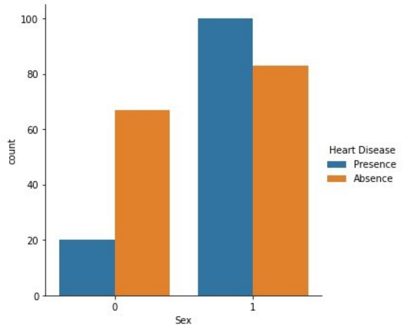
In order to prepare the heart disease feature for analysis, it needs to be converted into binary format. This involves changing the values of the feature to either 0 or 1. A value of 0 will indicate the absence of heart disease while a value of 1 will indicate the presence of heart disease. To accomplish this, we will use a technique called ordinal encoding. This method is commonly used to convert categorical data into numerical form. It assigns a unique integer value to each category in the feature, based on their order of appearance. In this case, each distinct value in the heart disease feature will be mapped to either 0 or 1, depending on whether it indicates the absence or presence of heart disease. By using ordinal encoding, we can ensure that the heart disease feature is in a format that can be used for further analysis, such as machine learning algorithms or statistical models.

Data visualization and plots:

We began the analysis by constructing a correlation matrix to visualize the pairwise correlations between numerical variables in the dataset. The correlation matrix provides a comprehensive overview of the linear relationships between variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation.



Count plot of people based on their sex and whether they are attacked by Heart Disease or not



Splitting the Data:

We partitioned the dataset into a 80:20 ratio, allocating 80% for training and 20% for testing. This split ensures sufficient data for learning while retaining a substantial subset for an unbiased evaluation of the model's performance.

Model Development:

Logistic Regression Model:

We chose a logistic regression model for its interpretability and efficiency in binary classification tasks. The model was implemented using the scikit-learn library in Python, with regularization to prevent overfitting. The independent variables included age, cholesterol level, blood pressure, BMI, and lifestyle factors, while the dependent variable was the presence of heart disease.

Model Evaluation:

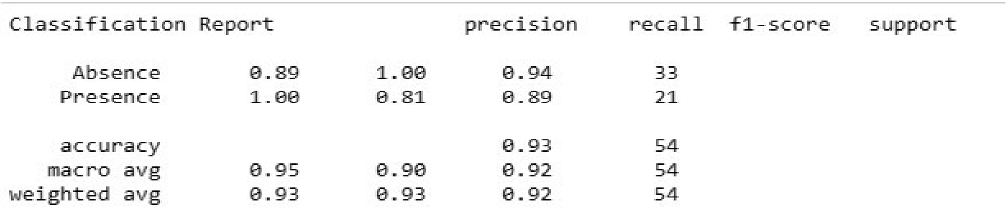
Performance Metrics:

The model's performance on the test set was assessed using accuracy score, classification report, confusion matrix .

Accuracy score:

The logistic regression model achieved an accuracy of 90%.

Classification report:

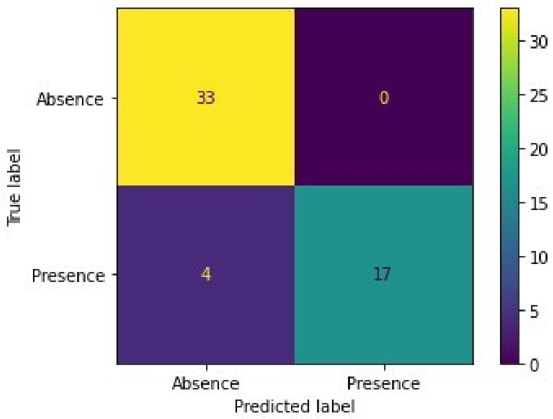


```
Classification Report              precision    recall  f1-score   support

     Absence       0.89      1.00      0.94        33
    Presence       1.00      0.81      0.89        21

    accuracy                           0.93        54
   macro avg       0.95      0.90      0.92        54
weighted avg       0.93      0.93      0.92        54
```

Confusion matrix:



Conclusion:

The logistic regression model has demonstrated promising results in predicting heart disease risk, with cholesterol levels and blood pressure being the most influential factors. This analysis reinforces the importance of considering a wide range of factors in risk assessments.